Learning with tree tensor networks: complexity estimates and model selection

Bertrand Michel and Anthony Nouy*

July 2, 2020

Abstract

In this paper, we propose and analyze a model selection method for tree tensor networks in an empirical risk minimization framework. Tree tensor networks, or tree-based tensor formats, are prominent model classes for the approximation of high-dimensional functions in numerical analysis and data science. They correspond to sum-product neural networks with a sparse connectivity associated with a dimension partition tree T, widths given by a tuple r of tensor ranks, and multilinear activation functions (or units). The approximation power of these model classes has been proved to be near-optimal for classical smoothness classes. However, in an empirical risk minimization framework with a limited number of observations, the dimension tree T and ranks r should be selected carefully to balance estimation and approximation errors. In this paper, we propose a complexity-based model selection strategy à la Barron, Birgé, Massart. Given a family of model classes, with different trees, ranks and tensor product feature spaces, a model is selected by minimizing a penalized empirical risk, with a penalty depending on the complexity of the model class. After deriving bounds of the metric entropy of tree tensor networks with bounded parameters, we deduce a form of the penalty from bounds on suprema of empirical processes. This choice of penalty yields a risk bound for the predictor associated with the selected model. For classical smoothness spaces, we show that the proposed strategy is minimax optimal in a least-squares setting. In practice, the amplitude of the penalty is calibrated with a slope heuristics method. Numerical experiments in a least-squares regression setting illustrate the performance of the strategy for the approximation of multivariate functions and univariate functions identified with tensors by tensorization (quantization).

1 Introduction

Typical tasks in statistical learning include the estimation of a regression function or of posterior probabilities for classification (supervised learning), or the estimation of the probability distribution of a random variable from samples of the distribution (unsupervised learning). These approximation tasks can be formulated as a minimization problem of a risk functional $\mathcal{R}(f)$ whose minimizer $f^\star$ is the target (or oracle) function, and such that $\mathcal{R}(f) - \mathcal{R}(f^\star)$ measures some discrepancy between the function f and $f^\star$. The risk is usually defined as

$$\mathcal{R}(f) = \mathbb{E}(\gamma(f, Z)),$$

with $Z = (X, Y)$ for supervised learning or $Z = X$ for unsupervised learning, and where γ is a contrast function. For supervised learning, the contrast γ is usually chosen as $\gamma(f, (x, y)) = \ell(y, f(x))$ where $\ell(y, f(x))$ measures some discrepancy between y and the prediction $f(x)$ for a given realization (x, y) of (X, Y). In practice, given i.i.d. realizations $(Z_1, \ldots, Z_n)$ of Z, an approximation $\hat{f}_n^M$ is obtained by the minimization of an empirical risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \gamma(f, Z_i)$$

*Centrale Nantes, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France

over a subset of functions M, also called a model class or hypothesis set. Assuming that the risk admits a minimizer f^M over M, the error $\mathcal{R}(\widehat{f}_n^M) - \mathcal{R}(f^\star)$ can be decomposed into two contributions: an approximation error $\mathcal{R}(f^M) - \mathcal{R}(f^\star)$ which quantifies the best we can expect from the model class M, and an estimation error $\mathcal{R}(\widehat{f}_n^M) - \mathcal{R}(f^M)$ which is due to the use of a limited number of observations. For a given model class, a first problem is to understand how these errors behave under some assumptions on the target function. When considering an increasing sequence of model classes, the approximation error decreases but the estimation error usually increases. Then strategies are required for the selection of a particular model class.

In many applications, the target function $f^\star(x)$ is a function of many variables $x = (x_1, \ldots, x_d)$. For applications in image or signal classification, x may be an image (with d the number of pixels or patches) or a discrete time signal (with d the number of time instants) and $f^\star(x)$ provides a label to a particular input x. For applications in computational science, the target function may be the solution of a high-dimensional partial differential equation, a parameter-dependent equation or a stochastic equation. In all these applications, when d is large and when the number of observations is limited, one has to rely on suitable model classes M of moderate complexity that exploit specific structures of the target function $f^\star$ and yield an approximation $\widehat{f}_n^M$ with low approximation and estimation errors. Typical examples of model classes include additive functions $f_1(x_1) + \cdots + f_d(x_d)$, sums of multiplicative functions $\sum_{k=1}^m f_1^k(x_1) \cdots f_d^k(x_d)$, projection pursuit $f_1(w_1^T x) + \cdots + f_m(w_m^T x)$, or feed-forward neural networks $\sigma_L \circ f_L \circ \ldots \circ \sigma_1 \circ f_1(x)$ where the f_k are affine maps and the σ_k are given nonlinear functions.

In this paper, we consider the class of functions in tree-based tensor format, or tree tensor networks. These model classes are well-known approximation tools in numerical analysis and computational physics and have also been more recently considered in statistical learning. They are particular cases of feed-forward neural networks with an architecture given by a dimension partition tree and multilinear activation functions (see [26, 14]). For an overview of these tools, the reader is referred to the monograph [22] and the surveys [30, 6, 25, 12, 13]. Some results on the approximation power of tree tensor networks can be found in [32, 20, 5] for multivariate functions, or in [24, 23, 2, 3] for tensorized (or quantized) functions. A tree-based tensor format is a set of functions

$$M_r^T(\mathcal{H}) = \{f \in \mathcal{H} : \operatorname{rank}_\alpha(f) \leq r_\alpha, \alpha \in T\},$$

where T is a dimension partition tree over $\{1, \ldots, d\}$, $r = (r_\alpha) \in \mathbb{N}^{|T|}$ is a tuple of integers and $\mathcal{H}$ is a finite dimensional tensor space of multivariate functions (e.g., polynomials, splines), which is a tensor product feature space. A function f in $M_r^T(\mathcal{H})$ have a α-rank $\operatorname{rank}_\alpha(f)$ bounded by r_α, that means it admits a representation

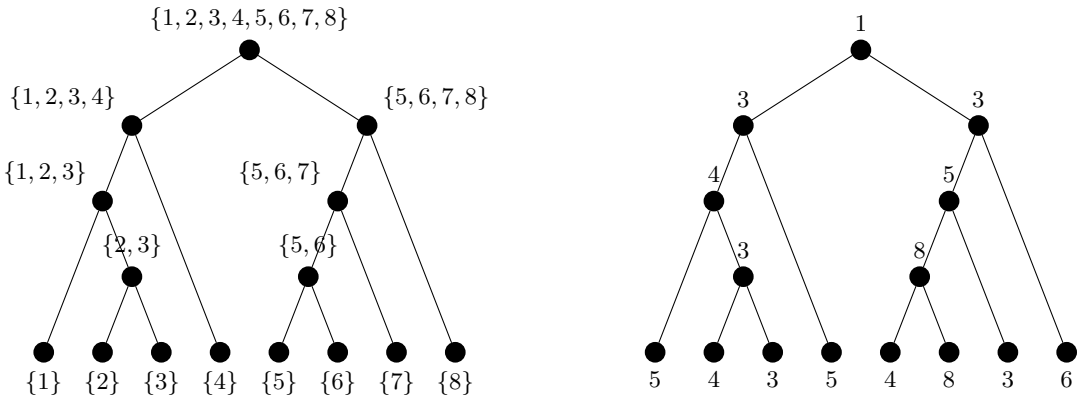
$$f(x) = \sum_{k=1}^{r_\alpha} g_k^\alpha(x_\alpha) h_k^{\alpha^c}(x_{\alpha^c})$$

for some functions g_k^α and $h_k^{\alpha^c}$ of complementary groups of variables. The function f admits a representation as a composition of multilinear functions. For instance, for the dimension tree of Figure 1a,

$$\begin{aligned} f(x) = f^{1,\ldots,8}\Big(&f^{1,2,3,4}\big(f^{1,2,3}(f^1(\phi^1(x_1)), f^{2,3}(f^2(\phi^2(x_2)), f^3(\phi^3(x_3)))), f^4(\phi^4(x_4))\big), \\ &f^{5,6,7,8}\big(f^{5,6,7}(f^{5,6}(f^5(\phi^5(x_5)), f^6(\phi^6(x_6))), f^7(\phi^7(x_7))), f^8(\phi^8(x_8))\big)\Big) \end{aligned}$$

where $\phi^\nu(x_\nu) \in \mathbb{R}^{n_\nu}$ is a vector of n_ν features in the variable x_ν, and f^α is a multilinear map with values in $\mathbb{R}^{r_\alpha}$. It corresponds to the neural network illustrated on Figure 2.

The main contribution of the paper is a complexity-based strategy for the selection of a model class in an empirical risk minimization framework. Given a family of model classes $M_m = M_{r_m}^{T_m}(\mathcal{H}_m)$, $m \in \mathcal{M}$, associated with different trees T_m, ranks r_m and background approximation spaces $\mathcal{H}_m$, and given the



(a) Dimension tree T.

(b) r_α, $\alpha \in T$.

Figure 1: Dimension tree T over $\{1, \ldots, 8\}$ (a) and ranks $r = (r_\alpha)_{\alpha \in T}$ (b).

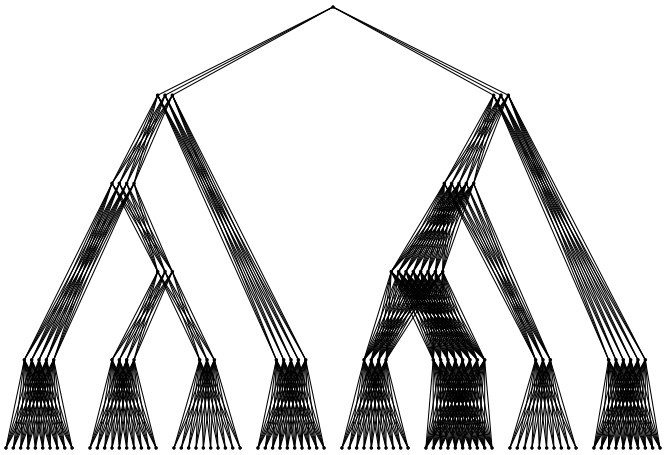


Figure 2: Neural network corresponding to the format $M_r^T(\mathcal{H})$ with the tree T and ranks r of Figure 1, and $n_\nu = 10$ features per variable.

corresponding predictors $\hat{f}_m$ that minimize the empirical risk, we propose a strategy to select a particular model $\hat{m}$ with a guaranteed performance. For that purpose, we make use of the model selection approach of Barron, Birgé and Massart (see [28] for a general introduction to the topic) where $\hat{m}$ is obtained by minimizing a penalized empirical risk

$$\widehat{\mathcal{R}}_n(\hat{f}_m) + \mathrm{pen}(m)$$

with a penalty function $\mathrm{pen}(m)$ derived from complexity estimates of the model classes M_m, of the form $\mathrm{pen}(m) \sim O(\sqrt{C_m/n})$ (up to logarithmic terms) in a general setting, or of the form $\mathrm{pen}(m) \sim O(C_m/n)$ (again up to logarithmic terms) in a bounded least-squares setting where faster convergence rates can be obtained. In particular, we find that our strategy is minimax adaptive over Sobolev spaces. In practice, the penalty is taken of the form $\mathrm{pen}(m) = \lambda\sqrt{C_m/n}$ (or $\mathrm{pen}(m) = \lambda C_m/n$ in a bounded regression setting), where λ is calibrated with the slope heuristics method proposed in [10]. The family of models can be generated by adaptive learning algorithms such as the ones proposed in [18, 17].

Note that our method is a ℓ_0 type approach. Convex regularization methods would be an interesting alternative route to follow. A straightforward convexification of tensor formats consists in using the sum of nuclear norms of unfoldings (see, e.g., [33] for Tucker format) but this is known to be far from optimal from a statistical point of view [31]. A convex regularization method based on the tensor nuclear norm has

been proposed for the Tucker format, or shallow tensor network, which comes with theoretical guarantees (see [35]). However, there is no straightforward extension of this approach to general tree tensor networks.

The outline of the paper is as follows. In Section 2, we describe the model class of tree tensor networks (or tree-based tensor formats) in the case of vector-valued functions, which generalizes the classical definition for real-valued functions [22, 16] and allows considering applications such as multiclass classification. In Section 3, we provide estimates of the metric and bracketing entropies in L^p spaces for tree tensor networks M_m with bounded parameters. In Section 4, we derive bounds for the estimation error in a classical empirical risk minimization framework. These bounds are derived from concentration inequalities for empirical processes. In Section 5, we present the complexity-based model selection approach and we derive risk bounds for particular choices of penalty, first in a general setting and then in the bounded least-squares setting. Then we present the practical aspects of the approach, which includes the slope heuristics method for penalty calibration and the exploration strategies for the generation of a sequence of model classes and associated predictors. Finally in Section 6, we present some numerical experiments that validate the proposed model selection strategy.

2 Tree tensor networks

We consider functions $f(x) = f(x_1, \ldots, x_d)$ defined on a product set $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ and with values in $\mathbb{R}^s$. Typically, $\mathcal{X}_\nu$ is a subset of $\mathbb{R}$ or $\mathbb{R}^{d_\nu}$ but it could be a set of more general objects (sequences, functions, graphs...).

2.1 Tensor product feature space

For each $\nu \in \{1, \ldots, d\}$, we introduce a finite-dimensional space $\mathcal{H}_\nu$ of functions defined on $\mathcal{X}_\nu$. We let $\{\phi^\nu_{i_\nu} : i_\nu \in I^\nu\}$ be a basis of $\mathcal{H}_\nu$, with $I^\nu = \{1, \ldots, n_\nu\}$. The functions $\phi^\nu_{i_\nu}(x_\nu)$ may be polynomials, splines, wavelets, kernel functions, or more general functions that extract n_ν features from a given input $x_\nu \in \mathcal{X}_\nu$. We let $\phi^\nu : \mathcal{X}_\nu \to \mathbb{R}^{n_\nu}$ be the associated *feature map* defined by $\phi^\nu(x_\nu) = (\phi^\nu_1(x_\nu), \ldots, \phi^\nu_{n_\nu}(x_\nu))^T \in \mathbb{R}^{n_\nu}$. The functions $\phi_i(x) = \phi^1_{i_1}(x_1) \ldots \phi^d_{i_d}(x_d)$, $i \in I = I^1 \times \ldots \times I^d$, form a basis of the tensor product space $\mathcal{H} = \mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_d$. A function $f \in \mathcal{H}$ admits a representation

$$f(x) = \sum_{i \in I} a_i \phi_i(x) = \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} a_{i_1,\ldots,i_d} \phi^1_{i_1}(x_1) \ldots \phi^d_{i_d}(x_d), \tag{1}$$

where $a \in \mathbb{R}^I = \mathbb{R}^{n_1 \times \ldots \times n_d}$ is an algebraic tensor (or multi-dimensional array) of size $n_1 \times \ldots \times n_d$. The map ϕ from $\mathcal{X}$ to $\mathbb{R}^I$ which associates to x the elementary tensor $\phi(x) = \phi^1(x_1) \otimes \ldots \otimes \phi^d(x_d) \in \mathbb{R}^I$ defines a *tensor product feature map*.

A function f defined on $\mathcal{X}$ with values in $\mathbb{R}^s$ whose components f_k ($1 \leq k \leq s$) are in $\mathcal{H}$ is identified with an element of the product space $\mathcal{H}^s$, which is itself identified with the space $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_d \otimes \mathbb{R}^s$ of tensors of order $d + 1$.

2.2 Tree-based ranks and related tensor formats

For any $\alpha \subset \{1, \ldots, d\} := D$, we introduce the tensor space $\mathcal{H}_\alpha = \bigotimes_{\nu \in \alpha} \mathcal{H}_\nu$ of functions defined on $\mathcal{X}_\alpha = \times_{\nu \in \alpha} \mathcal{X}_\nu$, and for $x \in \mathcal{X}$, we let $x_\alpha = (x_\nu)_{\nu \in \alpha} \in \mathcal{X}_\alpha$ denote the group of variables α. We denote by $\alpha^c = D \setminus \alpha$. We use the conventions $\mathcal{H}_\emptyset = \mathbb{R}$ and $\mathcal{H}_D = \mathcal{H}$.

Definition 2.1. *The α-rank of a function $f : \mathcal{X} \to \mathbb{R}^s$, denoted* $\mathrm{rank}_\alpha(f)$*, is the minimal integer r_α such that*

$$f(x) = \sum_{k=1}^{r_\alpha} g^\alpha_k(x_\alpha) h^{\alpha^c}_k(x_{\alpha^c}) \tag{2}$$

for some functions $g_k^\alpha : \mathcal{X}_\alpha \to \mathbb{R}$ *and* $h_k^\alpha : \mathcal{X}_{\alpha^c} \to \mathbb{R}^s$.

The above definition generalizes the classical notion of α-rank for vector-valued functions. It coincides with the classical notion of α-rank when $f : \mathcal{X} \to \mathbb{R}^s$ is seen as a real-valued function of $s+1$ variables defined on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d \times \{1, \ldots, s\}$. A function $f \in \mathcal{H}^s$ admits a representation (2) with functions $g_k^\alpha \in \mathcal{H}_\alpha$ and $h_k^{\alpha^c} \in \mathcal{H}^s_{\alpha^c}$. For $f \neq 0$, we have $\text{rank}_\emptyset(f) = 1$ and $1 \leq \text{rank}_D(f) \leq s$.

We let T be a *dimension partition tree* over D, with root D and leaves $\{\nu\}$, $1 \leq \nu \leq d$. For a node $\alpha \in T$, we denote by $S(\alpha)$ the set of children of α. For any node α, we have either $S(\alpha) = \emptyset$ (for leaf nodes) or $S(\alpha) \geq 2$ (for interior nodes). We denote by $\mathcal{L}(T)$ the set of leaves of T, and by $\mathcal{I}(T) = T \setminus \mathcal{L}(T)$ its interior nodes. For an interior node $\alpha \in \mathcal{I}(T)$, $S(\alpha)$ forms a partition of α. The T-rank of a function f is the tuple $\text{rank}_T(f) = (\text{rank}_\alpha(f))_{\alpha \in T}$. The number of nodes of a dimension partition tree over D is bounded as $|T| \leq 2d - 1$.

Given a tuple $r \in \mathbb{N}^{|T|}$ we introduce the model class $M_r^T(\mathcal{H}^s)$ of functions in $\mathcal{H}^s$ with T-rank bounded by r,

$$M_r^T(\mathcal{H}^s) = \{f \in \mathcal{H}^s : \text{rank}_T(f) \leq r\}.$$

A function $f \in M_r^T(\mathcal{H}^s)$ admits the representation

$$f(x) = \sum_{k_D=1}^{r_D} c_{k_D} g_{k_D}^D(x)$$

where the c_{k_D} are vectors in $\mathbb{R}^s$ and where the functions $g_{k_D}^D \in \mathcal{H}$ are defined recursively. For any interior node $\alpha \in \mathcal{I}(T)$, the functions $g_{k_\alpha}^\alpha$ admit the representation

$$g_{k_\alpha}^\alpha(x_\alpha) = \sum_{\substack{1 \leq k_\beta \leq r_\beta \\ \beta \in S(\alpha)}} C^\alpha_{k_\alpha, (k_\beta)_{\beta \in S(\alpha)}} \prod_{\beta \in S(\alpha)} g^\beta_{k_\beta}(x_\beta),$$

where $C^\alpha \in \mathbb{R}^{r_\alpha \times (\times_{\beta \in S(\alpha)} r_\beta)}$. For a leaf node $\alpha \in \mathcal{L}(T)$, the functions $g_{k_\alpha}^\alpha \in \mathcal{H}_\alpha$ admit the representation

$$g_{k_\alpha}^\alpha(x_\alpha) = \sum_{i_\alpha \in I^\alpha} C^\alpha_{k_\alpha, i_\alpha} \phi^\alpha_{i_\alpha}(x_\alpha).$$

We let $C^\emptyset$ denote the matrix whose columns are the vectors c_{k_D}. We introduce the tree $T^\star = T \cup \emptyset$ and we use the conventions $r_\emptyset = s$ and $S(\emptyset) = D$. A function f in $M_r^T(\mathcal{H}^s)$ therefore admits an explicit representation

$$f_k(x) = \sum_{\substack{i_\alpha \in I^\alpha \\ \alpha \in \mathcal{L}(T)}} \sum_{\substack{1 \leq k_\beta \leq r_\beta \\ \beta \in T}} C^\emptyset_{k, k_D} \prod_{\alpha \in T \setminus \mathcal{L}(T)} C^\alpha_{k_\alpha, (k_\beta)_{\beta \in S(\alpha)}} \prod_{\alpha \in \mathcal{L}(T)} C^\alpha_{k_\alpha, i_\alpha} \prod_{\alpha \in \mathcal{L}(T)} \phi^\alpha_{i_\alpha}(x_\alpha) \tag{3}$$

where the set of parameters $(C^\alpha)_{\alpha \in T^\star}$ form a tree network of tensors, and $C^\alpha \in \mathbb{R}^{\{1,\ldots,r_\alpha\} \times I^\alpha} := \mathbb{R}^{K^\alpha}$, where $I^\alpha = \{1, \ldots, r_D\}$ for $\alpha = \emptyset$, $I^\alpha = \times_{\beta \in S(\alpha)} \{1, \ldots, r_\beta\}$ for $\alpha \in \mathcal{I}(T)$ or $I^\alpha = \{1, \ldots, n_\alpha\}$ for $\alpha \in \mathcal{L}(T)$. We let $R_{\mathcal{H},T,r}$ be the map which associates to a set of tensors $(C^\alpha)_{\alpha \in T^\star}$ the function $f = R_{\mathcal{H},T,r}((C^\alpha)_{\alpha \in T^\star})$ defined by (3), so that

$$M_r^T(\mathcal{H}^s) = \{f = R_{\mathcal{H},T,r}((C^\alpha)_{\alpha \in T^\star}) : C^\alpha \in \mathbb{R}^{K^\alpha}, \alpha \in T^\star\}.$$

From the representation (3), we obtain the following

Lemma 2.2. *The map* $R_{r,T,\mathcal{H}}$ *is a multilinear map from the product space* $\bigtimes_{\alpha \in T^\star} \mathbb{R}^{K^\alpha}$ *to* $\mathcal{H}^s$.

Remark 2.3. *If* $r_D = s$, *the parameter* $C^\emptyset \in \mathbb{R}^{s \times s}$ *can be chosen as the identity matrix, so that the parameters of a function in* $M_r^T(\mathcal{H}^s)$ *are reduced to the set of tensors* $(C^\alpha)_{\alpha \in T}$. *This includes the classical case of tree-based tensor formats for real-valued functions* ($s = r_D = 1$). *In this situation, we let* $T^\star = T$.

2.3 Tree tensor networks as compositions of multilinear functions

A function f in $M_r^T(\mathcal{H}^s)$ admits a representation in terms of compositions of multilinear functions. For a given $\alpha \in T$, we let $g^\alpha(x_\alpha) = (g^\alpha_{k_\alpha}(x_\alpha))_{1 \leq k_\alpha \leq r_\alpha} \in \mathbb{R}^{r_\alpha}$. The matrix $C^\emptyset \in \mathbb{R}^{s \times r_D}$ is linearly identified with a linear map $f^\emptyset$ from $\mathbb{R}^{r_D}$ to $\mathbb{R}^s$. Therefore, a function f in $M_r^T(\mathcal{H}^s)$ admits the representation

$$f(x) = f^\emptyset(g^D(x)).$$

For any $\alpha \in \mathcal{I}(T)$, the tensor C^α can be linearly identified with a multilinear map

$$f^\alpha : \bigtimes_{\beta \in S(\alpha)} \mathbb{R}^{r_\beta} \to \mathbb{R}^{r_\alpha}$$

defined by

$$f^\alpha_{k_\alpha}((z^\beta)_{\beta \in S(\alpha)}) = \sum_{\substack{1 \leq k_\beta \leq r_\beta \\ \beta \in S(\alpha)}} C^\alpha_{k_\alpha, (k_\beta)_{\beta \in S(\alpha)}} \prod_{\beta \in S(\alpha)} z^\beta_{k_\beta}$$

for $z^\beta \in \mathbb{R}^{r_\beta}$. Therefore, g^α admits the representation

$$g^\alpha(x_\alpha) = f^\alpha((g^\beta(x_\beta)_{\beta \in S(\alpha)}). \tag{4}$$

For a leaf node $\alpha \in \mathcal{L}(T)$, the tensor C^α can be linearly identified with a linear map $f^\alpha : \mathbb{R}^{n_\alpha} \to \mathbb{R}^{r_\alpha}$, and

$$g^\alpha(x_\alpha) = f^\alpha(\phi^\alpha(x_\alpha)). \tag{5}$$

Therefore, a function f in $M_r^T(\mathcal{H}^s)$ can be parametrized by a tree network of linear or multilinear maps $\boldsymbol{f} = (f^\alpha)_{\alpha \in T^\star}$ (identified with the tree tensor network $(C^\alpha)_{\alpha \in T^\star}$).

We denote by F^α the space of linear maps from $\mathbb{R}^{r_D}$ to $\mathbb{R}^s$ for $\alpha = \emptyset$, the space of multilinear maps from $\bigtimes_{\beta \in S(\alpha)} \mathbb{R}^{r_\beta}$ to $\mathbb{R}^{r_\alpha}$ for $\alpha \in \mathcal{I}(T)$, or the space of linear maps from $\mathbb{R}^{n_\alpha}$ to $\mathbb{R}^{r_\alpha}$ for a leaf node $\alpha \in \mathcal{L}(T)$. We denote by

$$F_{T,r} := \bigtimes_{\alpha \in T^\star} F^\alpha$$

the parameter space and by $\mathcal{R}_{\mathcal{H},T,r}$ the representation map which associates to a network $\boldsymbol{f} = (f^\alpha)_{\alpha \in T^\star} \in F_{T,r}$ the function f. Then

$$M_r^T(\mathcal{H}^s) = \{\mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f}) : \boldsymbol{f} \in F_{T,r}\}.$$

Since F^α is linearly identified with $\mathbb{R}^{K^\alpha}$ for all $\alpha \in T^\star$, we deduce the following property from Lemma 2.2.

Lemma 2.4. *The map $\mathcal{R}_{\mathcal{H},T,r}$ is a multilinear map from the product space $F_{T,r} = \bigtimes_{\alpha \in T^\star} F^\alpha$ to the space of functions defined on $\mathcal{X}$.*

2.4 Representation complexity

When interpreting a tensor (or function) network $\boldsymbol{f} \in F_{T,r}$ as a neural network, a classical measure of complexity if the number of neurons, which is the sum of ranks r_α, $\alpha \in T^\star$. This leads to a first measure of complexity of a function $f = \mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f})$ defined by

$$\mathrm{compl}_N(\boldsymbol{f}) = \sum_{\alpha \in T^\star} r_\alpha.$$

From an approximation or statistical perspective, a more natural measure of complexity for a function $f \in M_r^T(\mathcal{H}^s)$ is its representation complexity, that is the dimension of the corresponding parameter space $F_{T,r}$, or the number of weights of the corresponding sum-product neural network. We let $N_\alpha = \dim(F^\alpha)$, with $N_\alpha = s r_D$ for $\alpha = \emptyset$, $N_\alpha = r_\alpha n_\alpha$ for $\alpha \in \mathcal{L}(T)$ and $N_\alpha = r_\alpha \prod_{\beta \in S(\alpha)} r_\beta$ for $\alpha \in T^\star \setminus \mathcal{L}(T)$. Then the representation complexity of a function $f = \mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f})$ is

$$\mathrm{compl}_C(\boldsymbol{f}) := C(T, r, \mathcal{H}^s) = \sum_{\alpha \in T^\star} N_\alpha = s r_D + \sum_{\alpha \in \mathcal{I}(T)} r_\alpha \prod_{\beta \in S(\alpha)} r_\beta + \sum_{\alpha \in \mathcal{L}(T)} r_\alpha n_\alpha. \tag{6}$$

Remark 2.5. *If $r_D = s$, the function $f^{\emptyset} : \mathbb{R}^s \to \mathbb{R}^s$ can be taken as the identity map, so that the parameters of $M_r^T(\mathcal{H}^s)$ are reduced to the set of functions $\boldsymbol{f} = (f^\alpha)_{\alpha \in T}$. In this case, we let $T^\star = T$, and the complexity is*

$$\mathrm{compl}_C(\boldsymbol{f}) := C(T, r, \mathcal{H}^s) = \sum_{\alpha \in T} N_\alpha = \sum_{\alpha \in \mathcal{I}(T)} r_\alpha \prod_{\beta \in S(\alpha)} r_\beta + \sum_{\alpha \in \mathcal{L}(T)} r_\alpha n_\alpha. \tag{7}$$

Another measure of complexity of $f = \mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f})$ can be defined as

$$\mathrm{compl}_S(\boldsymbol{f}) = \sum_{\alpha \in T^\star} \|f^\alpha\|_{\ell^0}, \tag{8}$$

where $\|f^\alpha\|_{\ell^0}$ is the number of non-zero entries in the tensor C^α associated with the multilinear map f^α. This measure of complexity takes into account a possible sparsity in tensors or in the corresponding sum-product neural network. We note that $\mathrm{compl}_S(\boldsymbol{f}) \leq \mathrm{compl}_C(\boldsymbol{f})$. These different measures of complexity lead to the definition of different approximation tools and corresponding approximation classes, see [2, 3] for tensor networks, and [19] for similar results on ReLU or RePU neural networks.

2.5 Normalized parametrization

A function $f \in M_r^T(\mathcal{H}^s)$ admits infinitely many equivalent parametrizations. From the multilinearity of the representation map $\mathcal{R}_{\mathcal{H},T,r}$ (see Lemma 2.4), it is clear that the model class $M_r^T(\mathcal{H}^s)$ is a cone, i.e. $cM_r^T(\mathcal{H}^s) \subset M_r^T(\mathcal{H}^s)$ for any $c \in \mathbb{R}$, and that given some norms $\|\cdot\|_{F^\alpha}$ on the spaces F^α, $\alpha \in T^\star$, we have

$$M_r^T(\mathcal{H}^s) = \{cf : c \in \mathbb{R}, f \in M_r^T(\mathcal{H}^s)_1\},$$

where $M_r^T(\mathcal{H}^s)_1$ are elements of $M_r^T(\mathcal{H}^s)$ with bounded parameters, defined by

$$M_r^T(\mathcal{H}^s)_1 = \{f = \mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f}) : \boldsymbol{f} = (f^\alpha)_{\alpha \in T^\star} \in F_{T,r} :, \|f^\alpha\|_{F^\alpha} \leq 1, \alpha \in T^\star\}. \tag{9}$$

3 Metric entropy of tree tensor networks

We assume that the sets $\mathcal{X}_\nu$ are equipped with finite measures μ_ν, for all $\nu \in D = \{1, \ldots, d\}$, and the set $\mathcal{X}$ is equipped with the product measure $\mu = \mu_1 \otimes \ldots \otimes \mu_d$. For $1 \leq p \leq \infty$, we consider the space $L^p_\mu(\mathcal{X}; \mathbb{R}^s)$ of measurable functions defined on $\mathcal{X}$ with values in $\mathbb{R}^s$, with bounded norm $\|\cdot\|_{p,\mu}$ defined by

$$\|f\|_{p,\mu}^p = \int_{\mathcal{X}} \|f(x)\|_p^p d\mu(x) \quad \text{for } 1 \leq p < \infty, \quad \text{or} \quad \|f\|_{\infty,\mu} = \mu\text{-}\operatorname*{ess\,sup}_{\mathcal{X}} |f|.$$

We also consider the space $L^\infty(\mathcal{X}; \mathbb{R}^s)$ of functions defined on $\mathcal{X}$ with values in $\mathbb{R}^s$, with bounded norm

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

In the following, we denote by $L^\lambda(\mathcal{X}; \mathbb{R}^s)$ the space $L^p_\mu(\mathcal{X}; \mathbb{R}^s)$ equipped with the norm $\|\cdot\|_{p,\mu}$ when $\lambda = (p, \mu)$ or the space $L^\infty(\mathcal{X}; \mathbb{R}^s)$ equipped with the norm $\|\cdot\|_\infty$ when $\lambda = \infty$. If $\mathcal{H}_\nu \subset L^\lambda(\mathcal{X}_\nu)$ for all $\nu \in D$, then $\mathcal{H} \subset L^\lambda(\mathcal{X})$ and $\mathcal{H}^s \subset L^\lambda(\mathcal{X}; \mathbb{R}^s)$.

3.1 Continuity of the parametrization

We here study the continuity properties of the representation map $\mathcal{R}_{\mathcal{H},T,r}$ as a map from $F_{T,r} = \times_{\alpha \in T^\star} F^\alpha$ to $\mathcal{H}^s \subset L^\lambda(\mathcal{X}; \mathbb{R}^s)$, with $\lambda = (p, \mu)$ or $\lambda = \infty$. We consider norms $\|\cdot\|_{F^\alpha}$ on space F^α, $\alpha \in T^\star$, and the product norm $\|\cdot\|_F$ over $F_{T,r}$ defined by

$$\|(f^\alpha)_{\alpha \in T^\star}\|_{F_{T,r}} = \max_{\alpha \in T^\star} \|f^\alpha\|_{F^\alpha}.$$

From the multilinearity of $\mathcal{R}_{\mathcal{H},T,r}$ (Lemma 2.4), we easily deduce the following property.

Lemma 3.1. *Assuming $\mathcal{H} \subset L^\lambda(\mathcal{X})$, with either $\lambda = (p,\mu)$ or $\lambda = \infty$, the multilinear map $\mathcal{R}_{\mathcal{H},T,r}$ from $F_{T,r}$ to $\mathcal{H}^s \subset L^\lambda(\mathcal{X};\mathbb{R}^s)$ is continuous and such that for all $f = \mathcal{R}_{\mathcal{H},T,r}((f^\alpha)_{\alpha\in T^\star})$ in $M_r^T(\mathcal{H}^s)$,*

$$\|f\|_\lambda \leq L_\lambda \prod_{\alpha\in T^\star} \|f^\alpha\|_{F^\alpha}$$

for some constant $L_\lambda < \infty$ independent of f defined by

$$L_\lambda = \sup_{f=\mathcal{R}_{\mathcal{H},T,r}((f^\alpha)_{\alpha\in T^\star})} \frac{\|f\|_\lambda}{\prod_{\alpha\in T^\star}\|f^\alpha\|_{F^\alpha}}. \tag{10}$$

We denote by $B(F^\alpha)$ the unit ball of F^α and by $B(F_{T,r})$ the unit ball of F. The set $M_r^T(\mathcal{H}^s)_1$ defined by (9) is such that

$$M_r^T(\mathcal{H}^s)_1 = \mathcal{R}_{\mathcal{H},T,r}(B(F_{T,r})). \tag{11}$$

We then deduce that the map $\mathcal{R}_{\mathcal{H},T,r}$ is Lipschitz continuous on the set $M_r^T(\mathcal{H}^s)_1$.

Lemma 3.2. *Assuming $\mathcal{H} \subset L^\lambda(\mathcal{X})$, with either $\lambda = (p,\mu)$ or $\lambda = \infty$, for all $f = \mathcal{R}_{\mathcal{H},T,r}(\boldsymbol{f})$ and $\tilde{f} = \mathcal{R}_{\mathcal{H},T,r}(\tilde{\boldsymbol{f}})$ in $M_r^T(\mathcal{H}^s)_1$,*

$$\|f-\tilde{f}\|_\lambda \leq L_\lambda \sum_{\alpha\in T^\star} \|f^\alpha - \tilde{f}^\alpha\|_{F^\alpha} \leq L_\lambda |T^\star| \|\boldsymbol{f}-\tilde{\boldsymbol{f}}\|_{F_{T,r}}.$$

Proof. Denoting by $\{\alpha_1,\dots,\alpha_K\}$ the elements of $T^\star$, we have $f - \tilde{f} = \sum_{k=1}^K \mathcal{R}_{\mathcal{H},T,r}(\tilde{f}^{\alpha_1},\cdots,f^{\alpha_k} - \tilde{f}^{\alpha_k},\cdots,f^{\alpha_K})$. Them from Lemma 3.1, we obtain

$$\|f-\tilde{f}\|_\lambda \leq L_\lambda \sum_{k=1}^K \|f^{\alpha_k}-\tilde{f}^{\alpha_k}\|_{F^{\alpha_k}} \prod_{i<k}\|\tilde{f}^{\alpha_i}\|_{F^{\alpha_i}} \prod_{i>k}\|f^{\alpha_i}\|_{F^{\alpha_i}}, \tag{12}$$

and we conclude by noting that $\|f^\alpha\|_{F^\alpha} \leq 1$ and $\|\tilde{f}^\alpha\|_{F^\alpha} \leq 1$ for all $\alpha \in T^\star$. □

3.2 Metric entropy

The metric entropy $H(\epsilon, K, \|\cdot\|_X)$ of a compact subset K of a normed vector space $(X, \|\cdot\|_X)$ is defined as

$$H(\epsilon, K, \|\cdot\|_X) = \log N(\epsilon, K, \|\cdot\|_X),$$

with $N(\epsilon, K, \|\cdot\|_X)$ the covering number of K, which is the minimal number of balls of radius ϵ (for $\|\cdot\|_X$) necessary to cover K. We have the following result on the metric entropy of tensor networks with bounded parameters.

Proposition 3.3. *Assuming that $\mathcal{H} \subset L^\lambda(\mathcal{X})$, with either $\lambda = \infty$ or $\lambda = (p,\mu)$, $1 \leq p \leq \infty$, the metric entropy of the model class*

$$M_r^T(\mathcal{H}^s)_R = \{cf : c \in \mathbb{R}, |c| \leq R, f \in M_r^T(\mathcal{H}^s)_1\} \tag{13}$$

in $L^\lambda(\mathcal{X};\mathbb{R}^s)$ is such that

$$H(\epsilon, M_r^T(\mathcal{H}^s)_R, \|\cdot\|_\lambda) \leq C(T,r,\mathcal{H}^s)\log(3\epsilon^{-1}RL_\lambda|T^\star|).$$

Proof. The covering number of the unit ball $B(F^\alpha)$ of the N_α-dimensional space F^α is such that $N(\epsilon, B(F^\alpha), \|\cdot\|_{F^\alpha}) \leq (3\epsilon^{-1})^{N_\alpha}$. Then the unit ball $B(F_{T,r})$ of the product space $F_{T,r}$ equipped with the product topology has a covering number $N(\epsilon, B(F_{T,r}), \|\cdot\|_{F_{T,r}}) \leq \prod_{\alpha\in T^\star} N(\epsilon, B(F^\alpha), \|\cdot\|_{F^\alpha}) \leq (3\epsilon^{-1})^{C(T,r,\mathcal{H}^s)}$ with $C(T,r,\mathcal{H}^s) = \sum_{\alpha\in T^\star} N_\alpha$. From the Lipschitz continuity of $\mathcal{R}_{\mathcal{H},T,r}$ on $M_r^T(\mathcal{H}^s)_1$ (Lemma 3.2), we deduce that $N(\epsilon, M_r^T(\mathcal{H}^s)_1, \|\cdot\|_\lambda) \leq (3\epsilon^{-1}L_\lambda|T^\star|)^{C(T,r,\mathcal{H}^s)}$, from which we deduce that $N(\epsilon, M_r^T(\mathcal{H}^s)_R, \|\cdot\|_\lambda) \leq (3\epsilon^{-1}RL_\lambda|T^\star|)^{C(T,r,\mathcal{H}^s)}$, which ends the proof. □

If f_1 and f_2 are two functions from $L^p_\mu(\mathcal{X};\mathbb{R}^s)$, the collection of functions $f \in L^p_\mu(\mathcal{X};\mathbb{R}^s)$ such that $f_1 \leq f \leq f_2$ almost everywhere is denoted by $[f_1, f_2]$ and called a bracket with extremities f_1 and f_2. The diameter of the bracket $[f_1, f_2]$ for the norm $\|\cdot\|_{p,\mu}$ is given by $\|f_2 - f_1\|_{p,\mu}$. The bracketing number $N_{[]}(\epsilon, K, \|\cdot\|_{p,\mu})$ of a set K is defined as the minimal number of brackets with diameters less than ϵ which are necessary to cover K. The corresponding bracketing entropy is defined as

$$H_{[]}(\epsilon, K, \|\cdot\|_{p,\mu}) := \log N_{[]}(\epsilon, K, \|\cdot\|_{p,\mu}).$$

Lemma 3.4. *For any $1 \leq p \leq \infty$ and any compact set K in $L^p(\mathcal{X};\mathbb{R}^s)$,*

$$H_{[]}(\epsilon, K, \|\cdot\|_{p,\mu}) \leq H(\frac{\epsilon}{2}\mu(\mathcal{X})^{-1/p}, K, \|\cdot\|_{\infty,\mu}),$$

where $\mu(\mathcal{X})$ is the mass of the measure μ, and $\mu(\mathcal{X})^{-1/p} = 1$ for $p = \infty$.

Proof. Let $\gamma = \frac{\epsilon}{2}\mu(\mathcal{X})^{-1/p}$ and let $\mathcal{N}$ be a γ-net of K for the norm $\|\cdot\|_{\infty,\mu}$ with cardinal $N(\gamma, K, \|\cdot\|_{\infty,\mu})$. Then for any $f \in K$, there exists a $\tilde{f} \in \mathcal{N}$ such that $\|f - \tilde{f}\|_{\infty,\mu} \leq \gamma$, which implies that f is in the bracket $[\tilde{f}-\gamma, \tilde{f}+\gamma]$ with diameter $\|2\gamma\|_{p,\mu} = 2\gamma\mu(\mathcal{X})^{1/p} = \epsilon$. Then the collection of brackets $\{[\tilde{f}-\gamma, \tilde{f}+\gamma] : \tilde{f} \in \mathcal{N}\}$ with diameters ϵ covers K, which implies $N_{[]}(\epsilon, K, \|\cdot\|_{p,\mu}) \leq N(\gamma, K, \|\cdot\|_{\infty,\mu})$, which ends the proof. □

From Proposition 3.3 and Lemma 3.4, we directly deduce the following result.

Proposition 3.5. *For any $1 \leq p \leq \infty$, the set $M_r^T(\mathcal{H}^s)_R$ defined in (13) has a bracketing entropy*

$$H_{[]}(\epsilon, M_r^T(\mathcal{H}^s)_R, \|\cdot\|_{p,\mu}) \leq C(T, r, \mathcal{H}^s)\log(6\epsilon^{-1}\mu(\mathcal{X})^{1/p}RL_{\infty,\mu}|T^\star|),$$

with $\mu(\mathcal{X})^{-1/p} = 1$ for $p = \infty$.

3.3 A particular choice of norms

Assume that $\mathcal{H} \subset L^\lambda(\mathcal{X})$, with either $\lambda = (p, \mu)$ or $\lambda = \infty$. The continuity constant L_λ of the map $\mathcal{R}_{\mathcal{H},T,r}$ defined by (10) depends on λ, the norms on F^α, the chosen basis for $\mathcal{H}$ and also on the measure μ when $\lambda = (p, \mu)$. We here introduce a particular choice of norms and basis functions which allows to bound the continuity constant L_λ. We consider on the space $F^\emptyset$ of linear maps from $\mathbb{R}^{r_D}$ to $\mathbb{R}^s$ the norm (with $p = \infty$ when $\lambda = \infty$)

$$\|f^\emptyset\|_{F^\emptyset} = \max_{z \in \mathbb{R}^{r_D}} \frac{\|f^\emptyset(z)\|_p}{\|z\|_p},$$

which coincides with the classical matrix p-norm. For any interior node $\alpha \in \mathcal{I}(T)$, we introduce a norm $\|\cdot\|_{F^\alpha}$ over the space F^α of multilinear maps $f^\alpha : \bigtimes_{\beta \in S(\alpha)} \mathbb{R}^{r_\beta} \to \mathbb{R}^{r_\alpha}$, defined by

$$\|f^\alpha\|_{F^\alpha} = \max_{(z_\beta)_{\beta \in S(\alpha)} \in \bigtimes_{\beta \in S(\alpha)} \mathbb{R}^{r_\beta}} \frac{\|f^\alpha((z_\beta)_{\beta \in S(\alpha)})\|_p}{\prod_{\beta \in S(\alpha)} \|z_\beta\|_p}.$$

For a leaf node $\alpha \in \mathcal{L}(T)$, we introduce a norm $\|\cdot\|_{F^\alpha}$ over the space F^α of linear maps $f^\alpha : \mathbb{R}^{n_\alpha} \to \mathbb{R}^{r_\alpha}$, defined by

$$\|f^\alpha\|_{F^\alpha} = \max_{z_\alpha \in \mathbb{R}^{n_\alpha}} \frac{\|f^\alpha(z_\alpha)\|_p}{\|z_\alpha\|_p}. \tag{14}$$

We assume that for any $\nu \in D$, the feature map $\phi^\nu : \mathcal{X}_\nu \to \mathbb{R}^{n_\nu}$ is such that $\|\phi^\nu\|_\lambda = 1$. For $\lambda = (\infty, \mu)$ (resp. $\lambda = \infty$), that means that basis functions $\phi^\nu_{i_\nu}(x_\nu)$ have a unit norm in $L^{\infty,\mu}(\mathcal{X}_\nu)$ (resp. $L^\infty(\mathcal{X}_\nu)$). For $p < \infty$, that means that $\sum_{i=1}^{n_\nu} \|\phi^\nu_i\|^p_{p,\mu} = 1$, which can be obtained by rescaling basis functions so that $\|\phi^\nu_i\|_{p,\mu} = n_\nu^{-1/p}$.

Proposition 3.6. *Assume $\mathcal{H} \subset L^\lambda(\mathcal{X})$, with either $\lambda = (p, \mu)$ or $\lambda = \infty$. With the above choice of norms and normalization of basis functions (with $p = \infty$ when $\lambda = \infty$), the continuity constant L_λ defined by (10) is such that $L_\lambda \leq 1$, and for all $1 \leq q \leq p$, $L_{q,\mu} \leq \mu(\mathcal{X})^{1/q-1/p}L_\lambda \leq \mu(\mathcal{X})^{1/q-1/p}$.*

Proof. See Appendix A.1. □

4 Risk bounds for empirical risk minimization

In this section, we analyze the estimation error for tree tensor networks obtained by empirical risk minimization. We consider as fixed the approximation space $\mathcal{H}$, the tree T and the rank $r \in \mathbb{N}^{|T|}$. We assume that $\mathcal{H} \subset L^{\infty,\mu}(\mathcal{X})$, with $\mathcal{X}$ equipped with a finite measure μ. We consider the model class $M_r^T(\mathcal{H}^s)_R := M$ of tree tensor networks with bounded parameters, with the norms defined in Section 3.3 for $\lambda = (\infty, \mu)$ ($p = \infty$). We denote by $C_M = C(T, r, \mathcal{H}^s)$ the representation complexity of M defined by (6) (or (7) when $r_D = s$). We consider a risk

$$\mathcal{R}(f) = \mathbb{E}(\gamma(f, Z)),$$

where Z is a random variable taking values in $\mathcal{Z}$ and where $\gamma : \mathbb{R}^{\mathcal{X}} \times \mathcal{Z} \to \mathbb{R}$ is some contrast function. The minimizer of the risk over measurable functions defined on $\mathcal{X}$ is the target function $f^\star$. For f random (depending on the data), $\mathbb{E}(\gamma(f, Z))$ shall be understood as an expectation $\mathbb{E}_Z(\gamma(f, Z))$ w.r.t. Z (conditional to the data).

Example 4.1. *For supervised learning, we consider a random variable $Z = (X, Y)$, with Y a random variable with values in $\mathbb{R}^s$, X a $\mathcal{X}$-valued random variable with probability law μ. The contrast is chosen as $\gamma(f, (x, y)) = \ell(y, f(x))$ with ℓ a loss function measuring a discrepancy between y and the prediction $f(x)$.*

Example 4.2. *For the problem of estimating the probability distribution of a random variable X, we consider $Z = X$ and $s = 1$.*

Given the model class M, we denote by f^M a minimizer over M of the risk $\mathcal{R}$, and by $\hat{f}_n^M$ a minimizer over M of the empirical risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^n \gamma(f, Z_i),$$

which is seen as an empirical process over M. We introduce the excess risk

$$\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}(f^\star).$$

The excess risk for the estimator $\hat{f}_n^M$ satisfies

$$\mathcal{E}(\hat{f}_n^M) = \mathcal{E}(f^M) + \mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M), \tag{15}$$

where $\mathcal{E}(f^M)$ is the best approximation error in M and $\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M)$ is the estimation error. Using the optimality of $\hat{f}_n^M$, we obtain that the estimation error satisfies

$$\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M) \leq \widehat{\mathcal{R}}_n(f^M) - \mathcal{R}(f^M) - \widehat{\mathcal{R}}_n(\hat{f}_n^M) + \mathcal{R}(\hat{f}_n^M) := \bar{\mathcal{R}}_n(f^M) - \bar{\mathcal{R}}_n(\hat{f}_n^M), \tag{16}$$

where $\bar{\mathcal{R}}_n(f)$ is the centered empirical process

$$\bar{\mathcal{R}}_n(f) = \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) = \frac{1}{n} \sum_{i=1}^n \gamma(f, Z_i) - \mathbb{E}(\gamma(f, Z)). \tag{17}$$

To obtain bounds of the estimation error, it remains to quantify the fluctuations of the centered empirical process $\bar{\mathcal{R}}_n(f)$.

4.1 Concentration inequalities for empirical processes

We here apply classical results to control the fluctuations of the supremum of the empirical process $\bar{\mathcal{R}}_n(f)$ over the model class M.

Assumption 4.3 (Bounded contrast)**.** *Assume that γ is uniformly bounded over $M \times \mathcal{Z}$, i.e.*

$$|\gamma(f, Z)| \leq B \tag{18}$$

holds almost surely for all $f \in M$, with B a constant independent of f.

The above assumption yields a classical concentration inequality for the empirical process $\bar{\mathcal{R}}_n(f)$.

Lemma 4.4. *Under assumption 4.3, we have that*

$$\mathbb{P}(\bar{\mathcal{R}}_n(f) > \epsilon B) \vee \mathbb{P}(\bar{\mathcal{R}}_n(f) < -\epsilon B) \leq e^{-n\frac{\epsilon^2}{2}} \tag{19}$$

holds for all $f \in M$.

Proof. We have $\widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) = \frac{1}{n}\sum_{i=1}^n A_i^f - \mathbb{E}(A^f)$, where the $A_i^f = \gamma(f, Z_i)$ are i.i.d. copies of the random variable $A^f = \gamma(f, Z)$. From Assumption 4.3, we have that $|A^f| \leq B$ almost surely, so that A^f is subgaussian with parameter B^2 and the result simply follows from Hoeffding's inequality. □

A stronger assumption is required to obtain a uniform concentration inequality for the empirical process $\bar{\mathcal{R}}_n(f)$ over M.

Assumption 4.5. *Assume that $\gamma(\cdot, Z)$ is Lipschitz continuous over $M \subset L^{\infty,\mu}(\mathcal{X}; \mathbb{R}^s)$, i.e.*

$$|\gamma(f, Z) - \gamma(g, Z)| \leq \mathcal{L}\|f - g\|_{\infty,\mu} \tag{20}$$

holds almost surely for all $f, g \in M$, with $\mathcal{L}$ a constant independent of f and g.

Lemma 4.6. *Under Assumptions 4.3 and 4.5, we have that*

$$\mathbb{P}(\sup_{f\in M} \bar{\mathcal{R}}_n(f) > 2\epsilon B) \vee \mathbb{P}(\inf_{f\in M} \bar{\mathcal{R}}_n(f) < -2\epsilon B) \leq N_{\frac{\epsilon B}{2\mathcal{L}}} e^{-\frac{n\epsilon^2}{2}}, \tag{21}$$

where $N_{\frac{\epsilon B}{2\mathcal{L}}} = N(\frac{\epsilon B}{2\mathcal{L}}, M, \|\cdot\|_{\infty,\mu})$ is the covering number of M at scale $\frac{\epsilon B}{2\mathcal{L}}$, and

$$\log N_{\frac{\epsilon B}{2\mathcal{L}}} \leq C_M \log\left(6\mathcal{L}B^{-1}R|T^\star|\epsilon^{-1}\right).$$

Proof. See Appendix A.2. □

Lemma 4.7. *Under Assumptions 4.3 and 4.5,*

$$\mathbb{E}(\sup_{f\in M} |\bar{\mathcal{R}}_n(f)|) \leq 4B\sqrt{C_M}\sqrt{\frac{2\log((\beta \vee e)\sqrt{n})}{n}}.$$

with $\beta = 6\mathcal{L}B^{-1}R|T^\star|$.

Proof. See Appendix A.2. □

4.2 Risk bounds for the minimizer of the empirical risk

From the properties of the centered empirical process, we can now derive upper bounds of the estimation error in probability and in expectation.

Proposition 4.8. *Under Assumptions 4.3 and 4.5, the estimation error satisfies*

$$\mathbb{P}(\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M) > 2\epsilon B) \leq 2e^{C_M \log(\beta\epsilon^{-1}) - \frac{n\epsilon^2}{2}},$$

where $\beta = 6\mathcal{L}B^{-1}R|T^\star|$. Moreover,

$$\mathbb{E}(\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M)) \leq 4B\sqrt{C_M}\sqrt{\frac{2\log((\beta \vee e)\sqrt{n})}{n}},$$

and thus

$$\mathbb{E}(\mathcal{E}(\hat{f}_n^M)) \leq \mathcal{E}(f^M) + 4B\sqrt{C_M}\sqrt{\frac{2\log((\beta \vee e)\sqrt{n})}{n}}.$$

Proof. See Appendix A.2. □

Proposition 4.9. *Under Assumptions 4.3 and 4.5, for any $t > 0$, with probability larger than $1-\exp(-t)$,*

$$\sup_{f\in M} -\bar{\mathcal{R}}_n(f) \leq 4B\sqrt{C_M}\sqrt{\frac{2\log(6\mathcal{L}B^{-1}R|T^\star|\sqrt{n})}{n}} + 2B\sqrt{\frac{t}{2n}}. \tag{22}$$

Moreover, with probability larger than $1-\exp(-t)$,

$$\mathcal{E}(\hat{f}_n^M) \leq \mathcal{E}(f^M) + 8B\sqrt{C_M}\sqrt{\frac{2\log(6\mathcal{L}B^{-1}R|T^\star|\sqrt{n})}{n}} + 4B\sqrt{\frac{t}{2n}}. \tag{23}$$

Proof. See Appendix A.2. □

Example 4.10 (Least-squares bounded regression)**.** *We consider the least-squares regression setting with $\gamma(f,Z) = \|Y-f(X)\|_{\ell^2}^2$. Let μ be the distribution of X. The excess risk $\mathcal{E}(f) = \mathcal{R}(f)-\mathcal{R}(f^\star) = \|f-f^\star\|_{2,\mu}^2$ admits $f^\star(x) = \mathbb{E}(Y|X=x)$ as a minimizer. We assume that $\|Y\|_{\ell^\infty} \leq R$ almost surely. For all $f \in M$, we have $\gamma(f,Z) \leq s\|Y-f(X)\|_{\ell^\infty}^2 \leq 2s(\|Y\|_{\ell^\infty}^2 + \|f\|_\infty^2)$, so that $0 \leq \gamma(f,Z) \leq B$ almost surely, with $B = 4sR^2$. Also, it holds almost surely*

$$\begin{aligned}
|\gamma(f,Z)-\gamma(g,Z)| &= |(2Y-f(X)-g(X), f(X)-g(X))_{\ell^2}| \\
&\leq \|2Y-f(X)-g(X)\|_{\ell^1}\|f(X)-g(X)\|_{\ell^\infty} \\
&\leq s(2\|Y\|_{\ell^\infty} + \|g\|_{\infty,\mu} + \|f\|_{\infty,\mu})\|f-g\|_{\infty,\mu}.
\end{aligned}$$

Then for all $f,g \in M$, $|\gamma(f,Z)-\gamma(g,Z)| \leq \mathcal{L}\|f-g\|_{\infty,\mu}$ with $\mathcal{L} = 4sR$. The constant β from Proposition 4.8 is $\beta = 6|T^\star|$.

Example 4.11 (L^2 density estimation)**.** *We consider the estimation of the probability law ν of X. Assuming that ν admits a density $f^\star$ with respect to the measure μ, and assuming $f^\star \in L^2_\mu(\mathcal{X})$, we consider the contrast $\gamma(f,x) = \|f\|_{2,\mu}^2 - 2f(x)$, so that $\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}(f^\star) = \|f-f^\star\|_{2,\mu}^2$ admits $f^\star$ as a minimizer. We assume that μ is a finite measure on $\mathcal{X}$ and that $f^\star$ is uniformly bounded by R. Then $|\gamma(f,X)| \leq B$ almost surely with $B = R(\mu(\mathcal{X})R+2)$. Also, for all $f,g \in M$, we have almost surely*

$$\begin{aligned}
|\gamma(f,X)-\gamma(g,X)| &= |\|f\|_{2,\mu}^2 - \|g\|_{2,\mu}^2 - 2(f(X)-g(X))| \\
&\leq |\int (f-g)(f+g)d\mu| + 2\|f-g\|_{\infty,\mu} \\
&\leq (\|f+g\|_{1,\mu}+2)\|f-g\|_{\infty,\mu} \\
&\leq \mathcal{L}\|f-g\|_{\infty,\mu}
\end{aligned}$$

with $\mathcal{L} = 2(\mu(\mathcal{X})R+1)$. Since $1/R \leq \mathcal{L}/B \leq 2/R$, the constant β from Proposition 4.8 is such that $6|T^\star| \leq \beta \leq 12|T^\star|$.

4.3 Improved risk bounds for least squares contrasts

In this section, we provide an improved excess risk bound in the specific case of least squares contrasts. Our results come from Talagrand Inequalities and generic chaining bounds ; we follow the presentation given in the book of [27]. The excess risk bound given below strongly relies on the link between the excess risk and the variance of the excess loss (see Inequality (**??**) in the proof of Proposition 4.12), as explained in Chapter 5 of [27] and Chapter 8 in [28].

Let γ be either the least squares contrast in the bounded regression setting (as described in Example 4.10, with $s=1$), or the least squares contrast for density estimation (as described in Example 4.11). In particular, note that in the regression setting it is assumed that $\|Y\|_{\ell^\infty} \leq R$ almost surely.

As before, we consider the model class $M = M_r^T(\mathcal{H})_R$ of tree tensor networks with bounded parameters. Contrary to the two previous subsections, it is now assumed that $\mathcal{H} \subset L^\infty(\mathcal{X})$ equipped with the norm $\|\cdot\|_\infty$ and we still use the normalization of the parameters with $\lambda = \infty$ ($p = \infty$) introduced in Section 3.3. Note that $L^\infty(\mathcal{X}) \subset L^{\infty,\mu}(\mathcal{X})$, where μ is the distribution of the X_i's in the regression setting (see Example 4.10) or the reference measure for density estimation (see Example 4.11). In particular, in this setting $\|f\|_{\infty,\mu} \leq \|f\|_\infty < \infty$ for any $f \in \mathcal{H}$.

Proposition 4.12. *Under the previous assumptions, there exists an absolute constant $\mathcal{A}$ and a constant κ such that for any $\varepsilon \in (0,1]$ and any $t > 0$, with probability at least $1 - \mathcal{A}\exp(-t)$, it holds*

$$\mathcal{E}(\hat{f}_n^M) \leq (1+\varepsilon)\mathcal{E}(f^M) + \frac{\kappa R^2}{n}\left[\frac{a_T C_M}{\varepsilon^2}\log^+\left(\frac{n\varepsilon^2}{a_T C_M}\right) + \frac{t}{\varepsilon}\right] \tag{24}$$

where $a_T = 1 + \log^+\left(\frac{3|T^\star|}{4e}\right)$, and κ depends on linearly on $\mu(\mathcal{X})$[1]. Then by integrating according to t, we obtain that for any $\varepsilon \in (0,1]$,

$$\mathbb{E}\mathcal{E}(\hat{f}_n^M) \leq (1+\varepsilon)\mathcal{E}(f^M) + \frac{\kappa R^2}{n}\left[\frac{a_T C_M}{\varepsilon^2}\log^+\left(\frac{n\varepsilon^2}{a_T C_M}\right) + \frac{\mathcal{A}}{\varepsilon}\right].$$

Proof. See Appendix A.2.1. □

Note that the term a_T is upper bounded by a term of the order of $\log d$ because $|T^*| \leq 2d$. Thus the constants in the risk bound (24) does not explode with the dimension d in regression. Note however that in density estimation, the constant κ depends linearly on the mass $\mu(\mathcal{X})$ of the reference measure, which may grow exponentially with d.

5 Model selection for tree tensor networks

We now consider a family of approximation spaces $\mathcal{H}_m = \mathcal{H}_m^s \subset L^{\infty,\mu}(\mathcal{X})$, $m \in \mathcal{M}$, with $\mathcal{X}$ equipped with a finite measure μ, as in Sections 4.1 and 4.2. Let $(M_m)_{m\in\mathcal{M}}$ be a given family of tree tensor networks with $M_m = M_{r_m}^{T_m}(\mathcal{H}_m)_R$ and where the parameters are bounded according to the norms defined in Section 3.3 for $\lambda = (\infty, \mu)$ ($p = \infty$). Each model m has a particular tree T_m, a rank r_m, an approximation space $\mathcal{H}_m$, and a radius R. We denote by $C_m = C(T_m, r_m, \mathcal{H}_m)$ the corresponding representation complexity. For some $m \in \mathcal{M}$, we let f_m be a minimizer of the risk over M_m,

$$f_m \in \arg\min_{f\in M_m} \mathcal{R}(f),$$

and $\hat{f}_m$ be a minimizer of the empirical risk over M_m,

$$\hat{f}_m \in \arg\min_{f\in M_m} \widehat{\mathcal{R}}_n(f).$$

At this stage of the procedure, we have at hand a family of predictors $\hat{f}_m$ and our goal is to provide a strategy for selecting a good predictor. To this aim, we make use of the model selection approach of Barron, Birgé and Massart. More precisely, we adapt a general theorem from [28] to our problem. Similar model selection strategies can be found in [34, 21, 11], see also [9] for an application to the selection of principal curves.

Given some penalty function $\text{pen} : \mathcal{M} \to \mathbb{R}^+$, we define $\hat{m}$ as the minimizer over $\mathcal{M}$ of the criterion

$$\text{crit}(m) := \widehat{\mathcal{R}}_n(\hat{f}_m) + \text{pen}(m), \tag{25}$$

and we finally select the predictor $\hat{f}_{\hat{m}}$.

[1] With $\mu(\mathcal{X}) = 1$ for regression.

Assumption 5.1. *We consider a family of positive weights $(x_m)_{m\in\mathcal{M}}$ over the family of models such that*

$$\Sigma = \sum_{m\in\mathcal{M}} \exp(-x_m) < \infty.$$

This assumption and the choice of the weights is the discussed further in Section 5.3.

5.1 A general model selection for tree tensor network

We follow a standard strategy that corresponds to the so-called *Vapnik's structural minimization of the risk method* (see for instance [28, Section 8.2]) to choose the penalty function and derive a risk bound for the estimator selected by the criterion (25). By definition of $\hat{m}$, for any $m \in \mathcal{M}$,

$$\mathcal{R}_n(\hat{f}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \mathcal{R}_n(\hat{f}_m) + \mathrm{pen}(m) \leq \mathcal{R}_n(f_m) + \mathrm{pen}(m).$$

Therefore,

$$\mathcal{R}_n(\hat{f}_{\hat{m}}) \leq \mathcal{R}_n(f_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m})$$

and thus

$$\mathcal{R}(\hat{f}_{\hat{m}}) + \bar{\mathcal{R}}_n(\hat{f}_{\hat{m}}) \leq \mathcal{R}(f_m) + \bar{\mathcal{R}}_n(f_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}),$$

where $\bar{\mathcal{R}}_n(f)$ is the centered empirical process defined in (17). We finally derive the following upper bound on the excess risk

$$\mathcal{E}(\hat{f}_{\hat{m}}) \leq \mathcal{E}(f_m) + \bar{\mathcal{R}}_n(f_m) - \bar{\mathcal{R}}_n(\hat{f}_{\hat{m}}) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \tag{26}$$

We now provide a risk bound for a model selection strategy based on the criterion (25) with a suitable choice of penalty.

Theorem 5.2. *Under Assumptions 4.3, 4.5 and 5.1, if the penalty is such that*

$$\mathrm{pen}(m) \geq \lambda_m \sqrt{\frac{C_m}{n}} + 2B\sqrt{\frac{x_m}{2n}}, \tag{27}$$

with

$$\lambda_m = 4B\sqrt{2\log(6\mathcal{L}B^{-1}R|T_m|^\star\sqrt{n})},$$

then the estimator $\hat{f}_{\hat{m}}$ selected according to the criterion (25) *satisfies the following risk bound*

$$\mathbb{E}(\mathcal{E}(\hat{f}_{\hat{m}})) \leq \inf_{m\in\mathcal{M}} \{\mathcal{E}(f_m) + \mathrm{pen}(m)\} + B\Sigma\sqrt{\frac{\pi}{2n}}.$$

Proof. See Appendix A.3. □

Theorem 5.2 gives a strong justification for using a penalty proportional to $\sqrt{C_m/n}$, at least for not too large family of models. However, it is known that the Vapnik's structural minimization of the risk may lead to suboptimal rates of convergence. For instance, in the bounded regression setting, it is known that a penalty proportional to the VapnikChervonenkis dimension (typically in $O(C_m/n)$ leads to minimax rates of convergence in various setting (see for instance Chapter 12 in [21]) whereas Vapnik's structural minimization of the risk (typically with penalty in $O(\sqrt{C_m/n})$) is too pessimistic to provide fast rates of convergence. Note that the approach of [21] is based on a truncation strategy which is not easy to calibrate in practice. In the next section, we give an improved model selection result for least squares inference.

5.2 Oracle inequalities for least squares inference on tree tensor networks

In this section, we give an improved model selection result for least squares inference based on Proposition 4.12. This corresponds to the approach presented in Sections 8.3 and 8.4 of [28] or in Section 6.3 of [27].

We consider least squares density estimation and least squares bounded regression ($s = 1$) in the same framework as Section 4.3: we now consider a family of approximation spaces $\mathcal{H}_m \subset L^\infty(\mathcal{X})$ with $s = 1$ and equipped with the norm $\|\cdot\|_\infty$. We use the same normalization of the parameters with $p = \infty$ ($\lambda = \infty$) as introduced in Section 3.3. As before we consider a family of tree tensor networks $(M_m)_{m\in\mathcal{M}}$ where each model $M_m = M_{r_m}^{T_m}(\mathcal{H}_m)_R$ has a particular tree T_m, a rank r_m, an approximation space $\mathcal{H}_m$, and a radius R.

Theorem 5.3. *In the setting of Proposition 4.12 and under Assumption 5.1, there exists numerical constants K_1 and K_2 and K_3 such that if the penalty satisfies*

$$\mathrm{pen}(m) = K_1 R^2 \left[\frac{a_m C_m}{n\varepsilon^2}\log\frac{n\varepsilon^2}{a_m C_m} + \frac{x_m}{n\varepsilon}\right]$$

with $a_m = 1+\log^+\left(\frac{3|T_m^\star|}{4e}\right)$, then the estimator $\hat{f}_{\hat{m}}$ selected according to the penalized criterion (25) *satisfies the following oracle inequality*

$$\mathbb{E}\mathcal{E}(\hat{f}_{\hat{m}}) \le \frac{1+\varepsilon}{1-\varepsilon}\inf_{m\in\mathcal{M}}\left\{\mathcal{E}(f_m) + K_2 R^2\left[\frac{a_m C_m}{n\varepsilon^2}\log\frac{n\varepsilon^2}{a_m C_m} + \frac{x_m}{n\varepsilon}\right]\right\} + \frac{K_3 R^2 \Sigma}{n}\frac{1+\varepsilon}{\varepsilon(1-\varepsilon)}. \tag{28}$$

Proof. The proof is adapted from Theorem 6.5 in [27], see Appendix A.3. □

This theorem provides an improved oracle inequality bound with a penalty in $\frac{C_m}{n}$, up to logarithmic terms. In Section 5.4, we will derive adaptive optimal rates of convergence (in the minimax sense) from this model selection result. In Section 6 we illustrate how to calibrate the penalty in practice using the slope heuristics method.

5.3 Choosing the weights in the penalty function

The weights x_m represent the price to pay for the richness of the model collection, when there are many models with the same complexity C_m. A typical choice for the weights is $x_m = x(C_m)$ with a weight function x such that

$$x(c) \ge \beta c + \log(N_c),$$

where $N_c = |\{m \in \mathcal{M} \,:\, C_m = c\}|$ is the number of models with complexity c, and β some positive constant. With such a choice, $\Sigma = \sum_{m\in\mathcal{M}}\exp(-x_m) = \sum_{c\ge 1} N_c \exp(-x(c)) \le (e^\beta - 1)^{-1}$, so that Assumption 5.1 is satisfied. With such a weight function, if the model collection is not too rich, the weight x_m is comparable to or smaller than the complexity C_m.

We restrict the following analysis to the case where the approximation space is fixed: $\mathcal{H}_m = \mathcal{H}^s$ for any $m \in \mathcal{M}$ and we only consider binary trees, for which $|T_m| = 2d - 1$.

We first assume that the binary tree T is fixed and we need to upper bound the number N_c of models having the complexity c to define the weights. According the definition of the representation complexity given in Section 2.4, a format with complexity c satisfies

$$c = \sum_{\alpha\in T^\star} N_\alpha = sr_D + \sum_{\alpha\in\mathcal{I}(T)} r_\alpha \prod_{\beta\in S(\alpha)} r_\beta + \sum_{\alpha\in\mathcal{L}(T)} r_\alpha n_\alpha. \tag{29}$$

The number of triplets of integers (k_1, k_2, k_3) such that the product $k_1k_2k_3$ is less than an integer q_α is clearly less than q_α^2. So, the number of formats such that $N_\alpha = q_\alpha$ for any $\alpha \in T^\star$ is less than

$$\prod_{\alpha\in T^\star} q_\alpha^2 \leq \left[\left(\prod_{\alpha\in T^\star} q_\alpha\right)^{1/|T^\star|}\right]^{2|T^\star|} \leq \left[\frac{1}{|T^\star|}\sum_{\alpha\in T^\star} q_\alpha\right]^{2|T^\star|} \leq \left[\frac{c}{|T^\star|}\right]^{2|T^\star|}.$$

Moreover, the number of tuple of integers $(q_\alpha)_{\alpha\in T^\star}$ satisfying $\sum_{\alpha\in T^\star} q_\alpha = c$ is $\binom{c+|T^\star|}{c}$. For a fixed binary tree, the number N_c of all possible formats of complexity c is thus such that

$$N_c \leq \binom{c+|T^\star|}{c}\left[\frac{c}{|T^\star|}\right]^{2|T^\star|}.$$

Using the inequality

$$\log\binom{k}{\ell} \leq \ell(1+\log\frac{k}{\ell}), \tag{30}$$

and the fact that $|T^\star| \leq C_m$ for any model m in the collection, we obtain

$$\log(N_c) \leq c(1+\log(\frac{c+|T^\star|}{c})) + 2|T^\star|\log(\frac{c}{|T^\star|}) \leq c(1+\log(2)) + 4d\log(c) \lesssim c.$$

Then for a given binary tree T, we finally take a weight function

$$x(c) = \eta c \tag{31}$$

for some $\eta > 0$. In the situation where all the formats of the collection rely on a same tree T, using the weight function given in (31), Theorem 5.3 shows that we can use a penalty proportional to C_m.

Leaving aside the computational aspects for the moment (see Section 5.6), we now consider the situation where the formats of the collection rely on several possible trees T. The number of binary dimension partition trees (or full binary trees) with d leaves is the Catalan number $\frac{1}{d}\binom{2d-2}{d-1}$. The number N_c of possible formats of complexity c based on all possible binary dimension partition trees with d leaves is thus such that

$$N_c \leq \frac{1}{d}\binom{2d-2}{d-1}\binom{c+2d}{c}\left[\frac{c}{2d}\right]^{4d}.$$

Using again Inequality (30) and the fact that $|T^\star| = 2d \leq C_m$ for any model m in the collection, we obtain $N_c \leq d(1+\log(2)) + c(1+\log(2)) + 4d\log(c) \lesssim c$ and we finally propose the weight function

$$x(c) = \eta c \tag{32}$$

for some $\eta > 0$. In the situation where a large number of trees has been explored, we still can use penalties proportional to the format complexity C_m.

5.4 Approximation and minimax rates of tree tensor networks

For each dimension $\nu \in \{1, \ldots, d\}$, we consider approximation tools $(\mathcal{H}_{\nu,p_\nu})_{p_\nu\in\mathbb{N}}$ for functions of the variable x_ν, and we let $(\mathcal{H}_p)_{p\in\mathbb{N}^d}$ be the corresponding approximation tool for multivariate functions, where $\mathcal{H}_p = \mathcal{H}_{1,p_1} \otimes \ldots \otimes \mathcal{H}_{d,p_d}$.

For adaptive methods in p and r (with fixed tree T), we define an approximation tool

$$\Phi = (\Phi_c)_{c\in\mathbb{N}}, \quad \Phi_c = \{f = \mathcal{R}_{\mathcal{H}_p,T,r}(\boldsymbol{f}) : \boldsymbol{f} \in F_{T,r} : r \in \mathbb{N}^r, p \in \mathbb{N}^d, \mathrm{compl}(\boldsymbol{f}) \leq c\},$$

where $\mathrm{compl}(\boldsymbol{f})$ is a measure of complexity of the network $\boldsymbol{f}$, and Φ_c is the set of functions with associated network with complexity less than c.

For tree adaptive methods, we define the sets Φ_c as

$$\Phi_c = \{f = \mathcal{R}_{\mathcal{H}_p,T,r}(\boldsymbol{f}) : \boldsymbol{f} \in F_{T,r} : T \in \mathcal{T}, r \in \mathbb{N}^r, p \in \mathbb{N}^d, \mathrm{compl}(\boldsymbol{f}) \leq c\},$$

where $\mathcal{T}$ is a collection of possible dimension trees.

The best approximation error by a tensor network with complexity less than c is defined by

$$e_c(f^\star) = \inf_{f \in \Phi_c} \mathcal{R}(f) - \mathcal{R}(f^\star).$$

Then given a growth function $\gamma : \mathbb{N} \to \mathbb{N}$, an approximation class for tree tensor networks can be defined as the set

$$\mathcal{A}^\gamma = \{f^\star : \sup_{c \geq 1} \gamma(c) e_c(f^\star) < \infty\},$$

which corresponds to functions that can be approximated with tree tensor networks with a convergence in $O(\gamma(c)^{-1})$.

The approximation class $\mathcal{A}^\gamma$ depends on the measure of complexity of the network, and on wether or not tree adaptation is considered. Natural measures of complexity of a network $\boldsymbol{f}$ are the representation complexity $\mathrm{compl}_C(\boldsymbol{f}) = C(T, r, \mathcal{H}_p)$ or the sparse representation complexity $\mathrm{compl}_S(\boldsymbol{f})$ (see Section 2.4).

When considering the complexity measure $\mathrm{compl} = \mathrm{compl}_C$, we easily derive from Theorem 5.2 or 5.3 upper bounds on the rates of convergence of our model selection procedure for functions in $\mathcal{A}^\gamma$ by balancing the penalty term and the approximation term in the risk bounds.

Next we provide examples that shows that minimax rates can be achieved by tensor networks for classical smoothness classes. In all examples, we consider a least-squares setting with real valued functions ($s = 1$), where $\mathcal{R}(f) - \mathcal{R}(f^\star)$ is the squared L^2 norm of $f - f^\star$.

5.4.1 Multivariate functions

Sobolev spaces of multivariate functions. Consider a function $f^\star$ in the Sobolev space H^r, $r \in \mathbb{N}$, of functions on $(0,1)^d$ or the d-dimensional torus $\mathbb{T}^d$, and optimal approximation tools $(\mathcal{H}_{\nu,p_\nu})_{p_\nu \in \mathbb{N}}$ for univariate Sobolev functions (e.g., splines or trigonometric polynomials). For any fixed tree T, and when considering the representation complexity measure compl_C, we have $e_c(f^\star) = O(c^{-\frac{2r}{d}})$(see, e.g., [5]), and therefore H^r is included in $\mathcal{A}^\gamma$ with $\gamma(c) = c^{\frac{2r}{d}}$. In the setting of Theorem 5.3, for $f^\star$ in the Sobolev space H^r over $(0,1)^d$, and when considering the family of all possible formats, we find that the rate of convergence of $\hat{f}_{\hat{m}}$ is of order $n^{-\frac{2r}{2r+d}} \log(n)^{\frac{2r}{2r+d}}$ which is known to be the minimax rate of convergence over H^r (up to the logarithmic term). Our model selection procedure (with variable p and r) therefore achieves minimax rates for Sobolev spaces of any order, and is thus minimax adaptive to the regularity over Sovolev spaces.

Sobolev spaces of multivariate functions with dominating mixed smoothness. Consider a function $f^\star$ in the mixed Sobolev space H^r_{mix}, $r \in \mathbb{N}$, on the d-dimensional torus $\mathbb{T}^d$, and optimal approximation tools $(\mathcal{H}_{\nu,p_\nu})_{p_\nu \in \mathbb{N}}$ for univariate Sobolev functions on $\mathbb{T}$ (e.g., trigonometric polynomials). For a fixed binary tree T, when considering the complexity measure compl_C, we have $e_c(f^\star) = O(c^{-\frac{2r}{3}} \log(c)^{dr})$ (see [32, 5]), and therefore, the space H^r_{mix} is included in $\mathcal{A}^\gamma$ with $\gamma(c) = c^{\frac{2r}{3}} \log(c)^{-dr}$. In the bounded regression framework of Theorem 5.3, our model selection procedure shows a rate of convergence upper bounded by $n^{-\frac{2r}{3+2r}} (\log n)^{rd}$. To our knowledge, the minimax rates of convergence over mixed Sobolev spaces are unknown for regression. However, the results of [29] for Gaussian white noise model as well as the results of [1] for density estimation suggest that these rates should be of the order of $n^{-\frac{2r}{1+2r}}$, up to a logarithmic term. In fact, the minimax rate can not be obtained by our strategy since the rate of approximation error in $O(c^{-\frac{2r}{3}})$ (up to logarithmic terms) is not the optimal rate of convergence which is in $O(c^{-2r})$ (up to logarithmic terms), the latter rate being achieved by hyperbolic cross approximation [15].

An optimal rate should probably be achieved with tree tensor networks by further exploiting sparsity in the tensors, and using the corresponding measure of complexity compl_S. Indeed, optimal approximation rates should be obtained by shallow tensor networks (associated with a trivial tree) with a sparse tensor C^D with $O(c)$ non zero entries, and a sparsity pattern based on hyperbolic crosses. Then noting that such a shallow network (which is a canonical tensor format with rank $O(c)$) can be encoded within a tree tensor network with sparse tensors and the same overall complexity compl_S in $O(c)$, minimax rates (up to log terms) should probably be obtained for mixed Sobolev space H^r_{mix} for any tree T, when combined with an estimate of the metric entropy of sets Φ_c with the complexity measure compl_S.

5.4.2 Univariate functions

Tree tensor networks can be used for the approximation of univariate functions after identification of a function $f \in L^2(0,1)$ with an order-d tensor (or d-variate function) in $\mathbb{R}^2 \otimes \ldots \otimes \mathbb{R}^2 \otimes L^2(0,1) := (\mathbb{R}^2)^{\otimes d} \otimes L^2(0,1)$, see Section 6.1 or [2] for a more general setting. By considering an approximation subspace S in $L^2(0,1)$, say $S = \mathbb{P}_m$, we define a tensor subspace $(\mathbb{R}^2)^{\otimes d} \otimes \mathbb{P}_m = \mathcal{H}_{d,m}$, which is isometrically identified with the space of univariate splines of degree m over a uniform partition of $[0,1]$ into 2^d intervals.

An approximation tool is then defined by considering tensor networks in the tensor spaces $\mathcal{H}_{d,m}$ with variable d and fixed m. In [2], the authors consider tensor networks associated with linear trees, that is the tensor train format (or equivalently, recurrent sum-product neural networks). The variable d setting can be interpreted as the tree adaptive setting presented above, where the family of trees $\mathcal{T} = \{T_d : d \in \mathbb{N}\}$, with T_d the linear tree over $\{1, \ldots, d\}$ with interior nodes $\{1, \ldots, \nu\}$, $2 \leq \nu \leq d$.

The following results are based on results from [3, Main results 3.1, 3.2 and 3.4] for Sobolev, Besov or analytic functions.

Sobolev spaces of univariate functions. For functions $f^\star$ in the Sobolev space H^r of univariate functions on $(0,1)$, and when considering the complexity measure compl_C, the approximation error $e_c(f^\star) = O(c^{-2r})$ achieves the best possible approximation rate[2], that is H^r is included in A^γ with $\gamma(r) = n^{2r}$ for any $r \in \mathbb{N}$. Together with Theorem 5.3, we find that $\hat{f}_{\hat{m}}$ achieves a convergence in $n^{-\frac{2r}{2r+1}}$ (up to logarithmic term). This shows that our model selection procedure (with variable d and fixed m, in particular $m = 0$) achieves minimax rates (up to logarithmic terms) for Sobolev spaces of any order r (without adapting the degree m to the regularity of $f^\star$).

Besov spaces. Near optimal approximation rates are also obtained for Besov spaces of univariate functions on $(0,1)$. More presicely, consider a function $f^\star$ in the Besov space $B^\alpha_{\tau,\tau}$, with $\alpha > 0$ and $\tau = (r + 1/2)^{-1}$ the Sobolev embedding number.

When considering the complexity measure compl_C, we have $B^\alpha_{\tau,\tau} \subset A^\gamma$ with $\gamma(c) = c^{\alpha-\epsilon}$ for arbitrary $\epsilon > 0$ and all $\alpha > 0$ (which is close to half of the rate obtained with optimal approximation tools, e.g. free knot splines). Together with Theorem 5.3, we find that $\hat{f}_{\hat{m}}$ achieves a convergence in $n^{-\frac{\alpha-\epsilon}{\alpha+1}}$ (up to logarithmic term), which are close (but not equal to) minimax rates in $n^{-\frac{2\alpha}{2\alpha+1}}$ (up to log terms).

Note that when considering the complexity measure compl_S, we show $B^\alpha_{\tau,\tau} \subset A^\gamma$ with $\gamma(c) = c^{2\alpha-\epsilon}$ for arbitrary $\epsilon > 0$, which is arbitrarily close to optimal approximation rates. Therefore, a strategy taking into account sparsity of tensors could be able to achieve rates arbitrarily close to minimax rates for Besov spaces $B^\alpha_{\tau,\tau}$ of arbitrary smoothness α (without the need of adapting m to the regularity of $f^\star$).

Analytic functions. For a function $f^\star$ analytic on an open interval containing $[0,1]$ and when considering the complexity measure compl_C, the approximation error converges exponentially fast as $e_c(f^\star) = O(\rho^{-c^{2/3}})$ for some $\rho > 1$. That means $f^\star \in A^\gamma$ with $\gamma(c) = \rho^{c^{2/3}}$. Together with Theorem 5.3, we find that $\hat{f}_{\hat{m}}$

[2] also obtained by other tools such as splines of degree greater than r-1

achieves a convergence in $n^{-1}\log(n)$ (up to logarithmic term). This is known to be the minimax rate for nonparametric estimation of analytic densities [8].

5.5 Slope heuristics for penalty calibration

The aim of the slope heuristics method proposed by Birgé and Massart [10] is precisely to calibrate penalty function for model selection purposes. See [7] and [4] for a general presentation of the method. This method has shown very good performances and comes with mathematical guarantees in various settings among other for non parametric Gaussian regression with i.i.d. error terms, see [10, 4] and references therein. The slope heuristics have several versions (see [4]).

The aim is to tune the constant λ in a penalty of the form $\mathrm{pen}(m) = \lambda \mathrm{pen}_{\mathrm{shape}}(m)$ where $\mathrm{pen}_{\mathrm{shape}}$ is a known penalty shape. Let $\hat{m}(\lambda)$ be the model selected by penalized criterion with constant λ:

$$\hat{m}(\lambda) \in \mathrm{argmin}_{m\in\mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\hat{f}_m) + \lambda \mathrm{pen}_{\mathrm{shape}}(m) \right\}.$$

Let C_m denote the complexity of the model. The complexity jump algorithm consists of the following steps:

1. Compute the function $\lambda \mapsto \hat{m}(\lambda)$,
2. Find the constant $\hat{\lambda}^{cj} > 0$ that corresponds to the highest jump of the function $\lambda \mapsto C_{\hat{m}(\lambda)}$,
3. Select the model $\hat{m} = \hat{m}(2\hat{\lambda}^{cj})$ such that

$$\hat{m} \in \arg\min_{m\in\mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\hat{f}_m) + 2\hat{\lambda}^{cj} \mathrm{pen}_{\mathrm{shape}}(m) \right\}.$$

5.6 Exploration strategy

The exploration of all possible model classes $M_r^T(\mathcal{H}^s)$ with a complexity bounded by some c is intractable since the number of such models is exponential in the number of variables d. Therefore, strategies should be introduced to propose a set of candidate model classes M_m, $m \in \mathcal{M}$.

In practice, a possible approach is to rely on adaptive learning algorithms from [18] (see also [17]) that generate predictors $\hat{f}_m$ (minimizing the empirical risk) in a sequence of model classes.

5.6.1 Fixed tree

For a fixed tree T, the proposed algorithm generates a sequence of model classes $M_m = M_{r_m}^T(\mathcal{H}_m^s)$ with increasing ranks r_m, $m \geq 1$, by successively increasing the α-ranks for nodes α associated with the highest (estimated) truncation errors

$$\inf_{\mathrm{rank}_\alpha(f)\leq r_{m,\alpha}} \mathcal{R}(f) - \mathcal{R}(f^\star).$$

For each m, the background approximation space is taken as $\mathcal{H}_m := \mathcal{H}_{p_m} = \mathcal{H}_{1,p_{m,1}} \otimes \ldots \otimes \mathcal{H}_{d,p_{m,d}}$, where for each dimension $\nu \in \{1,\ldots,d\}$, $(\mathcal{H}_{\nu,k})_{k\in\mathbb{N}}$ is a given approximation tool (e.g., polynomials, wavelets). Exploring all possible tuples p_m is again a combinatorial problem. The algorithm proposed in [18, 17] relies on a validation approach for the selection of a particular tuple. Note that a complexity-based model selection method could also be considered for the selection of a tuple p_m.

5.6.2 Variable tree

Although the set of possible dimension trees over $\{1, \ldots, d\}$ is finite, exploring this whole set of dimension trees is intractable for high and even moderate d. In [18], a stochastic algorithm has been proposed for optimizing the dimension tree for the compression of a tensor. This tree optimization algorithm has been combined with the rank-adaptive strategy discussed above. The resulting algorithm generates a sequence of predictors in tree tensor networks associated with different trees. In the numerical experiments, we use this learning algorithm with tree adaptation to generate a set of candidate trees. Then the learning algorithm with rank adaptation but fixed tree is used with each of these trees.

6 Numerical experiments

In this section, we illustrate the proposed model selection approach for supervised learning problems in a least-squares regression setting. Y is a real-valued random variable ($s = 1$) defined by

$$Y = f^\star(X) + \epsilon$$

where ϵ is independent of X and has zero mean and standard deviation $\gamma\sigma(f^\star(X))$. The parameter γ therefore controls the noise level in relative precision.

For a given training sample, we use the learning strategies described in Section 5.6 that generate a sequence of predictors $\hat{f}_m$, $m \in \mathcal{M}$, associated with a certain collection of models $\mathcal{M}$ (which depends on the training sample). Given a set of predictors $\hat{f}_m$, $m \in \mathcal{M}$, we denote by $\hat{m}^\star$ the index of the model that minimizes the risk over $\mathcal{M}$, i.e.

$$\hat{m}^\star \in \arg\min_{m \in \mathcal{M}} \mathcal{R}(\hat{f}_m).$$

The model $\hat{m}^\star$ is the oracle model in $\mathcal{M}$ for a given training sample.

We also denote by $\hat{m}(\lambda)$ the model such that

$$\hat{m}(\lambda) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\hat{f}_m) + \lambda \mathrm{pen}_{\mathrm{shape}}(m) \right\},$$

where $\mathrm{pen}_{\mathrm{shape}}(m) = C_m/n$, and by $\hat{m} = \hat{m}(2\hat{\lambda}^{cj})$ the model selected by our model selection strategy, where $\hat{\lambda}^{cj}$ is calibrated with the complexity jump algorithm (see Section 5.5).

We consider two different types of problems: the approximation of univariate functions defined on $(0, 1)$, identified with a multivariate function through tensorization (Section 6.1), and the approximation of multivariate functions defined on a subset of $\mathbb{R}^d$ (Section 6.2).

For a given function f, the risk $\mathcal{R}(f)$ is evaluated using a sample of size 10^5 independent of the training sample. Statistics of complexities and risks (such as the expected complexity $\mathbb{E}(C_{\hat{m}})$ or the expected risk $\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$) are computed using 20 different training samples.

6.1 Tensorized function

Here we consider tree tensor networks for the approximation of a univariate function in $L^2(0, 1)$, see Section 5.4.2 and [2, 3] for a general presentation. A function f defined on $(0, 1)$ can be linearly identified with a function $\boldsymbol{f} = \mathcal{T}_l(f)$ of $l + 1$ variables defined on $\{0, 1\}^l \times (0, 1)$ such that

$$f(x) = \mathcal{T}_l(f)(i_0, \ldots, i_{l-1}, y) \quad \text{for} \quad x = 2^{-l}(\sum_{k=0}^{l-1} i_k 2^k + y).$$

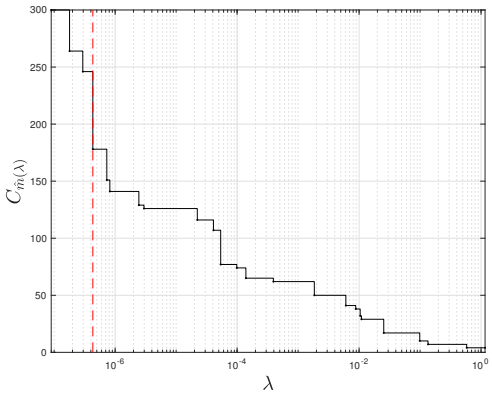
The map $\mathcal{T}_l$ is called the tensorization map at level l. This allows to isometrically identify the space $L^2(0,1)$ with the tensor space $\mathbb{R}^2 \otimes \ldots \otimes \mathbb{R}^2 \otimes L^2(0,1)$ of order $d = l+1$. Then we consider the approximation space $\mathcal{H}_l = \mathbb{R}^2 \otimes \ldots \otimes \mathbb{R}^2 \otimes \mathbb{P}_0$ of d-variate functions $\boldsymbol{f}(i_0, \ldots, i_l, y)$ independent of the variable y. The space $\mathcal{H}_l$ is linearly identified with the space of piecewise constant functions on the uniform partition of $(0,1)$ into 2^l intervals. Then we consider model classes $M_{l,T,r} = \{f : \mathcal{T}_l(f) \in M_r^T(\mathcal{H}_l)\}$, which are piecewise constant functions whose tensorized version $\mathcal{T}_l(f)$ is in a particular tree-based tensor format.

In the following experiments, for each $l \in \{1, \ldots, 12\}$, we consider a fixed linear binary tree T (with interior nodes $\{1, \ldots, k\}$, $1 \leq k \leq l+1$) and use the rank adaptive learning algorithm (Section 5.6.1) to produce a sequence of 25 approximations with increasing ranks.

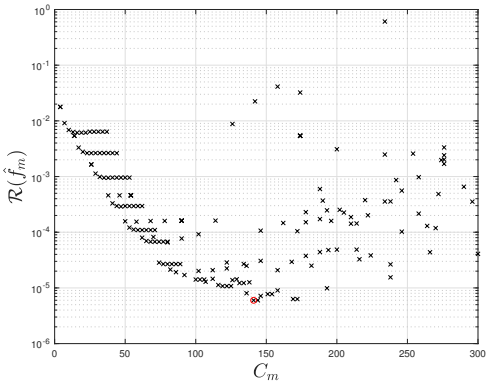
Three functions $f^\star(x)$ are considered. The first function $f^\star(x) = \sqrt{x}$ is analytic on the open interval $(0,1)$ and its derivative has a singularity at zero. The second function $f^\star(x) = \frac{1}{1+x}$ is analytic on a larger interval including $[0,1]$. The third function is in the Sobolev space $H^2(0,1)$. For all functions, the proposed model selection approach shows a very good performance. It selects with high probability a model with a risk very close to the risk of the oracle $\hat{f}_{\hat{m}^\star}$.

6.1.1 Tensorized function $f^\star(x) = \sqrt{x}$

We consider the function $f^\star(x) = \sqrt{x}$ which is analytic on the open interval $(0,1)$, with a singular derivative at zero. We observe on Figures 3 and 4 that the model selection approach selects a model close to optimal for different sample size n and noise level. Tables 1 and 2 show expectations of complexities and errors for the selected estimator and illustrate the very good performance of the approach when compared to the oracle.



(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

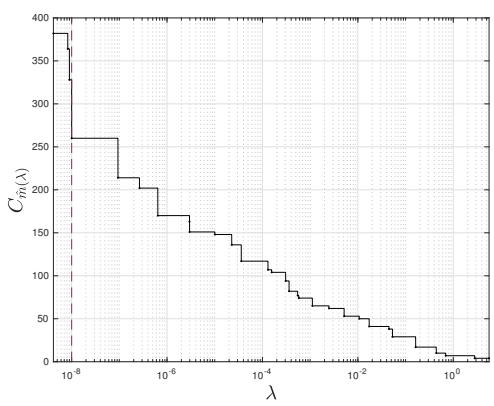


(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

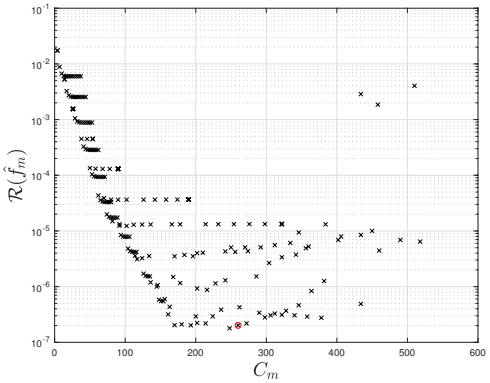
Figure 3: Slope heuristics for the tensorized function $f^\star(x) = \sqrt{x}$ with $n = 200$ and $\gamma = 0.001$.

n	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
100	123.2	91.6	1.6e-05	5.0e-05
200	163.8	165.0	3.0e-06	5.1e-06
500	182.2	182.6	9.2e-07	1.2e-06
1000	190.2	228.5	7.1e-07	1.4e-06

Table 1: Expectation of complexities and risks of the model selected by the slope heuristics, with the function $f^\star(x) = \sqrt{x}$ and different values of n and $\gamma = 0.001$.



(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).



(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

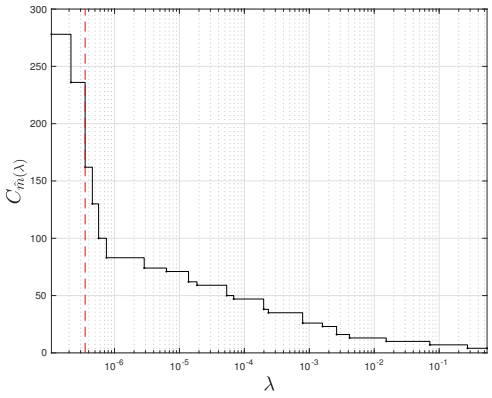
Figure 4: Slope heuristics for the tensorized function $f^\star(x) = \sqrt{x}$ with $n = 1000$ and $\gamma = 0.0001$.

γ	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
10^{-3}	190.2	228.5	7.1e-07	1.4e-06
10^{-4}	242.8	251.4	1.5e-07	2.1e-07
10^{-5}	219.8	267.4	1.3e-07	2.4e-07
0	218.6	258.6	1.1e-07	2.1e-07

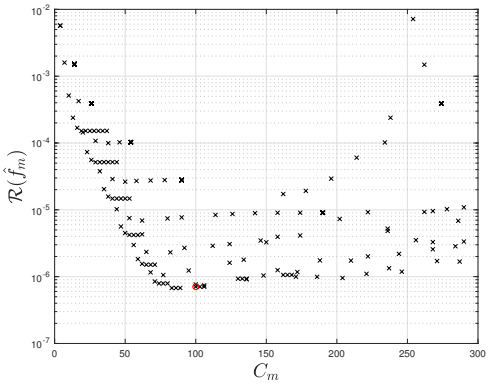
Table 2: Expectation of complexities and risks of the model selected by the slope heuristics, with the function $f^\star(x) = \sqrt{x}$ and different values of γ and $n = 1000$.

6.1.2 Tensorized function $f^\star(x) = \frac{1}{1+x}$.

We consider the function $f^\star(x) = \frac{1}{1+x}$ which is analytic on the interval $(-1, \infty)$ including $[0, 1]$. Figures 5 and 6 illustrate the good behaviour of the model selection approach for different sample size and noise level. Tables 3 and 4 show expectations of complexities and errors for the selected estimator and illustrate again the very good performance of the approach when compared to the oracle.

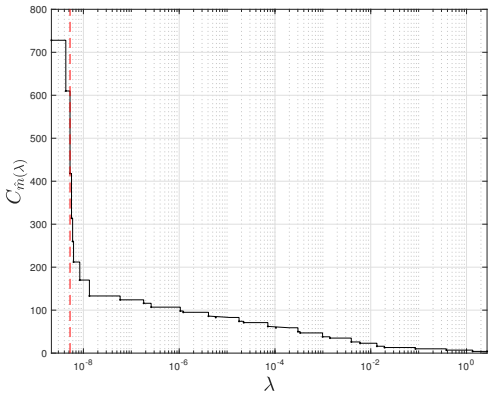


(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

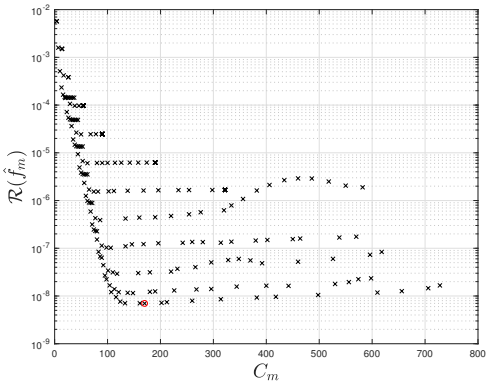


(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

Figure 5: Slope heuristics for the tensorized function $f^\star(x) = \frac{1}{1+x}$ with $n = 200$ and $\gamma = 0.001$.



(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).



(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

Figure 6: Slope heuristics for the tensorized function $f^\star(x) = \frac{1}{1+x}$ with $n = 1000$ and $\gamma = 0.0001$.

n	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
100	88.0	83.0	9.3e-07	1.0e-06
200	97.3	92.8	6.4e-07	6.6e-07
500	92.9	124.4	5.8e-07	6.9e-07
1000	108.4	107.5	5.3e-07	5.3e-07

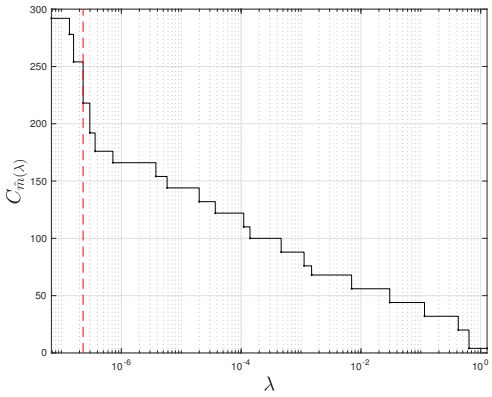
Table 3: Expectation of complexities and risks of the model selected by the slope heuristics, with the function $f^\star(x) = \frac{1}{1+x}$, different values of n and $\gamma = 0.001$.

γ	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
10^{-3}	108.4	107.5	5.3e-07	5.3e-07
10^{-4}	159.3	151.1	6.9e-09	6.9e-09
10^{-5}	152.0	182.2	1.6e-09	1.9e-09
0	156.8	155.8	1.6e-09	1.6e-09

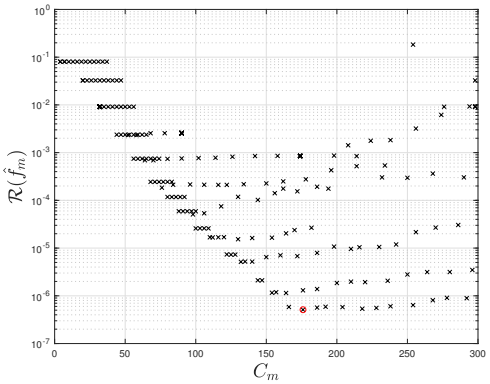
Table 4: Expectation of complexities and risks of the model selected by the slope heuristics, with the function $f^\star(x) = \frac{1}{1+x}$, different values of γ and $n = 1000$.

6.1.3 Tensorized function $f^\star(x) = g(g(x))^2$ with $g(x) = 1 - 2|x - \frac{1}{2}|$.

We consider the function $f^\star(x) = g(g(x))^2$ with $g(x) = 1-2|x-\frac{1}{2}|$, which is in the Sobolev space $H^2(0,1)$. Figures 7b and 8 illustrate again the good behaviour of the model selection approach for different sample size and noise level. And Tables 3 and 4 again illustrate again the very good performance (in expectation) for the selected estimator of the approach when compared to the oracle.

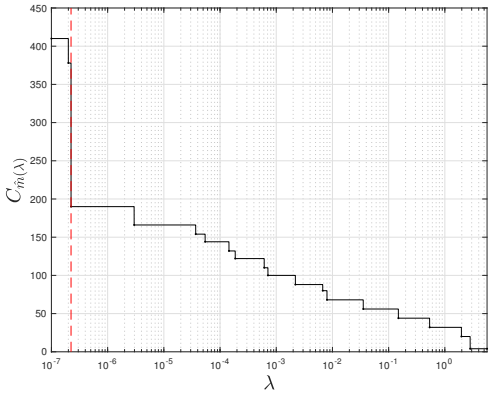


(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

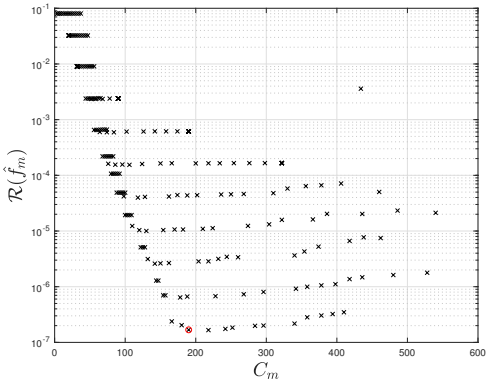


(b) Points $(C_m, \mathcal{R}(\hat{f}_m))$, $m \in \mathcal{M}$, and selected model (red).

Figure 7: Slope heuristics for the tensorized function $f^\star(x) = (g(g(x)))^2$ with $n = 200$ and $\gamma = 0.001$.



(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).



(b) Points $(C_m, \mathcal{R}(\hat{f}_m))$, $m \in \mathcal{M}$, and selected model (red).

Figure 8: Slope heuristics for the tensorized function $f^\star(x) = (g(g(x)))^2$ with $n = 1000$ and $\gamma = 0.0001$.

n	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
200	176.4	181.6	6.3e-07	1.6e-06
500	188.2	198.8	3.9e-07	4.1e-07
1000	196.6	233.8	3.2e-07	3.5e-07

Table 5: Expectation of complexities and risks for the function $f^\star(x) = (g(g(x)))^2$, different values of n and $\gamma = 0.001$.

γ	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
10^{-3}	196.6	233.8	3.2e-07	3.5e-07
10^{-4}	195.8	205.8	1.7e-07	1.7e-07
10^{-5}	191.0	226.6	1.7e-07	1.8e-07
0	194.0	232.6	1.7e-07	1.9e-07

Table 6: Expectation of complexities and risks of the model selected by the slope heuristics, with the function $f^\star(x) = (g(g(x)))^2$, different values of γ and $n = 1000$.

6.2 Multivariate functions

6.2.1 Corner peak function

We consider the function

$$f^\star(X) = \frac{1}{1 + \sum_{\nu=1}^{d} \nu^{-2} X_\nu}$$

with $d = 10$, where the $X_\nu \sim U(0,1)$ are i.i.d. uniform random variables. The function $f^\star$ is analytic on $[0,1]^d$. We use the fixed balanced binary tree T of Figure 9. Figures 10 and 11 illustrate the very good behaviour of the model selection approach for a sample size $n = 1000$ and noise level $\gamma = 0.001$, where the best model appears to be always selected. In Tables 7 and 8, we observe that the expectation of complexities and errors for the selected estimator (for different values of n and γ), which are of the are of the same order as for the oracle.

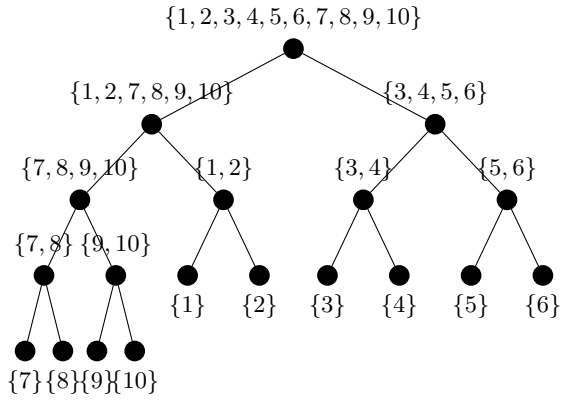
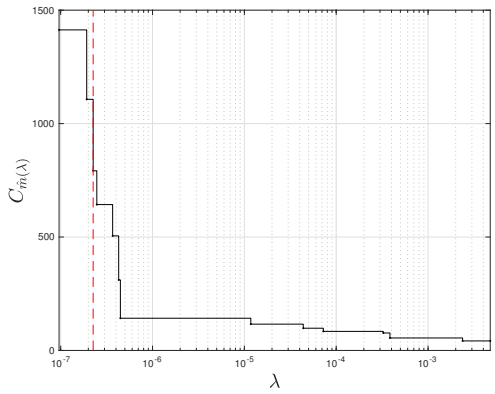


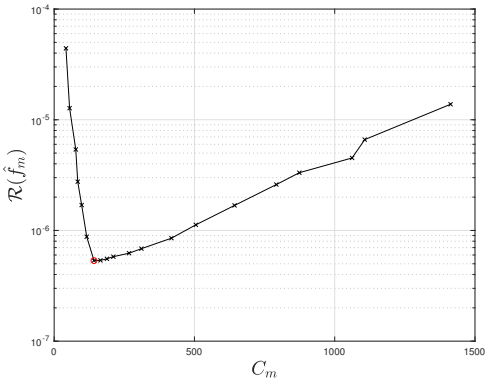
Figure 9: Corner peak function. Dimension tree T.

n	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
100	124.1	73.7	2.1e-06	1.1e-05
500	286.7	291.3	9.8e-11	1.0e-10
1000	286.2	293.8	6.6e-11	6.7e-11

Table 7: Expectation of complexities and risks selected by the slope heuristics, with the Corner peak function, different values of n and $\gamma = 10^{-5}$.

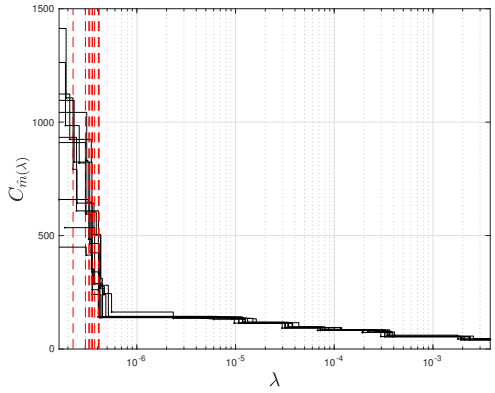


(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

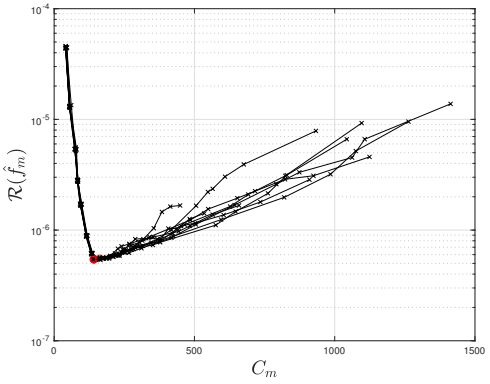


(b) Function $C_m \mapsto \mathcal{R}(\hat{f}_m)$ and selected model (red).

Figure 10: Slope heuristics for the Corner peak function with $n = 1000$ and $\gamma = 0.001$.



(a) Functions $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).



(b) Functions $C_m \mapsto \mathcal{R}(\hat{f}_m)$ and selected model (red).

Figure 11: Slope heuristics for the Corner peak function with $n = 1000$ and $\gamma = 0.001$, superposition of 10 different samples.

γ	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
10^{-2}	95.5	79.8	5.4e-05	5.5e-05
10^{-3}	143.1	143.1	5.4e-07	5.4e-07
10^{-4}	223.2	193.7	5.9e-09	6.0e-09
10^{-5}	286.2	293.8	6.6e-11	6.7e-11
0	598.7	538.4	2.5e-15	1.8e-14

Table 8: Expectation of complexities and risks selected by the slope heuristics, with the Corner peak function, different values of γ and $n = 1000$.

6.2.2 Borehole function

We consider the function

$$g(U_1, \ldots, U_8) = \frac{2\pi U_3(U_4 - U_6)}{(U_2 - \log(U_1))(1 + \frac{2U_7U_3}{(U_2 - \log(U_1))U_1^2U_8} + \frac{U_3}{U_5})}$$

which models the water flow through a borehole as a function of 8 independent random variables $U_1 \sim \mathcal{N}(0.1, 0.0161812)$, $U_2 \sim \mathcal{N}(7.71, 1.0056)$, $U_3 \sim U(63070, 115600)$, $U_4 \sim U(990, 1110)$, $U_5 \sim U(63.1, 116)$, $U_6 \sim U(700, 820)$, $U_7 \sim U(1120, 1680)$, $U_8 \sim U(9855, 12045)$. Then we consider the function

$$f^\star(X_1, \ldots, X_d) = g(g_1(X_1), \ldots, g_8(X_8)),$$

where g_ν are functions such that $U_\nu = g_\nu(X_\nu)$, with $X_\nu \sim \mathcal{N}(0, 1)$ for $\nu \in \{1, 2\}$, and $X_\nu \sim U(-1, 1)$ for $\nu \in \{3, \ldots, 8\}$. Function $f^\star$ is thus defined on $\mathcal{X} = \mathbb{R}^2 \times [-1, 1]^6$. As univariate approximation tools, we use polynomial spaces $\mathcal{H}_{\nu,p_\nu} = \mathbb{P}_{p_\nu}(\mathcal{X}_\nu)$, $\nu \in D$.

We use the exploration strategy described in Section 5.6.1. More precisely, we first run a learning algorithm with tree adaptation from an initial binary tree drawn randomly, with $n = 100$ samples. The learning algorithm visited the 9 trees plotted in Figure 12. Then for each of these trees, we start a learning algorithm with fixed tree and rank adaptation. Figures 13 to 15 illustrate the behaviour of the model selection strategy for different sample size n. Table 9 shows the expectation of complexities and risks. The model selection approach shows very good performances, except for very small training size $n = 100$, where the approach selects a model rather far from the optimal one (in terms of expected risk and complexity).

n	$\mathbb{E}(C_{\hat{m}^\star})$	$\mathbb{E}(C_{\hat{m}})$	$\mathbb{E}(\mathcal{R}(\hat{f}_{m^\star}))$	$\mathbb{E}(\mathcal{R}(\hat{f}_{\hat{m}}))$
100	132.1	63.4	6.9e-06	9.3e-04
200	149.7	156.0	3.0e-08	1.1e-07
500	144.7	178.2	1.0e-08	1.8e-08
1000	154.1	194.2	8.3e-09	1.2e-08

Table 9: Borehole function. Expectation of complexities and risks. $\gamma = 10^{-6}$, different n.

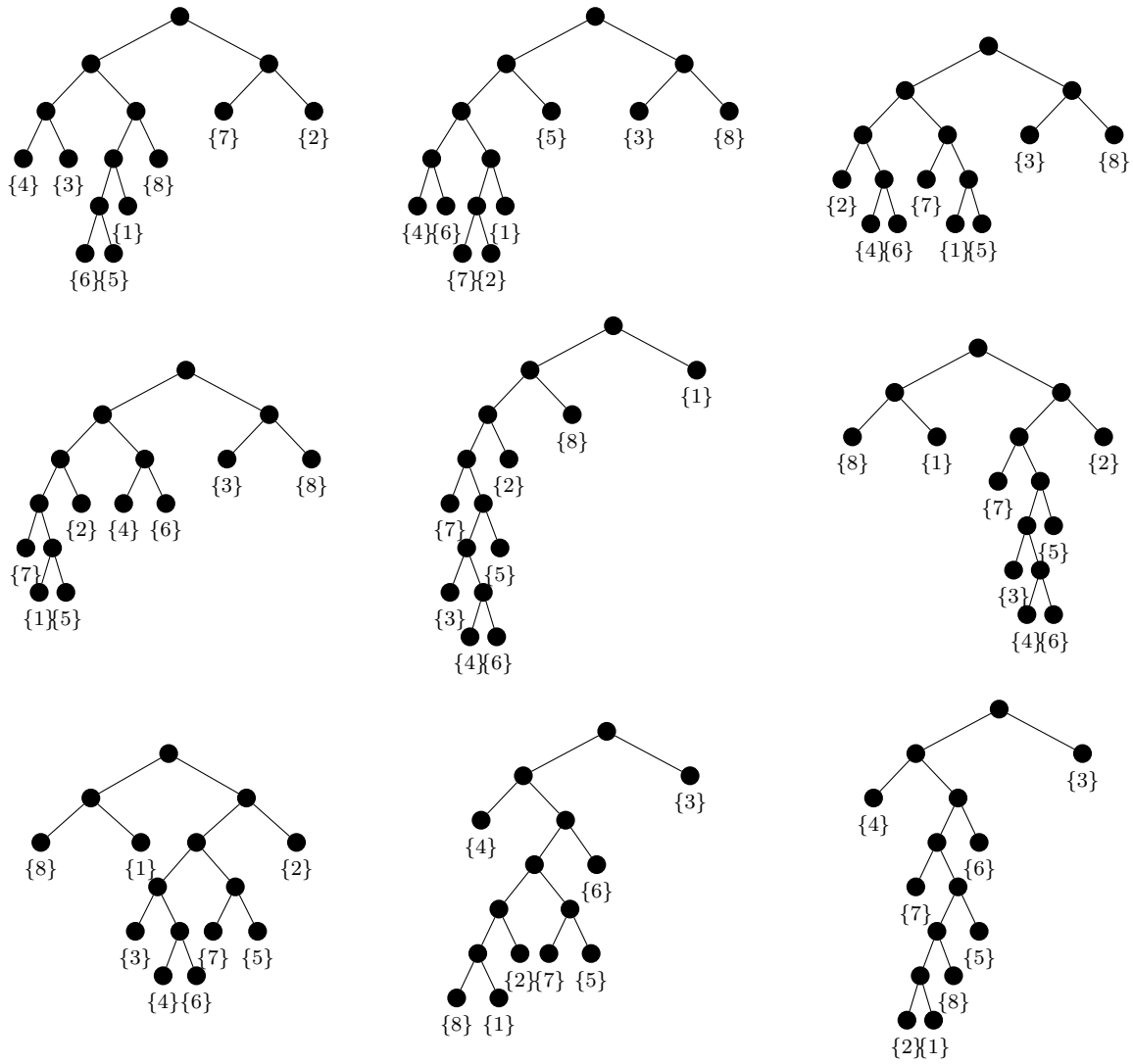
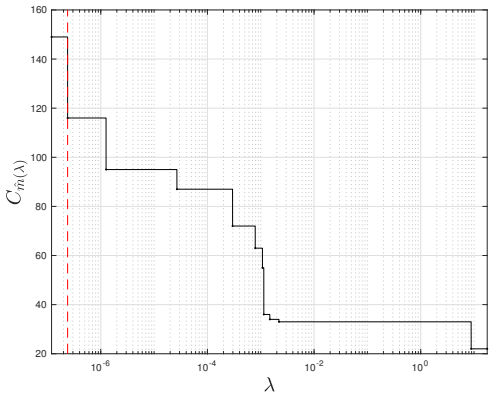
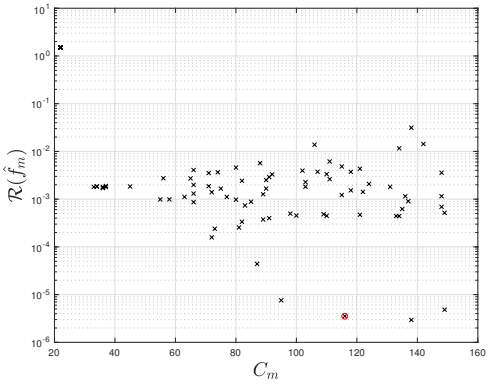


Figure 12: Borehole function. The path of 9 trees generated by the tree adaptive learning algorithm.

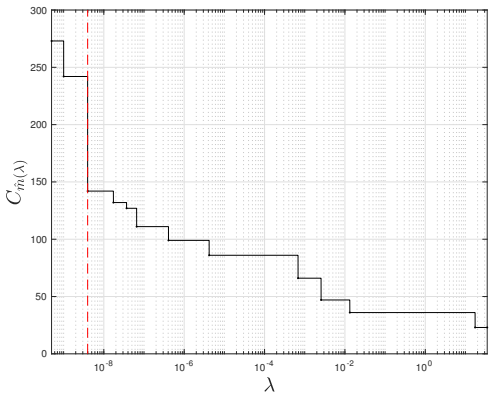


(a) Functions $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

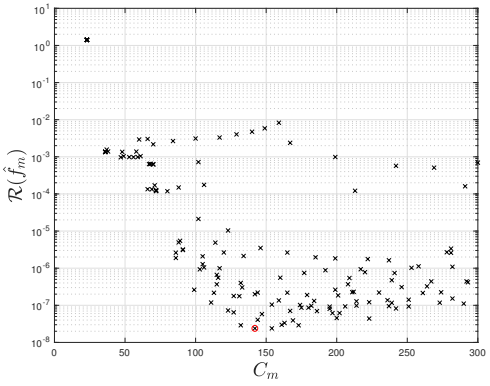


(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

Figure 13: Slope heuristics for Borehole function with $n = 100$ and $\gamma = 10^{-6}$.

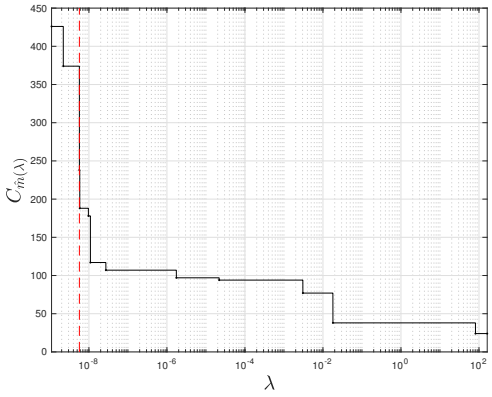


(a) Functions $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).

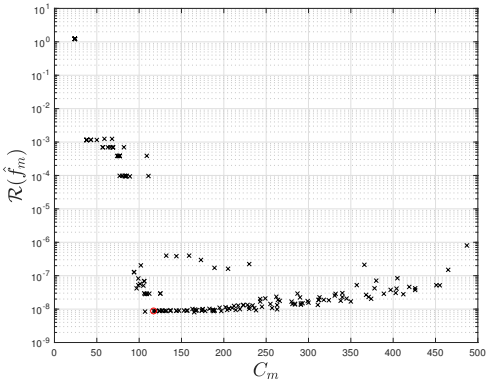


(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

Figure 14: Slope heuristics for Borehole function with $n = 200$ and $\gamma = 10^{-6}$.



(a) Function $\lambda \mapsto C_{\hat{m}(\lambda)}$, λ^{cj} (red).



(b) Points $(C_m, \mathcal{R}(\hat{f}_m)$, $m \in \mathcal{M}$, and selected model (red).

Figure 15: Slope heuristics for Borehole function with $n = 1000$ and $\gamma = 10^{-6}$.

References

[1] Nathalie Akakpo. Multivariate intensity estimation via hyperbolic wavelet selection. *Journal of Multivariate Analysis*, 161:32–57, 2017.

[2] Mazen Ali and Anthony Nouy. Approximation with tensor networks. part I: Approximation spaces. *arXiv e-prints, arxiv:2007.00118*, 2020.

[3] Mazen Ali and Anthony Nouy. Approximation with tensor networks. part II: Approximation rates for smoothness classes. *arXiv e-prints, arxiv:2007.00128*, 2020.

[4] Sylvain Arlot. Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*, 2019.

[5] M. Bachmayr, A. Nouy, and R. Schneider. Approximation power of tree tensor networks for compositional functions, 2020.

[6] M. Bachmayr, R. Schneider, and A. Uschmajew. Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. *Foundations of Computational Mathematics*, pages 1–50, 2016.

[7] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[8] Eduard Belitser et al. Efficient estimation of analytic density under random censorship. *Bernoulli*, 4(4):519–543, 1998.

[9] Gérard Biau and Aurélie Fischer. Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58(3):1924–1939, 2012.

[10] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[11] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[12] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, and D. Mandic. Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429, 2016.

[13] A. Cichocki, A.-H. Phan, Q. Zhao, N. Lee, I. Oseledets, M. Sugiyama, and D. Mandic. Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives. *Foundations and Trends® in Machine Learning*, 9(6):431–673, 2017.

[14] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory*, pages 698–728, 2016.

[15] Dinh Dũng, Vladimir Temlyakov, and Tino Ullrich. *Hyperbolic cross approximation*. Springer, 2018.

[16] A. Falcó, W. Hackbusch, and A. Nouy. Tree-based tensor formats. *SeMA Journal*, 2018.

[17] E. Grelier, A. Nouy, and R. Lebrun. Learning high-dimensional probability distributions using tree tensor networks. *ArXiv e-prints*, arXiv:1912.07913, 2019.

[18] Erwan Grelier, Anthony Nouy, and Mathilde Chevreuil. Learning with tree-based tensor formats. *arXiv e-prints*, arXiv:1811.04455, 2018.

[19] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *arXiv e-prints*, arXiv:1905.01208, 2019.

[20] Michael Griebel and Helmut Harbrecht. Analysis of tensor approximation schemes for continuous functions. *arXiv e-prints arXiv:1903.04234* , 2019.

[21] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[22] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42 of *Springer series in computational mathematics*. Springer, Heidelberg, 2012.

[23] Vladimir Kazeev, Ivan Oseledets, Maxim Rakhuba, and Christoph Schwab. Qtt-finite-element approximation for multiscale problems i: model problems in one dimension. *Advances in Computational Mathematics*, 43(2):411–442, 2017.

[24] Vladimir Kazeev and Christoph Schwab. Approximation of singularities by quantized-tensor fem. *PAMM*, 15(1):743–746, 2015.

[25] B. Khoromskij. Tensors-structured numerical methods in scientific computing: Survey on recent advances. *Chemometrics and Intelligent Laboratory Systems*, 110(1):1–19, 2012.

[26] Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. In *International Conference on Learning Representations*, 2018.

[27] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole dEté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[28] P. Massart. *Concentration Inequalities and Model Selection*, volume Lecture Notes in Mathematics 1896. Springer-Verlag, 2007.

[29] Michael Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statistica Sinica*, 10:399–431, 2000.

[30] A. Nouy. Low-rank methods for high-dimensional approximation and model order reduction. In P. Benner, A. Cohen, M. Ohlberger, and K. Willcox, editors, *Model Reduction and Approximation: Theory and Algorithms*. SIAM, Philadelphia, PA, 2017.

[31] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[32] R. Schneider and A. Uschmajew. Approximation rates for the hierarchical tensor format in periodic sobolev spaces. *Journal of Complexity*, 30(2):56–71, 2014. Dagstuhl 2012.

[33] Marco Signoretto, Lieven De Lathauwer, and Johan AK Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

[34] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[35] Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

A Proofs

A.1 Proofs of Section 3

Proof of Proposition 3.6. Let $f = \mathcal{R}_{\mathcal{H},T,r}((f^\alpha)_{\alpha\in T^\star})$ and let $\lambda = (p,\mu)$, $1 \le p \le \infty$, or $\lambda = \infty$ (with $p = \infty$ when $\lambda = \infty$). For $x \in \mathcal{X}$, we first note that

$$\|f(x)\|_\lambda = \|f^\emptyset(g^D(x))\|_\lambda \le \|f^\emptyset\|_{F^\emptyset}\|g^D(x)\|_\lambda.$$

Then for any interior node $\alpha \in \mathcal{I}(T)$, we have

$$\|g^\alpha(x_\alpha)\|_\lambda = \|f^\alpha((g^\beta(x_\beta))_{\beta\in S(\alpha)})\|_\lambda \le \|f^\alpha\|_{F^\alpha} \prod_{\beta\in S(\alpha)} \|g^\beta(x_\beta)\|_\lambda,$$

and for any leaf node $\alpha \in \mathcal{L}(T)$,

$$\|g^\alpha(x_\alpha)\|_\lambda = \|f^\alpha(\phi^\alpha(x_\alpha))\|_\lambda \le \|f^\alpha\|_{F^\alpha}\|\phi^\alpha(x_\alpha)\|_\lambda.$$

We deduce that

$$\|f(x)\|_\lambda \le \prod_{\alpha\in T^\star} \|f^\alpha\|_{F^\alpha} \prod_{1\le\nu\le d} \|\phi^\nu(x_\nu)\|_\lambda,$$

and therefore, since μ is a product measure and from the particular normalization of functions ϕ^ν, we obtain

$$\|f\|_\lambda \le \prod_{\alpha\in T^\star} \|f^\alpha\|_{F^\alpha} \prod_{1\le\nu\le d} \|\phi^\nu\|_\lambda = \prod_{\alpha\in T^\star} \|f^\alpha\|_{F^\alpha},$$

which proves that $L_\lambda \le 1$. Finally for $1 \le q \le p$, we note that

$$\mu(\mathcal{X})^{1/p-1/q}\|f\|_{q,\mu} \le \|f\|_{p,\mu} \le \|f\|_\infty,$$

which yields $L_{q,\mu} \le \mu(\mathcal{X})^{1/q-1/p}L_\lambda$. □

A.2 Proofs of Section 4

Proof of Lemma 4.6. Let $\gamma = \frac{\epsilon B}{2\mathcal{L}}$ and let $\mathcal{N}$ be a γ-net of M for the $\|\cdot\|_{\infty,\mu}$-norm, with cardinal $N_{\frac{\epsilon B}{2\mathcal{L}}}$. Using Lemma 4.4 and a union bound argument, we obtain

$$\mathbb{P}(\sup_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(g) > \epsilon B) \vee \mathbb{P}(\inf_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(g) < -\epsilon B) \le N_{\frac{\epsilon B}{2\mathcal{L}}} e^{-\frac{n\epsilon^2}{2}}.$$

For any $f \in M$, there exists a $g \in \mathcal{N}$ such that $\|f-g\|_{\infty,\mu} \le \gamma$. Noting that

$$\bar{\mathcal{R}}_n(f) = \bar{\mathcal{R}}_n(g) + \widehat{\mathcal{R}}_n(f) - \widehat{\mathcal{R}}_n(g) + \mathcal{R}(g) - \mathcal{R}(f),$$

we deduce from Assumption 4.5 that

$$\bar{\mathcal{R}}_n(f) \le \bar{\mathcal{R}}_n(g) + 2\mathcal{L}\|f-g\|_{\infty,\mu} \le \sup_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(g) + \epsilon B,$$

and

$$\bar{\mathcal{R}}_n(f) \ge \bar{\mathcal{R}}_n(g) - 2\mathcal{L}\|f-g\|_{\infty,\mu} \ge \inf_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(g) - \epsilon B.$$

This implies that

$$\mathbb{P}(\sup_{f\in M} \bar{\mathcal{R}}_n(f) > 2\epsilon B) \le \mathbb{P}(\sup_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(f) > \epsilon B),$$

and

$$\mathbb{P}(\inf_{f\in M} \bar{\mathcal{R}}_n(f) < -2\epsilon B) \le \mathbb{P}(\inf_{g\in\mathcal{N}} \bar{\mathcal{R}}_n(f) < -\epsilon B),$$

which yields (21). The bound on $N_{\frac{\epsilon B}{2\mathcal{L}}}$ directly follows from Proposition 3.3 and Proposition 3.6. □

Proof of Lemma 4.7. We have

$$\begin{aligned}\mathbb{E}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|) &= \int_0^\infty \mathbb{P}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|>t)dt \\ &= 2B\int_0^\infty \mathbb{P}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|>2\epsilon B)d\epsilon.\end{aligned}$$

Let $\beta = 6\mathcal{L}B^{-1}R|T^\star|$. Then, according to Lemma 4.6, for any $\delta > 0$,

$$\begin{aligned}\mathbb{E}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|) &\le 2B\left[\delta + \int_\delta^\infty 2(\beta\epsilon^{-1})^{C_M}e^{-n\frac{\epsilon^2}{2}}d\epsilon\right], \\ &= 2B\left[\delta + 2\beta^{C_M}\int_{n\delta^2/2}^\infty \left(\frac{2u}{n}\right)^{-C_M/2} e^{-u}\frac{1}{\sqrt{2nu}}du\right] \\ &\le 2B\left[\delta + 2n^{-1}\beta^{C_M}\delta^{-C_M-1}e^{-n\delta^2/2}\right],\end{aligned}$$

By taking

$$\delta = \sqrt{\frac{2C_M}{n}\log((\beta\vee e)\sqrt{n})},$$

we have

$$\begin{aligned}n^{-1}\beta^{C_M}\delta^{-C_M-1}e^{-n\delta^2/2} &= n^{-1}\beta^{C_M}\delta^{-C_M-1}(\beta\vee e)^{-C_M}n^{-\frac{C_M}{2}} \\ &\le \delta^{-C_M-1}n^{-\frac{C_M}{2}-1} \\ &= \delta(\delta^2 n)^{-\frac{C_M}{2}-1} \\ &= \delta(2C_M\log((\beta\vee e)\sqrt{n}))^{-\frac{C_M}{2}-1} \\ &\le \delta\end{aligned}$$

where we have used the fact that $2C_M\log((\beta\vee e)\sqrt{n}) \ge 1$. Then

$$\mathbb{E}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|) \le 4B\delta,$$

which concludes the proof. □

Proof of Proposition 4.8. Starting from (16), we obtain

$$\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M) \le \bar{\mathcal{R}}_n(f^M) - \bar{\mathcal{R}}_n(\hat{f}_n^M) \le \sup_{f\in M}\bar{\mathcal{R}}_n(f) - \inf_{f\in M}\bar{\mathcal{R}}_n(f).$$

Then using Lemma 4.6, we deduce

$$\mathbb{P}(\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M) > 2\epsilon B) \le \mathbb{P}(\sup_{f\in M}\bar{\mathcal{R}}_n(f) > \epsilon B) + \mathbb{P}(\inf_{f\in M}\bar{\mathcal{R}}_n(f) < -\epsilon B) \le 2N_{\frac{\epsilon B}{2\mathcal{L}}}e^{-\frac{n\epsilon^2}{2}},$$

with $\log N_{\frac{\epsilon B}{2\mathcal{L}}} \le C_M\log(\beta\epsilon^{-1})$. In the same way for the expectation bound, we have

$$\mathbb{E}(\mathcal{R}(\hat{f}_n^M) - \mathcal{R}(f^M)) \le \mathbb{E}(\bar{\mathcal{R}}_n(f^M) - \bar{\mathcal{R}}_n(\hat{f}_n^M)) \le \mathbb{E}(\sup_{f\in M}|\bar{\mathcal{R}}_n(f)|)$$

and the result directly follows from Lemma 4.7. □

Proof of Proposition 4.9. The two inequalities come from standard application of the bounded difference Inequality, see for instance Theorem 5.1 in [28]. The bounded difference inequality applied to $\sup_{f\in M} -\bar{\mathcal{R}}_n(f) = \sup_{f\in M} \mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)$ gives that with probability larger than $1-\exp(-t)$,

$$\sup_{f\in M} -\bar{\mathcal{R}}_n(f) \quad \leq \quad \mathbb{E}(\sup_{f\in M} -\bar{\mathcal{R}}_n(f)) + 2B\sqrt{\frac{t}{2n}}.$$

Inequality 22 directly derives from this inequality and Lemma 4.7. Next, Inequality (16) gives that

$$\mathcal{E}(\hat{f}_n^M) \leq \mathcal{E}(f^M) + 2\sup_{f\in M} |\bar{\mathcal{R}}_n(f)|.$$

We finally prove the risk bound (23) by applying the bounded difference Inequality to $\sup_{f\in M} |\bar{\mathcal{R}}_n(f)|$ and Lemma 4.7 again. □

A.2.1 Proof of Proposition 4.12

The proof follows the presentation of [27]. The least-squares contrast γ corresponds either to the regression contrast or the density estimation contrast. Under the assumptions of the proposition, in both frameworks the oracle function satisfies $\|f^\star\|_\infty \leq R$.

• We first prove the proposition in the case where $R=1$, by assuming for the moment that $M = M_1 = M_r^T(\mathcal{H}^s)_1$. For the regression framework, it is also assumed for the moment that $\|Y\|_{\ell^\infty} \leq 1$ almost surely. Note that we also have $\|f^\star\|_\infty \leq 1$.

• For the least-squares regression contrast (see Example 4.10), we have $\gamma(f,Z) = \|Y-f(X)\|_{\ell^2}^2$. For all $f\in M_1$, it gives $\gamma(f,Z) \leq \|Y-f(X)\|_{\ell^\infty}^2 \leq 2(\|Y\|_{\ell^\infty}^2 + \|f\|_\infty^2)$, so that $0\leq\gamma(f,Z)\leq B$ almost surely, with $B=4$. The distribution of the random variable X is denoted μ. Then

$$\begin{aligned}
\mathbb{E}((\gamma(f,Z)-\gamma(f^\star,Z))^2) &= \mathbb{E}\left[(f^\star(X)-f(X))\,(2Y-f(X)-f^\star(X))\right]^2 \\
&= \mathbb{E}\left[(f^\star(X)-f(X))\,(2(Y-f^\star(X))+f^\star(X)-f(X))\right]^2 \\
&= \mathbb{E}\left[(f^\star(X)-f(X))\,(2(Y-f^\star(X))\right]^2 + \mathbb{E}\left[f^\star(X)-f(X)\right]^4 \\
&\leq (4\|Y-f^\star(X)\|_{\ell^\infty}^2 + \|f^\star-f\|_\infty^2)\|f-f^\star\|_{2,\mu}^2 \\
&\leq 8\|f-f^\star\|_{2,\mu}^2 = 2B\|f-f^\star\|_{2,\mu}^2,
\end{aligned}$$

where the last inequality has been obtained using $\|Y-f^\star(X)\|_{\ell^\infty} = \|Y-\mathbb{E}(Y|X)\|_{\ell^\infty} \leq 1$. Let $\gamma_1 = \gamma/B$. We have $0\leq\gamma_1\leq 1$ and the normalized excess risk satisfies

$$\mathcal{E}_1(f) := \mathbb{E}\left[\gamma_1(f,Z)-\gamma_1(f^\star,Z)\right] = \frac{1}{B}\|f-f^\star\|_{2,\mu}^2 = \frac{1}{B}\mathcal{E}(f)$$

and

$$\mathbb{E}([\gamma_1(f,Z)-\gamma_1(f^\star,Z)]^2) \leq D\|f-f^\star\|_{2,\mu}^2$$

with $D = \frac{2}{B} = \frac{1}{2}$.

• We now consider the density estimation framework with $\gamma(f,x) = \|f\|_{2,\mu}^2 - 2f(x)$. According to Example 4.11, $|\gamma(f,X)| \leq B = \mu(\mathcal{X}) + 2$. The excess risk satisfies $\mathcal{E}(f) = \|f^\star - f\|_{2,\mu}^2$ and

$$\begin{aligned}
\mathbb{E}([\gamma(f,Z)-\gamma(f^\star,Z)])^2 &= \mathbb{E}(\left[\|f\|_{2,\mu}^2 - \|f^\star\|_{2,\mu}^2 + 2(f^\star(X)-f(X))\right]^2) \\
&\leq (\|f\|_{2,\mu}^2 - \|f^\star\|_{2,\mu}^2)^2 + 4(\|f\|_{2,\mu}^2 - \|f^\star\|_{2,\mu}^2)\langle f^\star - f, f^\star\rangle_{2,\mu} + 4\|f-f^\star\|_{2,\mu}^2 \\
&= (\|f\|_{2,\mu}^2 - \|f^\star\|_{2,\mu}^2)(\|f\|_{2,\mu}^2 - \|f^\star\|_{2,\mu}^2 + 4\langle f^\star - f, f^\star\rangle_{2,\mu}) + 4\|f-f^\star\|_{2,\mu}^2 \\
&= \langle f-f^\star, f+f^\star\rangle_{2,\mu}\langle f-f^\star, f-3f^\star\rangle_{2,\mu} + 4\|f-f^\star\|_2^2 \\
&= \langle f-f^\star, f+f^\star\rangle_{2,\mu}\|f-f^\star\|_2^2 - \langle f-f^\star, f+f^\star\rangle_{2,\mu}\langle f-f^\star, 2f^\star\rangle_{2,\mu} + 4\|f-f^\star\|_{2,\mu}^2.
\end{aligned}$$

We have $\langle f - f^\star, f + f^\star \rangle_{2,\mu} \leq \|f\|_{2,\mu}^2 \leq \mu(\mathcal{X})$, $\|f^\star\|_{1,\mu} = 1 \leq \mu(\mathcal{X})^{1/2}$ and $\|f^\star\|_{2,\mu}^2 \leq \|f^\star\|_{\infty,\mu}\|f^\star\|_{1,\mu} \leq 1$. Then

$$\begin{aligned}\mathbb{E}((\gamma(f,Z) - \gamma(f^\star,Z))^2) &\leq (\mu(\mathcal{X}) + 2\|f + f^\star\|_{2,\mu}\|f^\star\|_{2,\mu} + 4)\|f - f^\star\|_{2,\mu}^2 \\ &\leq (\mu(\mathcal{X}) + 2\mu(\mathcal{X}) + 2 + 4)\|f - f^\star\|_{2,\mu}^2 \\ &= 3(\mu(\mathcal{X}) + 2)\|f - f^\star\|_{2,\mu}^2 \\ &= 3B\|f - f^\star\|_{2,\mu}^2.\end{aligned}$$

Let $\gamma_1 = \frac{1}{2B}(\gamma + B)$. Then $0, \leq \gamma_1(f, X) \leq 1$ almost surely for any $f \in M_1$. Moreover,

$$\mathcal{E}_1(f) := \mathbb{E}\left[\gamma_1(f,Z) - \gamma_1(f^\star,Z)\right] = \frac{1}{2B}\mathcal{E}(f)$$

and

$$\mathbb{E}((\gamma_1(f,Z) - \gamma_1(f^\star,Z))^2) \leq D\|f - f^\star\|_{2,\mu}^2$$

with $D = \frac{3}{4B} \leq \frac{3}{12}$, where we have used $\mu(\mathcal{X}) \geq 1$.

• For $\delta > 0$, we introduce

$$\omega_n(\delta) = \omega_n(M_1, f^\star, \delta) = \mathbb{E} \sup_{f \in M_1 \,|\, \|f - f^{M_1}\|_{2,\mu}^2 \leq \delta/D} \left| \frac{1}{n}\sum_{i=1}^n \gamma_1(f, Z_i) - \mathbb{E}(\gamma_1(f,Z)) \right|$$

Following [27] (Section 4.1 p.57), we introduce the sharp transformation $\sharp$ of the function ω:

$$\omega_n^\sharp(\varepsilon) = \inf\left\{ \delta > 0 \,,\, \sup_{\sigma \geq \delta} \frac{\omega_n(\sigma)}{\sigma} \leq \varepsilon \right\}.$$

According to Proposition 4.1 in [27], there exists absolute constants κ_1 and $\mathcal{A}$ such that for any $\varepsilon \in (0,1]$ and any $t > 0$, with probability at least $1 - \mathcal{A}\exp(-t)$,

$$\mathcal{E}_1(\hat{f}_n^{M_1}) \leq (1+\varepsilon)\mathcal{E}_1(f^{M_1}) + \frac{1}{D}\omega_n^\sharp\left(\frac{\varepsilon}{\kappa_1 D}\right) + \frac{\kappa_1 D}{\varepsilon}\frac{t}{n}. \tag{33}$$

The sharp transformation is monotonic: if $\Psi_1 \leq \Psi_2$ then $\Psi_1^\sharp \leq \Psi_2^\sharp$ (see Appendix A.3 in [27]). Thus it remains to find an upper bound on the sharp transformation of an upper bound on ω_n.

• We use standard symmetrization and contraction arguments for Rademacher variables. The Rademacher process indexed by the class M_1 is defined by

$$\mathrm{Rad}_n(f) = \frac{1}{n}\sum_{i=1} n\varepsilon_i f(X_i)$$

where the ε_i's are i.i.d. Rademacher random variables (that is, ε_i takes the values $+1$ and -1 with probability $1/2$ each) independent of the X_i's. By the symmetrization Inequality (see for instance Theorem 2.1 in [27]),

$$\omega_n(\delta) \leq 2\mathbb{E} \sup_{f \in M_1 \,|\, \|f - f^{M_1}\|_{2,\mu}^2 \leq \delta/D} \left|\mathrm{Rad}_n\left(\gamma(f,\cdot) - \gamma(f^{M_1},\cdot)\right)\right|.$$

We introduce the function

$$\Psi_n(\delta) = \mathbb{E} \sup_{f \in M_1 \,|\, \|f - f^{M_1}\|_{2,\mu}^2 \leq \delta} \left|\mathrm{Rad}_n(f - f^{M_1})\right|.$$

For bounded regression, using the contraction Lemma with Lipschitz constant equal to 2, (see for instance Theorem 2.3 in [27]),

$$\omega_n(\delta) \quad \leq \quad 8\Psi_n(\delta/D).$$

In the density estimation setting, we have $\gamma(f,Z) = \|f\|_{2,\mu}^2 - 2f(Z)$ and since the fluctuations of a constant function are obviously zero, we obtain

$$\begin{aligned}\omega_n(\delta) &\leq 4\mathbb{E} \sup_{f\in M_1 \,|\, \|f-f^{M_1}\|_{2,\mu}^2 \leq \delta/D} \left|\mathrm{Rad}_n(f-f^{M_1})\right| \\ &\leq 4\Psi_n(\delta/D). \end{aligned} \tag{34}$$

• We now introduce the subset of the L^2 ball centered at f^{M_1}

$$M_1(\delta, f^\star) = \{f - f^{M_1} \,|\, f \in M_1\,, \|f-f^{M_1}\|_{2,\mu}^2 \leq \delta\}.$$

In the density estimation setting, the empirical measure is ν_n. We also denote by ν_n the empirical measure in the regression setting (take $\nu = \mu$). The constant function $F = 2$ is an envelop for $M_1(\delta, f^\star)$ and $\|F\|_{2,\nu_n} = 2$. According to Proposition 3.3,

$$H\left(\varepsilon, M_1(\delta, f^\star), \|\cdot\|_{2,\nu_n}\right) \quad \leq \quad C_M \log\left(\frac{3|T^\star|L_{2,\nu_n}}{\varepsilon}\right) \mathbb{1}_{\varepsilon\leq 2} \quad \nu^{\otimes n}\text{-almost surely,}$$

where L_{2,ν_n} is defined by (10) for the measure ν_n and for $p = 2$. According to Proposition 3.6, L_{2,ν_n} satisfies

$$L_{2,\nu_n} \leq \sqrt{\nu_n(\mathcal{X})} L_{\infty,\nu_n} = L_{\infty,\nu_n}.$$

Here it is assumed that $\mathcal{H} \subset L^\infty(\mathcal{X})$ equipped with the norm $\|\cdot\|_\infty$. According to Proposition 3.6 we have $L_{\infty,\nu_n} \leq L_\infty \leq 1$, thus $L_{2,\nu_n} \leq 1$ and

$$\begin{aligned} H\left(\varepsilon, M_1(\delta, f^\star), \|\cdot\|_{2,\nu_n}\right) &\leq C_M \left[\log\left(\frac{4e}{\varepsilon}\right) + \log^+\left(\frac{3|T^\star|}{4e}\right)\right] \mathbb{1}_{\varepsilon\leq 4} \\ &\leq C_M \left[1 + \log^+\left(\frac{3|T^\star|}{4e}\right)\right] \log\left(\frac{4e}{\varepsilon}\right) \mathbb{1}_{\varepsilon\leq 4} \\ &\leq C_M\, a_T\, h\left(\frac{2}{\varepsilon}\right) \end{aligned}$$

with $a_T = 1 + \log^+\left(\frac{3|T^\star|}{4e}\right)$ and $h(u) := \log(2eu)\,\mathbb{1}_{u\geq\frac{1}{2}}$. We are now in position to apply Theorem A.1, which is given at the end of this section. We can take $\sigma^2 = \delta$ in Theorem A.1 because $\mathbb{E}_\nu(g^2(X)) \leq \delta$ for $g \in M_1(\delta, f^\star)$. Thus, there exists an absolute constant $\kappa_2 > 0$ such that

$$\Psi_n(\delta) \quad \leq \quad \kappa_2 \left[\sqrt{\frac{\delta}{n} C_M a_T h\left(\frac{2}{\sqrt{\delta}}\right)} \vee \left(\frac{2}{n} C_M a_T h\left(\frac{2}{\sqrt{\delta}}\right)\right)\right].$$

For regression, it can be easily checked that (see also Example 3 p.80 in [27])

$$\Psi_n^\sharp(\varepsilon) \leq \kappa_2 \frac{C_M a_T}{\varepsilon^2 n} \log\left(\frac{16e^2\varepsilon^2}{\kappa_2 C_M a_T}\right).$$

Similar calculations hold for density estimation. Together with Inequalities (33) and (34), and according to the properties of the sharp transformation (see Appendix A.3 in [27]), it gives that with probability at

least $1-\mathcal{A}\exp(-t)$,

$$\begin{aligned}\mathcal{E}_1(\hat{f}_n^{M_1}) &\leq (1+\varepsilon)\mathcal{E}_1(f^{M_1}) + \frac{1}{D}\left(8\Psi_n\left(\frac{\cdot}{D}\right)\right)^\sharp\left(\frac{\varepsilon}{\kappa_1 D}\right) + \frac{\kappa_1 D}{\varepsilon}\frac{t}{n} \\ &\leq (1+\varepsilon)\mathcal{E}_1(f^{M_1}) + (8\Psi_n)^\sharp\left(\frac{\varepsilon}{\kappa_1}\right) + \frac{\kappa_1 D}{\varepsilon}\frac{t}{n} \\ &\leq (1+\varepsilon)\mathcal{E}_1(f^{M_1}) + \kappa_3\frac{a_T C_M}{\varepsilon^2 n}\log\left(\frac{\kappa_4\varepsilon^2}{a_T C_M}\right) + \frac{\kappa_1 D}{\varepsilon}\frac{t}{n},\end{aligned}$$

where κ_3 and κ_4 are absolute constants. This completes the proof for $R=1$, by rewriting the risk bound for the excess risk $\mathcal{E} = B\mathcal{E}_1$.

• We now consider the more general situation where $M = M_r^T(\mathcal{H}^s)_R$ with $R \geq 1$. We first consider regression. We now assume that $\|Y\|_{\ell^\infty} \leq R$ almost surely. Let $f^\star$, f^M and $\hat{f}^M$ defined as in Section 4 for the observations $Z_1, \ldots, Z_n$. We consider the least squares regression problem for the normalized data $(X_1, Y_1/R), \ldots, (X_n, Y_n/R)$ with the functional set M_1. For this problem the oracle $f_1^\star$ satisfies $f_1^\star = f^\star/R$, the best approximation f^{M^1} on M_1 satisfies $f^{M^1} = f^M/R$ and the least squares estimator $\hat{f}^{M^1}$ also satisfies $\hat{f}^{M^1} = \hat{f}^M/R$. The risk bound (24) is valid for the normalized data (with $R=1$) and it directly gives (24) for $R \geq 1$. The same method applies for proving the risk bound in the density estimation case.

A.2.2 An adaptation of Theorem 3.12 in [27]

We consider the same framework as in [27]. We observe $X_1, \ldots, X_n$ according to the distribution ν and let ν_n be the empirical measure. Let $\mathcal{F}$ be a function space. Assume that the functions in $\mathcal{F}$ are uniformly bounded by a constant U and let $F \leq U$ denote a measurable envelop of $\mathcal{F}$. We assume that σ^2 is a number such that

$$\sup_{f\in\mathcal{F}} \mathbb{E}_\nu f^2 \leq \sigma^2 \leq \|F\|_{2,\nu}.$$

Let $h : [0,\infty) \mapsto [0,\infty)$ be a regularly varying function of exponent $0 \leq \alpha < 2$, strictly increasing for $u \geq 1/2$ and such that $h(u) = 0$ for $0 \leq u < 1/2$.

Next result is an adaptation of Theorem 3.12 in [27] which provides a better control on the constant $\kappa_h > 0$ when multiplying the metric entropy function by a constant. In particular in this version the constant $\kappa_h > 0$ depends only on h and not on c.

Theorem A.1 (Theorem 3.12 in [27])**.** *Let $c > 0$. If, for all $\varepsilon > 0$ and $n \geq 1$,*

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,\nu_n}) \leq ch\left(\frac{\|F\|_{2,\nu_n}}{\varepsilon}\right) \quad \nu^{\otimes n}\text{-almost surely,}$$

then there exists a constant $\kappa_h > 0$ that depends only on h such that

$$\mathbb{E}\sup_{f\in\mathcal{F}}|R_n(f)| \leq \kappa_h\left[\frac{\sigma}{\sqrt{n}}\sqrt{ch\left(\frac{\|F\|_{2,\nu}}{\sigma}\right)} \vee \frac{U}{n}ch\left(\frac{\|F\|_{2,\nu}}{\varepsilon}\right)\right].$$

Proof. The proof of Theorem 3.12 of [27] starts by applying Theorem 3.11 of [27]. As in [27] we assume without loss of generality that $U = 1$. In our context it gives

$$E := \mathbb{E}\sup_{f\in\mathcal{F}}|R_n(f)| \leq C\sqrt{c}n^{-1/2}\mathbb{E}\int_0^{2\sigma_n}\sqrt{h\left(\frac{\|F\|_{2,\nu_n}}{\varepsilon}\right)}d\varepsilon$$

where $\sigma_n = \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i)^2$ and where C is an universal numerical constant. By following the lines of the proof of [27], we find that E satisfies the following inequation

$$E \leq \sqrt{c}\kappa_{h,1}n^{-1} + \sqrt{c}\kappa_{h,2}n^{-1/2}\sigma\sqrt{h\left(\frac{\|F\|_{2,\nu}}{\sigma}\right)} + \sqrt{c}\kappa_{h,3}n^{-1/2}\sqrt{E}\sqrt{h\left(\frac{\|F\|_{2,\nu}}{\sigma}\right)}$$

where $\kappa_{h,1}$, $\kappa_{h,2}$ and $\kappa_{h,3}$ are positive numerical constants which only depends on the function h (see the proof of Koltchinskii for the expression of these three constants). Solving this inequation completes the proof. □

A.3 Proofs of Section 5

Proof of Theorem 5.2. According to Inequality (22) of Proposition 4.9, for any $t > 0$ and any $m \in \mathcal{M}$, one has with probability larger than $1 - \exp(-t)$,

$$\sup_{f \in M_m} -\bar{\mathcal{R}}_n(f) \leq \lambda_m \sqrt{\frac{C_m}{n}} + 2B\sqrt{\frac{t}{2n}}.$$

Then with probability larger than $1 - \sum_{m \in \mathcal{M}} \exp(-x_m - t) = 1 - \Sigma \exp(-t)$, it holds

$$-\bar{\mathcal{R}}_n(\hat{f}_{\hat{m}}) \leq \sup_{f \in M_{\hat{m}}} -\bar{\mathcal{R}}_n(f) \leq \lambda_{\hat{m}} \sqrt{\frac{C_{\hat{m}}}{n}} + 2B\sqrt{\frac{t + x_{\hat{m}}}{2n}},$$

which together with (26) implies that

$$\mathcal{E}(\hat{f}_{\hat{m}}) \leq \mathcal{E}(f_m) + \bar{\mathcal{R}}_n(f_m) + \lambda_{\hat{m}} \sqrt{\frac{C_{\hat{m}}}{n}} + 2B\sqrt{\frac{x_{\hat{m}}}{2n}} - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m) + 2B\sqrt{\frac{t}{2n}}$$

holds for all $m \in \mathcal{M}$. Then, with the condition (27) on the penalty function, the upper bound

$$\mathcal{E}(\hat{f}_{\hat{m}}) \leq \mathcal{E}(f_m) + \bar{\mathcal{R}}_n(f_m) + \mathrm{pen}(m) + 2B\sqrt{\frac{t}{2n}}$$

holds for all $m \in \mathcal{M}$ simultaneously, with probability larger than $1 - \Sigma \exp(-t)$. Next, integrating with respect to t gives

$$\mathbb{E}\left[0 \vee \left(\mathcal{E}(\hat{f}_{\hat{m}}) - \mathcal{E}(f_m) - \bar{\mathcal{R}}_n(f_m) - \mathrm{pen}(m)\right)\right] \leq 2B\Sigma\sqrt{\frac{2\pi}{n}\frac{1}{4}}.$$

Finally, since $\bar{\mathcal{R}}_n(f_m)$ has zero mean, for any $m \in \mathcal{M}$,

$$\mathbb{E}(\mathcal{E}(\hat{f}_{\hat{m}})) \leq \mathcal{E}(f_m) + \mathrm{pen}(m) + B\Sigma\sqrt{\frac{\pi}{2n}},$$

and we conclude by taking the infimum over $m \in \mathcal{M}$. □

Proof of Theorem 5.3. The proof is adapted from Theorem 6.5 in [27], which corresponds to a alternative statement of Theorem 8.5 in [28]. We follow the lines of Section 6.3 in [27] (p.107-108).

We first consider the case $R = 1$ and we consider the normalized contrast γ_1 and the normalized risk $\mathcal{E}_1$ as for the proof of Proposition 4.12. We have shown that

$$\mathbb{E}\left[\gamma_1(f, Z) - \gamma_1(f^\star, Z)\right]^2 \leq D\|f - f^\star\|_2^2$$

where D does not depend on the model M_m. Next, it has also been shown in the proof of Proposition 4.12, that for $\varepsilon \in (0, 1]$,

$$\omega_n^\sharp(\varepsilon) \leq \kappa \frac{a_m C_M}{n\varepsilon^2} \log^+\left(\frac{n\varepsilon^2}{a_m C_M}\right)$$

with $a_m = 1 + \log^+\left(\frac{3|T_m^\star|}{4e}\right)$ and where κ is an absolute constant. We consider the penalized criterion (25) with a penalty of the form

$$\mathrm{pen}(m) = \kappa_1 \frac{a_m C_m}{n\varepsilon^2} \log^+ \frac{n\varepsilon^2}{a_m C_m} + \kappa_2 \frac{x_m}{n\varepsilon}.$$

Theorem 6.5 of [27] can be applied here with $\bar{\delta}_n^\varepsilon(m) = \tilde{\delta}_n^\varepsilon(m) = \hat{\delta}_n^\varepsilon(m) = \kappa \frac{a_m C_m}{n\varepsilon^2} \log \frac{n\varepsilon^2}{a_m C_m} + K \frac{x_m + t}{n\varepsilon}$ (and thus $p_m = 0$ in the theorem) and we also note that for any $t > 0$ the penalty can be rewritten

$$\mathrm{pen}(m) = K_1 \left[\frac{a_m C_m}{n\varepsilon^2} \log \frac{n\varepsilon^2}{a_m C_m} + \frac{x_m + t}{n\varepsilon} \right].$$

Finally, according to Theorem 6.5 in [27], there exist numerical constants K_1, K_2 and K_3 such that for any $t > 0$,

$$P\left(\mathcal{E}_1(\hat{f}_{\hat{m}}) \leq \frac{1+\varepsilon}{1-\varepsilon} \inf_{m \in \mathcal{M}} \left\{ \mathcal{E}_1(\hat{f}_{\hat{m}}) + K_2 \left[\frac{a_m C_m}{n\varepsilon^2} \log \frac{n\varepsilon^2}{a_m C_m} + \frac{x_m + t}{n\varepsilon} \right] \right\} \right) \leq K_3 \sum_{m \in \mathcal{M}} \exp(-t - x_m)$$

Under Assumption 5.1, we easily derive the oracle bound (28) by rewriting it for the constrast γ and then by integrating this probability bound with respect to t. This bound generalizes to the case $R \geq 1$ as in the proof of Proposition 4.12 □