Workshop "Apprentissage et simulation en grande dimension",
Airbus Group, June 24-26, 2019

Deep tensor networks

Part IV: Unsupervised learning

Anthony Nouy

Centrale Nantes, Laboratoire de Mathématiques Jean Leray

Unsupervised learning

We consider the problem of the estimation of the probability distribution ρ of a high-dimensional random vector $X = (X_1, \ldots, X_d)$ from samples of the distribution.

The distribution is characterized by a density

$$f(x_1, \ldots, x_d)$$

with respect to a measure μ.

In high dimension, this requires using suitable model classes (or hypothesis sets).

Typical model classes

- multiplicative models

$$f^1(x_1) \dots f^d(x_d),$$

  which corresponds to the hypothesis that the components of X are independent,

- general multiplicative models

$$\prod_{\alpha \in T} f^\alpha(x_\alpha),$$

  where T is a collection of subsets $\alpha \subset \{1, \dots, d\}$ and x_α denotes the corresponding group of variables. This includes graphical models,

- mixture models

$$\sum_{k=1}^{K} \gamma_k f_k(x),$$

  with $\sum_{k=1}^{K} \gamma_k = 1$, and the f_k in suitable model classes (possibly of different types). For example, a mixture of multiplicative models takes the form

$$\sum_{k=1}^{K} \gamma_k f_k^1(x_1) \dots f_k^d(x_d).$$

Outline

Outline

Representation of probabilistic models in tensor formats

We discuss the representation in tensor formats of the distribution of a random variable $X = (X_1, \ldots, X_d)$ with density $f(x)$ with respect to some measure μ with support $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_d$ in $\mathbb{R}^d$.

This includes the case of a discrete random variable taking values in a subset of a finite or countable set $\mathcal{X}$, with measure $\rho = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x)\delta_x$, by letting

$$f(x) := \mathbb{P}(X = x) \quad \text{and} \quad \mu := \sum_{x \in \mathcal{X}} \delta_x,$$

and $f \in \mathbb{R}^{\mathcal{X}}$ is identified with an algebraic tensor $\mathbb{R}^{\#\mathcal{X}_1 \times \ldots \times \#\mathcal{X}_d}$.

Representation of mixtures

Assume X is a mixture of m random variables $Z^i = (Z_1^i, \ldots, Z_d^i)$ with weights γ_i, $1 \leq i \leq m$, such that $\sum_{i=1}^m \gamma_i = 1$. Then

$$f(x) = \sum_{i=1}^{m} \gamma_i f^i(x).$$

where f^i is the density of Z^i with respect to μ.

For any $\alpha \subset D$,

$$\text{rank}_\alpha(f) \leq \sum_{i=1}^{m} \text{rank}_\alpha(f^i),$$

and therefore, for any tree T,

$$\text{rank}_T(f) \leq \sum_{i=1}^{m} \text{rank}_T(f^i).$$

Assuming that Z^i has independent components, we have $f^i(x) = f_i^1(x_1) \ldots f_i^d(x_d)$ with $\text{rank}_T(f^i) = (1, \ldots, 1)$, and therefore, $\text{rank}_T(f) \leq (m, \ldots, m)$.

Representation of Markov processes

Consider a discrete time Markov process $X = (X_1, \ldots, X_d)$ whose density is given by

$$f(x) = f_{d|d-1}(x_d|x_{d-1}) \ldots f_{2|1}(x_2|x_1) f_1(x_1),$$

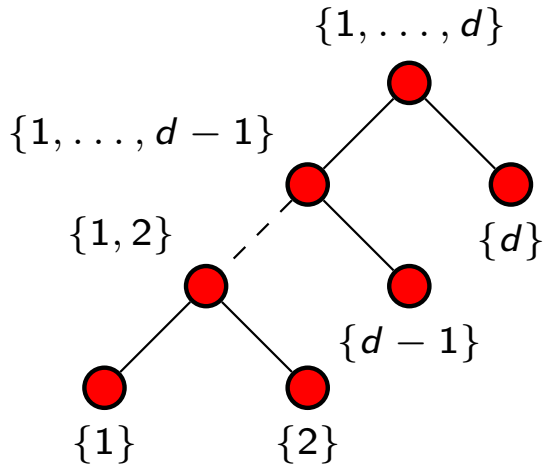
where f_1 is the density of X_1 and $f_{i|i-1}(\cdot|x_{i-1})$ is the density of X_i knowing $X_{i-1} = x_{i-1}$.

Let m_i be the rank of the bivariate function $(t, s) \mapsto f_{i|i-1}(t|s)$, $i = 2, \ldots, d$. We note that

- $\text{rank}_{\{1\}}(f) = \text{rank}(f_{2|1}) = m_2$,
- $\text{rank}_{\{d\}}(f) = \text{rank}(f_{d|d-1}) = m_d$
- for $2 \leq \nu \leq d-1$, $\text{rank}_{\{\nu\}}(f) \leq \text{rank}(f_{\nu|\nu-1})\,\text{rank}(f_{\nu+1|\nu}) = m_\nu m_{\nu+1}$.
- for $1 < \nu < d$, $\text{rank}_{\{1,\ldots,\nu\}}(f) = \text{rank}(f_{\nu+1|\nu})$.

Representation of Markov processes

Consider the linear tree $T = \{\{1,\dots,d\},\{1\},\dots,\{d\},\{1,2\},\dots,\{1,\dots,d-1\}\}$. We have $\text{rank}_T(f) \leq \{1, m_2, m_2m_3, \dots, m_{d-1}m_d, m_d, m_3, \dots, m_d\}$.



Letting $m = \max_i m_i$, we then have

$$\text{rank}_\alpha(f) \leq m^2 \text{ for the leaves } \alpha \text{ and}$$
$$\text{rank}_\alpha(f) \leq m^2 \text{ for internal nodes } \alpha,$$

and a complexity in $O(m^4)$ for the representation in the tree-based format.

Note that the choice of tree is here crucial. By considering a permutation $T_\sigma = \{\sigma(\alpha) : \alpha \in T\}$ of T, we may observe an exponential dependance in d for the T_σ-rank.

Graphical models

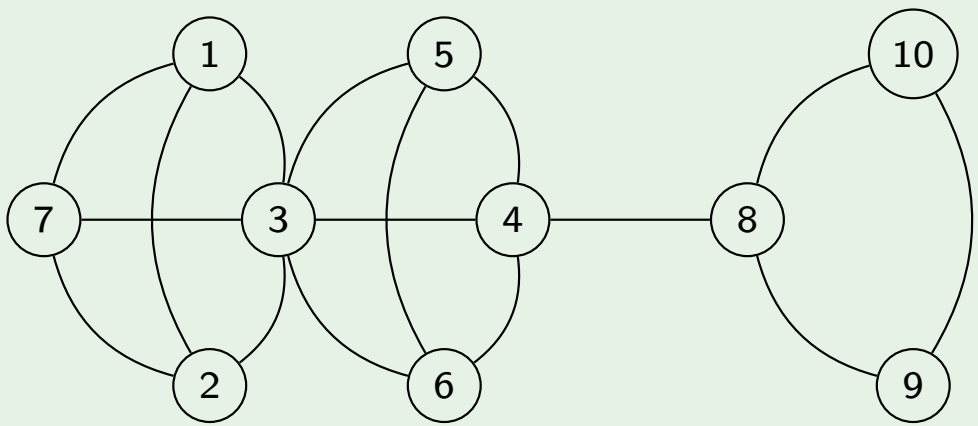
Let us consider a graphical model with a density of the form

$$f(x) = \prod_{\beta \in \mathcal{C}} g_\beta(x_\beta)$$

where $\mathcal{C} \subset 2^D$ represents the cliques of a graph G with nodes $\{1\}, \ldots, \{d\}$.

Example

Graph G with cliques $\mathcal{C} = \{\{1,2,3,7\}, \{3,4,5,6\}, \{4,8\}, \{8,9,10\}\}$



and corresponding density

$$f(x) = g_{1,2,3,7}(x_1, x_2, x_3, x_7) g_{3,4,5,6}(x_3, x_4, x_5, x_6) g_{4,8}(x_4, x_8) g_{8,9,10}(x_8, x_9, x_{10})$$

Graphical models

Consider $\alpha \subset D$. Note that if $\alpha \in \mathcal{C}$,

$$\mathrm{rank}_\alpha(g_\alpha) = 1.$$

Also, for a clique $\beta \in \mathcal{C}$ such that either $\beta \subset \alpha$ or $\beta \subset \alpha^c$,

$$\mathrm{rank}_\alpha(g_\beta) = 1.$$

Then let $\mathcal{C}_\alpha$ be the set of cliques that intersects both α and α^c,

$$\mathcal{C}_\alpha = \{\beta \in \mathcal{C} : \beta \cap \alpha \neq \emptyset, \beta \cap \alpha^c \neq \emptyset\}.$$

Since $\mathcal{C} \setminus \mathcal{C}_\alpha = \{\beta \in \mathcal{C} : \beta \subset \alpha^c \text{ or } \beta \subset \alpha\}$, we have

$$\mathrm{rank}_\alpha(f) = \mathrm{rank}_\alpha(\prod_{\beta \in \mathcal{C}_\alpha} g_\beta) \leq \prod_{\beta \in \mathcal{C}_\alpha} \mathrm{rank}_\alpha(g_\beta).$$

Assuming that the α-rank of all functions g_β are bounded by m, we have

$$\mathrm{rank}_\alpha(f) \leq m^{\#\mathcal{C}_\alpha}.$$

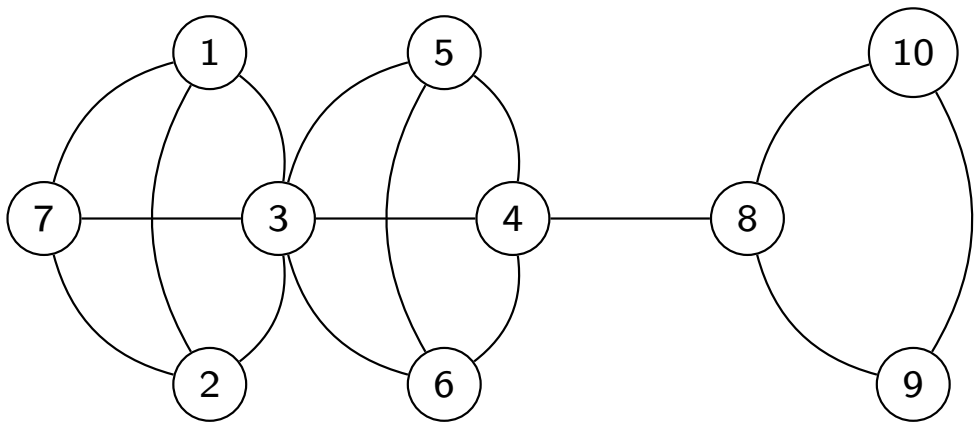
For the representation of f in tree-based format, a tree T should be chosen such that $\#\mathcal{C}_\alpha$ is small for all $\alpha \in T$.

Graphical models

As an example, assume that the discrete random variables X_ν taking 5 possible instances $\{1, \ldots, 5\}$, and consider the graphical model

$$f(x) = g_{1,2,3,7}(x_1, x_2, x_3, x_7) g_{3,4,5,6}(x_3, x_4, x_5, x_6) g_{4,8}(x_4, x_8) g_{8,9,10}(x_8, x_9, x_{10})$$

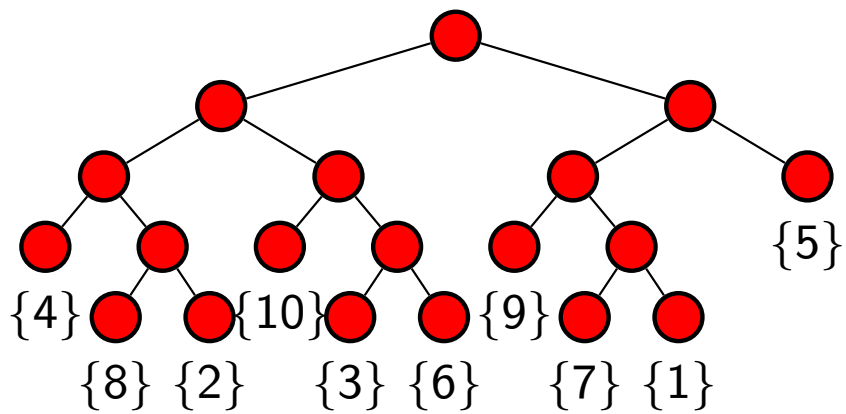
f is identified with a tensor of size $5^{10} = 9,765,625$. The entries of the tensors g_α are generated randomly.



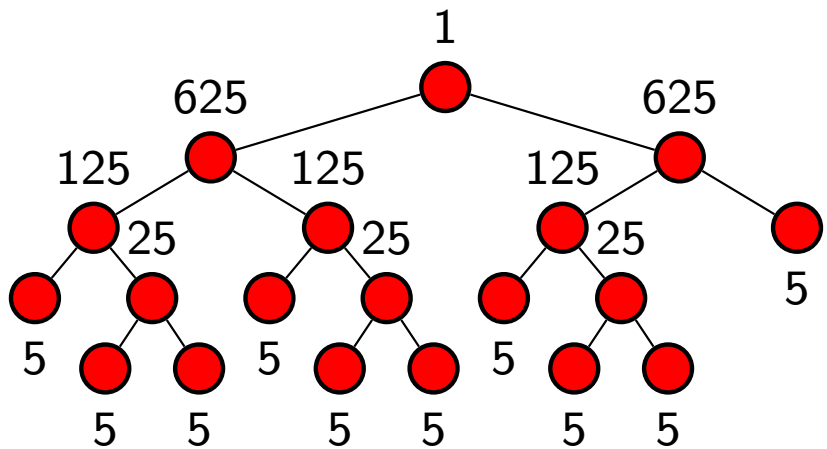
Graphical models

We first consider a random binary tree and compute a representation of the graphical model at precision 10^{-13} in the corresponding tree-based format.

We observe a storage complexity of $10,595,875$, higher than the storage of the full tensor.

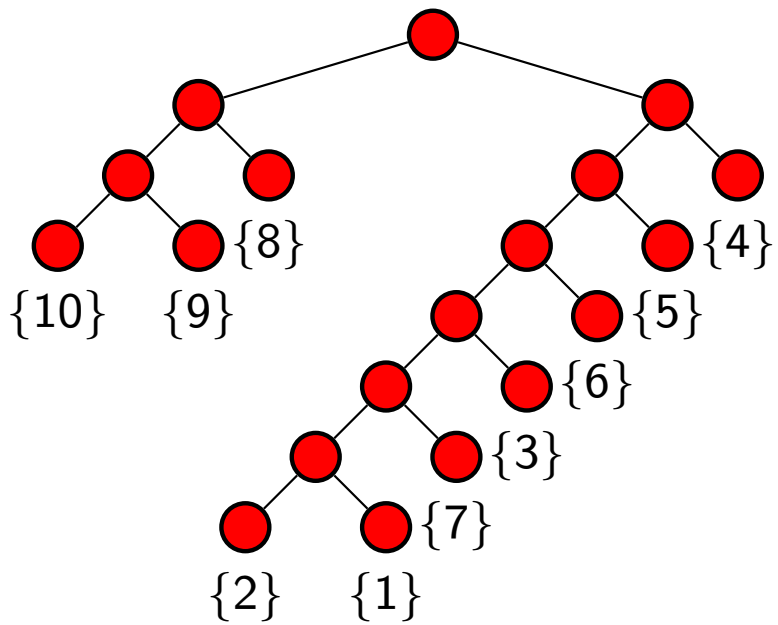


(a) Dimension tree.

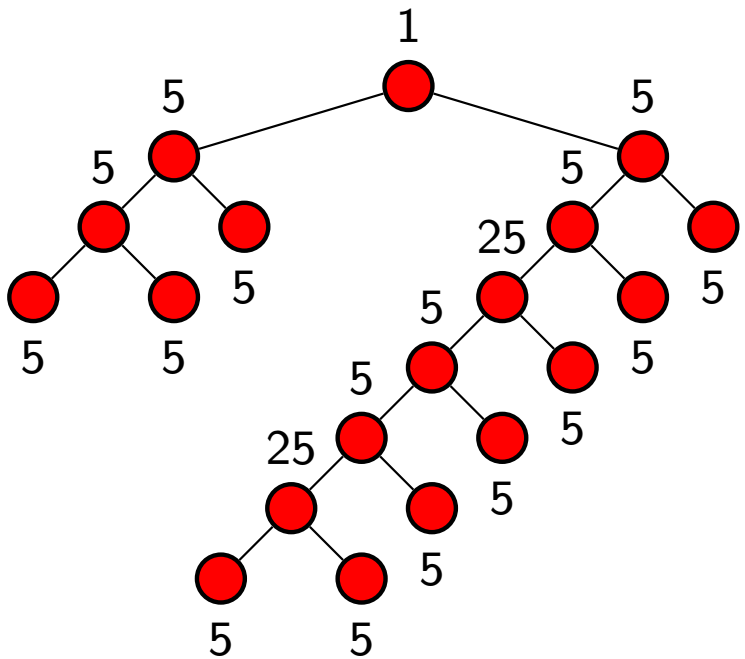


(b) Representation rank.

We now run a tree optimization algorithm and we obtain a new representation (at the same precision) with a storage complexity of $3,275$.



(c) Optimized dimension tree.



(d) Representation rank.

Representation of cumulative distribution function

The density is related to the cumulative distribution function by

$$F(x) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} f(t_1, \dots, t_d) d\mu_1(t_1) \dots d\mu_d(t_d).$$

If the density admits a representation

$$f(x) = \sum_{i \in I} a_i \prod_{\nu=1}^{d} \phi_{i_\nu}^{\nu}(x_\nu),$$

then

$$F(x) = \sum_{i \in I} a_i \prod_{\nu=1}^{d} \psi_{i_\nu}^{\nu}(x_\nu), \quad \text{with } \psi_{i_\nu}^{\nu}(x_\nu) = \int_{-\infty}^{x_\nu} \phi_{i_\nu}^{\nu}(t) d\mu_\nu(t).$$

Then if f has an exact representation in tree-based format $\mathcal{T}_r^T$, then F also has an exact representation in the same format $\mathcal{T}_r^T$, and

$$\text{rank}_T(F) \leq \text{rank}_T(f)$$

.

Representation of copulas

By Sklar's theorem, the cumulative distribution function F of $(X_1, \ldots, X_d)$ can be written

$$F(x) = C(F_1(x_1), \ldots, F_d(x_d)),$$

where the function C is a copula of $(X_1, \ldots, X_d)$.

If the copula admits a representation

$$C(u) = \sum_{i \in I} a_i \prod_{\nu=1}^{d} \phi_{i_\nu}^{\nu}(u_\nu),$$

then

$$F(x) = \sum_{i \in I} a_i \prod_{\nu=1}^{d} \phi_{i_\nu}^{\nu}(F_\nu(x_\nu)).$$

Then if C admits an exact representation in tree-based format $\mathcal{T}_r^T$, F also admits an exact representation in $\mathcal{T}_r^T$.

Outline

Contrast function and risk

We consider the approximation of a probability distribution of a random vector X from independent samples of this distribution.

We assume that the distribution of X has a density f with respect to a measure μ with support $\mathcal{X}$.

We introduce a contrast function $\gamma : L^0_\mu(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}$ and the associated risk functional $\mathcal{R} : L^0_\mu(\mathcal{X}) \to \mathbb{R}$ defined by

$$\mathcal{R}(v) = \mathbb{E}(\gamma(v, X)) = \int \gamma(v, x) d\mu(x),$$

such that the minimizer of $\mathcal{R}$ over the set of μ-measurable functions is the density f.

Empirical risk minimization

Given independent samples $\{x_i\}_{i=1}^n$ of X, an approximation $\hat{f}_n^F$ of the density is then obtained by minimizing the empirical risk (a statistical estimation of the risk)

$$\hat{\mathcal{R}}_n(v) = \frac{1}{n} \sum_{i=1}^{n} \gamma(v, x_i)$$

over a certain model class F, here the class of functions in tree-based tensor format.

Maximum likelihood estimation

Choosing the contrast function as $\gamma(v, x) = -\log(v(x))$ leads to

$$\mathcal{R}(v) = \int -\log(v(x)) f(x) d\mu(x) = \int (-\log(f(x)) - \log(\frac{v(x)}{f(x)})) f(x) d\mu(x)$$

so that

$$\mathcal{R}(v) = \mathcal{R}(f) + D_{\mathsf{KL}}(f \| v),$$

with $D_{\mathsf{KL}}(f \| v)$ the Kullback-Leibler divergence between f and v.

The empirical risk

$$\hat{\mathcal{R}}_n(v) = -\frac{1}{n} \sum_{i=1}^{n} \log(v(x_i))$$

corresponds to the log-likelihood.

L^2 density estimation

With the contrast function as

$$\gamma(v, x) = \|v\|^2_{L^2_\mu} - 2v(x),$$

we have

$$\mathcal{R}(v) = \mathcal{R}(f) + \|f - v\|^2_{L^2_\mu},$$

so that the minimization of $\mathcal{R}(v)$ is equivalent to the minimization of the distance (in L^2_μ norm) between v and the density f.

The empirical risk is

$$\hat{\mathcal{R}}_n(v) = \|v\|^2_{L^2_\mu} - \frac{2}{n}\sum_{i=1}^{n} v(x_i).$$

Note that for discrete random variables, $\|v\|^2_{L^2_\mu} = \sum_{x\in\mathcal{X}} v(x)^2$ which coincides with the Frobenius norm when identifying v with a tensor in $\mathbb{R}^{\mathcal{X}}$.

L^2 density estimation

Consider as a model class a finite-dimensional space F of L^2_μ with an orthonormal basis $\{\varphi_k\}_{k=1}^m$.

The minimizer f^F of the risk is given by the orthogonal projection onto F

$$f^F(x) = \sum_{k=1}^m a_k \varphi_k(x), \quad a_k = (\varphi_k, f)_{L^2_\mu} = \int \varphi_k(x) f(x) d\mu(x) = \mathbb{E}(\varphi_k(X)).$$

The empirical risk for $v = \sum_{k=1}^m a_k \varphi_k(x) = \varphi(x)^T a$ is

$$\hat{\mathcal{R}}_n(v) = \|a\|_2^2 - \frac{2}{n} \sum_{i=1}^n \varphi(x)^T a,$$

so that the minimizer f_n^F of the empirical risk is given by

$$f_n^F(x) = \sum_{k=1}^m \hat{a}_k \varphi_k(x), \quad \hat{a}_k = \frac{1}{n} \sum_{i=1}^n \varphi_k(x_i).$$

Tree-based formats

Let V_ν be a subspace of $L^2_{\mu_\nu}(\mathcal{X}_\nu)$ and let us consider the tensor space

$$V = V^1 \otimes \ldots \otimes V^d,$$

a subspace of $L^2_\mu(\mathcal{X})$.

We let $\{\phi^\nu_{i_\nu} : i \in I^\nu\}$ be a basis of V_ν, typically polynomials, wavelets...

We consider the model class of tree-based tensors

$$\mathcal{T}_r^T = \{v \in V : \mathrm{rank}_T(v) \leq r\}$$

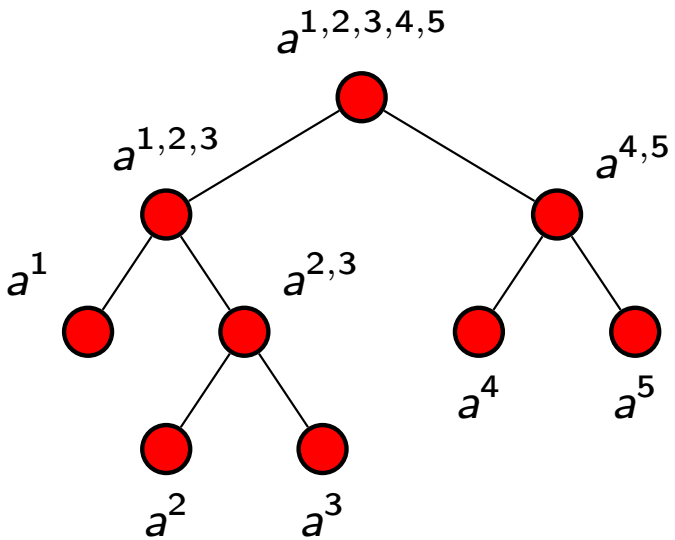
where T is a dimension partition tree and r a tuple of ranks.

Tree-based formats

A function v in $\mathcal{T}_r^T$ admits the following multilinear parametrization

$$v(x) = \sum_{\substack{i_\alpha \in I^\alpha \\ \alpha \in \mathcal{L}(T)}} \sum_{\substack{1 \le k_\beta \le r_\beta \\ \beta \in T}} \prod_{\alpha \in T \setminus \mathcal{L}(T)} a^\alpha_{(k_\beta)_{\beta \in S(\alpha)}, k_\alpha} \prod_{\alpha \in \mathcal{L}(T)} a^\alpha_{i_\alpha, k_\alpha} \phi^\alpha_{i_\alpha}(x_\alpha) = \Psi(x)((a^\alpha)_{\alpha \in T})$$

where each parameter a^α is in a tensor space $\mathbb{R}^{K^\alpha}$ and $\Psi(x)$ is a multilinear map.



For a given α, the partial map

$$a^\alpha \in \mathbb{R}^{K^\alpha} \mapsto \Psi(\cdot)((a^\alpha)_{\alpha \in T})$$

is linear and can be identified with a tensor $\Psi^\alpha(x) \in \mathbb{R}^{K^\alpha}$ such that

$$\Psi(\cdot)((a^\alpha)_{\alpha \in T}) = \sum_{k \in K^\alpha} \Psi^\alpha_k(x) a^\alpha_k := (\Psi^\alpha(x), a^\alpha)_{\ell_2}.$$

Learning with tree-based formats

The empirical risk minimization problem over the model class $\mathcal{T}_r^T$

$$\min_{(a^\alpha)_{\alpha \in T}} \frac{1}{n} \sum_{i=1}^{n} \gamma(\Psi(\cdot)((a^\alpha)_{\alpha \in T}), x_i)$$

can be solved using an alternating minimization algorithm, solving at each step

$$\min_{a^\alpha} \frac{1}{n} \sum_{i=1}^{n} \gamma((\Psi^\alpha(\cdot), a^\alpha)_{\ell_2}, x_i)$$

for fixed a^β, $\beta \in T \setminus \{\alpha\}$.

L^2 density estimation with tree-based formats

For a given α, it is possible to obtain a representation of a tensor v in $\mathcal{T}^T$ such that

$$v(x) = \sum_{k \in K^\alpha} \Psi_k^\alpha(x) a_k^\alpha,$$

where $\{\Psi_k^\alpha(x)\}_{k \in K^\alpha}$ form an orthonormal system in L^2_μ.

The tensor a^α minimizing the empirical risk therefore admits an explicit expression

$$a^\alpha = \frac{1}{n} \sum_{i=1}^{n} \Psi^\alpha(x_i).$$

Exploiting sparsity

Sparsity in the parameters can be easily exploited. For a given subset $A \subset K^\alpha$, the optimization of the empirical risk over the set of tensors

$$\{a^\alpha \in \mathbb{R}^{K^\alpha} : a_k^\alpha = 0 \text{ for } k \notin A\},$$

yields a solution

$$a_k^\alpha = \frac{1}{n} \sum_{i=1}^n \Psi^\alpha(x_i)_k \mathbf{1}_{k \in A}.$$

The associated risk can be evaluated by cross-validation using the leave-one estimate

$$\mathcal{R}_n^{loo}(v) = \frac{-n^2}{(1-n)^2} \|v\|^2 + \frac{2n-1}{n(n-1)^2} \sum_{i=1}^n \sum_{k \in A} (\Psi_k^\alpha(x_i))^2$$

In practice, we consider a collection of candidate sets $A_1, \ldots, A_M$ in K^α and retain the one that minimizes the cross-validation estimate.

Rank adaptation

We apply the strategy introduce in the supervised learning setting.

We construct a sequence of approximations in tree-based format

$$u^m \in \mathcal{T}_{r^m}^T$$

with increasing ranks

$$\begin{cases} r_\alpha^{m+1} = r_\alpha^m + 1 & \text{if } \alpha \in T_m^\theta \\ r_\alpha^{m+1} = r_\alpha^m & \text{if } \alpha \notin T_m^\theta \end{cases}$$

where T_m^θ is a subset of nodes in T which is selected such that

$$T_m^\theta = \left\{ \alpha : \varepsilon_{r_\alpha^m}^\alpha(\tilde{u}) \geq \theta \max_{\beta \in T} \varepsilon_{r_\beta^m}^\beta(\tilde{u}) \right\}$$

where $\varepsilon_{r_\beta^m}^\beta(\tilde{u})$ is an estimation of the truncation error

$$\varepsilon_{r_\alpha^m}^\alpha(u) = \min_{\text{rank}_\alpha(v) \leq r_\alpha^m} \mathcal{R}(v) - \mathcal{R}(u)$$

with $\tilde{u}$ obtained by a correction of the current approximation u^m.

Rank and tree adaptation

For rank and tree adaptation, we apply the same strategy as in the supervised learning setting.

We start with an initial tree T and learn an approximation $v \in \mathcal{T}_r^T$ with rank $r = (1, ..., 1)$. Then we repeat the following steps:

- compute a correction of v to obtain a better approximation $\tilde{u}$, compute the truncation errors

$$\eta_\alpha = \min_{\mathrm{rank}_\alpha(v) \leq r_\alpha^m} \mathcal{R}(v) - \mathcal{R}(\tilde{u})$$

  and increase by one the ranks r_α for

$$\alpha \in T_\theta = \left\{ \alpha : \eta_\alpha \geq \theta \max_{\beta \in T} \eta_\beta \right\}$$

- learn an approximation v in the format $\mathcal{T}_r^T$ (with a good initialization),
- optimize the tree for reducing the storage complexity of v (using a stochastic algorithm): if a better tree T' is found, change the representation of v and set $r = \mathrm{rank}_{T'}(v)$ and $T = T'$.

Illustration: truncated normal distribution

We consider truncated normal distribution with zero mean and covariance matrix Σ. Its support is $\mathcal{X} = \times_{\nu=1}^{6}[-5\sigma_\nu, 5\sigma_\nu]$, with $\sigma_\nu^2 = \Sigma_{\nu\nu}$, and its density (with respect to Lebesgue measure) is

$$f(x)\mathrm{d}\mu(x) \sim \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)\mathbf{1}_{x\in\mathcal{X}},$$

We consider polynomial approximation spaces $V^\nu = \mathbb{P}_{50}(\mathcal{X}_\nu)$.

Illustration: truncated normal distribution

Consider

$$\Sigma = \begin{pmatrix} 2 & 0 & 0.5 & 1 & 0 & 0.5 \\ 0 & 1 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}.$$

After a permutation $(3, 6, 1, 4, 2, 5)$ of its rows and columns, it comes the matrix

$$\begin{pmatrix} 2 & 1 & 0.5 & 0 & 0 & 0 \\ 1 & 2 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}$$

(X_1, X_3, X_4, X_6) and (X_2, X_5) are independent, as well as X_4 and (X_3, X_6), so that

$$f(x) = f_{1,3,4,6}(x_1, x_3, x_4, x_6) f_{2,5}(x_2, x_5) = f_{4,1}(x_4, x_1) f_{1,3,6}(x_1, x_3, x_6) f_{2,5}(x_2, x_5)$$

.

Illustration: truncated normal distribution

n	Risk $\times 10^{-2}$	L^2-error	T	$C(T, r)$
10^2	$[-5.50, 119]$	$[0.53, 4.06]$	Fig. (a)	$[311, 311]$
10^3	$[-7.29, -5.93]$	$[0.22, 0.47]$	Fig. (b)	$[311, 637]$
10^4	$[-7.60, -6.85]$	$[0.11, 0.33]$	Fig. (c)	$[521, 911]$
10^5	$[-7.68, -7.66]$	$[0.04, 0.07]$	Fig. (c)	$[911, 1213]$
10^6	$[-7.70, -7.69]$	$[0.01, 0.01]$	Fig. (c)	$[1283, 1546]$

Table: Ranges over 10 trials

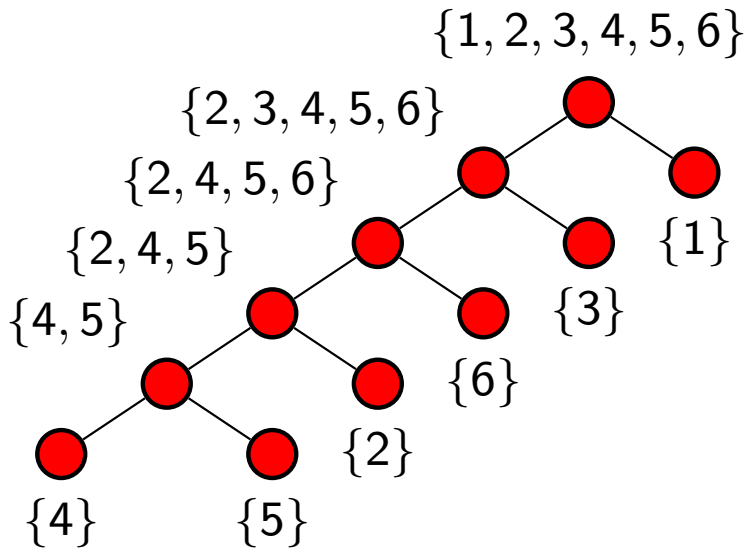


Figure: (a) Best tree over 10 trials for $n = 10^2$