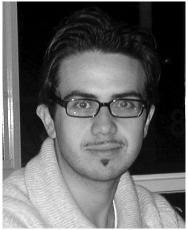
# Distributed multi-agent Gaussian regression via finite-dimensional approximations

Gianluigi Pillonetto Luca Schenato Damiano Varagnolo

LULEÅ UNIVERSITY OF TECHNOLOGY



# Roadmap

- Gaussian regression
- Karhunen-Loève expansions
- Statistical bounds

# Function estimation

$$f : \mathcal{X} \to \mathbb{R} \tag{1}$$

$$y_m = f\left(x_m\right) + \nu_m \qquad m = 1, \dots, M \tag{2}$$

$$x_m \sim \mu(\mathcal{X}) \;\; \text{i.i.d.} \qquad \nu_m \sim \mathcal{N}\left(0, \sigma_\nu^2\right) \qquad m = 1, \dots, M \tag{3}$$

$$\{x_m\}_{m=1}^{M} \qquad \{\nu_m\}_{m=1}^{M} \qquad \text{mutually independent} \tag{4}$$

*Problem: estimate $f$ starting from $\{x_m, y_m\}$*

# Function estimation – parametric approach

$$y_m = f\left(x_m; \theta\right) + \nu_m \qquad \textit{(known structure or set of alternative structures)} \tag{5}$$

# Function estimation – parametric approach

$$y_m = f\left(x_m; \theta\right) + \nu_m \qquad \textit{(known structure or set of alternative structures)} \tag{5}$$

Least squares (classic approach):

$$\theta^* = \arg\min_{\widetilde{\theta} \in \Theta} \sum_m \left( y_m - f\left(x_m; \widetilde{\theta}\right) \right)^2 \tag{6}$$

# Function estimation – parametric approach

$$y_m = f\left(x_m; \theta\right) + \nu_m \qquad \textit{(known structure or set of alternative structures)} \tag{5}$$

Least squares (classic approach):

$$\theta^* = \arg\min_{\widetilde{\theta}\in\Theta} \sum_m \left(y_m - f\left(x_m; \widetilde{\theta}\right)\right)^2 \tag{6}$$

Potential problems:

- non-convexity
- model order selection

# Nonparametric approach *(cast the problem as a Gaussian regression)*

$$f \sim \mathcal{N}(0, K) \qquad\qquad K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{so that} \quad \mathbb{E}\left[f(x) f(x')\right] = K\left(x, x'\right) \tag{7}$$

# Nonparametric approach *(cast the problem as a Gaussian regression)*

$$f \sim \mathcal{N}(0, K) \qquad K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{so that} \quad \mathbb{E}\left[f(x) f(x')\right] = K\left(x, x'\right) \tag{7}$$

Examples:

- Brownian motion: $K\left(x, x'\right) = \min\left(x, x'\right) \qquad \mathcal{X} = [0, 1]$
- Radial basis: $K\left(x, x'\right) = \exp\left(-\left\|x - x'\right\|^2\right) \qquad \mathcal{X} \subseteq \mathbb{R}^m$

# Nonparametric approach *(cast the problem as a Gaussian regression)*

$$f \sim \mathcal{N}(0, K) \qquad K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{so that} \quad \mathbb{E}\left[f(x)f(x')\right] = K\left(x, x'\right) \tag{7}$$

$$x_m \sim \mu(\mathcal{X}) \;\; \text{i.i.d.} \qquad y_m = f\left(x_m\right) + \nu_m \qquad \nu_m \sim \mathcal{N}\left(0, \sigma_\nu^2\right) \qquad m = 1, \ldots, M \tag{8}$$

$$\{x_m\}_{m=1}^{M} \qquad \{\nu_m\}_{m=1}^{M} \qquad f \qquad \text{mutually independent} \tag{9}$$

# Maximum a posteriori estimator

$$\widehat{f}_{\mathrm{MAP}}(x) = \mathbb{E}\left[ f(x) \;\middle|\; \{x_m, y_m\} \right] \qquad \textit{(also MV)}$$

# Maximum a posteriori estimator

$$\begin{aligned} \widehat{f}_{\mathrm{MAP}}(x) &= \mathbb{E}\Big[ f(x) \;\Big|\; \{x_m, y_m\} \Big] && \textit{(also MV)} \\ &= \sum_{m=1}^{M} K(x, x_m) c_m && \textit{(a.k.a. regularization network)} \end{aligned} \tag{10}$$

# Maximum a posteriori estimator

$$\begin{aligned} \widehat{f}_{\mathrm{MAP}}(x) &= \mathbb{E}\left[ f(x) \,\middle|\, \{x_m, y_m\} \right] \qquad \textit{(also MV)} \\ &= \sum_{m=1}^{M} K(x, x_m) c_m \qquad \textit{(a.k.a. regularization network)} \end{aligned} \tag{10}$$

$$\begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} = H_{\mathrm{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} \tag{11}$$

$$H_{\mathrm{MAP}} := \left( \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1} \tag{12}$$

# Gaussian regression – practical issues associated to the MAP

$$\widehat{f}_{\mathrm{MAP}}(x) = \left[ K(x, x_1) \; \ldots \; K(x, x_M) \right] H_{\mathrm{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} \tag{13}$$

$$H_{\mathrm{MAP}} := \left( \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1} \tag{14}$$

computational cost $O\left(M^3\right)$

# How may we tackle the $O\left(M^3\right)$ computational cost issue?

$$H_{\mathrm{MAP}} := \left( \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1} \tag{15}$$

Typical approaches: low-rank / sparsification approximations

Smola & Schölkopf (2000)
Sparse greedy matrix approximations for machine learning

Quiñonero-Candela & Rasmussen (2005)
A unifying view of sparse approximate Gaussian process regression

Bach & Jordan (2005)
Predictive low-rank decompositions for kernel methods

Snelson & Ghahramani (2006)
Sparse Gaussian processes using pseudo inputs

Culis et al. (2006)
Learning low-rank kernel matrices

Zhang & Kwok (2010)
Clustered Nyström method for large scale manifold learning and dimension reduction

Ambikasaran et al. (2016)
Fast direct methods for Gaussian processes

# Our focus: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x)$$

# Our focus: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\text{interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\text{remainder}} \tag{16}$$

# Our focus: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{\text{=:interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{\text{=:remainder}} \tag{16}$$

$$\lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x') \qquad \lambda_1 \geq \lambda_2 \ldots > 0 \tag{17}$$

$$K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \qquad \int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij} \tag{18}$$

# Our focus: Karhunen-Loève expansions

$$f(x) = \sum_{e=1}^{+\infty} \alpha_e \phi_e(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{\text{=:interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{\text{=:remainder}} \tag{16}$$

$$\lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x') \qquad \lambda_1 \geq \lambda_2 \ldots > 0 \tag{17}$$

$$K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \qquad \int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij} \tag{18}$$

$$a_e \sim \mathcal{N}(0, \lambda_e), \ e = 1, \ldots, E \qquad b_e \sim \mathcal{N}(0, \lambda_{E+e}), \ e = 1, 2, \ldots \tag{19}$$

Zhu et al. (1998)

Gaussian regression and optimal finite dimensional linear models

# Our focus: Karhunen-Loève expansions

$$f(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{\text{=:interesting}} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{\text{=:remainder}} \qquad \lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x,x')\phi_e(x')d\mu(x') \tag{20}$$

$\implies$ first $E$ $\phi_e$'s $=$ best a-priori $E$-dimensional approximation in a MSE sense

# Our focus: Karhunen-Loève expansions

$$\boldsymbol{y} \coloneqq [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} \coloneqq [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} \coloneqq [a_1, \ldots, a_E]^T \quad \boldsymbol{b} \coloneqq [b_1, b_2, \ldots]^T \tag{21}$$

# Our focus: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \ldots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \ldots]^T \tag{21}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \ldots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \ldots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \ldots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \ldots \end{bmatrix} \tag{22}$$

# Our focus: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \dots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \dots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \dots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \dots]^T \tag{21}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \dots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \dots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \dots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \dots \end{bmatrix} \tag{22}$$

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{23}$$

# Our focus: Karhunen-Loève expansions

$$\boldsymbol{y} := [y_1, \ldots, y_M]^T \quad \boldsymbol{\nu} := [\nu_1, \ldots, \nu_M]^T \quad \boldsymbol{a} := [a_1, \ldots, a_E]^T \quad \boldsymbol{b} := [b_1, b_2, \ldots]^T \tag{21}$$

$$G := \begin{bmatrix} \phi_1(x_1) & \ldots & \phi_E(x_1) \\ \vdots & & \vdots \\ \phi_1(x_M) & \ldots & \phi_E(x_M) \end{bmatrix} \qquad Z := \begin{bmatrix} \phi_{E+1}(x_1) & \phi_{E+2}(x_1) & \ldots \\ \vdots & & \vdots \\ \phi_{E+1}(x_M) & \phi_{E+2}(x_M) & \ldots \end{bmatrix} \tag{22}$$

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{23}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} \widehat{\boldsymbol{a}} \qquad \widehat{\boldsymbol{a}} = H\boldsymbol{y} \qquad H := \left( \frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{24}$$

# Our focus: Karhunen-Loève expansions

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu} \tag{25}$$

$$\widehat{f}_E(x) := \begin{bmatrix} \phi_1(x) & \cdots & \phi_E(x) \end{bmatrix} H\boldsymbol{y} \qquad H := \left( \frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{26}$$

computational cost: $O(E^3)$

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

*our aim: bound the statistical performance of $\widehat{f}_E$ as a function of $E$*

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

*our aim: bound the statistical performance of $\widehat{f}_E$ as a function of $E$*

Key quantities:

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

*our aim: bound the statistical performance of $\widehat{f}_E$ as a function of $E$*

Key quantities:

- $\mathbb{E}\left[\left\|f - \widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right]$

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\| f(x) - \widehat{f}_E(x) \right\|^2\right]$?

*our aim: bound the statistical performance of* $\widehat{f}_E$ *as a function of* $E$

Key quantities:

- $\mathbb{E}\left[\left\| f - \widehat{f}_E \right\|^2 \mid \boldsymbol{x}\right]$
- $\mathbb{E}\left[\mathbb{E}\left[\left\| f - \widehat{f}_E \right\|^2 \mid \boldsymbol{x}\right] \mid \mathcal{E}\right]$ $\qquad \mathbb{P}\left[\mathcal{E}\right] \geq 1 - \alpha$

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

*our aim: bound the statistical performance of $\widehat{f}_E$ as a function of $E$*

Key quantities:

- $\mathbb{E}\left[\left\|f - \widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right]$
- $\mathbb{E}\left[\mathbb{E}\left[\left\|f - \widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right] \mid \mathcal{E}\right]$ $\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \alpha$

  $\alpha \in (0,1)$ = *desired confidence level, e.g.,* $\alpha = 0.01$ *or* $\alpha = 0.05$

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

*our aim: bound the statistical performance of $\widehat{f}_E$ as a function of $E$*

Key quantities:

- $\mathbb{E}\left[\left\|f - \widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right]$
- $\mathbb{E}\left[\mathbb{E}\left[\left\|f - \widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right] \mid \mathcal{E}\right]$ $\qquad \mathbb{P}\left[\mathcal{E}\right] \geq 1 - \alpha$

  $\alpha \in (0,1)$ = *desired confidence level, e.g.,* $\alpha = 0.01$ *or* $\alpha = 0.05$
- $k := \sup_{e \in \mathbb{N}, x \in \mathcal{X}} \left|\phi_e(x)\right|^2$
- $\varepsilon \in (0,1]$ = opportune distance index between $\dfrac{G^T G}{M}$ and $I$

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

in words

if $M$ is sufficiently[1] big w.r.t. $E$

then with at least probability $1 - \alpha$ the expected performance of $\widehat{f}_E$

is upper bounded by the following $\mathrm{Bnd}$, that is computable a-priori

$$\mathrm{Bnd} := \frac{kM}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e^2}{(\varepsilon M \lambda_e + \sigma_\nu^2)^2} \right) \left( \sum_{e=E+1}^{+\infty} \lambda_e \right) + \frac{\sigma_\nu^2}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e}{\varepsilon M \lambda_e + \sigma_\nu^2} \right) + \left( \sum_{e=E+1}^{+\infty} \lambda_e \right) \tag{27}$$

[1] With $k, \alpha$ and $\varepsilon$ influencing what “sufficiently” means in numbers

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x) - \widehat{f}_E(x)\right\|^2\right]$?

formally

# How well does $\widehat{f}_E$ perform w.r.t. $\mathbb{E}\left[\left\|f(x)-\widehat{f}_E(x)\right\|^2\right]$?

formally

if $E, M, k, \alpha, \varepsilon$ satisfy $1-\varepsilon+\varepsilon\log(\varepsilon) \geq \frac{Ek}{M}\log\left(\frac{E}{\alpha}\right)$ then

$$\mathbb{P}\left[\mathbb{E}\left[\mathbb{E}\left[\left\|f-\widehat{f}_E\right\|^2 \mid \boldsymbol{x}\right] \mid \mathcal{E}\right] \leq \mathrm{Bnd}\right] \geq 1-\alpha \tag{28}$$

with

$$\mathrm{Bnd} := \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e^2}{(\varepsilon M\lambda_e+\sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty}\lambda_e\right)+\frac{\sigma_\nu^2}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e}{\varepsilon M\lambda_e+\sigma_\nu^2}\right)+\left(\sum_{e=E+1}^{+\infty}\lambda_e\right) \tag{29}$$

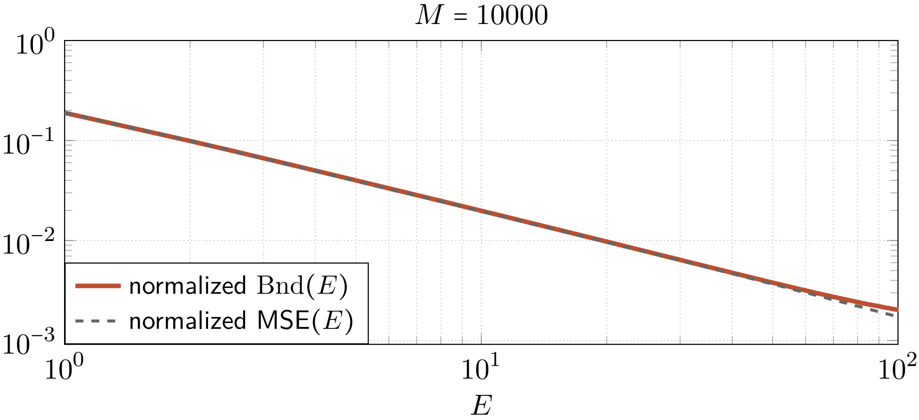
# What do we enable?

given $M$ and $K$, how big should $E$ be
so to have a certain expected statistical performance?

$$\mathrm{Bnd} := \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e^2}{(\varepsilon M\lambda_e+\sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty}\lambda_e\right)+\frac{\sigma_\nu^2}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e}{\varepsilon M\lambda_e+\sigma_\nu^2}\right)+\left(\sum_{e=E+1}^{+\infty}\lambda_e\right) \tag{30}$$

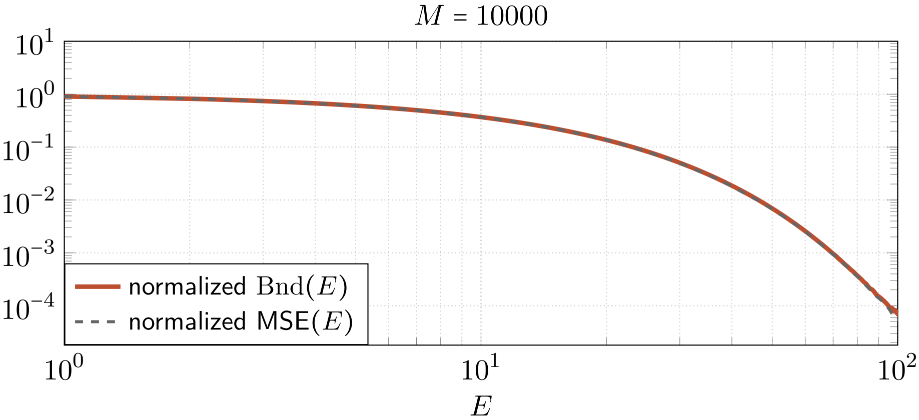
# How significant is $\mathrm{Bnd}$?

Case splines, i.e., $K(x,x') = \min\left(x,x'\right)$ $\qquad \mathcal{X} = [0,1]$ $\qquad \lambda_e = \dfrac{1}{(e\pi - \pi/2)^2}$

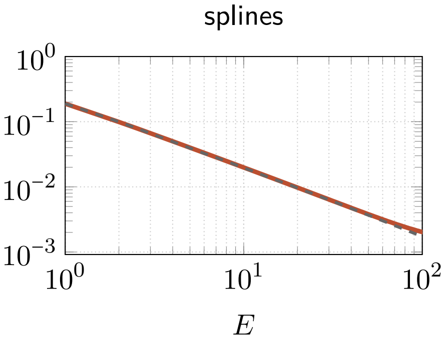


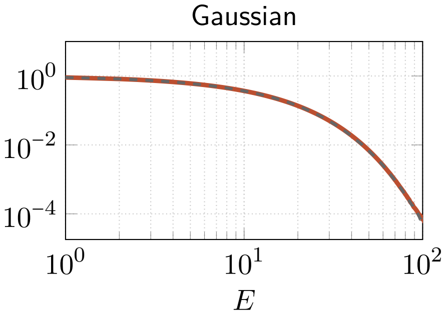
# How significant is $\mathrm{Bnd}$?

Case Gaussian, i.e., $K(x, x') = \exp\left(-\left\|x - x'\right\|^2\right)$ $\quad\mathcal{X} = [0, 1]\quad$ $\lambda_e = \exp(-0.1e)$



# What can we do with this result?


splines


Gaussian

*understand how big $E$ should be a priori so to obtain certain statistical performance*

# What do we enable?

PAMI extension: Distributed multi-agent Gaussian regression via finite-dimensional approximations

*strategies for distributedly tuning*
*the hyperparameters of distributed estimators*
*through the minimization of the bound*
*(both a priori and a posteriori)*

# Distributed multi-agent Gaussian regression via finite-dimensional approximations

Gianluigi Pillonetto Luca Schenato Damiano Varagnolo

LULEÅ UNIVERSITY OF TECHNOLOGY

damiano.varagnolo@ltu.se

entirely written in LATEX2ε using Beamer and TikZ