KERNEL METHODS AND THE CURSE OF DIMENSIONALITY

arXiv:1905.10843

Stefano Spigler

Jonas Paccolat, Mario Geiger, Matthieu Wyart

EPFL

SUPERVISED DEEP LEARNING

- Why and how does deep **supervised** learning work?

- Learn from examples: **how many** are needed?

- Typical tasks:
  - Regression (fitting functions)
  - Classification

LEARNING CURVES

- Performance is evaluated through the **generalization error** ϵ

- Learning curves decay with number of examples n, often as

$$\epsilon \sim n^{-\beta}$$

- β depends on the **dataset** and on the **algorithm**

Deep networks: $\beta \sim 0.07$-0.35 [Hestness et al. 2017]

We lack a theory for β for deep networks!

LINK WITH KERNEL LEARNING

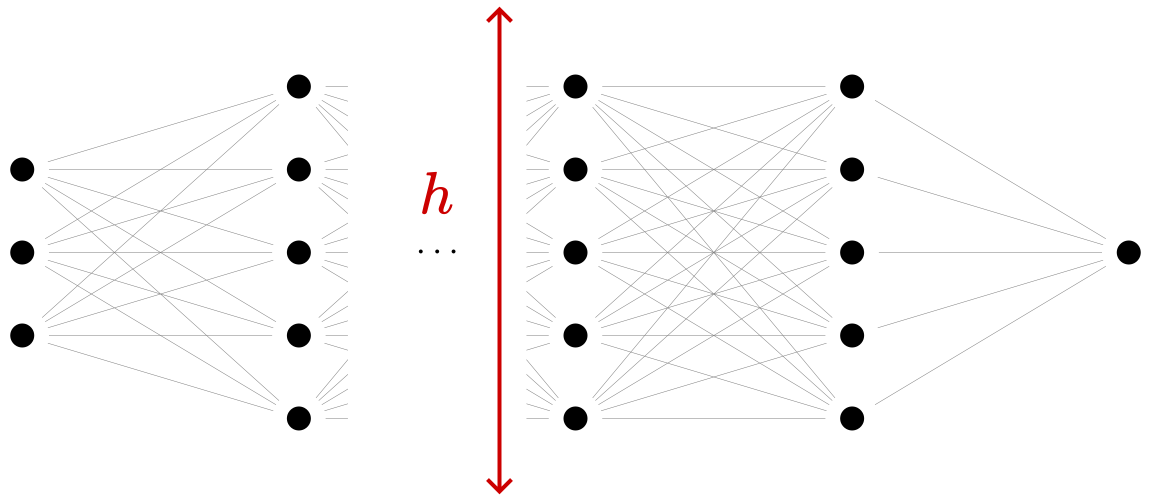
- Performance increases with **overparametrization**

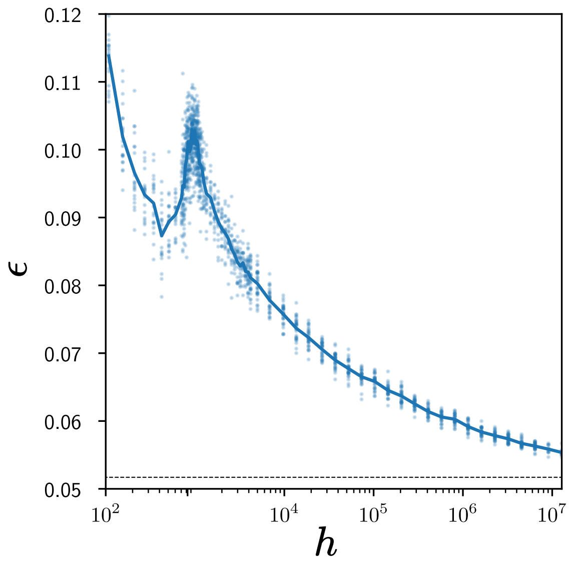
[Neyshabur et al. 2017, 2018, Advani and Saxe 2017]

[Spigler et al. 2018, Geiger et al. 2019, Belkin et al. 2019]

$\longrightarrow$ study the infinite-width limit!

[Mei et al. 2017, Rotskoff and Vanden-Eijnden 2018, Jacot et al. 2018, Chizat and Bach 2018, ...]





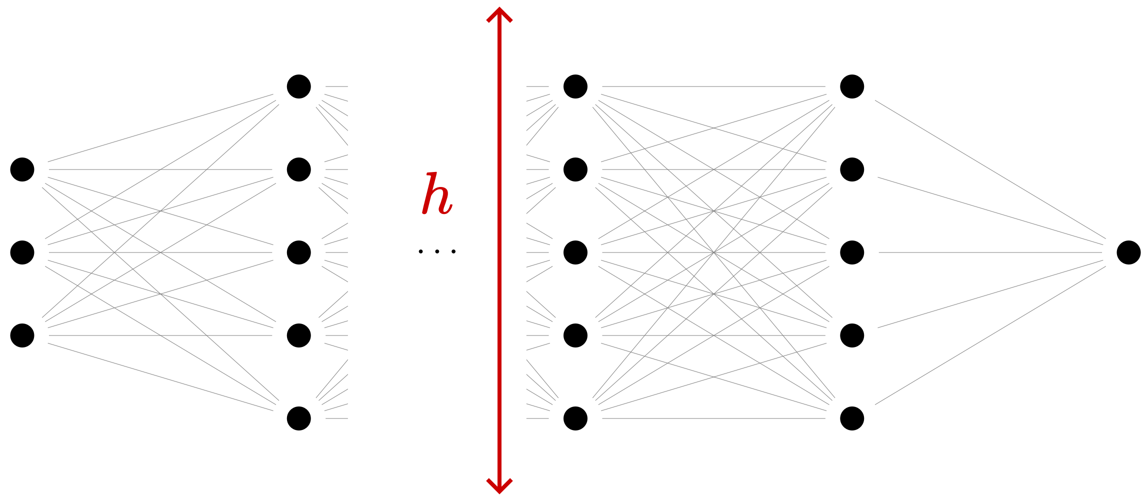
LINK WITH KERNEL LEARNING

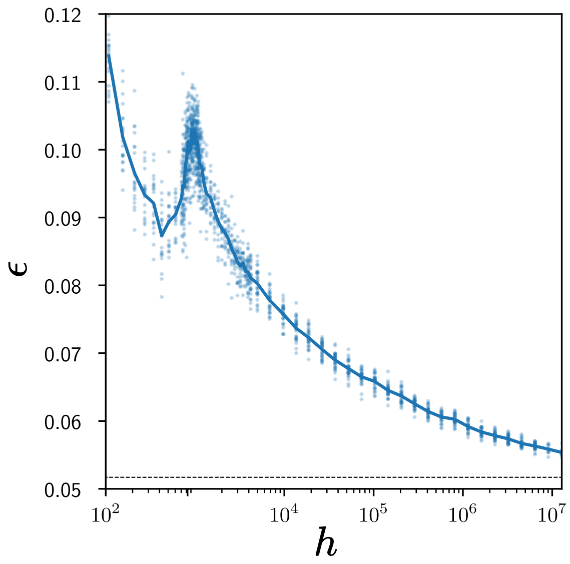
- Performance increases with **overparametrization**

[Neyshabur et al. 2017, 2018, Advani and Saxe 2017]
[Spigler et al. 2018, Geiger et al. 2019, Belkin et al. 2019]

$\longrightarrow$ study the infinite-width limit!

[Mei et al. 2017, Rotskoff and Vanden-Eijnden 2018, Jacot et al. 2018, Chizat and Bach 2018, ...]





- With a specific scaling, infinite-width limit $\rightarrow$ **kernel learning**

[Jacot et al. 2018]
Neural Tangent Kernel

(next slides)

What are the learning curves of kernels like?

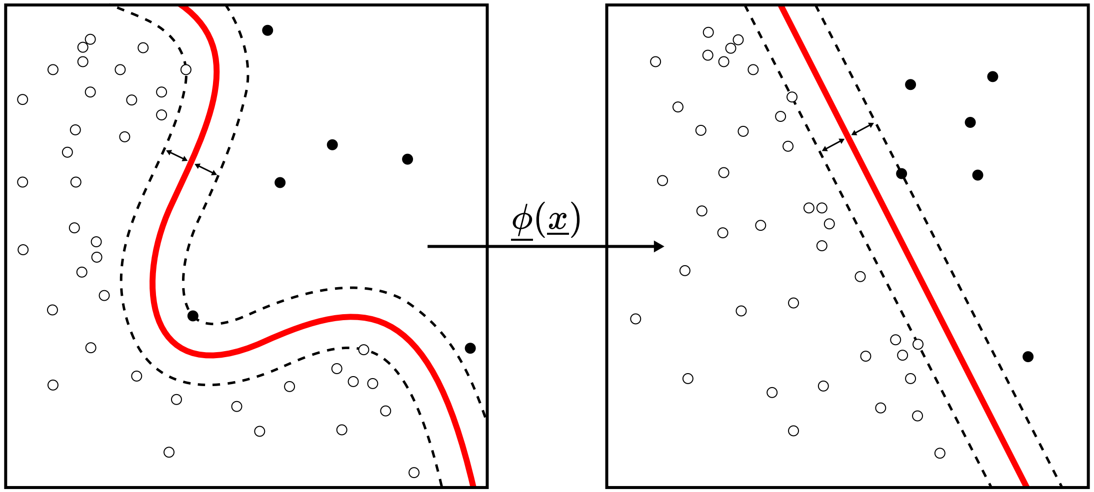
OUTLINE

- Very brief introduction to kernel methods
- Performance of kernels on real data
- Gaussian data: Teacher-Student regression
- Gaussian approximation: smoothness and effective dimension
- Dimensional reduction via invariants in the task

KERNEL METHODS

- Kernel methods learn non-linear functions or boundaries
- Map data to a **feature space**, where the problem is linear

data $\underline{x} \longrightarrow \underline{\phi}(\underline{x}) \longrightarrow$ use linear combination of features



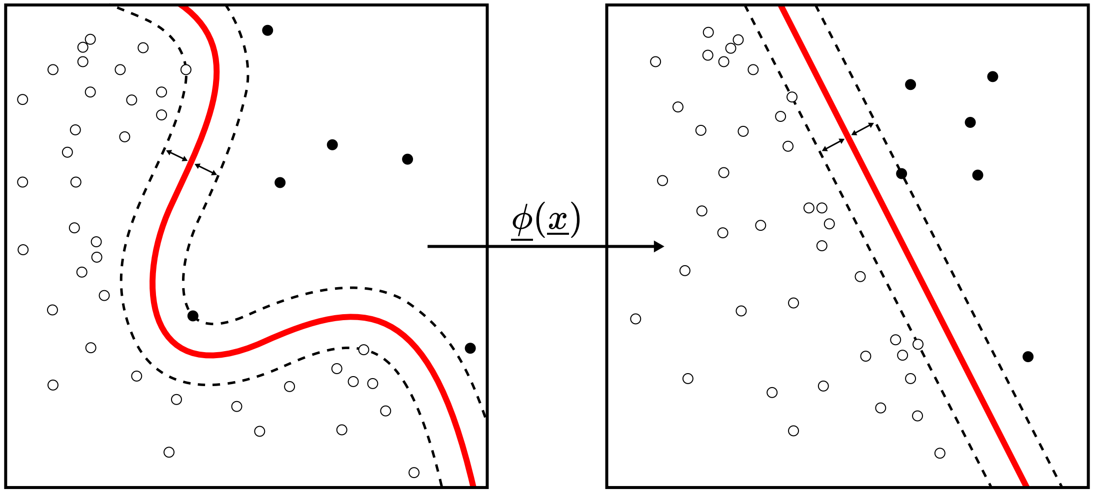
KERNEL METHODS

- Kernel methods learn non-linear functions or boundaries
- Map data to a **feature space**, where the problem is linear

data $\underline{x} \longrightarrow \underline{\phi}(\underline{x}) \longrightarrow$ use linear combination of features

only scalar products are needed: $\boxed{\underline{\phi}(\underline{x}) \cdot \underline{\phi}(\underline{x}')}$

kernel $K(\underline{x}, \underline{x}')$



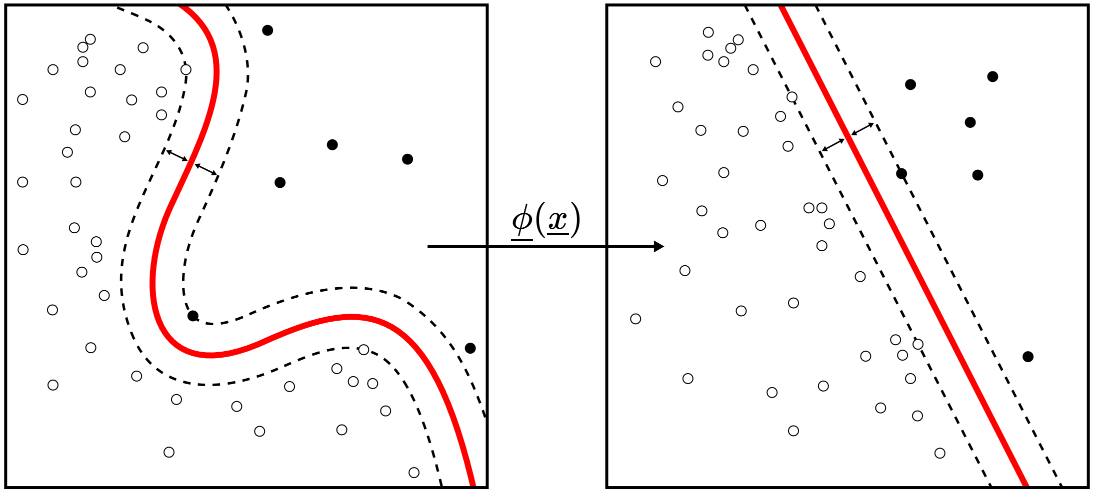
KERNEL METHODS

- Kernel methods learn non-linear functions or boundaries
- Map data to a **feature space**, where the problem is linear

data $\underline{x} \longrightarrow \underline{\phi}(\underline{x}) \longrightarrow$ use linear combination of features

only scalar products are needed: $\boxed{\underline{\phi}(\underline{x}) \cdot \underline{\phi}(\underline{x}')}$

kernel $K(\underline{x}, \underline{x}')$



Gaussian:

$$K(\underline{x}, \underline{x}') = \exp\left(-\frac{\|\underline{x}-\underline{x}'\|^2}{\sigma^2}\right)$$

Laplace:

$$K(\underline{x}, \underline{x}') = \exp\left(-\frac{\|\underline{x}-\underline{x}'\|}{\sigma}\right)$$

KERNEL REGRESSION

E.g. kernel regression:

- **Target function** $\underline{x}_\mu \to Z(\underline{x}_\mu), \;\; \mu = 1, \ldots, n$

KERNEL REGRESSION

E.g. kernel regression:

- **Target function** $\underline{x}_\mu \to Z(\underline{x}_\mu), \;\; \mu = 1, \dots, n$

- Build an estimator $\hat{Z}_K(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K(\underline{x}_\mu, \underline{x})$

KERNEL REGRESSION

E.g. kernel regression:

- **Target function** $\underline{x}_\mu \to Z(\underline{x}_\mu), \quad \mu = 1, \dots, n$
- Build an estimator $\hat{Z}_K(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K(\underline{x}_\mu, \underline{x})$
- Minimize training MSE $= \frac{1}{n} \sum_{\mu=1}^{n} \left[\hat{Z}_K(\underline{x}_\mu) - Z(\underline{x}_\mu) \right]^2$

KERNEL REGRESSION

E.g. kernel regression:

- **Target function** $\underline{x}_\mu \to Z(\underline{x}_\mu), \quad \mu = 1, \ldots, n$
- Build an estimator $\hat{Z}_K(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K(\underline{x}_\mu, \underline{x})$
- Minimize training MSE $= \frac{1}{n} \sum_{\mu=1}^{n} \left[\hat{Z}_K(\underline{x}_\mu) - Z(\underline{x}_\mu)\right]^2$
- Estimate the **generalization error** $\epsilon = \mathbb{E}_{\underline{x}} \left[\hat{Z}_K(\underline{x}) - Z(\underline{x})\right]^2$

REPRODUCING KERNEL HILBERT SPACE (RKHS)

A kernel K induces a corresponding Hilbert space $\mathcal{H}_K$ with norm

$$\|Z\|_K = \int \mathrm{d}\underline{x}\mathrm{d}\underline{y}\, Z(\underline{x})K^{-1}(\underline{x},\underline{y})Z(\underline{y})$$

where $K^{-1}(\underline{x},\underline{y})$ is such that

$$\int \mathrm{d}\underline{y}\, K^{-1}(\underline{x},\underline{y})K(\underline{y},\underline{z}) = \delta(\underline{x},\underline{z})$$

$\mathcal{H}_K$ is called the **Reproducing Kernel Hilbert Space** (RKHS)

PREVIOUS WORKS

Regression: **performance depends on the target function!**

PREVIOUS WORKS

Regression: **performance depends on the target function!**

d = dimension of the input space

- If only assumed to be **Lipschitz**, then $\beta = \frac{1}{d}$

Curse of dimensionality!

[Luxburg and Bousquet 2004]

PREVIOUS WORKS

Regression: **performance depends on the target function!**

d = dimension of the input space

- If only assumed to be **Lipschitz**, then $\beta = \frac{1}{d}$

Curse of dimensionality!

[Luxburg and Bousquet 2004]

- If assumed to be in the **RKHS**, then $\beta \geq \frac{1}{2}$ does not depend on d

[Smola et al. 1998, Rudi and Rosasco 2017]

PREVIOUS WORKS

Regression: **performance depends on the target function!**

d = dimension of the input space

- If only assumed to be **Lipschitz**, then $\beta = \frac{1}{d}$

Curse of dimensionality!

[Luxburg and Bousquet 2004]

- If assumed to be in the **RKHS**, then $\beta \geq \frac{1}{2}$ does not depend on d

[Smola et al. 1998, Rudi and Rosasco 2017]

- Yet, RKHS is a very strong assumption on the smoothness of the target function (see later on)

[Bach 2017]

REAL DATA AND ALGORITHMS

We apply kernel methods on

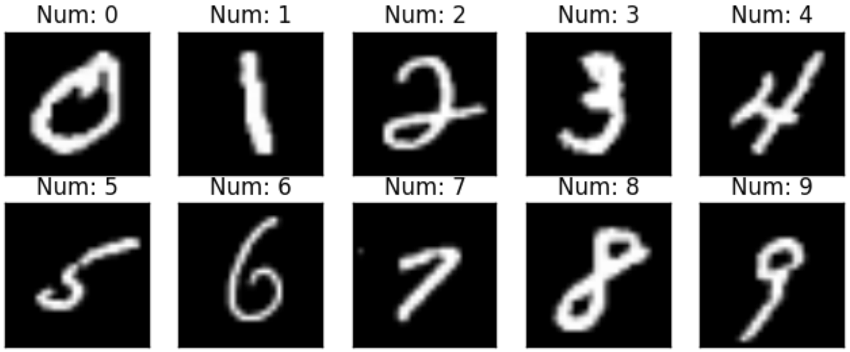
MNIST

2 classes: even/odd

70000 28x28 b/w pictures

$\downarrow$

dimension $d = 784$



CIFAR10

2 classes: first 5/last 5

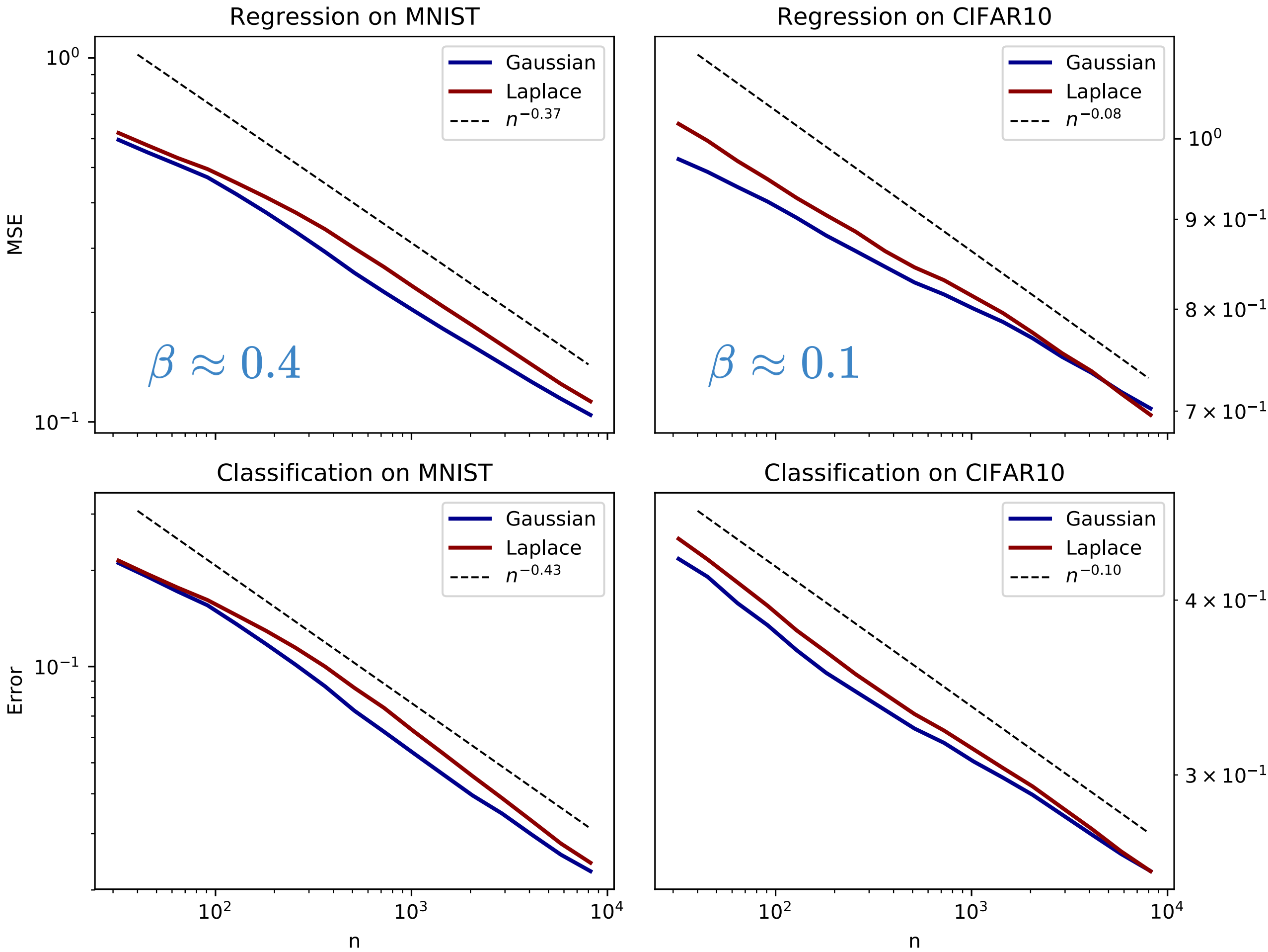
60000 32x32 RGB pictures

$\downarrow$

dimension $d = 3072$

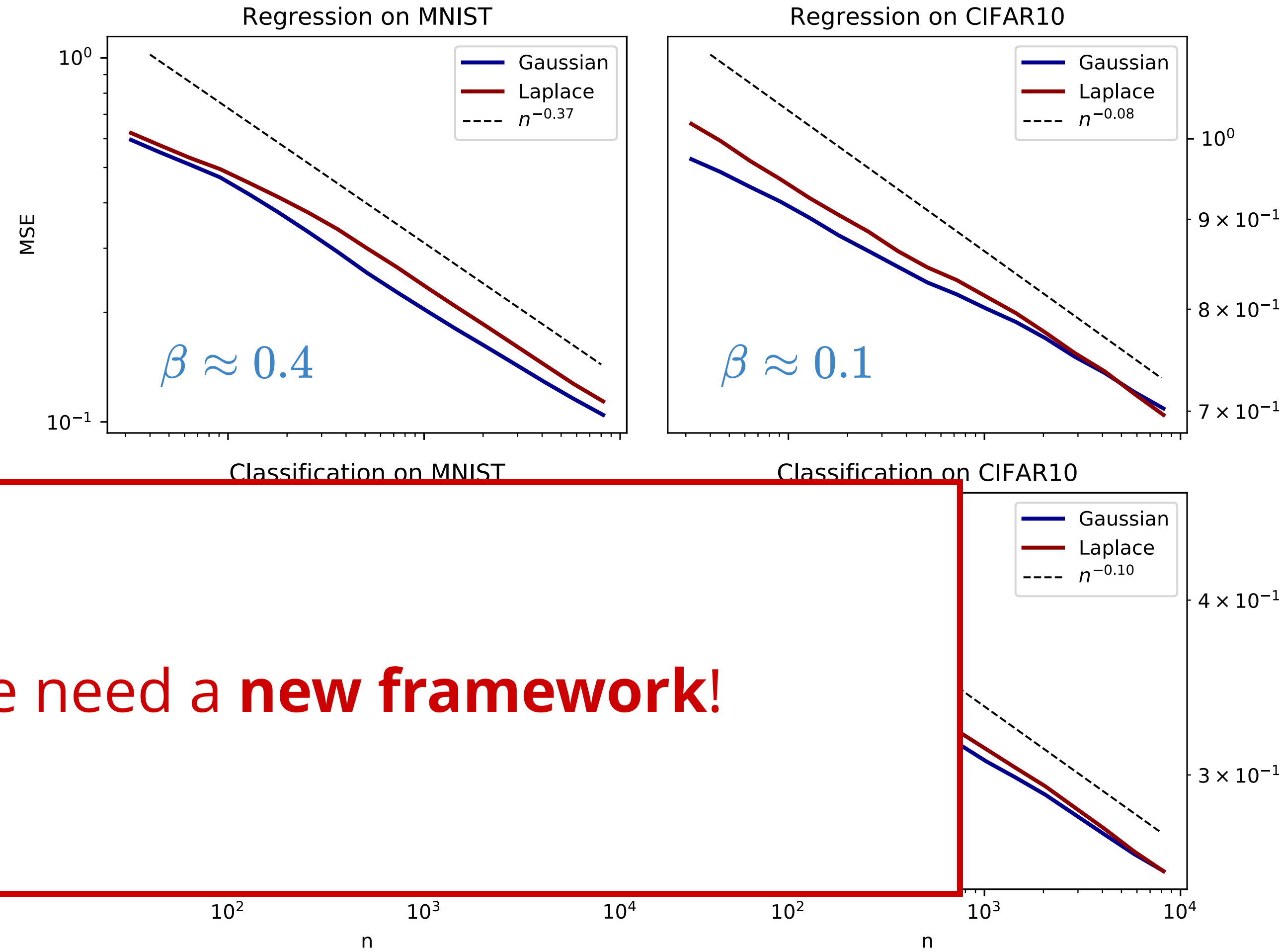
We perform
- **regression** $\longrightarrow$ kernel regression
- **classification** $\longrightarrow$ margin SVM

REAL DATA: EXPONENTS



- Same exponent for regression and classification
- Same exponent for Gaussian and Laplace kernel
- MNIST and CIFAR10 display exponents $\beta \gg \frac{1}{d}$ but $< \frac{1}{2}$

REAL DATA: EXPONENTS



We need a **new framework**!

- Same exponent for regression and classification
- Same exponent for Gaussian and Laplace kernel
- MNIST and CIFAR10 display exponents $\beta \gg \frac{1}{d}$ but $< \frac{1}{2}$

KERNEL TEACHER-STUDENT FRAMEWORK

- Controlled setting: **Teacher-Student regression**

KERNEL TEACHER-STUDENT FRAMEWORK

- Controlled setting: **Teacher-Student regression**

- Training data are sampled from a Gaussian Process:

$$Z_T(\underline{x}_1), \ldots, Z_T(\underline{x}_n) \sim \mathcal{N}(0, K_T)$$

$\underline{x}_\mu$ are random on a ***d*-dim hypersphere**

KERNEL TEACHER-STUDENT FRAMEWORK

- Controlled setting: **Teacher-Student regression**

- Training data are sampled from a Gaussian Process:

$$Z_T(\underline{x}_1), \ldots, Z_T(\underline{x}_n) \sim \mathcal{N}(0, K_T)$$

$\underline{x}_\mu$ are random on a ***d*-dim hypersphere**

$$\mathbb{E} Z_T(\underline{x}_\mu) = 0$$

$$\mathbb{E} Z_T(\underline{x}_\mu) Z_T(\underline{x}_\nu) = K_T(\|\underline{x}_\mu - \underline{x}_\nu\|)$$

KERNEL TEACHER-STUDENT FRAMEWORK

- Controlled setting: **Teacher-Student regression**

- Training data are sampled from a Gaussian Process:

$$Z_T(\underline{x}_1), \ldots, Z_T(\underline{x}_n) \ \sim \ \mathcal{N}(0, K_T)$$

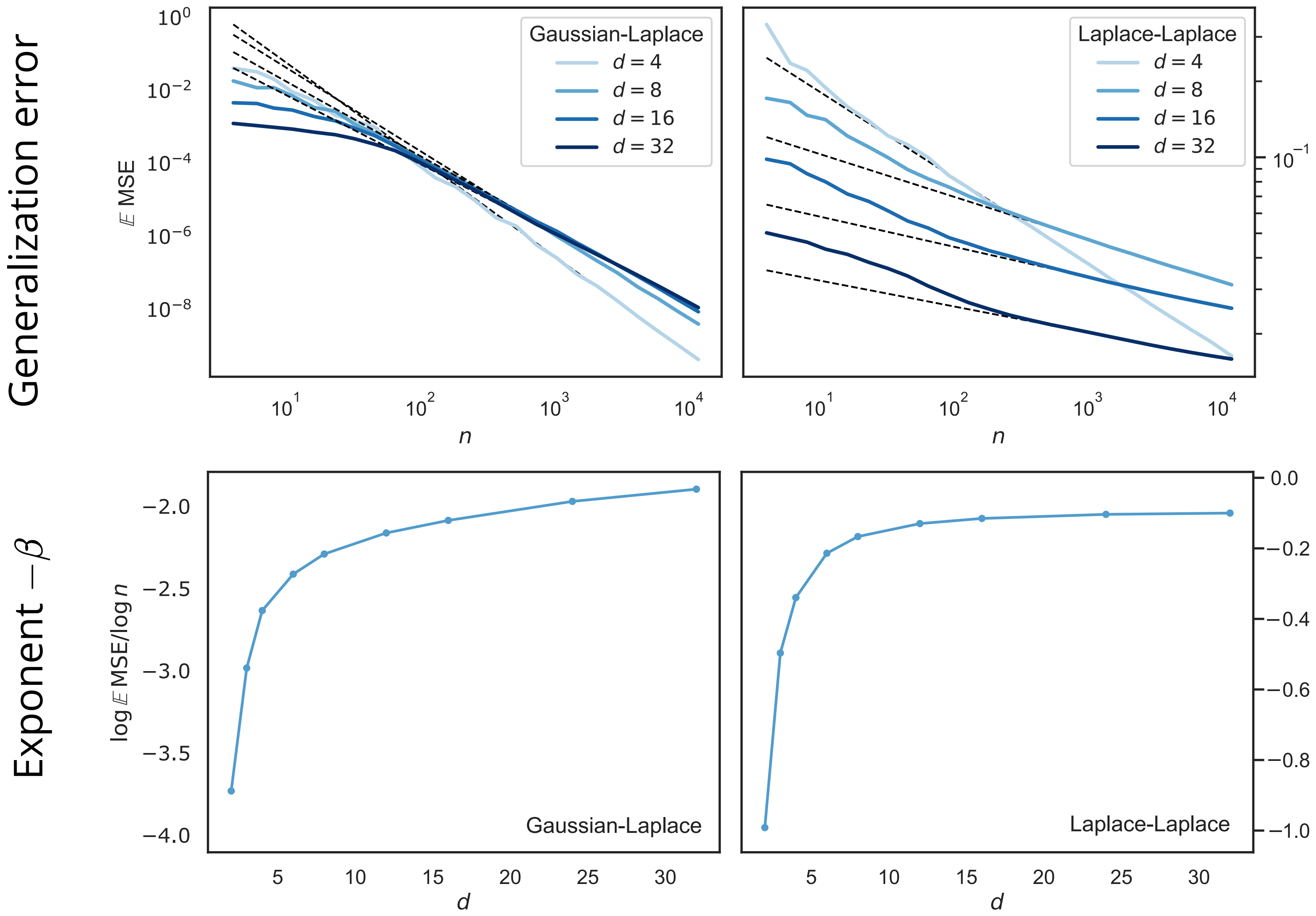
$\underline{x}_\mu$ are random on a ***d*-dim hypersphere**

$$\mathbb{E} Z_T(\underline{x}_\mu) = 0$$

$$\mathbb{E} Z_T(\underline{x}_\mu) Z_T(\underline{x}_\nu) = K_T(\|\underline{x}_\mu - \underline{x}_\nu\|)$$

- Regression is done with another kernel K_S

TEACHER-STUDENT: SIMULATIONS



Can we understand these curves?

TEACHER-STUDENT: REGRESSION

<u>Regression</u>:

$$\hat{Z}_S(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K_S(\underline{x}_\mu, \underline{x})$$

$$\text{Minimize} = \frac{1}{n}\sum_{\mu=1}^{n}\left[\hat{Z}_S(\underline{x}_\mu) - Z_T(\underline{x}_\mu)\right]^2$$

<u>Explicit solution</u>:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1}\underline{Z}$$

where

TEACHER-STUDENT: REGRESSION

<u>Regression</u>:

$$\hat{Z}_S(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K_S(\underline{x}_\mu, \underline{x})$$

$$\text{Minimize} = \frac{1}{n} \sum_{\mu=1}^{n} \left[\hat{Z}_S(\underline{x}_\mu) - Z_T(\underline{x}_\mu) \right]^2$$

<u>Explicit solution</u>:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1} \underline{Z}_T$$

where

$$\left\{ (\underline{k}_S(\underline{x}))_\mu = K_S(\underline{x}_\mu, \underline{x}) \right.$$

kernel overlap

TEACHER-STUDENT: REGRESSION

Regression:

$$\hat{Z}_S(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K_S(\underline{x}_\mu, \underline{x})$$

$$\text{Minimize} = \frac{1}{n} \sum_{\mu=1}^{n} \left[\hat{Z}_S(\underline{x}_\mu) - Z_T(\underline{x}_\mu)\right]^2$$

Explicit solution:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1} \underline{Z}_T$$

where

$$(\underline{k}_S(\underline{x}))_\mu = K_S(\underline{x}_\mu, \underline{x})$$

kernel overlap

$$(\mathbb{K}_S)_{\mu\nu} = K_S(\underline{x}_\mu, \underline{x}_\nu)$$

Gram matrix

TEACHER-STUDENT: REGRESSION

Regression:

$$\hat{Z}_S(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K_S(\underline{x}_\mu, \underline{x})$$

$$\text{Minimize} = \frac{1}{n}\sum_{\mu=1}^{n}\left[\hat{Z}_S(\underline{x}_\mu) - Z_T(\underline{x}_\mu)\right]^2$$

Explicit solution:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1} \underline{Z}_T$$

where

$$(\underline{k}_S(\underline{x}))_\mu = K_S(\underline{x}_\mu, \underline{x})$$

kernel overlap

$$(\mathbb{K}_S)_{\mu\nu} = K_S(\underline{x}_\mu, \underline{x}_\nu)$$

Gram matrix

$$(\underline{Z}_T)_\mu = Z_T(\underline{x}_\mu)$$

training data

TEACHER-STUDENT: REGRESSION

Regression:

$$\hat{Z}_S(\underline{x}) = \sum_{\mu=1}^{n} c_\mu K_S(\underline{x}_\mu, \underline{x})$$

$$\text{Minimize} = \frac{1}{n}\sum_{\mu=1}^{n} \left[\hat{Z}_S(\underline{x}_\mu) - Z_T(\underline{x}_\mu)\right]^2$$

Explicit solution:

$$\hat{Z}_S(\underline{x}) = \underline{k}_S(\underline{x}) \cdot \mathbb{K}_S^{-1} \underline{Z}_T$$

where

$$\begin{cases} (\underline{k}_S(\underline{x}))_\mu = K_S(\underline{x}_\mu, \underline{x}) \\ (\mathbb{K}_S)_{\mu\nu} = K_S(\underline{x}_\mu, \underline{x}_\nu) \\ (\underline{Z}_T)_\mu = Z_T(\underline{x}_\mu) \end{cases}$$

kernel overlap

Gram matrix

training data

Compute the generalization error ϵ and how it scales with n

$$\epsilon = \mathbb{E}_T \int \mathrm{d}^d\underline{x} \left[\hat{Z}_S(\underline{x}) - Z_T(\underline{x})\right]^2 \sim n^{-\beta}$$

TEACHER-STUDENT: THEOREM (1/2)

To compute the generalization error:

- We look at the problem in the **frequency domain**

- We assume that $\tilde{K}_S(\underline{w}) \sim \|\underline{w}\|^{-\alpha_S}$ and $\tilde{K}_T(\underline{w}) \sim \|\underline{w}\|^{-\alpha_T}$ as $\|\underline{w}\| \to \infty$

TEACHER-STUDENT: THEOREM (1/2)

To compute the generalization error:

- We look at the problem in the **frequency domain**

- We assume that $\tilde{K}_S(\underline{w}) \sim \|\underline{w}\|^{-\alpha_S}$ and $\tilde{K}_T(\underline{w}) \sim \|\underline{w}\|^{-\alpha_T}$ as $\|\underline{w}\| \to \infty$

E.g. Laplace has $\alpha = d + 1$ and Gaussian has $\alpha = \infty$

TEACHER-STUDENT: THEOREM (1/2)

To compute the generalization error:

- We look at the problem in the **frequency domain**

- We assume that $\tilde{K}_S(\underline{w}) \sim \|\underline{w}\|^{-\alpha_S}$ and $\tilde{K}_T(\underline{w}) \sim \|\underline{w}\|^{-\alpha_T}$ as $\|\underline{w}\| \to \infty$

  E.g. Laplace has $\alpha = d + 1$ and Gaussian has $\alpha = \infty$

- **SIMPLIFYING ASSUMPTION:** We take the n points $\underline{x}_\mu$ on a **regular** d-dim **lattice!**

TEACHER-STUDENT: THEOREM (1/2)

To compute the generalization error:

- We look at the problem in the **frequency domain**
- We assume that $\tilde{K}_S(\underline{w}) \sim \|\underline{w}\|^{-\alpha_S}$ and $\tilde{K}_T(\underline{w}) \sim \|\underline{w}\|^{-\alpha_T}$ as$\|\underline{w}\| \to \infty$

E.g. Laplace has $\alpha = d + 1$ and Gaussian has $\alpha = \infty$

- **SIMPLIFYING ASSUMPTION:** We take the n points $\underline{x}_\mu$ on a **regular** d-dim **lattice!**

(details: arXiv:1905.10843)

Then we can show that

for $n \gg 1$ $\quad \epsilon \sim n^{-\beta}$ with $\boxed{\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)}$

TEACHER-STUDENT: THEOREM (2/2)

$$\boxed{\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)}$$

- Large $\alpha \to$ fast decay at high freq $\to$ **indifference** to **local details**
- α_T is intrinsic to the **data** (T), α_S depends on the **algorithm** (S)
- If α_S is large enough, β takes the largest possible value $\frac{\alpha_T - d}{d}$

  (optimal learning)
- As soon as α_S is small enough, $\beta = \frac{2\alpha_S}{d}$

TEACHER-STUDENT: COMPARISON (1/2)

What is the prediction for our simulations?

$$\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)$$

- If Teacher=Student=Laplace $(\alpha_T = \alpha_S = d + 1)$

$$\beta = \frac{\alpha_T - d}{d} = \frac{1}{d}$$

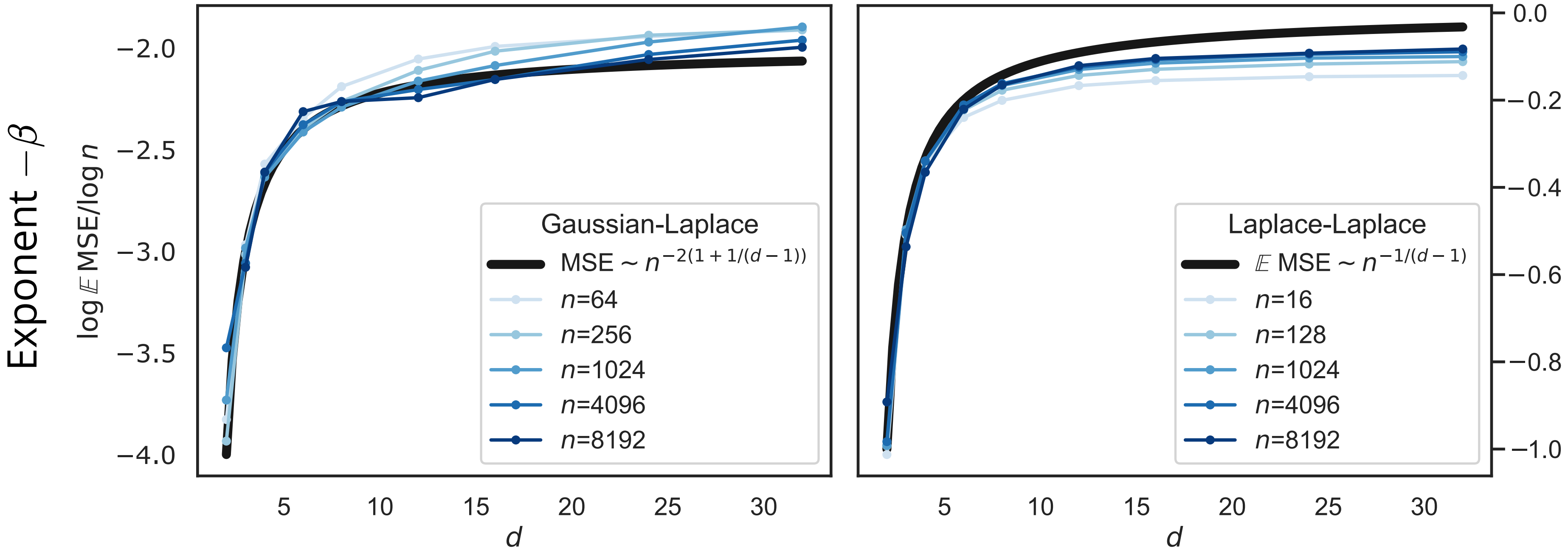
(curse of dimensionality!)

- If Teacher=Gaussian, Student=Laplace $(\alpha_T = \infty, \alpha_S = d + 1)$

$$\beta = \frac{2\alpha_S}{d} = 2 + \frac{2}{d}$$

TEACHER-STUDENT: COMPARISON (2/2)

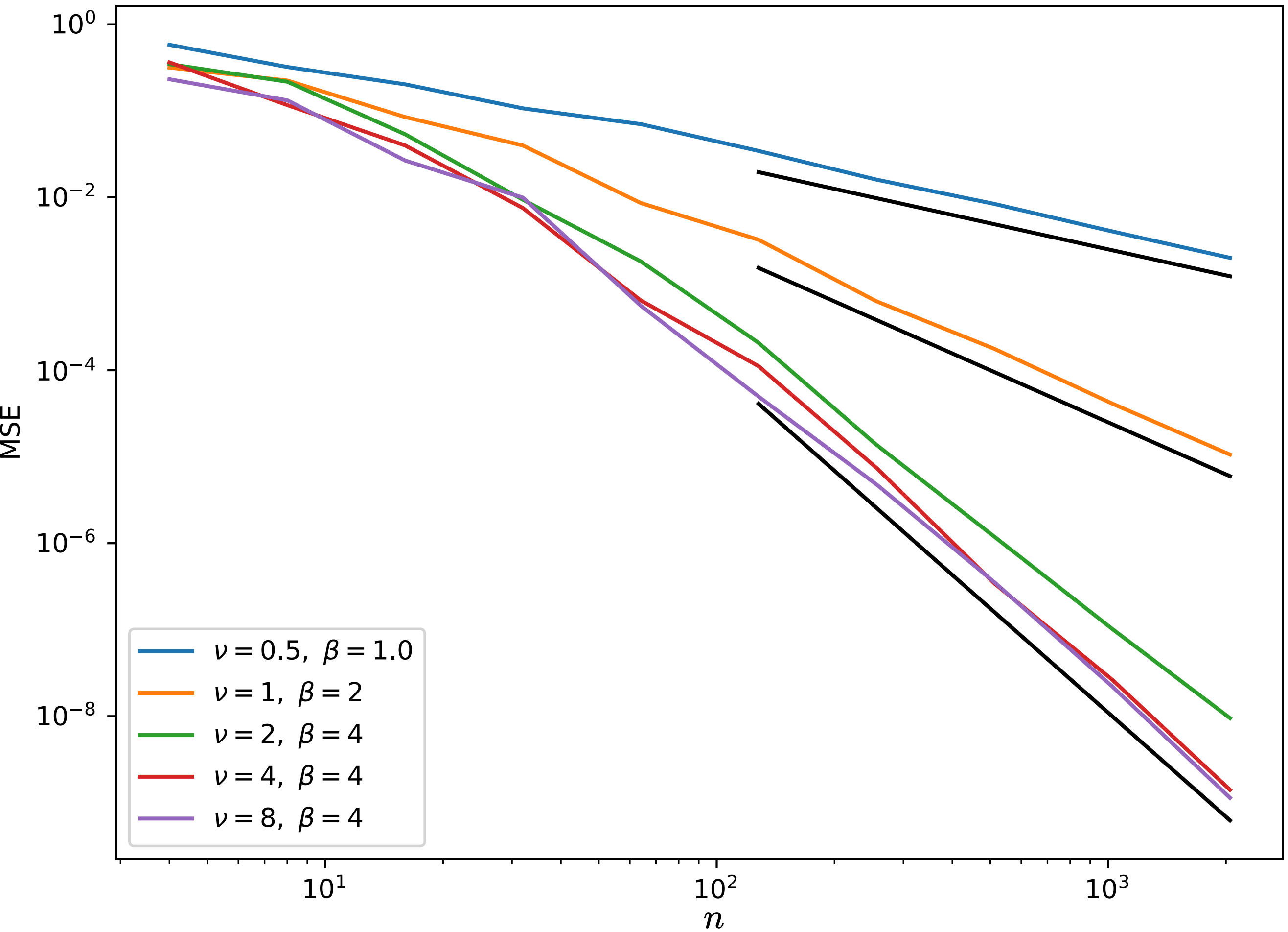
- Our result matches the numerical simulations
  (on hypersphere)
- There are finite size effects (small n)



TEACHER-STUDENT: MATÉRN TEACHER

Matérn kernels: $K_T(\underline{x}) = \frac{2^{1-\nu}}{\Gamma(\nu)} z^\nu \mathcal{K}_\nu(z),\quad z = \sqrt{2\nu}\frac{\|\underline{x}\|}{\sigma},\quad \alpha = d + 2\nu$

Laplace student, $K_S(\underline{x}) = \exp\left(-\frac{\|\underline{x}\|}{\sigma}\right)$



$d = 1$

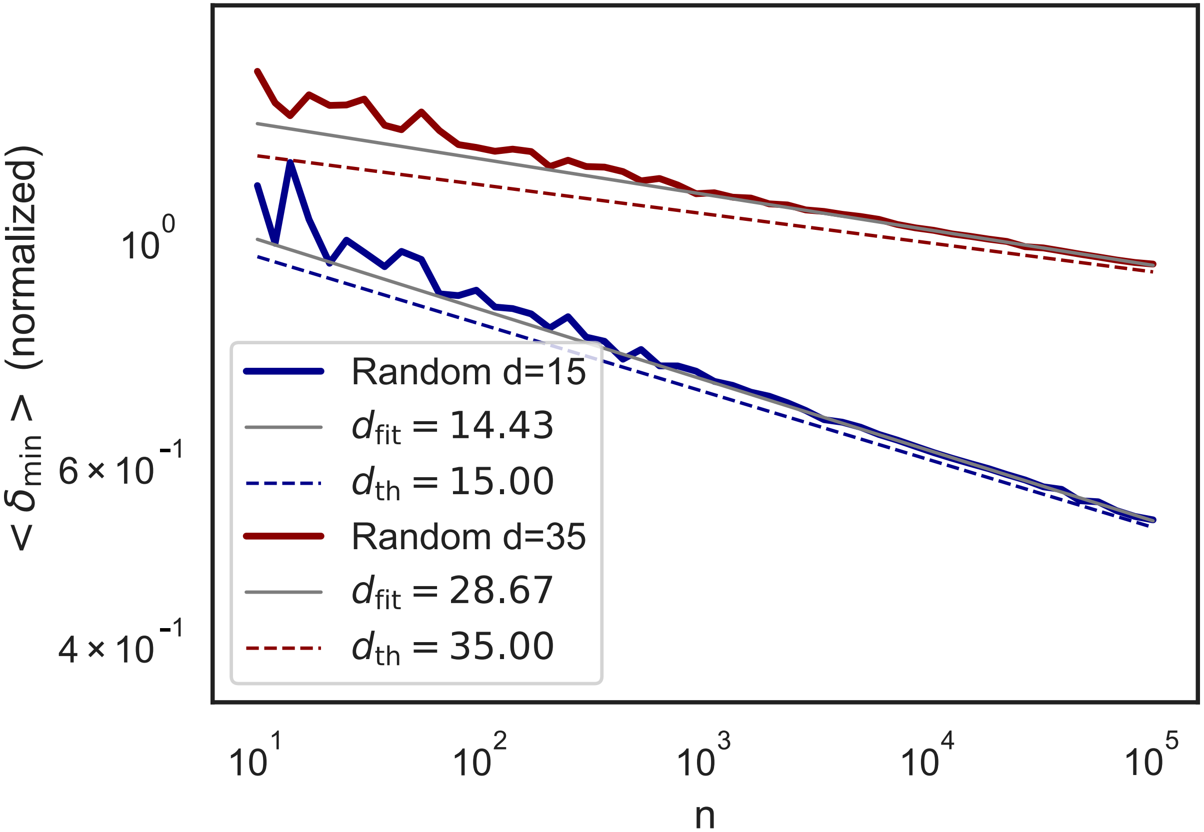
$\beta = \min(2\nu, 4)$

NEAREST-NEIGHBOR DISTANCE

Same result with points on *regular lattice* or *random hypersphere*?

What matters is how **nearest-neighbor distance $\boldsymbol{\delta}$** scales with n

(conjecture)



In both cases $\boxed{\delta \sim n^{\frac{1}{d}}}$

Finite size effects:
asymptotic scaling only
when n is large enough

BACK TOREAL DATA

What about real data?

$\longrightarrow$ *second order approximation* with a Gaussian process K_T:

does it capture some aspects?

BACK TO REAL DATA

What about real data?

$\longrightarrow$ *second order approximation* with a Gaussian process K_T:

does it capture some aspects?

- Gaussian processes are s-times (mean-square) differentiable,

$$s = \frac{\alpha_T - d}{2}$$

BACK TOREAL DATA

What about real data?

$\longrightarrow$ *second order approximation* with a Gaussian process K_T:

does it capture some aspects?

- Gaussian processes are s-times (mean-square) differentiable,

$$s = \frac{\alpha_T - d}{2}$$

- Fitted exponents are $\beta \approx 0.4$ (MNIST) and $\beta \approx 0.1$ (CIFAR10), regardless of the Student $\longrightarrow \beta = \frac{\alpha_T - d}{d}$ ☺

(since $\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)$ indep. of $\alpha_S \longrightarrow \beta = \frac{\alpha_T - d}{d}$)

BACK TOREAL DATA

What about real data?

$\longrightarrow$ *second order approximation* with a Gaussian process K_T:

does it capture some aspects?

- Gaussian processes are s-times (mean-square) differentiable,

$$s = \frac{\alpha_T - d}{2}$$

- Fitted exponents are $\beta \approx 0.4$ (MNIST) and $\beta \approx 0.1$ (CIFAR10), regardless of the Student $\longrightarrow \beta = \frac{\alpha_T - d}{d}$

(since $\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)$ indep. of $\alpha_S \longrightarrow \beta = \frac{\alpha_T - d}{d}$)

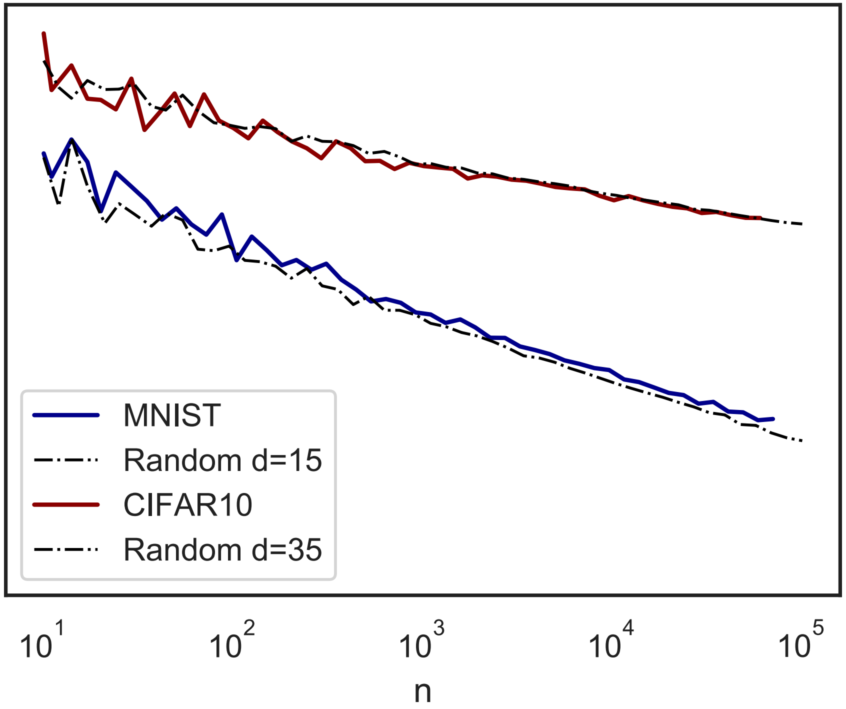
$\longrightarrow s = \frac{1}{2}\beta d$, $s \approx 0.2d \approx 156$ (MNIST) and $s \approx 0.05d \approx 153$ (CIFAR10)

This number is unreasonably large!

EFFECTIVE DIMENSION

- Measure NN-distance δ

- $\delta \sim n^{-\text{some exponent}}$

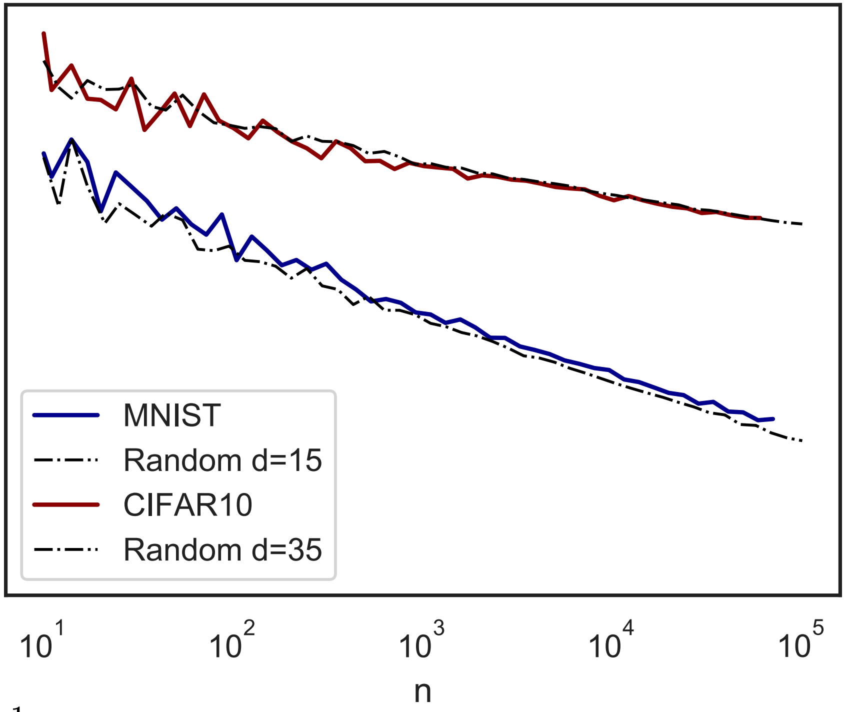


EFFECTIVE DIMENSION

- Measure NN-distance δ

- $\delta \sim n^{-\text{some exponent}}$

$$\downarrow$$



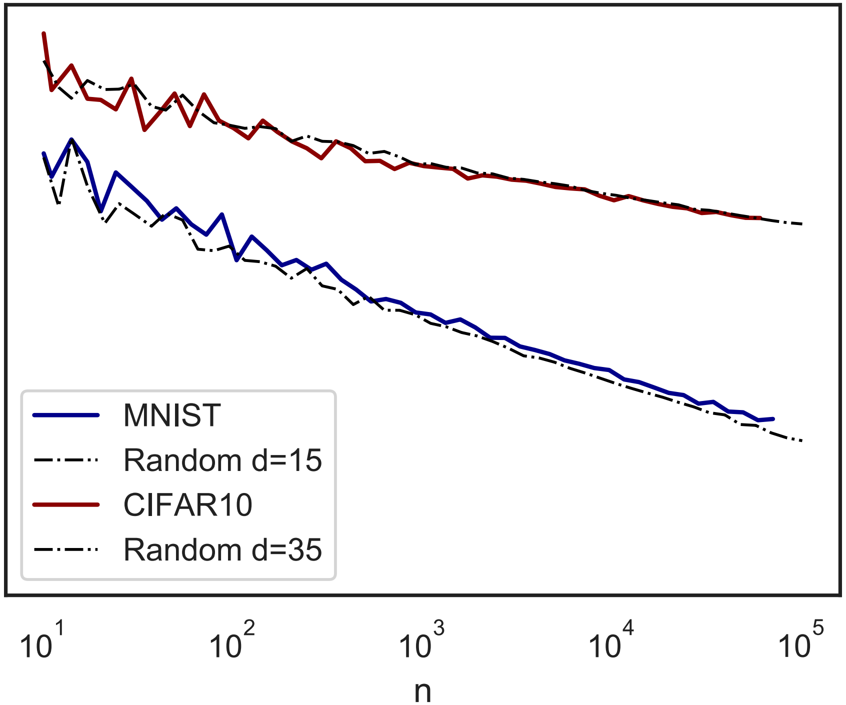
Define effective dimension as $\delta \sim n^{-\frac{1}{d_{\text{eff}}}}$

EFFECTIVE DIMENSION

- Measure NN-distance δ

- $\delta \sim n^{-\text{some exponent}}$

$\downarrow$



Define effective dimension as $\delta \sim n^{-\frac{1}{d_{\text{eff}}}}$

d_{eff} is much smaller

	β	d	d_{eff}	$s = \left\lfloor \frac{1}{2}\beta d_{\text{eff}} \right\rfloor$
MNIST	0.4	784	15	3
CIFAR10	0.1	3072	35	1

s is more reasonable!

CURSE OF DIMENSIONALITY (1/2)

- Loosely speaking, the (optimal) exponent is

$$\beta \approx \frac{\text{smoothness } \alpha_T - d = 2s}{\text{manifold dimension } d}$$

- To avoid the curse of dimensionality ($\beta \sim \frac{1}{d}$):
  - either the dimension of the manifold is small
  - or the data are extremely smooth

RKHS & SMOOTHNESS

- Indeed, what happens if we consider a field $Z_T(\underline{x})$ that
  - is an instance of a Teacher K_T (α_T)
  - lies in the RKHS of a Student K_S (α_S)

RKHS & SMOOTHNESS

- Indeed, what happens if we consider a field $Z_T(\underline{x})$ that
  - is an instance of a Teacher K_T (α_T)
  - lies in the RKHS of a Student K_S (α_S)

$$\mathbb{E}_T \|Z_T\|_{K_S} = \\ \mathbb{E}_T \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, Z_T(\underline{x}) K_S^{-1}(\underline{x},\underline{y}) Z_T(\underline{y}) = \\ \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, K_T(\underline{x},\underline{y}) K_S^{-1}(\underline{x},\underline{y}) < \infty \quad \Longrightarrow \quad \alpha_T > \alpha_S + d$$

RKHS & SMOOTHNESS

- Indeed, what happens if we consider a field $Z_T(\underline{x})$ that
  - is an instance of a Teacher K_T (α_T)
  - lies in the RKHS of a Student K_S (α_S)

$$\mathbb{E}_T \|Z_T\|_{K_S} = \mathbb{E}_T \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, Z_T(\underline{x}) K_S^{-1}(\underline{x},\underline{y}) Z_T(\underline{y}) = \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, K_T(\underline{x},\underline{y}) K_S^{-1}(\underline{x},\underline{y}) < \infty \implies \alpha_T > \alpha_S + d$$

$$K_S(\underline{0}) \propto \int \mathrm{d}\underline{w}\, \tilde{K}_S(\underline{w}) < \infty \implies \alpha_S > d$$

RKHS & SMOOTHNESS

- Indeed, what happens if we consider a field $Z_T(\underline{x})$ that
  - is an instance of a Teacher K_T (α_T)
  - lies in the RKHS of a Student K_S (α_S)

$$\mathbb{E}_T \|Z_T\|_{K_S} = \mathbb{E}_T \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, Z_T(\underline{x}) K_S^{-1}(\underline{x},\underline{y}) Z_T(\underline{y}) = \int \mathrm{d}^d\underline{x}\mathrm{d}^d\underline{y}\, K_T(\underline{x},\underline{y}) K_S^{-1}(\underline{x},\underline{y}) < \infty \quad \Longrightarrow \quad \alpha_T > \alpha_S + d$$

$$K_S(\underline{0}) \propto \int \mathrm{d}\underline{w}\, \tilde{K}_S(\underline{w}) < \infty \quad \Longrightarrow \quad \alpha_S > d$$

(it scales with d!)

Therefore the smoothness must be $s = \frac{\alpha_T - d}{2} > \frac{d}{2}$

$$\longrightarrow \beta > \frac{1}{2}$$

CURSE OF DIMENSIONALITY (2/2)

- Assume that the data are not smooth enough and live in d large

- **Dimensionality reduction** in the task rather than in the data?

- E.g. the n points $\underline{x}_\mu$ live in $\mathbb{R}^d$, but the target function is such that

$$Z_T(\underline{x}) = Z_T(\underline{x}_{\parallel}) \equiv Z_T(x_1, \ldots, x_{d_{\parallel}}), \quad d_{\parallel} < d$$

Similar setting studied in Bach 2017

- Can kernels understand the lower dimensional structure?

TASK INVARIANCE: KERNEL REGRESSION (1/2)

Theorem (informal formulation):

in the described setting with $d_{\|} \leq d$,

for $n \gg 1$ $\quad \epsilon \sim n^{-\beta} \quad$ with $\quad \boxed{\beta = \frac{1}{d}\min(\alpha_T - d, 2\alpha_S)}$

Regardless of $d_{\|}$!

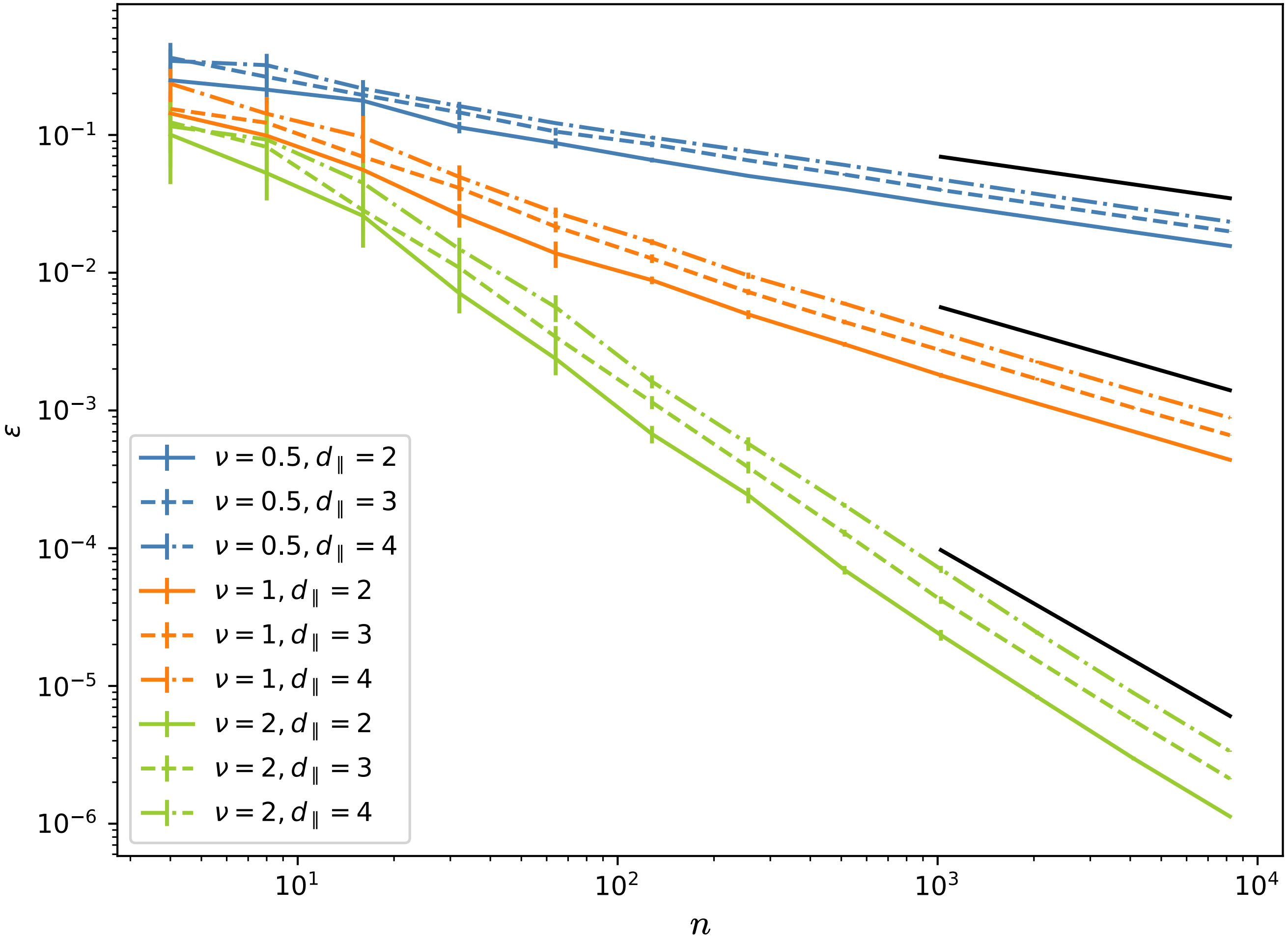
Similar result in Bach 2017

Two reasons contribute to this result:

- the nearest-neighbor distance always scales as $\delta \sim n^{-\frac{1}{d}}$
- $\alpha_T(d) - d$ only depends on the function $K_T(z)$ and not on d

TASK INVARIANCE: KERNEL REGRESSION (2/2)

Teacher = Matérn (with parameter ν), Student = Laplace, d=4



TASK INVARIANCE: CLASSIFICATION (1/2)

Classification with the **margin SVM** algorithm:

$$\hat{y}(\underline{x}) = \text{sign}\left[\sum_{\mu=1}^{n} c_\mu K\left(\frac{\|\underline{x}-\underline{x}^\mu\|}{\sigma}\right) + b\right]$$

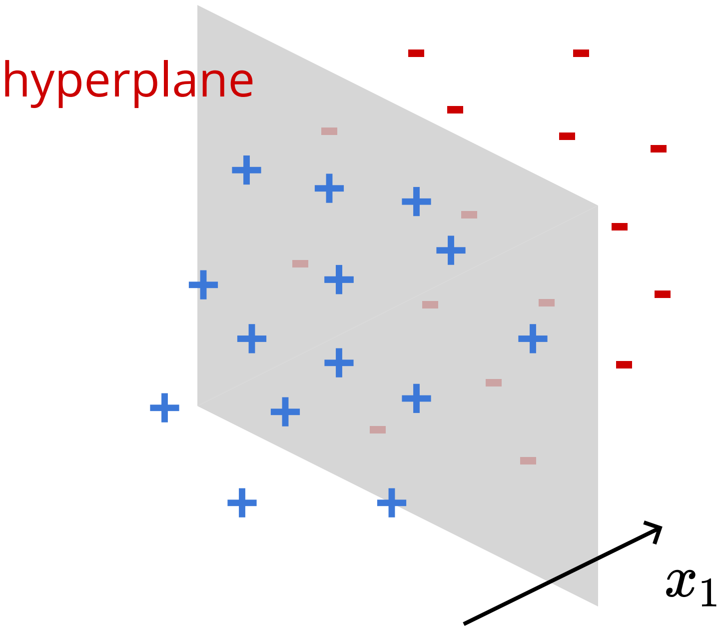
find $\{c_\mu\}, b$ by minimizing some function

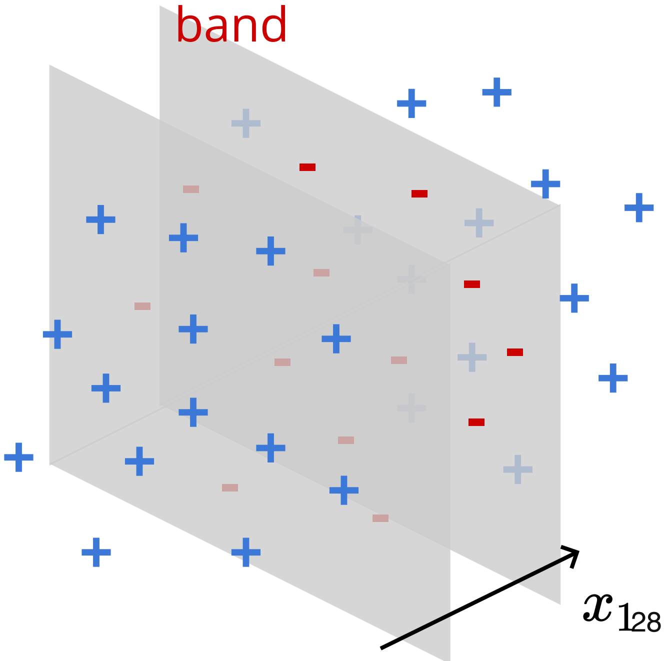
We consider a very simple setting:

- the label is $y(\underline{x}) = y(x_1) \longrightarrow d_\parallel = 1$

Non-Gaussian data!

$y(x_1)$:





TASK INVARIANCE: CLASSIFICATION (2/2)

Vary **kernel scale** $\sigma \longrightarrow$ **two regimes!**

- $\sigma \ll \delta$: then the estimator is tantamount to a **nearest-neighbor algorithm** $\longrightarrow$ curse of dimensionality $\beta = \frac{1}{d}$

- $\sigma \gg \delta$: important **correlations** in c_μ due to the **long-range kernel**. For the hyperplane with $d_\parallel = 1$ we find $\beta = \mathcal{O}(d^0)$!

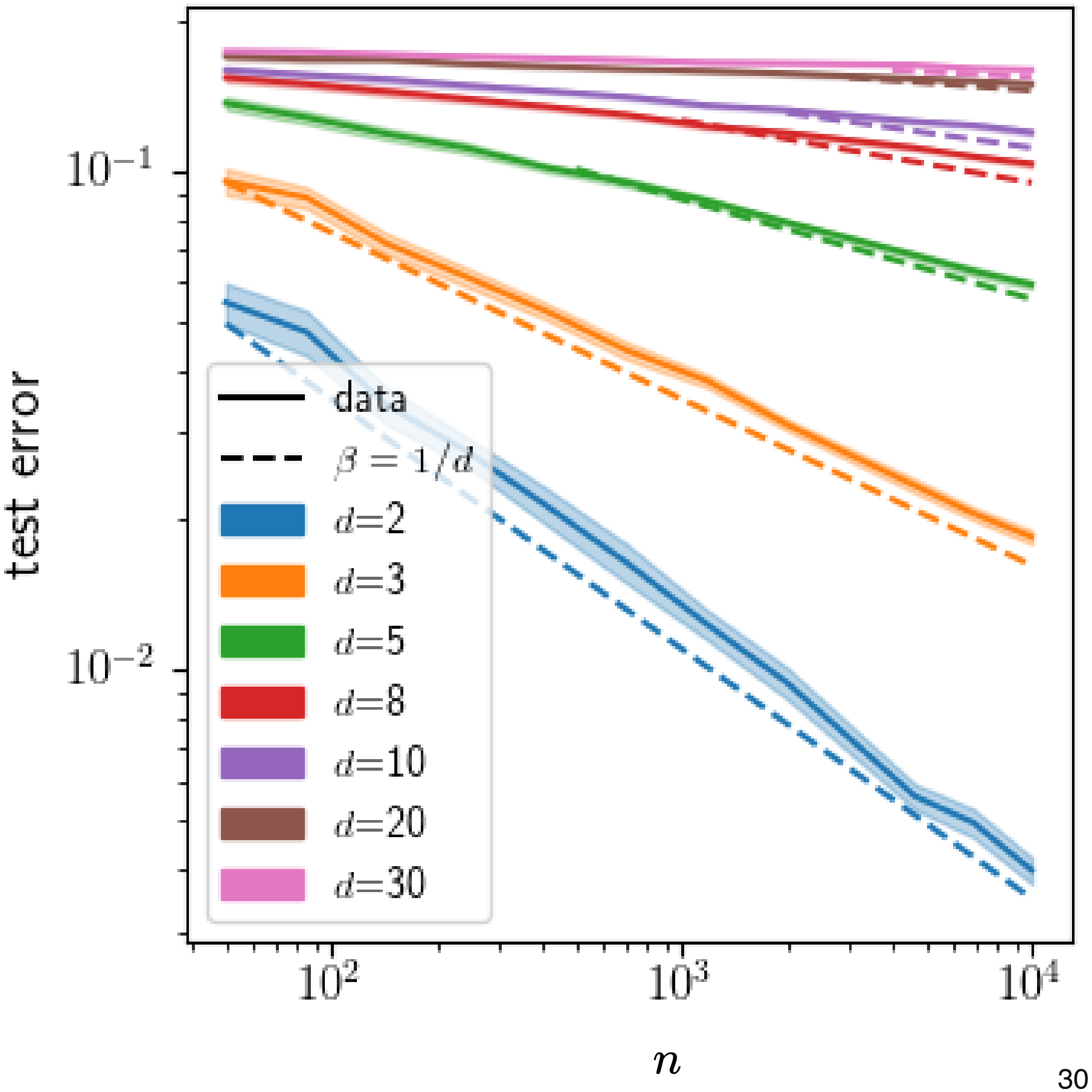
No curse of dimensionality!

THE NEAREST-NEIGHBOR LIMIT

hyperplane
interface

using a Laplace kernel
and
varying the dimension d:

$\beta = \frac{1}{d}$



KERNEL CORRELATIONS (1/2)

When $\sigma \gg \delta$ we can expand the kernel overlaps:

$$K\left(\frac{\|x-x^\mu\|}{\sigma}\right) \approx K(0) - \text{const} \times \left(\frac{\|x-x^\mu\|}{\sigma}\right)^{\xi}$$

(the exponent ξ is linked to the smoothness of the kernel)

We can derive some scaling arguments that lead to an exponent

$$\boxed{\beta = \frac{d+\xi-1}{3d+\xi-3}}$$

KERNEL CORRELATIONS (1/2)

When $\sigma \gg \delta$ we can expand the kernel overlaps:

$$K\left(\frac{\|\underline{x}-\underline{x}^{\mu}\|}{\sigma}\right) \approx K(0) - \text{const} \times \left(\frac{\|\underline{x}-\underline{x}^{\mu}\|}{\sigma}\right)^{\xi}$$

(the exponent ξ is linked to the smoothness of the kernel)

We can derive some scaling arguments that lead to an exponent

$$\boxed{\beta = \frac{d+\xi-1}{3d+\xi-3}}$$

Idea:

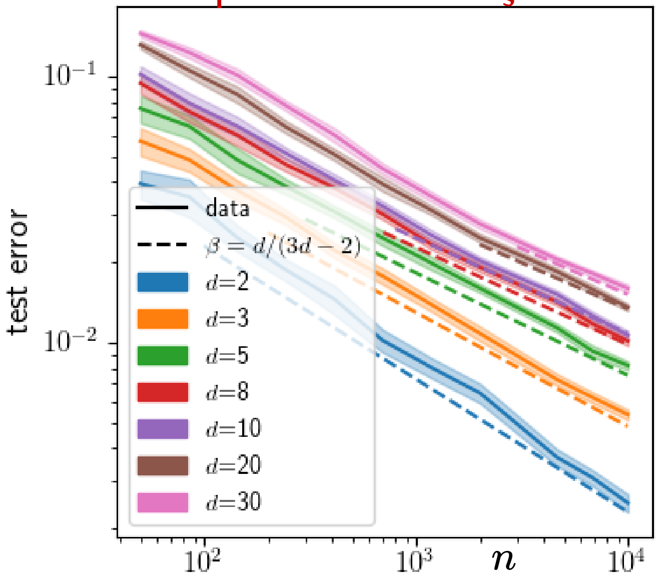
- support vectors ($c_\mu \neq 0$) are close to the interface
- we impose that the decision boundary has $\mathcal{O}(1)$ spatial fluctuations on a scale proportional to δ

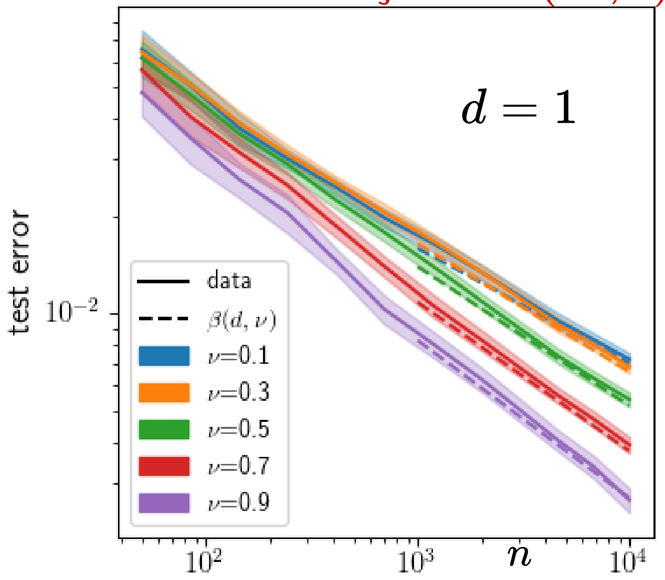
KERNEL CORRELATIONS (2/2)

Laplace kernel $\xi = 1$

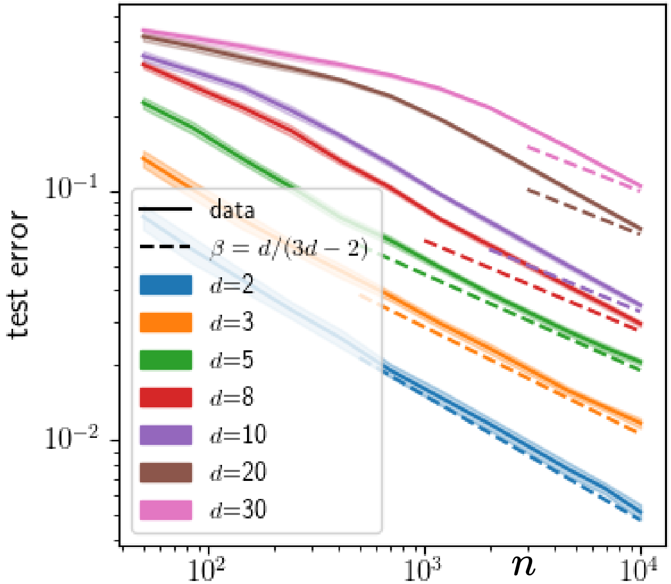
Matérn kernels $\xi = \min(2\nu, 2)$

hyperplane





band



$$\beta = \frac{d+\xi-1}{3d+\xi-3}$$

in all these cases!

KERNEL CORRELATIONS: HYPERSPHERE

What about other interfaces?

boundary = hypersphere:

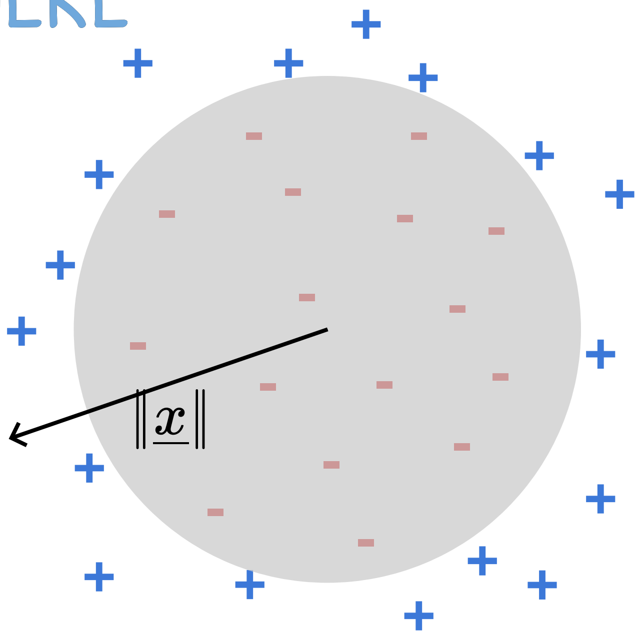
$$y(\underline{x}) = \mathrm{sign}(\|\underline{x}\| - R)$$

$(d_{\parallel} = 1)$

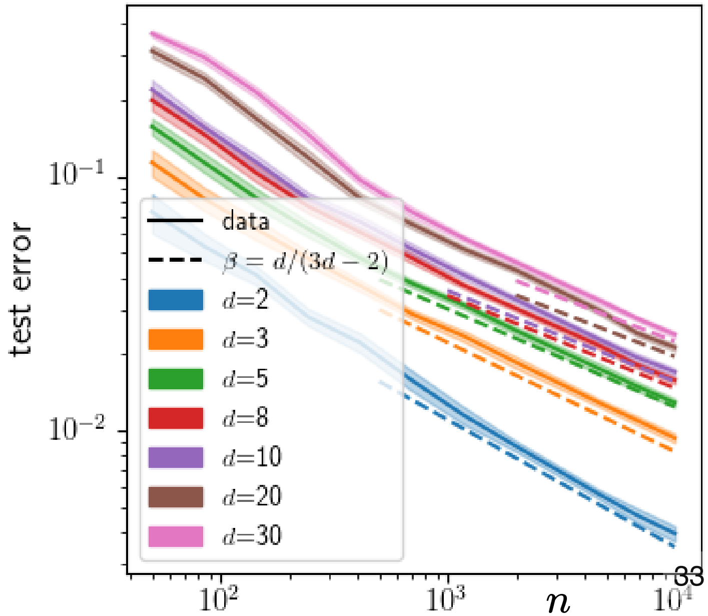
$$\beta = \frac{d+\xi-1}{3d+\xi-3}$$

(same exponent!)

(similar scaling arguments apply, provided $R \gg \delta$)



Laplace kernels ($\xi = 1$)



CONCLUSION

arXiv:1905.10843 + paper to be released soon!

- Learning curves of real data decay as **power laws** with exponents

$$\frac{1}{d} \ll \beta < \frac{1}{2}$$

- We introduce a **new framework** that links the exponent β to the degree of smoothness of Gaussian random data

- We justify how different kernels can lead to the same exponent β

- We show that the **effective dimension** of real data is $\ll d$. It can be linked to a (small) **effective smoothness** s

- We show that kernel regression is not able to capture invariants in the task, while kernel classification can

(in some regime and for **smooth interfaces**)