The feature space decomposition method

Applications to benign overfitting and kernel ridge regression

Guillaume Lecué
joint works with George Gavrilopoulos and Zong Shang

ESSEC, Paris

Vienna – September 2025

Feature space decomposition: idea 1/3

Consider N iid data in the linear regression model

$$Y_i = \langle X_i, \beta^* \rangle + \xi_i, i = 1, \dots, N$$

where $\mathbb{E}X_i = 0$, $\mathbb{E}X_i X_i^\top = \Sigma$, $\beta^* \in \mathbb{R}^p$ and $\mathbb{E}\xi_i = 0$ ind. of X_i.

Prediction = estimation

$$R(\hat{\beta}) - R(\beta^*) = \left\| \Sigma^{1/2}(\hat{\beta} - \beta^*) \right\|_2^2$$

where $R(\beta) = \mathbb{E}(Y - \langle X, \hat{\beta} \rangle)^2$

Our aim: obtain sharp bounds on $\left\| \Sigma^{1/2}(\hat{\beta} - \beta^*) \right\|_2^2$ for classical estimators $\hat{\beta}$ using the **feature space decomposition method**.

Feature space decomposition: idea 2/3

Assume that the spectrum of Σ is for $0 < \epsilon << 1$

$$\sigma_j = \begin{cases} 1 & \text{if } 1 \leq j \leq k \\ \epsilon & \text{if } k+1 \leq j \leq p. \end{cases}$$

then for $\hat{\beta} = \hat{\beta}_{1:k} + \hat{\beta}_{k+1:p}$

$$\begin{aligned} \left\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\right\|_2^2 &= \left\|\hat{\beta}_{1:k} - \beta^*_{1:k}\right\|_2^2 + \epsilon^2 \left\|\hat{\beta}_{k+1:p} - \beta^*_{k+1:p}\right\|_2^2 \\ &\leq \left\|\hat{\beta}_{1:k} - \beta^*_{1:k}\right\|_2^2 + 2\epsilon^2 \left(\left\|\hat{\beta}_{k+1:p}\right\|_2^2 + \left\|\beta^*_{k+1:p}\right\|_2^2\right) \end{aligned}$$

where

- $\beta \to \beta_{1:k}$ is the projection operator on $V_{1:k}$ = eigenspace of the top k eigenvectors of Σ
- $\beta \to \beta_{k+1:p}$ is the projection on $V_{k+1:p}$ = eigenspace of the last $p-k$ eigenvectors of Σ

Feature space decomposition: idea 3/3

$$\left\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\right\|_2^2 \leq \left\|\hat{\beta}_{1:k}-\beta^*_{1:k}\right\|_2^2 + 2\epsilon^2\left\|\hat{\beta}_{k+1:p}\right\|_2^2 + 2\epsilon^2\left\|\beta^*_{k+1:p}\right\|_2^2$$

Two ideas

- we don't expect $\hat{\beta}_{k+1:p}$ to estimate $\beta^*_{k+1:p} \Rightarrow$ we can use $V_{k+1:p}$ to do something else than estimation
- we only need to pay the cost of estimation (sample size, assumption on the model,..) only on the lower dimensional space $V_{1:k}$ instead of $\mathbb{R}^p$

signal alignement with top eigenvectors of the features

It will work

- under conditions on the spectrum of Σ
- if the signal β^* is mostly aligned with the top k eigenvectors of Σ

FSD: properties of the projected design matrices

Decomposition of the design matrix

$$\mathbb{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_N^\top \end{pmatrix} = \mathbb{X}_{1:k} + \mathbb{X}_{k+1:p}$$

where:

- $\mathbb{X}_{1:k}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}_{1:k}$ is the design matrix on the estimation part $V_{1:k}$ of the feature space
- $\mathbb{X}_{k+1:p} = \mathbb{X}\boldsymbol{\beta}_{k+1:p}$ is the design matrix on 'free' part $V_{k+1:p}$ of the feature space

We will need

- 'classical' properties for $\mathbb{X}_{1:k}$ required for estimation (control of some quadratic and multiplier processes)
- 'new properties' for $\mathbb{X}_{k+1:p}$: the Dvoretsky-Milman property

The Dvoretzky-Milman theorem

Theorem (Dvoretsky-Milman with Gaussian random matrix)

There are absolute constants $\kappa_{DM} \leq 1$ and c_1 such that the following holds. Let $\|\cdot\|$ be some norm on $\mathbb{R}^p$ and denote by B its unit ball and B^* its unit dual ball. Denote by $\mathbb{G} := \mathbb{G}^{(N\times p)}$, the $N \times p$ standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ Gaussian entries. Given any $0 < \epsilon \leq 1$. Assume that $N \leq \kappa_{DM}\epsilon^2 d_*(B)$. Then with probability at least $1 - \exp(-c_1\epsilon^2 d_*(B))$, for every $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$(1-\epsilon)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*) \leq \left\|\mathbb{G}^\top \boldsymbol{\lambda}\right\| \leq (1+\epsilon)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*)$$

where, the Dvoretsky-Milman dimension is

$$d_*(B) = \left(\frac{\ell_*(B^*)}{\operatorname{diam}(B^*, \ell_2^p)}\right)^2.$$

where $\ell_*(B^*) = \mathbb{E}\sup_{t\in B^*}\langle G, t\rangle = \mathbb{E}\,\|G\|$ for $G \sim \mathcal{N}(0, I_p)$.

Task in progress for stat.: Extend this result beyond the Gaussian case.

Application 1:
benign overfitting ot the minimum ℓ_2^d-norm interpolant estimator

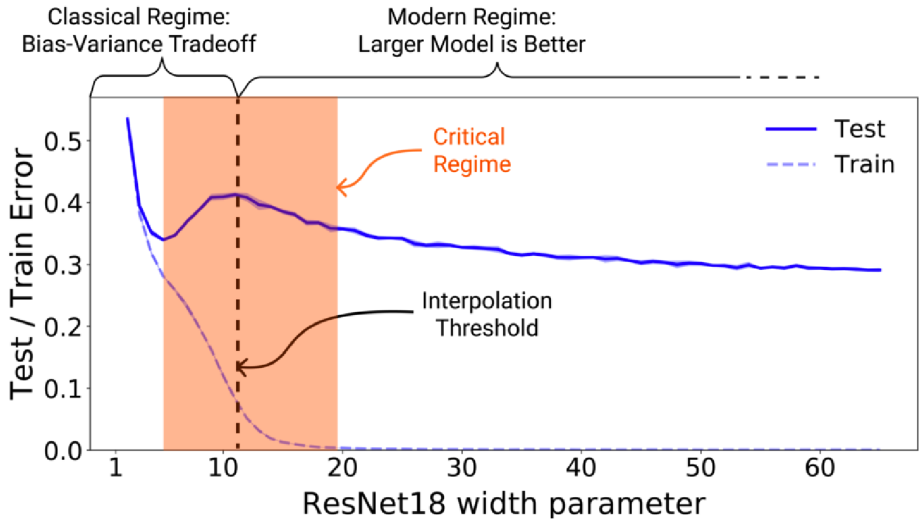
Bartlett, Long, Lugosi, Tsigler. *Benign Overfitting in Linear Regression*

Two surveys in **Acta Numerica**:

Bartlett, Montanari, Rahklin. *Deep learning: a statistical viewpoint'*

Belkin. *Fit without fear: remarkable mathematical phenomena of deep learning..'*

Double-descent and interpolant estimators



Nakkiran Et al..*Deep double descent*. 2021

The double descent phenomenon happens for over-parametrized models:

$$\text{number of parameters} >> \text{number of data}$$

and for interpolant estimators: $\boxed{\hat{f}(X_i) = Y_i, \forall i = 1, \ldots, N.}$

How is it possible that interpolant estimators generalize well?
Because the IE uses the free part of the feature space to interpolate the noise!

Which interpolant estimators appear in neural networks?

Idea 1: For large models there are many interpolant estimators: not all are good.
Idea 2: some algorithms used to train some wide neural networks tend to interpolant estimators.
Key idea: in general, algorithms used to train neural networks do not use an explicit regularization however they regularize implictly!

Implicit regularization / implicit bias toward smoothness

(informal) Proposition (implicit ℓ_2-regularization)

A SGD algorithm which interpolates the data at some point:

- stops its convergence, it becomes constant,
- equal to the interpolant estimator with the smallest ℓ_2^d-nom

Other algorithms are known to make implicit regularization
Vardi, Shamir, 'implicit regularization in ReLu networks with square loss..'
Gunasekar, Lee, Soudry, Srebro, 'Characterizing Implicit bias in terms'...

Understand explicit regularized IE in simple models

Linear regression model with (anisotropic) Gaussian design and independent Gaussian noise: $Y_i = \langle X_i, \beta^* \rangle + \xi_i$, $i = 1, \dots, N$ where $X_i \sim \mathcal{N}_p(0, \Sigma)$, $\beta^* \in \mathbb{R}^p$ and $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$ ind. of X_i. We note

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_N^\top \end{pmatrix} = \mathbb{G}^{(N \times p)} \Sigma^{1/2} \text{ and } \boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_N \end{pmatrix}$$

where $\mathbb{G}^{(N \times p)}$ is a $N \times p$ standard Gaussian matrix (i.i.d. $\mathcal{N}(0,1)$ entries). The minimum ℓ_2-norm interpolant estimator is

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ (\|\beta\|_2 : \mathbb{X}\beta = y)$$

Questions

Under which conditions the minimum ℓ_2-norm interpolant estimator $\hat{\beta}$ generalizes well in the Gaussian linear model?
Can we use the FSD to get sharp bounds for $\hat{\beta}$?

The FSD for the min-$\ell_2$norm IE

$$\mathbb{R}^p = V_{1:k^*} \otimes^{\perp} V_{k^*+1:d}$$

where

- $V_{1:k^*}$ is spanned by the k^* top singular vectors of Σ
- $V_{k^*+1:d}$ is spanned by the $p - k^*$ smallest ones

for

$$k^* = \min \left\{ k \in \{1, \ldots, p\} : N \lesssim \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{\sigma_{k+1}} \right\}$$

for the SVD, $\Sigma = \sum_{j=1}^{p} \sigma_j u_j u_j^\top$,

$$\Sigma_{1:k^*} = \sum_{j=1}^{k^*} \sigma_j u_j u_j^\top \text{ and } \Sigma_{k^*+1:p} = \sum_{j=k^*+1}^{p} \sigma_j u_j u_j^\top$$

We have $\mathbb{X} = \mathbb{G}^{(N\times p)}\Sigma = \mathbb{X}_{1:k^*} + \mathbb{X}_{k^*+1:p}$ where

$$\mathbb{X}_{1:k^*} = \mathbb{G}^{(N\times p)}\Sigma_{1:k^*} \text{ and } \mathbb{X}_{k^*+1:p} = \mathbb{G}^{(N\times p)}\Sigma_{k^*+1:p}.$$

Two geometrical tools on $V_{1:k^*}$ and $V_{k^*+1:p}$

★ **[RIP]** if $k^* \leq c_0 N$, then, w.h.p. for all $v \in V_{1:k^*}$,

$$\frac{1}{2}\left\|\Sigma_{1:k^*}^{1/2} v\right\|_2 \leq \frac{\left\|\mathbb{X}_{1:k^*} v\right\|_2}{\sqrt{N}} \leq \frac{3}{2}\left\|\Sigma_{1:k^*}^{1/2} v\right\|_2$$

and if $k^* \geq c_1 N$ this holds only on the cone $\left\{\left\|\Sigma_{1:k^*}^{1/2} v\right\|_2 \geq r^* \|v\|_2\right\}$.

★ **[Dvoretsky-Milman]** if

$$N \lesssim d^*(\Sigma_{k^*+1:p}^{-1/2} B_2^p) \sim \frac{\mathrm{Tr}(\Sigma_{k^*+1:p})}{\sigma_{k^*+1}}$$

then w.h.p. for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\frac{\sqrt{\mathrm{Tr}(\Sigma_{k^*+1:p})}}{2}\|\boldsymbol{\lambda}\|_2 \leq \left\|\mathbb{X}_{k^*+1:p}{}^\top \boldsymbol{\lambda}\right\|_2 \leq \frac{3\sqrt{\mathrm{Tr}(\Sigma_{k^*+1:p})}}{2}\|\boldsymbol{\lambda}\|_2$$

and so, for $A = \mathbb{X}_{k^*+1:p}^\top (\mathbb{X}_{k^*+1:p}\mathbb{X}_{k^*+1:p}^\top)^{-1}$,

$$\frac{1}{2\sqrt{\mathrm{Tr}(\Sigma_{k^*+1:p})}}\|\boldsymbol{\lambda}\|_2 \leq \|A\boldsymbol{\lambda}\|_2 \leq \frac{4}{\sqrt{\mathrm{Tr}(\Sigma_{k^*+1:p})}}\|\boldsymbol{\lambda}\|_2 .$$

Decomposition of the min-$\ell_2$norm IE

We have $\hat{\beta} = \hat{\beta}_{1:k^*} + \hat{\beta}_{k^*+1:p}$ where $\mathbb{X} = \mathbb{X}_{1:k^*} + \mathbb{X}_{k^*+1:p}$ and

$$\hat{\beta}_{1:k^*} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\|A(y - \mathbb{X}_{1:k^*}\beta)\|_2^2 + \|\beta\|_2^2 \right)$$

$$\underset{D.M.}{\sim} \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\|y - \mathbb{X}_{1:k^*}\beta\|_2^2 + \operatorname{Tr}(\Sigma_{k^*+1:p}) \|\beta\|_2^2 \right)$$

where $A = \mathbb{X}_{k^*+1:p}{}^\top (\mathbb{X}_{k^*+1:p}\mathbb{X}_{k^*+1:p}{}^\top)^{-1}$ and

$$\hat{\beta}_{k^*+1:p} = \underset{\beta}{\operatorname{argmin}} \left(\|\beta\|_2 : \mathbb{X}_{k^*+1:p}\beta = y - \mathbb{X}_{1:k^*}\hat{\beta}_{1:k^*} \right)$$

$$= A(y - \mathbb{X}_{1:k^*}\hat{\beta}_{1:k^*}).$$

$\hat{\beta} =$ $\underbrace{\hat{\beta}_{1:k^*}}$ + $\underbrace{\hat{\beta}_{k^*+1:p}}$

a 'ridge' estimator on $v_{1:k^*}$

min l2 IE of the residuals of $\hat{\beta}_{1:k^*}$

Decomposition of the excess risk (general case)

Excess risk decomposition:

$$\left\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\right\|_2^2 = \left\|\Sigma_{1:k^*}^{1/2}(\hat{\beta}_{1:k^*}-\beta^*_{1:k^*})\right\|_2^2 + \left\|\Sigma_{k^*+1:p}^{1/2}(\hat{\beta}_{k^*+1:p}-\beta^*_{k^*+1:p})\right\|_2^2$$

where

- $\left\|\Sigma_{1:k^*}^{1/2}(\hat{\beta}_{1:k^*}-\beta^*_{1:k^*})\right\|_2^2$ = estimation part: $\beta^*_{1:k^*}$ is estimated by the 'ridge' estimator $\hat{\beta}_{1:k^*}$
- $\left\|\Sigma_{k^*+1:p}^{1/2}(\hat{\beta}_{k^*+1:p}-\beta^*_{k^*+1:p})\right\|_2^2$ = price for overfitting

$$\left\|\Sigma_{k^*+1:p}^{1/2}(\hat{\beta}_{k^*+1:p}-\beta^*_{k^*+1:p})\right\|_2 \leq \left\|\Sigma_{k^*+1:p}^{1/2}\hat{\beta}_{k^*+1:p}\right\|_2 + \left\|\Sigma_{k^*+1:p}^{1/2}\beta^*_{k^*+1:p}\right\|_2$$

Conclusion 1: A large part of the space $\mathbb{R}^p$ (i.e. $V_{k^*+1:p}$) is used to interpolate the data.
Conclusion 2: Benign overfitting requires that the price for overfitting on $V_{k^*+1:p}$ does not harm the estimation property of $\hat{\beta}$.

Upper bound

Theorem (L. and Shang)

With probability at least $1 - \exp(-c_0 k^*)$,

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \lesssim \max\left\{\sigma_\xi\sqrt{\frac{k^*}{N}}, \left\|\Sigma_{1:k^*}^{-1/2}\boldsymbol{\beta}^*_{1:k^*}\right\|_2\left(\frac{\mathrm{Tr}(\Sigma_{k^*+1:p})}{N}\right),\right.$$
$$\left.\left\|\Sigma_{k^*+1:p}^{1/2}\boldsymbol{\beta}^*_{k^*+1:p}\right\|_2, \sigma_\xi\sqrt{\frac{N\mathrm{Tr}(\Sigma^2_{k^*+1:p})}{\mathrm{Tr}^2(\Sigma_{k^*+1:p})}}\right\}.$$

It improves on Tsigler, Bartlett. 'Benign overfitting in ridge regression':

- deviation from constant to exponentially small
- remove unecessary conditions thanks to Dvoretsky-Milman theorem
- extend the range of application
- proof based on the FSD

Matching lower bound

Theorem (L. and Shang)

If Σ is such that $k^* < N/4$ then

$$\mathbb{E}\left\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\right\|_2 \gtrsim \max\left\{\sigma_\xi\sqrt{\frac{k^*}{N}}, \left\|\Sigma_{1:k^*}^{-1/2}\beta^*_{1:k^*}\right\|_2\left(\frac{\mathrm{Tr}(\Sigma_{k^*+1:p})}{N}\right),\right.$$
$$\left.\left\|\Sigma_{k^*+1:p}^{1/2}\beta^*_{k^*+1:p}\right\|_2, \sigma_\xi\sqrt{\frac{N\mathrm{Tr}(\Sigma^2_{k^*+1:p})}{\mathrm{Tr}^2(\Sigma_{k^*+1:p})}}\right\}.$$

It improves on the Bayesian lower bounds from
Bartlett, Long, Lugosi, Tsigler. 'Benign overfitting in linear regression'
and
Tsigler, Bartlett. 'Benign overfitting in ridge regression'

Necessary and sufficient conditions for benign overfitting of the min-ℓ_2 interpolant estimator in linear regression

We say that **overfitting is benign for the min-ℓ_2 IE** when (Σ, β^*) is s.t.

A) estimation happens on a small dimension space k^*:

$$k^* = o(N)$$

B) $(\sigma_{k^*+1}, \ldots, \sigma_p)$ is 'well-spread' (i.e. it cannot be well approximated by a N-sparse vector):

$$N\mathrm{Tr}(\Sigma^2_{k^*+1:p}) = o\left(\mathrm{Tr}(\Sigma_{k^*+1:p})^2\right)$$

C) β^* is mostly supported on the eigenspace of the top k^* eigenvectors of Σ:

$$\left\|\Sigma^{1/2}_{k^*+1:p}\beta^*_{k^*+1:p}\right\|_2 = o(1)$$

D) $\sigma_{k^*} >> \sigma_{k^*+1}$:

$$\left\|\Sigma^{-1/2}_{1:k^*}\beta^*_{1:k^*}\right\|_2 \left(\frac{\mathrm{Tr}(\Sigma_{k^*+1:p})}{N}\right)^2 = o(1)$$

Extention to the heavy-tailed case

We obtain the same rate (but for a different deviation probability) when:

- $X = \Sigma^{1/2} Z$ where Z is:
  - symmetric with independent coordinates
  - there is some $\alpha \leq 2$ such that for all $2 \leq q \leq \log N$ and all $v \in \mathbb{R}^p$,
$$\left\| \langle X, v \rangle \right\|_{L_q} \leq C q^{1/\alpha} \left\| \langle X, v \rangle \right\|_{L_2}$$
- the noise ξ is
  - mean zero and independent of X
  - there is some $r > 4$ such that $\|\xi\|_{L_r} \leq C \|\xi\|_{L_2}$.

Application 2:
kernel ridge regression

Tsigler, Bartlett. *benign overfitting in ridge regression*

Mourtada, J. and Rosasco, L. *An elementary analysis of ridge regression with random design.*

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. *Linearized two-layers neural networks in high dimension.*

Liang and Rakhlin. *Just Interpolate: Kernel "Ridgeless" Regression Can Generalize*

Liang, Rakhlin, Zhai. *On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels*

Caron, Chrétien. *A finite sample analysis of the benign overfitting phenomenon for ridge function estimation*

Application to kernel ridge regression – 1/2

We have N iid data in the model

$$Y = f^*(X) + \xi = \left\langle \phi(X), f^* \right\rangle + \xi$$

where $f^* \in \mathcal{H}$ a RKHS with kernel $K : \Omega \times \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$ is compact and $\|K\|_\infty \leq 1$ and $\phi : x \in \Omega \to K(x, \cdot) \in \mathcal{H}$ is the feature map. The KRR with regularization parameter $\lambda \geq 0$:

$$\hat{f}_\lambda \in \operatorname*{argmin}_{f \in \mathcal{H}} \left(\sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

We have

$$\hat{f}_\lambda = \mathbb{X}_\phi^\top \left(\mathbb{X}_\phi \mathbb{X}_\phi^\top + \lambda I_N \right)^{-1} y.$$

where

$$\mathbb{X}_\phi = \begin{pmatrix} \phi(X_1)^\top \\ \vdots \\ \phi(X_N)^\top \end{pmatrix}, \text{ so that } \mathbb{X}_\phi f = \begin{pmatrix} \langle \phi(X_1), f \rangle_{\mathcal{H}} \\ \vdots \\ \langle \phi(X_N), f \rangle_{\mathcal{H}} \end{pmatrix} = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_N) \end{pmatrix}$$

Application to kernel ridge regression – 2/2

Theorem (Gavrilopoulos L. and Shang)

For $\Sigma = \mathbb{E}\phi(X) \otimes \phi(X)$, when $N \lesssim (\lambda + \mathrm{Tr}(\Sigma_{k^*+1:p}))/\sigma_{k^*+1}$,

$$\left\| \hat{f}_\lambda - f^* \right\|_{L_2} \lesssim \max \left\{ \sigma_\xi \sqrt{\frac{k^*}{N}}, \left\| \Sigma_{1:k^*}^{-1/2} \beta^*_{1:k^*} \right\|_2 \left(\frac{\lambda + \mathrm{Tr}(\Sigma_{k^*+1:p})}{N} \right), \right.$$
$$\left. \left\| \Sigma_{k^*+1:p}^{1/2} \beta^*_{k^*+1:p} \right\|_2, \sigma_\xi \frac{\sqrt{N \mathrm{Tr}(\Sigma^2_{k^*+1:p})}}{\lambda + \mathrm{Tr}(\Sigma_{k^*+1:p})} \right\}.$$

holds w.h.p. when the noise ξ is mean zero, independent of X and there is some $r > 4$ such that $\|\xi\|_{L_r} \leq C \|\xi\|_{L_2}$; $\phi(X)$ is such that $\exists \epsilon > 0$ s.t. for all $f \in \mathcal{H}$, $\|f(X)\|_{L_{4+\epsilon}} \leq C \|f(X)\|_{L_2}$; some conditions on the concentration of $\|\phi(X)\|_{\mathcal{H}}$...

Rem.: Matching lower bound in the Gaussian case; result true for all $\lambda \geq 0$; applications to the Gaussian equivalence conjecture and the multiple descents phenomena.

Decomposition of the KRR $\hat{f}_\lambda$

We have $\hat{f} = \hat{f}_{1:k} + \hat{f}_{k+1:\infty}$ where

$$\hat{f}_{1:k} \in \operatorname*{argmin}_f \left(\| Q\,(y - \mathbb{X}_{\phi,1:k} f) \|_{\mathcal{H}}^2 + \| f \|_{\mathcal{H}}^2 \right),$$

and $Q : \mathbb{R}^N \to \mathcal{H}_{k+1:\infty}$ is such that

$$Q^\top Q = \left(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N \right)^{-1} \underset{\textbf{D.M.}}{\sim} \left(\operatorname{Tr}(\Sigma_{k+1:\infty}) + \lambda \right)^{-1} I_N$$

hence

$$\hat{f}_{1:k} \approx \operatorname*{argmin}_f \left(\| y - \mathbb{X}_{\phi,1:k} f \|_2^2 + \left(\operatorname{Tr}(\Sigma_{k+1:\infty}) + \lambda \right) \| f \|_{\mathcal{H}}^2 \right),$$

hence $\hat{f}_{1:k}$ is a **"ridge" with tuning parameter** $\lambda + \operatorname{Tr}(\Sigma_{k+1:\infty})$ and

$$\begin{aligned} \hat{f}_{k+1:\infty} &= \mathbb{X}_{\phi,k+1:\infty}^\top \left(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N \right)^{-1} \left(y - \mathbb{X}_{\phi,1:k} \hat{f}_{1:k} \right) \\ &= \textbf{ridge} \text{ with parameter } \lambda \text{ for the residual } y - \mathbb{X}_{\phi,1:k} \hat{f}_{1:k} \end{aligned}$$

Application 3:
min-ℓ_q-norm interpolant estimators

Wang, Donhauser, Yang. *Tight bounds for minimum ℓ_1-norm interpolation of noisy data*

Koehler, Zhou, Sutherland, Srebro. *Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds, and Benign Overfitting*

FSD and decomposition of min ℓ_q interpolant estimators

N iid data (X_i, Y_i) in the linear model $Y = \langle X, \beta^* \rangle + \xi$. Let $q \geq 1$ and

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left(\|\beta\|_q : \mathbb{X}\beta = y \right).$$

We have $\hat{\beta} = \hat{\beta}_J + \hat{\beta}_{J^c}$ where

$$\hat{\beta}_J \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left(\|\mathcal{A}[y - \mathbb{X}_J \beta]\|_q^q + \|\beta\|_q^q \right)$$

where $\mathcal{A}[\mu] \in \operatorname*{argmin}_{\nu} \left(\|\nu\|_q : \mathbb{X}_{J^c} \nu = \mu \right)$ and

$$\hat{\beta}_{J^c} = \mathcal{A}[y - \mathbb{X}_J \hat{\beta}_J]$$

(ex. $q = 1$: BPDP); that is for a FSD

$$\mathbb{R}^p = \mathbb{R}^J \otimes^{\perp} \mathbb{R}^{J^c}$$

and $\mathbb{R}^J = \operatorname{span}(e_j : j \in J)$ and $\mathbb{R}^{J^c} = \operatorname{span}(e_j : j \in J^c)$ is adapted to the **canonical basis** $(e_j)_j$.

FSD and Dvorestky-Milman

D.M. for $\mathbb{X}_{J^c}^\top \Longrightarrow \mathcal{A}$ is isomorphic to the ℓ_2-norm:

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \ \|\boldsymbol{\lambda}\|_2 \, \ell_*(\Sigma_{J^c}^{1/2} B_q^p) \lesssim \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \lesssim \|\boldsymbol{\lambda}\|_2 \, \ell_*(\Sigma_{J^c}^{1/2} B_q^p)$$

$$\Longrightarrow \forall \boldsymbol{\mu} \in \mathbb{R}^N : \ \frac{\|\boldsymbol{\mu}\|_2}{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} \lesssim \|\mathcal{A}[\boldsymbol{\mu}]\|_q \lesssim \frac{\|\boldsymbol{\mu}\|_2}{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}$$

On the **estimation part** of the feature space $\hat{\boldsymbol{\beta}}$ behaves like

$$\hat{\boldsymbol{\beta}}_J \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\|\mathcal{A}[y - \mathbb{X}_J \beta]\|_q^q + \|\boldsymbol{\beta}\|_q^q \right)$$

$$\sim \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\|y - \mathbb{X}_J \beta\|_2^q + \ell_*(\Sigma_{J^c}^{1/2} B_q^p)^q \, \|\boldsymbol{\beta}\|_q^q \right)$$

ie a regularized ERM wrt the square loss function to the power q and ℓ_q^q regularization

- $q = 1$: square root LASSO of [Belloni, Chernozhukov and Wang]
- $q = 2$: ridge

Main result for the min ℓ_q, $1 < q < 2$, IE

Assume that $X = \Sigma^{1/2} Z$ where Σ is a diagonal matrix and Z has independent coordinates such that

- Z_{J^c} is $\mathcal{N}(0, I_{J^c})$
- Z_J is sub-gaussian;

if

$$|J| \lesssim N \lesssim \epsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)$$

then w.h.p. $\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2 \lesssim r(V_J, V_{J^c})$ and

$$\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c} - \boldsymbol{\beta}_{J^c}^*)\|_2 \lesssim \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2$$

$$+ (r(V_J, V_{J^c}) + \sigma_\xi)\left(\frac{\sqrt{N}\ell_*^{\frac{1}{q-1}}(\Sigma_{J^c}^{\frac{q}{2}} B_{\frac{2}{3-q}}^{J^c})}{\ell_*^{\frac{q}{q-1}}(\Sigma_{J^c}^{1/2} B_q^{J^c})} + \frac{N^{\frac{q}{2(q-1)}}\left(\operatorname{diam}(\Sigma_{J^c}^{\frac{q}{2}} B_{\frac{2}{3-q}}^{J^c})\right)^{\frac{1}{q-1}}}{\ell_*^{\frac{q}{q-1}}(\Sigma_{J^c}^{1/2} B_q^{J^c})}\right),$$

where, for some interpolation norm $\|\cdot\|$, $r(V_J, V_{J^c})$ equals

$$\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 + \sqrt{\epsilon_1}\sigma_\xi + \sqrt{\frac{|J|}{N}}\sigma_\xi + \sigma_\xi \frac{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)}{N^{q/2}}\left\|\boldsymbol{\beta}_J^* \odot |\boldsymbol{\beta}_J^*|^{\odot(q-2)}\right\|.$$

- Similar result for $q > 2$ under a weaker moment condition on Z_J.

Application 4 (work in progress): maximum margin interpolant estimator in classification

Cao, Gu, Belkin. *risk bounds for over-parametrized maximum margin classification*

Shamir. *The implicit bias of benign overfitting*

Stojanovic, Donhauser, Yang. *Tight bounds for maximum l1-margin classifiers*

The max-margin IE and its decomposition

The minimum ℓ_2-norm/max-margin interpolant (linear) classifier (aka hard margin support vectors machine) is

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\|\boldsymbol{\beta}\|_2 : \forall i \in [N],\ Y_i \langle X_i, \boldsymbol{\beta}\rangle \geq \mathbf{1}\right)$$
$$= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\|\boldsymbol{\beta}\|_2 : \mathbb{X}_y\boldsymbol{\beta} \succeq \mathbf{1}\right) = \hat{\boldsymbol{\beta}}_J + \hat{\boldsymbol{\beta}}_{J^c}$$

where $\mathbf{1} = (1, 1, \cdots, 1) \in \mathbb{R}^N$, $\mathbb{X}_y = [Y_1X_1|\cdots|Y_NX_N]^\top$,

$$\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{B}[\mathbf{1} - \mathbb{X}_y\hat{\boldsymbol{\beta}}_J].$$

where $\mathcal{B}[\boldsymbol{\mu}] \in \operatorname{argmin}_{\boldsymbol{\nu}} \left(\|\boldsymbol{\nu}\|_2 : \mathbb{X}_{y,J^c}\boldsymbol{\nu} \succeq \boldsymbol{\mu}\right)$,
$\mathbb{X}_{y,J^c} = [Y_1P_{J^c}X_1|\cdots|Y_NP_{J^c}X_N]^\top$ and

$$\hat{\boldsymbol{\beta}}_J \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\|\mathcal{B}[\mathbf{1} - \mathbb{X}_{y,J}\boldsymbol{\beta}]\|_2^2 + \|\boldsymbol{\beta}\|_2^2\right).$$

DM and the square hinge loss

D.M. for $\mathbb{X}_{y,J^c}^\top \Longrightarrow \mathcal{B}$ is isomorphic to the truncated ℓ_2-norm:

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2 \sqrt{\mathrm{Tr}(\Sigma_{J^c})} \lesssim \left\|\mathbb{X}_{y,J^c}^\top \boldsymbol{\lambda}\right\|_2 \lesssim \|\boldsymbol{\lambda}\|_2 \sqrt{\mathrm{Tr}(\Sigma_{J^c})}$$

$$\Rightarrow \forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|[\boldsymbol{\mu}]_+\|_2}{\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \lesssim \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \lesssim \frac{\|[\boldsymbol{\mu}]_+\|_2}{\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}$$

where $[\boldsymbol{\mu}]_+ = (\max(\mu_i, 0))_{i=1}^N$, and $\mathbb{X}_{y,J^c}^\top = [Y_1 P_{J^c} X_1 | \cdots | Y_N P_{J^c} X_N]$.

$$\hat{\boldsymbol{\beta}}_J \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\mathrm{argmin}} \left(\|\mathcal{B}[\mathbf{1} - \mathbb{X}_{y,J} \boldsymbol{\beta}_J]\|_2^2 + \|\boldsymbol{\beta}_J\|_2^2 \right)$$

$$\sim \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\mathrm{argmin}} \left(\|[\mathbf{1} - \mathbb{X}_y \boldsymbol{\beta}_J]_+\|_2^2 + \mathrm{Tr}(\Sigma_{J^c}) \|\boldsymbol{\beta}_J\|_2^2 \right)$$

is a regularized ERM for a the square hinge loss and ℓ_2^2 regularization and $\hat{\boldsymbol{\beta}}_{J^c}$ uses the free part of the FS to interpolate the data.

Excess $0-1$ risk decomposition

Setup: $(X, Y) \in \mathbb{R}^p \times \{-1, 1\}$,

$$f^*(x) = 2I(\eta(x) > 1/2) - 1 \text{ where } \eta(x) = \mathbb{P}[Y = 1|X = x].$$

$$\mathbb{P}\left(Y\hat{f}(X) < 0\right) - \mathbb{P}\left(Yf^*(X) < 0\right) \quad \text{0-1 excess risk}$$

$$= \mathbb{P}\left(Y\hat{f}(X) < 0\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0\right) \text{ error on the 'free part of the FS'}$$

$$+ \mathbb{P}\left(Y\hat{f}_J(X) < 0\right) - \mathbb{P}\left(Yf_J^*(X) < 0\right) \text{ error on the estimation part}$$

$$+ \mathbb{P}\left(Yf_J^*(X) < 0\right) - \mathbb{P}\left(Yf^*(X) < 0\right) \text{ approximation error}$$

where $\hat{f} = \hat{f}_J + \hat{f}_{J^c}$ for a FSD $\mathbb{R}^p = V_J \oplus^{\perp} V_{J^c}$.

Preliminary results in the Gaussian logistic model

$X \sim \mathcal{N}(0, \Sigma)$ and, for $\boldsymbol{\mu} \in \mathbb{R}^p$,

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + \exp(-2\langle \Sigma^{-1}\boldsymbol{\mu}, x\rangle)}.$$

The Bayes rule is $f^*(\cdot) = \mathrm{sign}(\langle \boldsymbol{\beta}^*, \cdot \rangle)$ for $\boldsymbol{\beta}^* = \Sigma^{-1}\boldsymbol{\mu}$.
If,

$$\dim(V_J) \lesssim N \lesssim \bar{\delta}^2 \frac{\mathrm{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}.$$

W.c.p. $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{\{0,1\}} \lesssim \|\boldsymbol{\beta}^*\|_2 (r(V_J, V_{J^c}))^2 + \bar{\delta}\sqrt{tP\ell_{\boldsymbol{\beta}_J^*}}$, when $\|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2 > r(V_J, V_{J^c})$ for

$$r(V_J, V_{J^c}) = P\ell_{\boldsymbol{\beta}_J^*}\sqrt{\frac{\dim(V_J)}{N}} + \frac{\mathrm{Tr}(\Sigma_{J^c})}{N}\|\boldsymbol{\beta}_J^*\| + \delta_3 t\sqrt{P\ell_{\boldsymbol{\beta}_J^*}}.$$

Rem.: best FSD for $V_J = \mathrm{span}(\boldsymbol{\beta}^*)$ (?)

Feature space decomposition: take away message

The FSD method can be used to

- decrease the cost of uniform convergence: we only estimate over the space $\mathbb{R}^J$

$$\mathbb{R}^p \longrightarrow \mathbb{R}^J$$

- understand new phenomenum like BO or revist old estimators thanks to the 'free part' of the feature space that allows estimators to do something else than estimation!

$$\text{freedom on } \mathbb{R}^{J^c}.$$

Thanks!

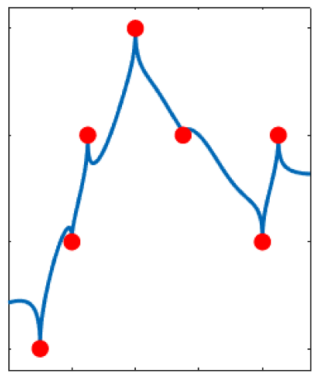
- G. Lecué and Z. Shang. *A geometrical viewpoint on the benign overfitting property of the minimum l2-norm interpolant estimator.* PTRF24
- G. Gavrilopoulos, G. Lecué and Z. Shang. *A Geometrical Analysis of Kernel Ridge Regression and its Applications.* AOS25
- G. Lecué and Z. Shang. *A Geometric Viewpoint on the Benign Overfitting Property of the Minimum ℓ_q-norm Interpolant Estimator in Regression and Classification Problems*. In preparation.

benign overfitting: a high dimensional phenomenon?

Not all consistent interpolant estimators need high dimensional spaces.

There are so far two types of interpolant estimators that generalize well:

- either you choose a minmax optimal estimator and you make small perturbation of it around the data points so that it interpolates well



Belkin, Rakhlin and Tsybakov. *Does data interpolation contradict statistical optimality?*

- or you look for a smooth estimator (even a linear one) and try to make it going through the data: in that case,

being smooth and interpolant requires space: $p >> N$.

What would do a statistician

↝ A statistician would not consider interpolant estimators (she/he would do some threshold in the direction of small singular values to avoid useless variance terms in directions where the signal $\boldsymbol{\beta}^*_{k^*+1:p}$ is not estimated)

↝ We study interpolant estimators because they appear naturally in deep learning and they perform very well for some tasks.

↝ In the end, we have enough space to interpolate and the generalization error $\left\|\Sigma^{1/2}\cdot\right\|_2$ does not put too much weights on this space so it does not harm to use that space to interpolate the data and that is what the min ℓ_2-norm interpolant estimator does.

mini-batch SGD interpolator tends to the min-ℓ_2 IE

Let a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and the mini-batch SGD algorithm

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \left(\frac{1}{|B_t|} \sum_{i \in B_t} \partial_1 \ell(\langle \theta^{(t)}, X_i \rangle, Y_i) X_i \right)$$

with $\theta^{(0)} = 0$.

★ For the square loss function: if at step T, $\hat{\theta}^{(T)}$ is interpolant (i.e. $\langle \hat{\theta}^{(T)}, X_i \rangle = Y_i, \forall i$) then for all $t \geq T, \hat{\theta}^{(t)} = \hat{\theta}^{(T)}$ and

$$\hat{\theta}^{(T)} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left(\|\theta\|_2 : \langle \theta, X_i \rangle = Y_i, \forall i \right) = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top)^{-1} y$$

(also true if we only assume that $\partial_1 \ell(u, u) = 0$).

★ For any loss function: if $\mathbb{X}\theta^{(T)} = y$ then it is equal to the min-ℓ_2-norm estimator.

The bias/variance approach in [TB] and [BLLT]

The minimum ℓ_2-norm IE has a closed form:

$$\hat{\boldsymbol{\beta}} = \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}y \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\boldsymbol{\beta}\|_2 : \mathbb{X}\boldsymbol{\beta} = y\right)$$

and so, for $y = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$, we have

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2 \lesssim \underbrace{\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)\boldsymbol{\beta}^*\right\|_2^2}_{\textit{bias}} + \underbrace{\left\|\Sigma^{1/2}\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\boldsymbol{\xi}\right\|_2^2}_{\textit{variance}}$$

Application: multiple descents

For $K(x,y) = h(\langle x,y\rangle/d)$ where h is a polynomial function.

Corollary (Gavrilopoulos L. and Shang)

Let $(f_d^*)_d$, $(\mu_d)_d$, $(\mathcal{H}_d)_d$ be a sequence of target functions, sub-Gaussian probability measures and RKHSs. As $N, d \to \infty$ with $\omega(d^\iota) \leq N \leq o(d^{\iota+1})$,

$$\left| \|\hat{f}_0 - f_d^*\|_{L_2(\mu_d)} - \|\Gamma_{>\iota}^{1/2} f_{>\iota}^*\|_{\mathcal{H}_d} \right| = o_{d,\mathbb{P}}(1)(\sigma_\xi + \|f_{>\iota}^*\|_{\mathcal{H}_d}).$$

- Improve the intervale from $\omega_d\left(d^\iota \log d\right) \leq N \leq O_d\left(d^{\iota+1-\delta_0}\right)$ to $\omega(d^\iota) \leq N \leq o(d^{\iota+1})$.
- The first sub-Gaussian result in multiple descent.
- We do not require the assumptions $\mathbb{E}h_d(X) = \mathbb{E}f_d^*(X) = 0$ as in [Ghorbani et al., 2021]
- Unification between [Liang et al., 2020] and [Ghorbani et al., 2021].

Application: multiple descents

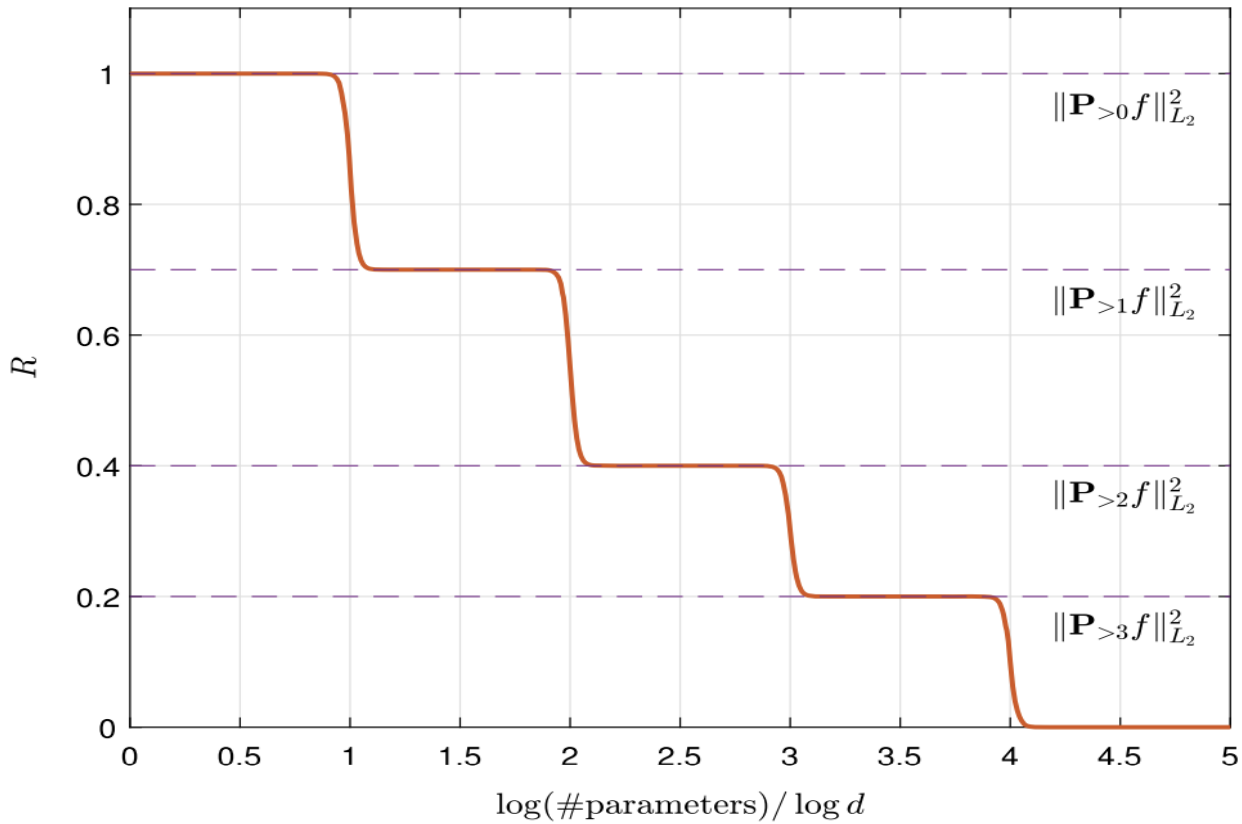


Figure: Taken from [Ghorbani et al., 2021]. Multiple descents caused by $\|\Gamma_{>\iota}^{1/2} f_{>\iota}^*\|_{\mathcal{H}_d}^2$. The effect of variance vanishes at limit.

Application: multiple descents

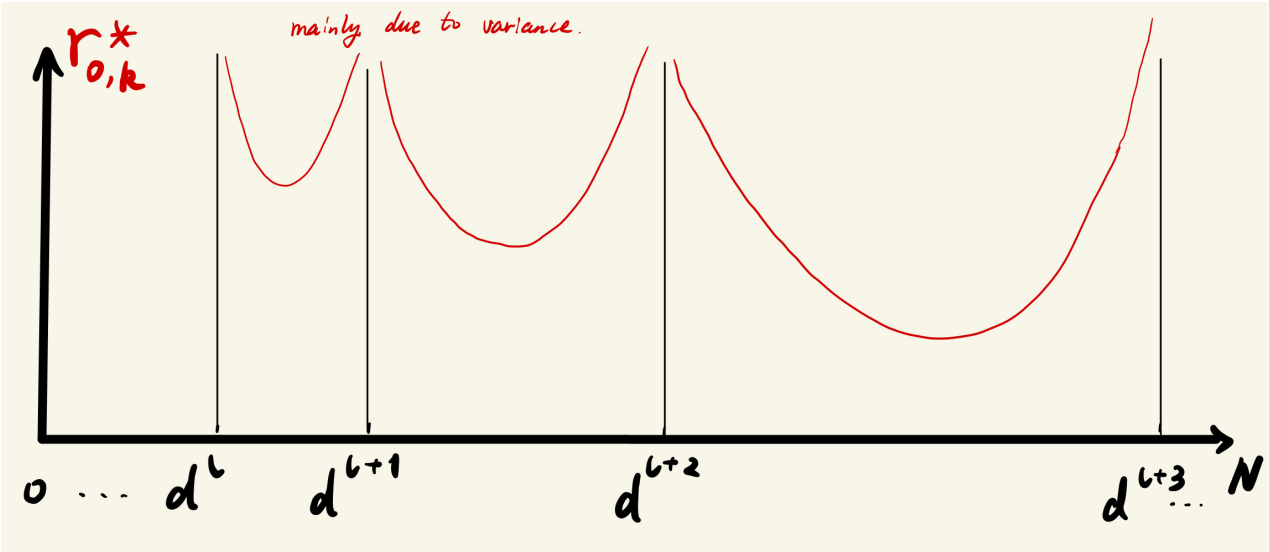


Figure: Non-asymptotic viewpoint. The descent and ascent are because of the variance $\sigma_\xi \sqrt{\frac{d^\iota}{N}}$ and $\sigma_\xi \sqrt{\frac{N}{d^{\iota+1}}}$, while the bottom of each valley is above $\|\Gamma_{>\iota}^{1/2} f^*_{>\iota}\|^2_{\mathcal{H}_d}$.