Implicit regularisation, Benign interpolation, and Over-parametrisation

Main references :

- M. Belkin, D. Hsu, S. Ma, S. Mandal. "Reconciling modern machine learning and the bias-variance trade-off. (2018)
- P. Bartlett, P. Long, G. Lugosi, A. Tsigler. "Benign overfitting in Linear Regression". (2019)
- T. Hastie, A. Montanari, S. Rosset, R. Tibshirani. "Surprises in high-dimensional ridgeless least-square interpolation". (2019)
- L. Chizat, F. Bach. "Implicit bias of gradient descent for wide 2-layer neural networks trained with the logistic loss" (2020)

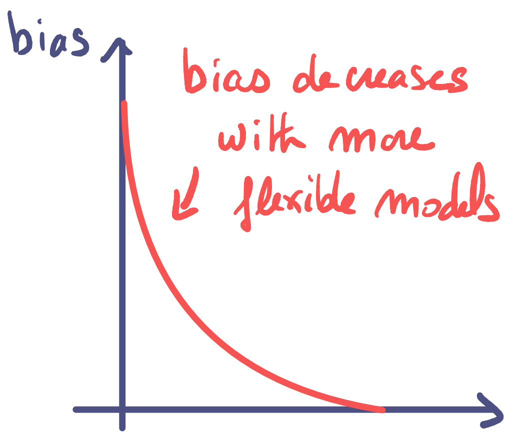
Lecture notes :

Lecture notes are available on the web page provided in the description of the video.

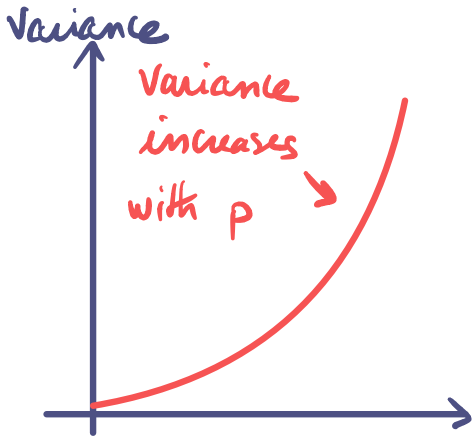
Double descent phenomenon

Textbook theory : risk = bias + variance

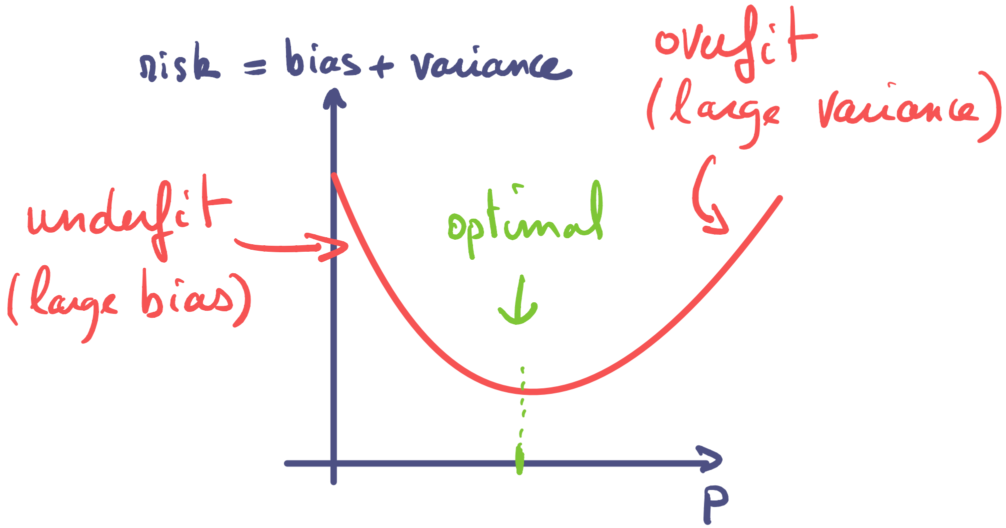
Let p = number of parameters in the model.



Decreasing bias

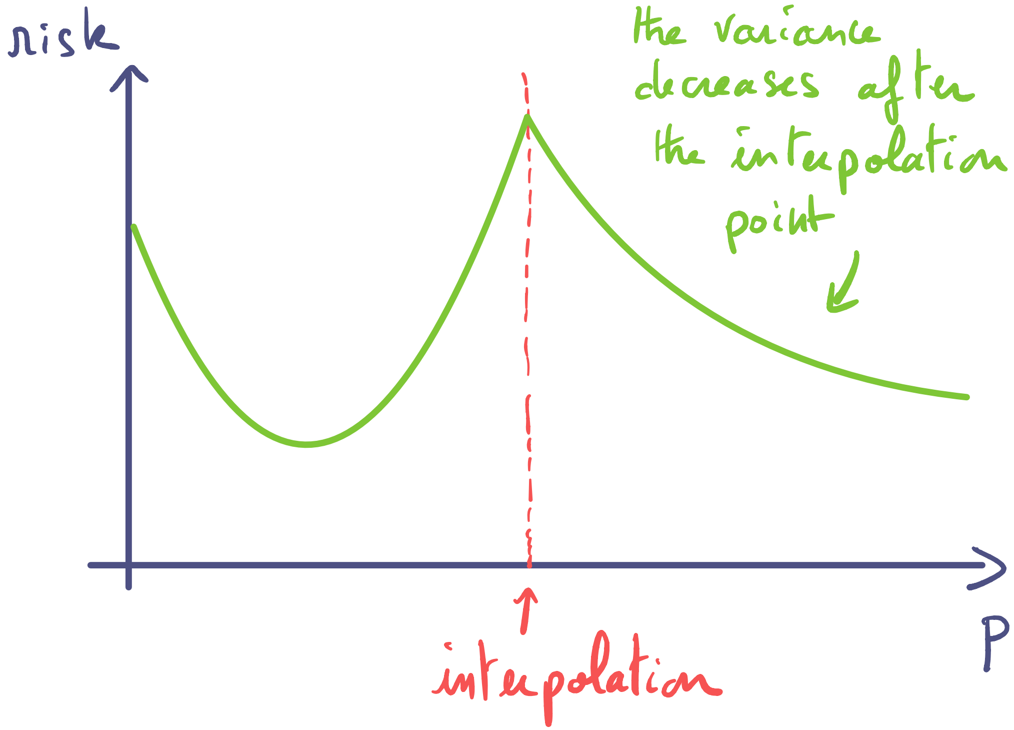


Increasing variance



Bias-variance trade-off

Empirical observations (neural networks)



Double descent phenomenon

Four parts

1. Implicit regularisation (a.k.a. implicit bias) of gradient descent
2. Benign interpolation: A) intuitions
3. Benign interpolation: B) mathematical analysis
4. Benign overparametrisation

1- Implicit regularization of Gradient Descent

a) Implicit regularisation for L.S.

- **Linear model**: $y_i = \langle x_i, \underbrace{\beta^*}_{\in \mathbb{R}^p} \rangle + \varepsilon_i$, $i=1,\dots,n$

with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$.

- **Notation**: in vectorial notation

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{=:Y} = \underbrace{\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}}_{=:X} \beta^* + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{=:\varepsilon}$$

- **Least-Square**: MLE estimation amounts to minimise the squares

$$\hat{\beta}^{MLE} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2 \quad \text{(LS)}$$

- **High-dimensional regime**:

Assume that $x_i \in \mathbb{R}^p$ with $p > n$.

then, $\dim(\ker(X)) = p - \operatorname{rank}(X) \geq p - n > 0$

$\Longrightarrow$ no unique solution to (LS)

- **Interpolation regime**: $p \geq n$

when in addition $\operatorname{rank}(X) = n$, then any solution $\hat{\beta}$ of (LS) fulfills

$Y = X\hat{\beta}$ i.e $y_i = x_i^T \hat{\beta}$ for $i=1,\dots,n$.

$\Longrightarrow$ $\hat{\beta}$ interpolates the learning data $(x_i, y_i)_{i=1,\dots,n}$.

$\Longrightarrow$ overfitting ?!?

- **Classical practice**: add a regularisation term in (LS) like $\lambda \|\beta\|^2$ in order to get a strictly convex objective function

- **Recent practice in N.N.**: apply GD on (LS) and use the solution selected by GD. What is this solution?

- G.D. on least-squares:

We apply G.D. on $\beta \to \|Y - X\beta\|^2$, started from $\hat\beta^0 = 0$:

$$\hat\beta^{t+1} \overset{GD}{=} \hat\beta^t - 2\eta X^T(X\hat\beta^t - Y)$$

$$\overset{\hat\beta^0=0}{=} \sum_{k=0}^{t} (I - 2\eta X^T X)^k \cdot 2\eta X^T Y$$

$$\xrightarrow[2\eta < |X|_{op}^{-2}]{t\to+\infty} (I - (I - 2\eta X^T X))^+ \cdot 2\eta X^T Y$$

$$= X^+ Y$$

Moore Penrose pseudo inverse (Appendix C)

So $\hat\beta^{LS} = X^+ Y$.

- Is there something special with this solution?

Lemma: In the interpolation regime, where $\mathrm{rank}(X) = n$, we have

$$\hat\beta^{LS} = \underset{Y = X\beta}{\mathrm{argmin}} \|\beta\|^2$$

regularisation of G.D.

Proof: - we first observe that

$X X^+ Y = Y$ so $Y = X\hat\beta^{LS}$

and any solution of $Y = X\beta$ can be decomposed as $\beta = \beta_0 + X^+ Y$, with $\beta_0 \in \ker(X)$

- Then, we notice that $\mathrm{range}(X^+) = \mathrm{range}(X^T)$ so we have

$$\mathbb{R}^p = \ker(X) \oplus \mathrm{range}(X^T) = \ker(X) \oplus \mathrm{range}(X^+)$$

and in particular $\|\beta\|^2 = \|\beta_0\|^2 + \|X^+ Y\|^2$.

So $X^+ Y = \underset{Y = X\beta}{\mathrm{argmin}} \|\beta\|^2$ $\square$

b) Implicit regularization for logistic

- When the labels $y_i \in \{-1,+1\}$, we can predict y by $\mathrm{sign}(\langle \hat\beta, x\rangle)$, where $\hat\beta$ is learnt by applying GD on the empirical logistic risk

$$\mathcal{L}(\beta) := \frac{1}{m}\sum_{i=1}^{m} \ell(-y_i \langle \beta, x_i\rangle)$$

with $\ell(-z) = \log(1+e^{-z})$

- Interpolation regime: when the learning data $(x_i, y_i)_{i=1,\dots,m}$ can be separated by an hyperplan.

- GD in interpolation regime:

Assume that at step t of GD, $\hat\beta^t$ perfectly classifies the training data

$$y_i \langle \hat\beta^t, x_i\rangle > 0, \quad \text{for } i=1,\dots,m.$$

- Since

$$\mathcal{L}(\hat\beta^t) = \frac{1}{m}\sum_{i=1}^{m} \ell\left(-\|\hat\beta^t\| \cdot y_i \left\langle \frac{\hat\beta^t}{\|\hat\beta^t\|}, x_i\right\rangle\right)$$

with $z \to \ell(-z)$ decreasing,

the loss $\mathcal{L}$ can be further decreased by sending $\|\hat\beta^t\| \to +\infty$.

- Since $\ell(-z) \overset{z\to+\infty}{\sim} e^{-z}$, we get

$$\mathcal{L}(\hat\beta^t) \sim \frac{1}{m}\sum_{i=1}^{m} \exp\left(-\|\hat\beta^t\|\, y_i \left\langle \frac{\hat\beta^t}{\|\hat\beta^t\|}, x_i\right\rangle\right)$$

when $\|\hat\beta^t\| \gg 1$

$$\approx \frac{N_{\min}}{m} \exp\left(-\|\hat\beta^t\| \cdot \min_{i=1..m} y_i \left\langle \frac{\hat\beta^t}{\|\hat\beta^t\|}, x_i\right\rangle\right)$$

which suggests that $u^t = \hat\beta^t / \|\hat\beta^t\|$ tends to solve the max-margin problem

$$(\mathrm{MM}) \quad \max_{\|u\|=1} \underbrace{\min_{i=1,\dots,m} y_i \langle u, x_i\rangle}_{\text{margin}},$$

which eventually occurs (Soudry et. al. 2017)

Theorem (informal)

The normalized solution $\hat{\beta}^t / \|\hat{\beta}^t\|$ of GD on the empirical logistic risk converges to the max-margin classifier (MM) when the data is linearly separable.

Take home message:

when interpolation is possible, GD selects some specific interpolating solutions

↗ minimal norm interpolating solutions in L^2 regression

↘ max-margin solution in logistic regression

⇝ implicit regularisation of GD.

These results can be generalized to more complex models like Neural Networks.

Example (Informal):

when training an homogeneous NN with GD and logistic loss:

if the normalized weights converge, then, at the limit, they solve the max-margin problem.

Next videos:

⇝ on the statistical benefit of this implicit regularisation for L.S.

2. Benign interpolation with high-dimensional input.

4/ Intuitions

Reminder:

- We consider the linear model:

$$y_i = \langle x_i, \underbrace{\beta^*}_{\in \mathbb{R}^p} \rangle + \varepsilon_i, \quad i=1,\dots,n$$

with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$. We set $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$.

- We compute $\hat\beta$ by minimizing the Least Square $\beta \to \|Y - X\beta\|^2$ by GD:

$\rightsquigarrow \hat\beta = X^+ Y$.

- Prediction: for $x \in \mathbb{R}^p$

$$\langle \hat\beta, x\rangle = \langle X^+ Y, x\rangle$$

$$\overset{Y = X\beta^* + \varepsilon}{=} \langle \underbrace{X^+X}_{= P_{X^T} \text{ (projection on } \mathrm{range}(X^T))} \beta^*, x\rangle + \langle X^+\varepsilon, x\rangle$$

$$= \underbrace{\langle \beta^*, x\rangle}_{\text{target}} - \underbrace{\langle (I - P_{X^T})\beta^*, x\rangle}_{\text{bias}} + \underbrace{\langle \varepsilon, (X^T)^+ x\rangle}_{\mathcal{N}(0, x^T(X^TX)^+ x) \text{ cond. on } X}$$

Remarks: $\underset{\mathrm{range}(X^T)}{\overset{\mathrm{range}(X^+)}{\|}}$

1) if $x \perp \mathrm{range}(X^T)$: $\langle \hat\beta, x\rangle = 0$

$\Rightarrow$ large bias, but no variance

2) $\mathrm{rank}(X) = n \ll p$:

$\|P_{X^T} x\| \ll \|x\|$ w.h.p. if $x \sim \mathcal{N}(0,\Sigma)$ with $\Sigma \simeq I_p$.

So only a small part of x is used for the prediction $\langle \hat\beta, x\rangle = \langle \hat\beta, P_{X^T} x\rangle$

Below we discuss two cases

a) isotropic case: $\Sigma = I_p$

b) spike model $\Sigma = I_k + \underbrace{\rho}_{\rho \ll 1} I_{p-k}$

Average prediction error :

- We have $\hat{\beta} = X^+ Y = \underbrace{X^+ X}_{P_{X^T}} \beta^* + X^+ \varepsilon$ with X^+X = orthogonal projection on $\mathrm{range}(X^T)$

- Let us compute the average prediction error, conditionally on the design X

$$R_X = \mathbb{E}\left[\langle x, \hat{\beta} - \beta^* \rangle^2 \mid X \right] \quad \text{where } x \sim \mathcal{N}(0, \Sigma) \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

$$R_X = \mathbb{E}_{x,\varepsilon}\left[\langle x, \hat{\beta} - \beta^* \rangle^2 \right] \overset{\hat{\beta} = P_{X^T}\beta^* + X^+\varepsilon}{=} \mathbb{E}_{x,\varepsilon}\left[\left(\langle (\underbrace{I - P_{X^T}}_{P_{\ker(X)}}) \beta^*, x \rangle - \langle X^+ \varepsilon, x \rangle \right)^2 \right]$$

$P_{\ker(X)}$ (projection on $\ker(X)$)

$$= \mathbb{E}_x\left[\beta^{*T} P_{\ker(X)}\, x x^T P_{\ker(X)} \beta^* \right] + \sigma^2\, \mathbb{E}_x\left[x^T \underbrace{X^+ (X^+)^T x}_{} \right] + \overset{\text{cross-term}}{0}$$

$= \langle x x^T, X^+ (X^+)^T \rangle_F$

$$= \underbrace{\beta^{*T} P_{\ker(X)}\, \Sigma\, P_{\ker(X)} \beta^*}_{B_X} + \underbrace{\sigma^2 \langle \Sigma, (X^T X)^+ \rangle_F}_{V_X}.$$

bias term $\qquad$ variance term

a) isotropic input : $\Sigma = I_p$

. Linear model : $y_i = \langle x_i, \beta^* \rangle + \varepsilon_i \, , \; i=1,\dots,n$

with $x_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$ and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

. singular values of random matrices : with $p > n$

Let $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$

$$= \sum_{k=1}^{n} \hat{\sigma}_k \hat{u}_k \hat{v}_k^T$$

we have :

$$\mathbb{E}[\|X\|_F^2] = \begin{cases} \sum_{i,j} \mathbb{E}[X_{ij}^2] = np \\ \sum_{k=1}^{n} \mathbb{E}[\hat{\sigma}_k^2] \leq n \, \mathbb{E}[\hat{\sigma}_1^2] \end{cases}$$

$\Rightarrow \mathbb{E}[\hat{\sigma}_1^2] \geq p$.

This suggests that $\hat{\sigma}_k$ scales like $\sqrt{p}$ when $p \gg n$.

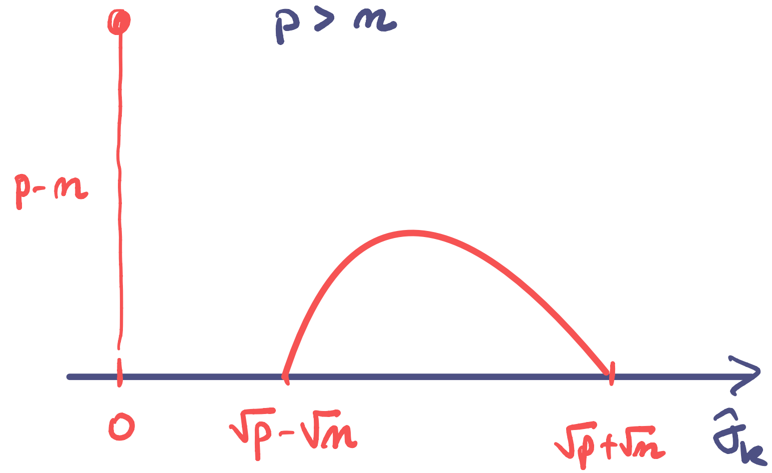
It eventually occurs :

Lemma 8.3 (Davidson & Szarek)

For $p > n$, we have for $k = 1, .., n$

$$\sqrt{p} - \sqrt{n} \leq \mathbb{E}[\hat{\sigma}_k] \leq \sqrt{p} + \sqrt{n}$$

Histogram of the $\hat{\sigma}_k$ for $p > n$



Furthermore, since $X \to \sigma_k(X)$ is 1-Lipschitz, by Gaussian concentration inequality, with high-probability

$$\hat{\sigma}_k \sim \sqrt{p} \quad \text{for } p \gg n \quad \forall k = 1 \ldots n$$

$$\Rightarrow X^T X \overset{p \gg n}{\simeq} p \cdot P_{X^T}$$

and $(X^T X)^+ \simeq \frac{1}{p} P_{X^T}$

[See Proposition 12.9 in the Lecture Notes]

Average prediction error :

$$\cdot\ R_X = \beta^{*T} P_{\ker(X)} \Sigma P_{\ker(X)} \beta^* + \sigma^2 \langle \Sigma, (X^TX)^+ \rangle_F$$

$$\underset{\Sigma = I_p}{=} \underbrace{\| P_{\ker(X)} \beta^* \|^2}_{= B_X} + \underbrace{\sigma^2 \, \mathrm{Tr}((X^TX)^+)}_{= V_X}$$

Variance :

$$\cdot\ V_X = \sigma^2 \, \mathrm{Tr}(X^TX)^+ = \sigma^2 \sum_{k=1}^{n} \frac{1}{\tilde{\sigma}_k^2}$$

$\tilde{\sigma}_k^2 \simeq p$ w.h. $\mathbb{P}_X$ when $p \gg n$

$$\approx \boxed{\sigma^2 \frac{n}{p}} \longrightarrow 0 \quad \text{when } p \gg n \quad !!!$$

$\rightsquigarrow$ In the interpolation regime ($p \geq n$), high-dimensional inputs kill the variance of G.D.-Least square

[See next video for the proof of this result]

<u>Discussion</u>: we observe that for a given $x \in \mathbb{R}^p$, with $\|x\|^2 = p$:

$\rightarrow$ $\displaystyle \mathrm{var}_\varepsilon \langle \hat\beta, x\rangle = \sigma^2 x^T (X^T X)^+ x = \sigma^2 x^T \Big(\sum_{k=1}^{n} \frac{1}{\hat\sigma_k^2} \hat v_k \hat v_k^T \Big) x = \frac{\sigma^2}{p} \underbrace{x^T P_{X^T} x}_{= \|P_{X^T} x\|^2}$

$\rightarrow$ if $x \in \mathrm{range}(X^T)$: $\displaystyle \mathrm{var}_\varepsilon \langle \hat\beta, x\rangle \overset{p \gg n}{\cong} \sigma^2 \frac{\|x\|^2}{p} \overset{\text{for } \|x\|^2 = p}{\approx} \sigma^2 \leftarrow$ strong overfit on range(X^T)

$\rightarrow$ but $\|P_{X^T} x\|^2 \ll \|x\|^2$ w.h. $\mathbb{P}_x$: due to the high-dimension of the input space, w.h. $\mathbb{P}_x$, a new x is almost not correlated with the learning points $x_1, \dots, x_n$ (columns of X^T), so

$$\mathrm{var}_\varepsilon \langle \hat\beta, x\rangle \cong \sigma^2 \frac{\|P_{X^T} x\|^2}{p} \quad \text{small}$$

$\rightarrow$ <u>in words</u>: we overfit on a small space $= \mathrm{span}\{x_1, \dots, x_n\}$ spanned by the learning inputs, but we underfit everywhere else

$\Rightarrow$ large bias and small variance for $p \gg n$!

Bias:

$$\mathbb{R}^p = \ker(X) \oplus \mathrm{range}(X^T)$$

- $$B_X = \| P_{\ker(X)} \beta^* \|^2 = \|\beta^*\|^2 - \| P_{X^T} \beta^* \|^2 \underset{\text{w.h. } \mathbb{P}_X}{\approx} \|\beta^*\|^2 \left(1 - \frac{n}{p}\right)$$

The larger p, the larger the bias B_X !!

$\mathrm{Range}(X^T) \overset{\text{unif.}}{\sim} \dim(n) \subset \mathbb{R}^p$

so $\| P_{X^T} \beta \|^2 \simeq \frac{n}{p} \|\beta\|^2$

Prediction risk:

- $$R_X = B_X + V_X \overset{p \gg n}{\approx} \underbrace{\|\beta^*\|^2}_{\text{large (mainly induced by the bias)}} + \underbrace{\frac{n}{p}\left(\sigma^2 - \|\beta^*\|^2\right)}_{\text{small since } n \ll p}$$

b) Anisotropic input (intuitions)

Prototypical example: $\Sigma = I_k + \underbrace{\rho}_{\ll 1} I_{>k} = \begin{array}{c} \overset{k}{\longleftrightarrow} \quad \overset{p-k}{\longleftrightarrow} \\ \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & \rho & \\ & & & & \ddots \\ & & & & & \rho \end{bmatrix} \end{array}$ with
- $k \ll m \ll p$
- $\rho \ll 1$

Then $\frac{1}{m} X^T X = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T \overset{k \ll m}{\simeq} \underbrace{I_k}_{\substack{\text{low dimensional} \\ \Rightarrow \text{ good estimation}}} +$ "something of rank $m-k$"

Trace $(\Sigma) = k + \rho(p-k)$

$$\simeq I_k + \rho \frac{p-k}{m-k} \underbrace{\hat{I}_{m-k}}_{\simeq \text{ projection of rank } m-k}$$

- So $X^T X \overset{k \ll m \ll p}{\simeq} m I_k + \rho p \hat{I}_{m-k}$ and $(X^T X)^+ \overset{k \ll m \ll p}{\simeq} \frac{1}{m} I_k + \frac{1}{\rho p} \hat{I}_{m-k}$. Hence:

$$V_X = \sigma^2 \langle (X^T X)^+, \Sigma \rangle_F \simeq \frac{\sigma^2}{m} \underbrace{\| I_k \|_F^2}_{= k} + \frac{\sigma^2}{p} \underbrace{\langle I_{>k}, \hat{I}_{m-k} \rangle_F}_{= \operatorname{Tr}(\hat{I}_{m-k}) = m-k \simeq m}$$

$$\simeq \sigma^2 \Big(\underbrace{\frac{k}{m}}_{\substack{\text{small due} \\ \text{to low rank } k}} + \underbrace{\frac{m}{p}}_{\substack{\text{small due} \\ \text{to high-dimensional input}}} \Big)$$

Next video:

$\rightsquigarrow$ we consider a general Σ

$\rightsquigarrow$ we prove (sharp) upper-bound on V_X

(recovering today's "results")

Benign interpolation with high-dimensional input.

B/ Mathematical analysis

Reminder:

- Learning data: $(x_i, y_i)_{i=1..n}$ with $x_i \sim \mathcal{N}(0,\Sigma)$ and $y_i = \langle x_i, \beta^* \rangle + \underbrace{\varepsilon_i}_{\overset{iid}{\sim}\mathcal{N}(0,\sigma^2)}$

- GD - least square: $\hat{\beta} = X^+ Y$

- Average prediction risk:

$$R_X := \mathbb{E}\left[\langle \hat{\beta} - \beta^*, x\rangle^2 \mid X\right]$$

$$= \underbrace{\beta^{*T} P_{\ker(X)} \Sigma P_{\ker(X)} \beta^*}_{B_X} + \underbrace{\sigma^2 \langle \Sigma, (X^T X)^+ \rangle_F}_{V_X}$$

bias term $\qquad$ variance term

$\rightsquigarrow$ to understand V_X, we need to understand $(X^T X)^+$.

a) <u>spectrum of XX^T</u> : $p > n$

- $\mathrm{spec}^*(XX^T) = \mathrm{spec}^*(X^TX)$
- $\mathrm{rank}(\underbrace{XX^T}_{n\times n}) = n$ for $p > n$

$\rightsquigarrow$ easier to analyse XX^T

. <u>Setting</u> : we assume (with no loss of generality) that $\Sigma = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$.

. We define :

$\rightarrow$ $k^* = \min \{ k : \lambda_{k+1} < \frac{\rho}{n} \sum_{j>k} \lambda_j \}$
with $\rho = \frac{1}{142}$

$\rightarrow$ $X = [\underbrace{X_\leq}_{k^*} \ \underbrace{X_>}_{p-k^*}]$ and $\Sigma = \begin{pmatrix} \Sigma_\leq & 0 \\ 0 & \Sigma_> \end{pmatrix}$

So that $XX^T = X_\leq X_\leq^T + X_> X_>^T$

Lemma (spec)

① if $k^* \leq \frac{n}{600}$, then with proba $\geq 1 - 2e^{-n/84}$

$$\left\| \Sigma_\leq^{-1/2} (X_\leq^T X_\leq) \Sigma_\leq^{-1/2} - n I_{k^*} \right\|_{op} \leq \frac{3n}{4}$$

② with proba $\geq 1 - 2e^{-n/84}$

$$\left| \mathrm{spec}^*(X_> X_>^T) - \sum_{j>k^*} \lambda_j \right| \leq \frac{3}{4} \sum_{j>k^*} \lambda_j$$

<u>Comments</u> : the above lemma shows that:

$\rightarrow$ the low dim. part $\frac{1}{n} X_\leq^T X_\leq$ remains "close" to $\Sigma_\leq$

$\rightarrow$ the low signal part $X_> X_>^T$ has a flat spectrum around $\sum_{j>k^*} \lambda_j$

To prove this lemma, we need a concentration bound on quadratic form of Gaussian vectors.

Theorem B.8. (Hanson-Wright)

Let $\varepsilon \sim \mathcal{N}(0, I_d)$ and $\alpha \in \mathbb{R}^d$.

There exists $\zeta \sim \mathcal{E}xp(1)$ such that

$$\sum_j \alpha_j \varepsilon_j^2 - \alpha^T \mathbb{1} \leq \sqrt{8\|\alpha\|^2 \zeta} \vee (8|\alpha|_\infty \zeta)$$

$$\leq \frac{|\alpha|_1}{4} + 8|\alpha|_\infty \zeta$$

Proof of H.W. (by Chernov method).

since $-\log(1-x) \leq x + x^2$ for $|x| \leq 1/2$,

we have for any $|\lambda| \leq 1/4$

$$\mathbb{E}\left[\exp(\lambda(\varepsilon_j^2 - 1))\right] = \frac{e^{-\lambda}}{(1-2\lambda)^{1/2}} \overset{(*)}{\leq} e^{2\lambda^2}.$$

So for any $|s| \leq 1/4|\alpha|_\infty$, we have

$$\mathbb{P}\left[\sum_j \alpha_j(\varepsilon_j^2 - 1) > t\right] \overset{\text{Chernov}}{\leq} e^{-st} \prod_j \underbrace{\mathbb{E}\left[e^{s\alpha_j(\varepsilon_j^2 - 1)}\right]}_{\overset{(*)}{\leq} \exp(2\alpha_j^2 s^2)}$$

$$\leq \exp(-st + 2\|\alpha\|^2 s^2)$$

Setting $s = \frac{t}{4\|\alpha\|^2} \wedge \frac{1}{4|\alpha|_\infty}$ we get

$$\mathbb{P}\left[\sum_j \alpha_j(\varepsilon_j^2 - 1) > t\right] \leq \exp\left(-\frac{1}{8}\left(\frac{t^2}{\|\alpha\|^2} \wedge \frac{t}{|\alpha|_\infty}\right)\right)$$

i.e. $\sum_j \alpha_j \varepsilon_j^2 - \alpha^T \mathbb{1} \leq \sqrt{8\|\alpha\|^2 \zeta} \vee (8|\alpha|_\infty \zeta)$.

Furthermore

$$\sqrt{8\|\alpha\|^2 \zeta} \leq 2\sqrt{\frac{|\alpha|_1}{4} \cdot |\alpha|_\infty 8\zeta} \overset{2ab \leq a^2 + b^2}{\leq} \frac{|\alpha|_1}{4} + 8|\alpha|_\infty \zeta$$

$\square$

To control the operator norm

$$|A|_{op} = \max_{\|u\|=1} \|Au\|$$

for A symmetric

$$= \max_{\|u\|=1} |\langle Au, u\rangle|$$

We use the following discretization lemma

Lemma 12.11-12: For any $\delta > 0$, there exists $\mathcal{N}_\delta \subset \{u \in \mathbb{R}^d : \|u\| = 1\}$ with cardinality $|\mathcal{N}_\delta| \leq \left(1 + \frac{2}{\delta}\right)^d$ such that, for any $A \in \mathbb{R}^{d \times d}$ symmetric

$$|A|_{op} \leq \frac{1}{1-2\delta} \max_{u \in \mathcal{N}_\delta} |\langle Au, u\rangle|$$

Proof: i) construction of $\mathcal{N}_\delta$: Let us set $\mathbb{S}_{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$, and pick iteratively

$u_1 \in \mathbb{S}_{d-1}$; $u_2 \in \mathbb{S}_{d-1} \setminus B(u_1, \delta)$; ...

$u_k \in \mathbb{S}_{d-1} \setminus \bigcup_{j \leq k-1} B(u_j, \delta)$; until impossible.

By construction, we have

(a) $\|u\| = 1$ and $\|u - u'\| > \delta$ $\forall u \neq u' \in \mathcal{N}_\delta$

(b) $\forall v \in \mathbb{S}_{d-1}$, $\exists u \in \mathcal{N}_\delta$ s.t. $\|u - v\| \leq \delta$

• From (a), we get that:

$$\bigsqcup_{u \in \mathcal{N}_\delta} B(u, \delta/2) \subset B(0, 1 + \delta/2)$$

so comparing the volumes

$$|\mathcal{N}_\delta| \cdot \left(\frac{\delta}{2}\right)^d V_d(1) \leq \left(1 + \frac{\delta}{2}\right)^d V_d(1)$$

$$\Rightarrow |\mathcal{N}_\delta| \leq \left(1 + \frac{2}{\delta}\right)^d$$

ii) approximation: from (b) we have

$$|A|_{op} = |\langle Au^*, u^*\rangle| \quad \text{with } u^* \in \mathbb{S}_{d-1}$$

with $u \in \mathcal{N}_\delta$, $\|u - u^*\| \leq \delta$

$$= |\langle Au, u\rangle + \langle A(u^* - u), u\rangle + \langle Au^*, u^* - u\rangle|$$

$$\leq |\langle Au, u\rangle| + |A|_{op}\delta + |A|_{op}\delta. \quad \square$$

We are ready to prove Lemma (Spec)

Proof of Lemma (Spec):

① Let us set $Z := X_\leq \Sigma_\leq^{-1/2} \in \mathbb{R}^{m \times k^*}$.

We have

. $Z_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$

. $Z^T Z = \Sigma_\leq^{-1/2} X_\leq^T X_\leq \Sigma_\leq^{-1/2}$

$$|\Sigma_{\leq}^{-1/2} X_{\leq}^T X_{\leq} \Sigma_{\leq}^{-1/2} - m I_{k^*}|_{op} = |Z^T Z - m I_{k^*}|_{op}$$

$$\leq 2 \max_{u \in \mathcal{N}_{1/4} \subset \mathbb{S}_{k^*-1}} |\langle (Z^T Z - m I_{k^*}) u, u \rangle|$$

$$= 2 \max_{u \in \mathcal{N}_{1/4}} | \|Zu\|^2 - m |$$

For any $u \in \mathbb{S}_{k^*-1}$ we have

$$[Zu]_j = \sum_k Z_{jk} u_k \overset{iid}{\sim} \mathcal{N}(0,1)$$

So $Zu \sim \mathcal{N}(0, I_m)$ and by Hanson-Wright

$\exists\, \xi_u, \xi'_u \sim \mathcal{E}xp(1)$ such that

$$|Z^T Z - m I_{k^*}|_{op} \leq 2 \max_{u \in \mathcal{N}_{1/4}} \left(\frac{m}{4} + 8 (\xi_u \vee \xi'_u) \right)$$

$\alpha_1 = \dots = \alpha_m = 1$

Furthermore,

union bound

$$\mathbb{P}\left[8 \max_{u \in \mathcal{N}_{1/4}} (\xi_u \vee \xi'_u) \geq \frac{m}{8} \right] \leq 2 \underbrace{|\mathcal{N}_{1/4}|}_{\leq 9^{k^*}} \underbrace{\mathbb{P}\left[\xi \geq \frac{m}{8.8}\right]}_{\leq e^{-m/64}}$$

for $k^* \leq \frac{m}{600}$ $\longrightarrow \leq 2 \exp(-\frac{m}{84})$

which proves (1)

(2) We set now $Z := X_> \Sigma_>^{-1/2} \in \mathbb{R}^{m \times (p-k^*)}$

With the same reasoning as before, we have

$$|X_> X_>^T - \mathrm{Tr}(\Sigma_>) I_m|_{op} = |Z \Sigma_> Z^T - \mathrm{Tr}(\Sigma_>) I_m|_{op}$$

$$\leq 2 \max_{u \in \mathcal{N}_{1/4} \subset \mathbb{S}_{m-1}} \left(\frac{\mathrm{Tr}(\Sigma_>)}{4} + 8 \lambda_{k^*+1} (\xi_u \vee \xi'_u) \right)$$

Furthermore,

$$\mathbb{P}\left[8 \lambda_{k^*+1} \max_{u \in \mathcal{N}_{1/4}} \xi_u \vee \xi'_u \geq \frac{\mathrm{Tr}(\Sigma_>)}{8} \right]$$

def. of k^*

$$\leq \mathbb{P}\left[\max_{u \in \mathcal{N}_{1/4}} \xi_u \vee \xi'_u \geq \frac{m}{64 \rho} \right]$$

union bound

$$\leq 2 \cdot 9^m \cdot \exp\left(-\frac{m}{64\rho}\right) \leq 2 \exp\left(-\frac{m}{84}\right)$$

$\frac{1}{\rho} = 142$

This proves (2).

$\square$

Remark: we have proved that

$$X^T X \simeq m \Sigma_{\leq} + \lambda^* \underbrace{\hat{I}_{m-k}}_{\text{random projector}}$$

where $\lambda^* = \sum_{j > k^*} \lambda_j$

b) Bounding the variance V_X:

We are ready to prove the main result.

Theorem (Bartlett et. al. 2019)

- $k^* = \min\{k : n\lambda_{k+1} \leq \rho \sum_{j>k} \lambda_j\}$ with $\rho = \frac{1}{142}$
- If $k^* \leq n/600$, then

$$V_X \leq 24\sigma^2 \left(\underbrace{\frac{k^*}{n}}_{\text{low. rank}} + \underbrace{\frac{n \sum_{j>k^*} \lambda_j^2}{(\sum_{j>k^*} \lambda_j)^2}}_{\text{H.D. regularization}} \right)$$

with $\mathbb{P}_X \geq 1 - 5\exp(-n/84)$.

Proof:

We have $V_X = \sigma^2 \langle (X^T X)^+, \Sigma \rangle$

Next lemma disentangles the low rank part and the H.D. part.

Lemma:

$$\boxed{\langle (X^T X)^+, \Sigma \rangle \leq \langle (X_\leq^T X_\leq)^+, \Sigma_\leq \rangle + \langle (X_>^T X_>)^+, \Sigma_> \rangle}$$

Proof: We have $(X^T X)^+ = X^T (X X^T)^{2+} X$

and $X X^T = X_\leq X_\leq^T + X_> X_>^T$. So

$$\langle (X^T X)^+, \Sigma \rangle = \langle (X_\leq X_\leq^T + X_> X_>^T)^{2+}, \underbrace{X \Sigma X^T}_{= X_\leq \Sigma_\leq X_\leq^T + X_> \Sigma_> X_>^T} \rangle$$

We have:

$X X^T \succeq X_\leq X_\leq^T$ and $X_> X_>^T$ so

$$\begin{cases} (X_\leq X_\leq^T)^{2+} \succeq (X X^T)^{2+} \text{ on range}(X_\leq) \\ (X_> X_>^T)^{2+} \succeq (X X^T)^{2+} \text{ on range}(X_>) \end{cases}$$

Hence

$$\langle (X^T X)^+, \Sigma \rangle \leq \langle (X_\leq X_\leq^T)^{2+}, X_\leq \Sigma_\leq X_\leq^T \rangle + \langle (X_> X_>^T)^{2+}, X_> \Sigma_> X_>^T \rangle$$

$$= \langle (X_\leq^T X_\leq)^+, \Sigma_\leq \rangle + \langle (X_>^T X_>)^+, \Sigma_> \rangle$$

$\square$

Low rank term : we have

$$\langle (X_{\leq}^T X_{\leq})^+, \Sigma_{\leq} \rangle_F \overset{k^* \leq m}{=} \langle (\Sigma_{\leq}^{-1/2} X_{\leq}^T X_{\leq} \Sigma_{\leq}^{-1/2})^{-1}, I_{k^*} \rangle_F$$

full rank

full rank

• From Lemma (Spec) 1 we have

$$\mathrm{Spec}\left(\Sigma_{\leq}^{-1/2} X_{\leq}^T X_{\leq} \Sigma_{\leq}^{-1/2}\right) \geq \frac{m}{4}$$

with proba $\geq 1 - 2\exp(-m/84)$

• On this event $\mathrm{Spec}\left((\Sigma_{\leq}^{-1/2} X_{\leq}^T X_{\leq} \Sigma_{\leq}^{-1/2})^{-1}\right) \leq \frac{4}{m}$

So $\langle (X_{\leq}^T X_{\leq})^+, \Sigma_{\leq} \rangle_F \leq k^* \cdot \frac{4}{m}$

with proba $\geq 1 - 2\exp(-\frac{m}{84})$.

H. D. term : Again from Lemma (Spec) 2

with proba $\geq 1 - 2\exp(-\frac{m}{84})$, we have

$$\mathrm{Sp}(X_> X_>^T) \subset \left[\tfrac{1}{4}\mathrm{Tr}(\Sigma_>), 2\,\mathrm{Tr}(\Sigma_>)\right].$$

Hence, on this event:

$$\langle (X_>^T X_>)^+, \Sigma_> \rangle_F = \langle (X_> X_>^T)^{2+}, X_> \Sigma_> X_>^T \rangle_F$$

$$\leq \left(\frac{4}{\mathrm{Tr}(\Sigma_>)}\right)^2 \langle I_m, X_> \Sigma_> X_>^T \rangle_F$$

It remains to evaluate the trace

$$\langle I_m, X_> \Sigma_> X_>^T \rangle_F = \langle I_m, Z \Sigma_>^2 Z \rangle_F$$

$$= \sum_{i=1}^m (Z \Sigma_>^2 Z^T)_{ii} = \sum_{i=1}^m \sum_{j=1}^{p-k^*} Z_{ij}^2 \lambda_{k^*+j}^2$$

H.W.

$$\leq \left(1+\tfrac{1}{4}\right) m \sum_{j>k^*} \lambda_j^2 + 8 \lambda_{k^*+1}^2 \Big\}$$

So with probability $\geq 1 - e^{-m/32}$, we have

$$\langle I_m, X_> \Sigma_> X_>^T \rangle_F \leq \frac{5m}{4} \sum_{j>k^*} \lambda_j^2 + \frac{m}{4} \underbrace{\lambda_{k^*+1}^2}_{\leq \sum_{j>k^*} \lambda_j^2}$$

$$\leq \frac{6}{4}\, m \sum_{j>k^*} \lambda_j^2$$

Hence, with proba $\geq 1 - 3\exp(-\frac{m}{84})$, we have

$$\langle (X_>^T X_>)^+, \Sigma_> \rangle_F \leq 24\, m \frac{\sum_{j>k^*} \lambda_j^2}{\left(\sum_{j>k^*} \lambda_j\right)^2}$$

$\square$

Remarks:

→ a matching lower bound shows that

$$V_X \simeq \sigma^2 \left(\frac{k^*}{n} + \frac{n \sum_{j>k^*} \lambda_j^2}{\left(\sum_{j>k^*} \lambda_j\right)^2} \right)$$

with $k^* = \min\{k : n\lambda_{k+1} \leq \rho \sum_{j>k} \lambda_j\}$, where $\rho = 1/142$

→ for $\Sigma = I_p$ and $p \gg n$, we have $k^* = 0$ and

$$V_X \simeq \sigma^2 \cdot \frac{np}{p^2} = \sigma \frac{n}{p}$$

⇝ We recover our guess

→ for $\Sigma = I_k + \alpha I_{\geq k}$ with $\alpha \ll 1$ and $k \ll n \ll p$,

we have $k^* = k$ and

$$V_X \simeq \sigma^2 \left(\frac{k}{n} + \frac{n\alpha^2(p-k)}{\alpha^2(p-k)^2} \right) \overset{k \ll p}{\simeq} \sigma^2 \left(\frac{k}{n} + \frac{n}{p} \right)$$

⇝ We recover our guess.

What about the bias B_X ?

It has been shown that

$$B_X \simeq \underbrace{\|\beta^*_{\leq}\|^2_{\Sigma_{\leq}^{-1}} \times \left(\frac{\sum_{j>k^*} \lambda_j}{n}\right)^2}_{\text{small bias for the low dimensional part}} + \underbrace{\|\beta^*_{>}\|^2_{\Sigma_{>}}}_{\text{high bias for the high-dimensional part.}}$$

- **Remark**:
  The bias for the low dimensional part $\simeq$ bias of ridge regression for $\beta^*_{\leq}$ with
  $$\lambda := \frac{1}{n} \sum_{j>k^*} \lambda_j .$$

- **Take home message**: in the interpolation regime, for the GD least square estimator $\hat{\beta}^{LS} = X^+ Y$
  - $\rightarrow$ the low dimensional part $\beta^*_{\leq}$ can be well estimated.
  - $\rightarrow$ the high-dimensional part $\beta^*_{>}$ is almost estimated by 0 $\rightarrow$ large bias, $\rightarrow$ small variance

3_ Benign Overparametrisation

Recap: so far we have seen that

→ GD induces an implicit regularization in the interpolation regime.

→ For linear regression, the interpolating estimator selected by G.D. on the least-square problem, is not suffering from an overly high variance in very high dimension, thanks to the implicit regularization of GD. But it may suffer from a large bias.

→ What does it tell us on "overparametrisation" (having much more parameters than the sample size)

In linear regression $y_i = \langle x_i, \underbrace{\beta}_{\in\mathbb{R}^p} \rangle + \varepsilon_i$

p is both → input dimension
→ number of parameters

⇝ we must look at a more complex model to disentangle over-parametrisation from high input dimension.

Is over-parametrisation new in statistics?

Certainly not! Think to classical non-parametric regression

$$y_i = f^*(x_i) + \varepsilon_i \;,\; i=1,..,n$$

where f^* is some (say) Sobolev function.

What is new then? In non-parametric statistics, we use some explicit regularisation

For example, cubic splines are solution to the regularized Least-Square problem

$$f^{spline} \in \underset{\int (f'')^2 < +\infty}{\operatorname{argmin}} \left\{ \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \underbrace{\int (f'')^2}_{\text{regularization by mean curvature}} \right\}$$

What is new in Neural Network practice is that some people use very complex over-parametrized models without any explicit regularization.

. Why does-it work?

Thanks to implicit regularization by G.D.

We will illustrate this on a simple example.

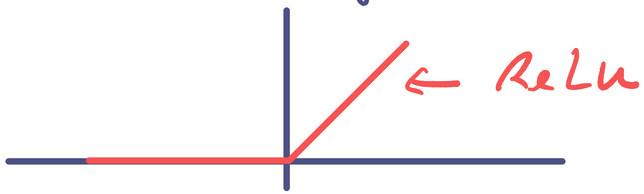
Over-parametrized 1-hidden layer Neural Network

$f_{\beta,\omega} : \mathbb{R}^d \to \mathbb{R}$, defined by

$$f_{\beta,\omega}(x) := \frac{1}{m} \sum_{j=1}^{m} \beta_j \varphi(\langle \omega_j, x \rangle)$$

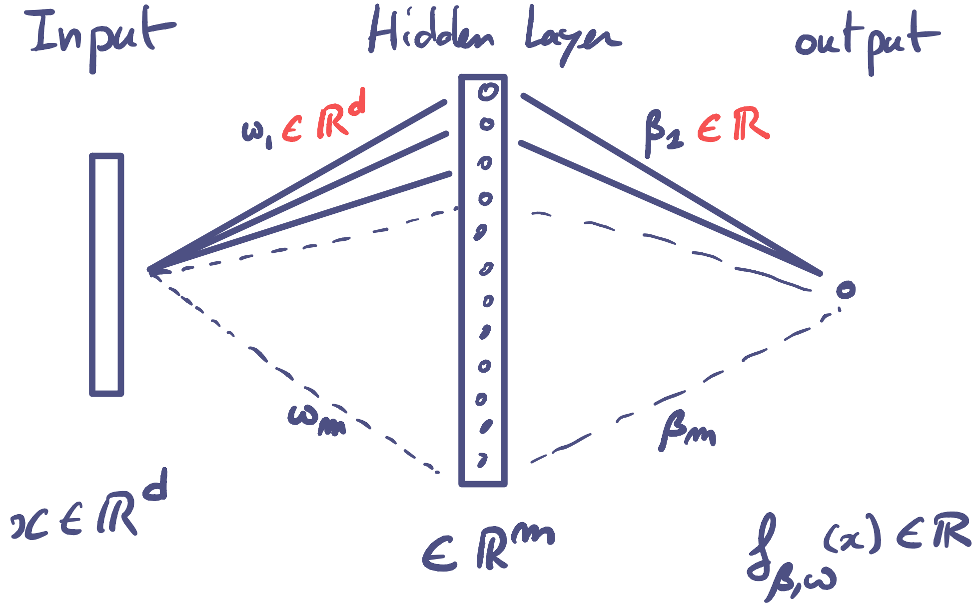
with

- φ = activation function e.g. ReLu



- m = number of hidden neurons
- $\theta = (\beta_j, \omega_j)_{j=1..m} \in \mathbb{R}^{m(d+1)}$
- overparametrization: $m \to \infty$

(with fixed input dimension)



learning: let $(x_i, y_i)_{i=1,..,n} \in (\mathbb{R}^d \times \{-1,1\})^n$ and ℓ be a convex loss function -

$(\hat{\beta}, \hat{\omega})$ are learnt from GD on the empirical loss

$$\mathcal{L}(\beta,\omega) = \frac{1}{n} \sum_{i=1}^{n} \ell(-y_i f_{\beta,\omega}(x_i))$$

Even if ℓ is convex, the function $\mathcal{L}$ is not convex

· over-parametrisation: $m \to +\infty$

the limit $m \to \infty$, corresponds to the parametrisation

$$f_\mu(x) := \int_{\beta,\omega} \beta\, \varphi(\langle \omega, x\rangle)\, d\mu(\underbrace{\beta,\omega}_{=\theta})$$

with $\mu \in \mathcal{P}(\mathbb{R}^{d+1})$

· Setting $\Phi(\theta, x) = \beta\, \varphi(\langle \omega, x\rangle)$

we have

$$f_\mu(x) = \langle \Phi(\cdot, x), \mu \rangle$$

linear!

where $\langle g, \mu \rangle := \int g(\theta)\, d\mu(\theta)$

$$\mathcal{L}(\mu) = \frac{1}{M} \sum_{i=1}^{m} \ell(-y_i f_\mu(x_i))$$

is convex!

⇝ here, overparametrisation helps for the optimisation landscape

⇝ is there a statistical price for it?

Example: φ = Relu, ℓ = logistic loss

· reparametrisation:

$$\Phi(\lambda\theta, x) = \lambda\beta\, \varphi(\langle \lambda\omega, x\rangle) = \lambda^2 \beta\, \varphi(\langle \omega, x\rangle) = \lambda^2 \Phi(\theta, x)$$

so with $\theta = \lambda u$, $\lambda > 0$ and $u \in S_d$

$$f_\mu(x) = \int_{\substack{\lambda>0\\ u \in S_d}} \Phi(\lambda u, x)\, d\mu(\lambda u)$$

$$= \int_{u \in S_d} \Phi(u, x) \underbrace{\int_{\lambda > 0} \lambda^2\, d\mu(\lambda u)}_{=: d\pi_\mu(u)\ \in \mathcal{M}_+(S_d)}$$

$$= \langle \Phi(\cdot, x), \pi_\mu \rangle$$

Overfitting: we can represent any Lipschitz function f by a f_μ. So if no point x_i has two different labels y_i in the learning data, then $\exists\, \mu$ such that

$$y_i = \operatorname{sign}(f_\mu(x_i)) \quad i=1..n \quad \text{(perfect fit)}$$

Max-margin: assume that at some stage t of G.D., $f_{\hat\mu^t}$ perfectly fit the data i.e for $i=1,\dots,n$

$$y_i f_{\hat\mu^t}(x_i) = \langle y_i \Phi(x_i,\cdot), \underbrace{\Pi_{\hat\mu^t}}_{=:\, \hat\pi^t} \rangle > 0.$$

Then, since $\ell(-z) = \log(1+e^{-z})$ decreases we can always decrease $\mathcal{L}(\hat\mu^t)$ by simply increasing the mass of $\hat\pi^t$

$\rightsquigarrow$ Consequence $|\hat\pi^t| \to +\infty$

$\rightsquigarrow$ since $\ell(-z) \simeq e^{-z}$ as $z \to +\infty$

$$\mathcal{L}(\hat\mu^t) \simeq \frac{1}{n}\sum_i \exp\left(-|\hat\pi^t| \left\langle y_i \Phi(x_i,\cdot), \frac{\hat\pi^t}{|\hat\pi^t|}\right\rangle\right)$$

$$\simeq \frac{N_{min}}{n} \exp\left(- \underbrace{|\hat\pi^t|}_{\to +\infty} \underbrace{\min_i \left\langle y_i \Phi(x_i,\cdot), \frac{\hat\pi^t}{|\hat\pi^t|}\right\rangle}_{\text{margin of } f_{\hat\pi^t/|\hat\pi^t|}}\right)$$

Guess: for t large

$$\frac{\hat\pi^t}{|\hat\pi^t|} \simeq \hat\nu \in \operatorname*{argmax}_{\nu \in \mathcal{P}(S_d)} \min_i y_i \underbrace{f_\nu(x_i)}_{\langle \Phi(x_i,\cdot), \nu\rangle}$$

Theorem (informal)

Under some mild conditions, the above guess holds true.

It is then possible to prove that the classifier $\hat{h}(x) := \operatorname{sign}(f_{\hat{\nu}}(x))$ has some nice statistical properties.

For example, it is able to adapt to low-dimensional structures...

(See Chizat and Bach 2020)

⇝ in this case

implicit regularisation of G.D.

\+ overparametrisation

$$\Downarrow$$

1) Nice optimisation landscape

2) Nice statistical behavior

Take home messages:

→ in the interpolation regime, G.D. has a regularizing effect by selecting some specific interpolating solutions

→ when the input space is very large, overfitting only occurs on domains which are rarely sampled. So overfitting does not harm prediction risk

→ over-parametrisation can be harmless, and even beneficial.

All these phenomena have to be better understood