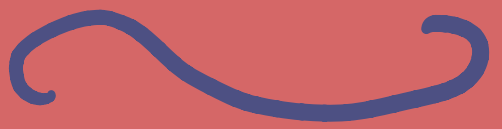
Implicit regularisation, Benign overfitting, and over-parametrisation



Lecture 5

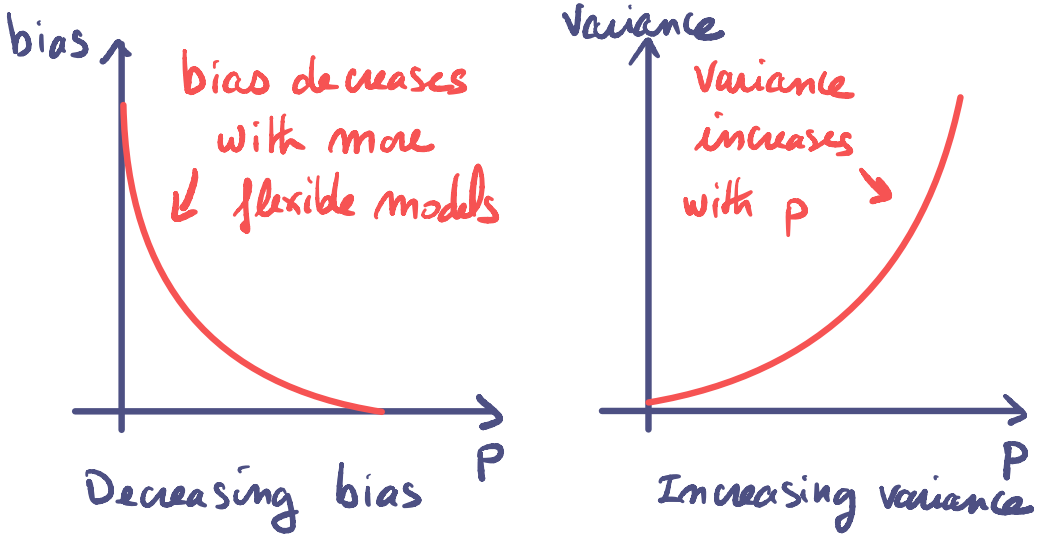
Main references for today:

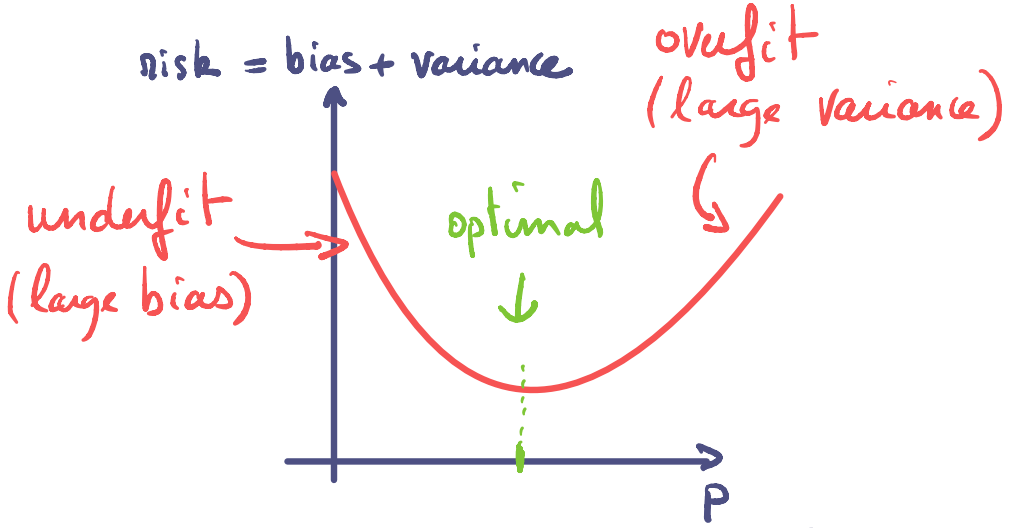
- M. Belkin, D. Hsu, S. Ma, S. Mandal. "Reconciling modern machine learning and the bias-variance trade-off. (2018)
- P. Bartlett, P. Long, G. Lugosi, A. Tsigler. "Benign overfitting in Linear Regression". (2019)
- T. Hastie, A. Montanari, S. Rosset, R. Tibshirani. "Surprises in high-dimensional ridgeless 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)

Double descent phenomenon

Textbook theory: risk = bias + variance

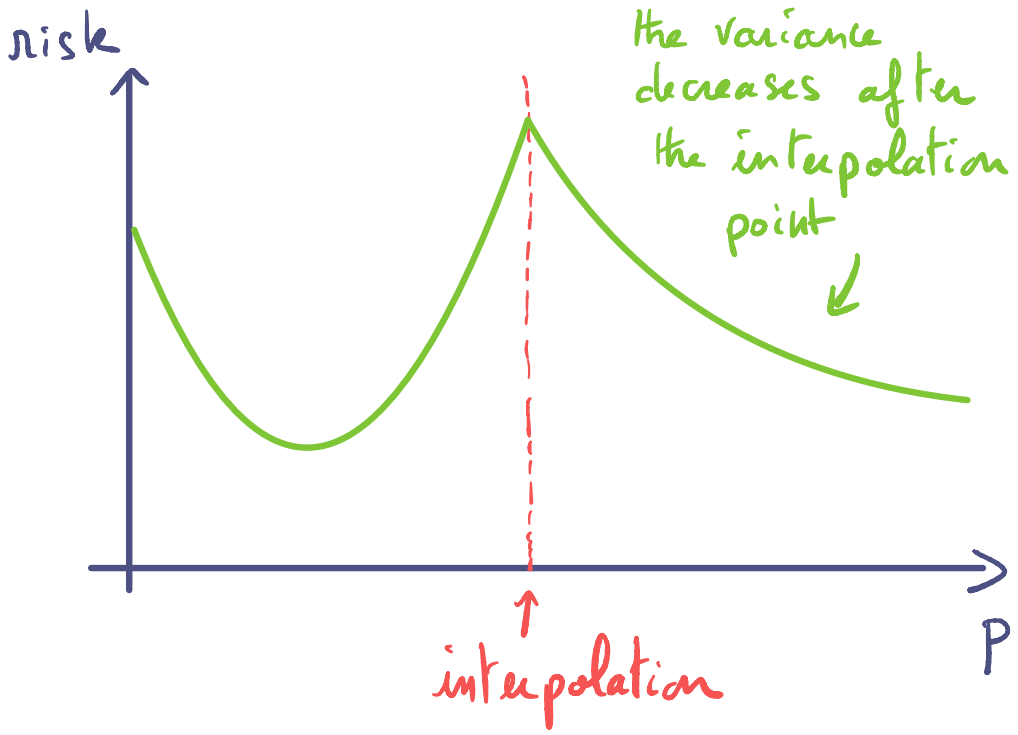
Let p = number of parameters in the model.





Bias-variance trade-off

Empirical observations (neural networks)



Double descent phenomenon

Today:

1. Implicit regularisation (a.k.a. implicit bias) of gradient descent
2. The magic of high-dimensional input: benign overfitting in high-dimensional (linear) regression
3. Benign overparametrisation

① Implicit regularisation of G.D.

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

- High-dimension: $x_i \in \mathbb{R}^p$ with $p > n$

- G.D. on least-squares:

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

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

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

$$\overset{t \to +\infty}{\longrightarrow} \left(I - (I - 2\eta X^T X)\right)^+ \cdot 2\eta X^T Y \qquad (2\eta < \|X\|_{op}^{-2})$$

$$= X^+ Y$$

↑ Moore Penrose pseudo inverse

So $\hat{\beta} = X^+ Y = \underset{Y = X\beta}{\operatorname{argmin}} \|\beta\|^2$

↑ if $\operatorname{rank}(X) = n$ (interpolation)

↑ regularisation of G.D.

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

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

$$\overset{Y = X\beta^* + \varepsilon}{=} \langle X^+ X \beta^*, x \rangle + \langle X^+ \varepsilon, x \rangle$$

$X^+X = P_{X^T}$ (projection on $\operatorname{range}(X^T)$)

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

Remarks: $\operatorname{range}(X^+) = \operatorname{range}(X^T)$

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

→ large bias, but no variance

2) $\operatorname{rank}(X) \leq n \ll p$:

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

② Benign overfitting with high-dimensional input

a) (Quasi-) isotropic input: $\Sigma = I_p$ (or $\Sigma \eqsim I_p$)

Variance analysis: $X \overset{\text{SVD}}{=} \sum_k \hat{\sigma}_k \hat{u}_k \hat{v}_k^T$

$$\text{Var}_\varepsilon \langle \hat{\beta}, x \rangle = \sigma^2 x^T (X^T X)^+ x$$

$$= \sigma^2 x^T \left(\sum_{k=1}^{n} \frac{1}{\hat{\sigma}_k^2} \hat{v}_k \hat{v}_k^T \right) x$$

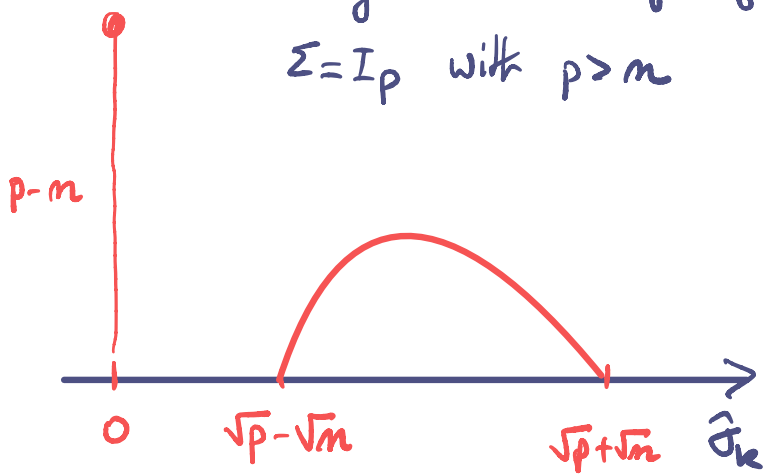
$\hat{\sigma}_n^2 \eqsim p$ w.h.p. when $p \gg n$

$$\eqsim \frac{\sigma^2}{p} \underbrace{x^T P_{X^T} x}_{= \| P_{X^T} x \|^2}$$

if $x \sim \mathcal{N}(0, I_p)$, $\| P_{X^T} x \|^2 \eqsim n$ w.h.p.

$$\eqsim \boxed{\sigma^2 \frac{n}{p}} \quad \text{with high-probability}$$

$\to 0$ when $p \gg n$!



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

Average prediction error:

$\hat\beta = P_{X^T}\beta^* + X^+\varepsilon$; $P_{\ker(X)}$ (projection on $\ker(X)$)

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

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

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

B_X ; V_X

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

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

    The larger p, the larger the bias B_X

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

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

  - $V_X = \sigma^2 \operatorname{Tr}(X^TX)^+ = \sigma^2 \sum_{k=1}^{n} \frac{1}{\tilde\sigma_k^2}$

    $\simeq \sigma^2 \frac{n}{p}$

so $R_X = B_X + V_X \simeq \|\beta^*\|^2 + \frac{n}{p}\left(\sigma^2 - \|\beta\|^2\right)$

$\Sigma \simeq I_p$: None complex formula, based on Stieltjes transform of Marchenko-Pastur distribution, but exhibits the same behavior.

Recap at this stage: ($\Sigma \simeq I_p$)

→ if $x \in \text{range}(X^T)$: $\text{var}_\varepsilon \langle \hat{\beta}, x \rangle = \sigma^2 \frac{\|x\|^2}{p} \simeq \sigma^2$ ← strong overfit on range(X^T)

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

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

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

b) Anisotropic case :

Prototypical example: $\Sigma = I_k + \rho I_{p-k} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \rho & \\ & & & \ddots \end{bmatrix}$ (k, $p-k$) with
- $k \ll n \ll p$
- $\rho \ll 1$

Then $\frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \simeq I_k + \rho \frac{p-k}{n-k} \hat{I}_{n-k}$

I_k: low dimensional $\Rightarrow$ good estimation

$\hat{I}_{n-k}$: $\simeq$ projection of rank $n-k$

$\frac{p-k}{n-k}$: Trace $= k + \rho(p-k)$

$k \ll n \ll p$

So $(X^T X)^+ \simeq \frac{1}{n} I_k + \frac{1}{\rho p} \hat{I}_{n-k}$ and

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

$$\simeq \sigma^2 \left(\underbrace{\frac{k}{n}}_{\text{small since small rank } k} + \underbrace{\frac{n}{p}}_{\text{small due to high-dimensional input}} \right)$$

General case : $\Sigma = \sum_{j=1}^{p} \lambda_j v_j v_j^T$

Theorem (informal)

. Define $k^* = \min\{k : n\lambda_{k+1} \le c \sum_{j>k} \lambda_j\}$ ($c < 1$)

. Then

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

Example : $\Sigma = I_{k^*} + \rho\, I_{n-k^*}$

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

$$\simeq \sigma^2 \left(\frac{k^*}{n} + \frac{n}{p} \right) \quad :)$$

$k^* \ll n$

Sketch of proof :

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

Set $\xi_j := \frac{1}{\sqrt{\lambda_j}} X v_j = \frac{1}{\sqrt{\lambda_j}} Z \Sigma^{1/2} v_j$

$= Z v_j \overset{iid}{\sim} N(0, I_n)$

$$\underbrace{XX^T}_{n \times n} = \sum_{j=1}^{p} X v_j v_j^T X^T = \sum_{j=1}^{p} \lambda_j \xi_j \xi_j^T$$

$$= \underbrace{S_{\le k^*}}_{\text{part related to } V_X = \text{span}\{v_1, \dots, v_{k^*}\}} + \underbrace{S_{>k^*}}_{\text{part related to } V_*^\perp}$$

Key Lemma :

$$\text{Sp}(S_{>k^*}) \simeq \lambda^* := \sum_{j>k^*} \lambda_j$$

$\rightsquigarrow$ delocalisation of small energy spectrum

Proof : For $\|v\| = 1$

$$v^T S_{>k^*} v = \sum_{j>k^*} \lambda_j \underbrace{(v^T \xi_j)^2}_{\overset{iid}{\sim} N(0,1)}$$

So

$$\max_{\|v\|=1} \left| v^T S_{>k^*} v - \underbrace{\sum_{j>k^*} \lambda_j}_{=\lambda^*} \right| = \max_{\|v\|=1} \left| \sum_{j>k^*} \lambda_j \underbrace{\left((v^T g_j)^2 - 1 \right)}_{\overset{iid}{\sim} N(0,1)} \right|$$

Hanson-Wright + ε-net discretisation $\longrightarrow$

$$\lesssim n \, |S_{>k^*}|_{op} + \sqrt{n \sum_{j>k^*} \lambda_j^2}$$

$$\lesssim \left(1 + \frac{1}{\alpha}\right) n \underbrace{|S_{>k^*}|_{op}}_{=\lambda_{k^*+1}} + \alpha \underbrace{\sum_{j>k^*} \lambda_j}_{=\lambda^*}$$

$$\lesssim \underbrace{c}_{\text{from } k^*} \left(1 + \frac{1}{\alpha}\right) \lambda^*$$

So

$$(1 - c') \lambda^* \lesssim sp(S_{>k^*}) \lesssim (1 + c') \lambda^*$$

where $c' \simeq c\left(1 + \frac{1}{\alpha}\right) + \alpha$ can be small if $c \ll 1$ and $\alpha = \sqrt{c}$.

□

For $S_{\leq k^*}$: similar arguments show that

$$sp\left(\frac{1}{n} S_{\leq k^*}\right) \simeq \{\lambda_1, \dots, \lambda_{k^*}\}$$

← low rank estimation part.

Consequence: $Sp(XX^T) \simeq \{n\lambda_1, \dots, n\lambda_{k^*}, \underbrace{\lambda^*, \dots, \lambda^*}_{n-k^* \text{ times}}\}$ and

$$Sp(X^TX) \simeq \{n\lambda_1, \dots, n\lambda_{k^*}, \underbrace{\lambda^*, \dots, \lambda^*}_{n-k^*}, \underbrace{0, \dots, 0}_{p-n}\}$$

So $\frac{1}{n} X^TX \simeq \underbrace{\Sigma_{\leq k^*}}_{\text{low rank part}} + \underbrace{\frac{\lambda^*}{n} \sum_{j=k^*+1}^{n} \hat{v}_j \hat{v}_j^T}_{\text{delocalised spectrum}}$ (compare to prototypical example)

$$V_X := \sigma^2 \langle \Sigma, (X^TX)^+ \rangle \simeq \frac{\sigma^2}{n} \langle \Sigma_{\leq k^*}, \Sigma_{\leq k^*}^+ \rangle + \frac{\sigma^2}{\lambda^*} \sum_{j=k^*+1}^{n} \hat{v}_j^T \Sigma_{>k^*} \hat{v}_j$$

$$\simeq \frac{\sigma^2}{n} k^* + \frac{\sigma^2}{\lambda^*} \sum_{l>k^*} \lambda_l \sum_{j=k^*+1}^{n} \underbrace{\hat{v}_j^T v_l v_l^T \hat{v}_j}_{= (\hat{v}_j^T v_l)^2 \simeq \frac{\lambda_l}{\lambda^*}}$$

since $\sum_l (\hat{v}_j^T v_l)^2 = 1$ and energy in direction v_l proportional to λ_l

$$\simeq \frac{k^*}{n} \sigma^2 + \frac{n - k^*}{(\lambda^*)^2} \sum_{l \geq k^*+1} \lambda_l^2$$

□

③ Benign over-parametrisation

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

p is both → input dimension
→ number of parameters

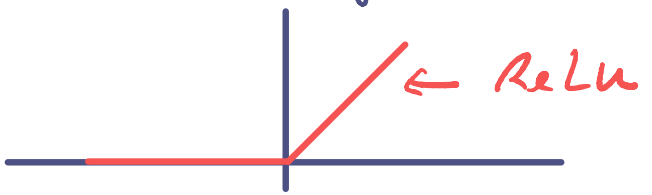
⇝ we must look at a more complex model to disentangle over-parametrisation from input dimension.

1-Hidden Layer Neural Network

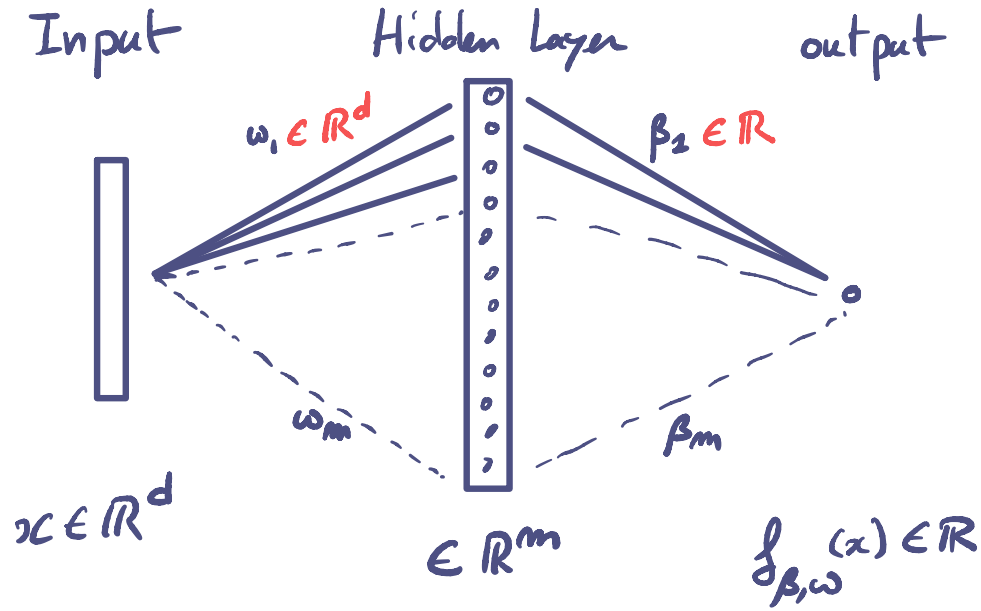
$$f_{\beta,\omega}(x) := \frac{1}{m} \sum_{j=1}^{m} \beta_j \varphi(\langle \omega_j, x \rangle)$$

with

- φ = activation function e.g. ReLu



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



learning: let $(x_i, y_i)_{i=1,\dots,n} \in (\mathbb{R}^d \times \{-1, 1\})^n$ 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 <u>not</u> 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

$$f_\mu(x) = \langle \Phi(\cdot, x), \mu\rangle \quad \text{(linear!)}$$

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

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

is convex !

$\rightsquigarrow$ here, over-parametrisation helps for the optimisation landscape

$\rightsquigarrow$ is there a statistical price for it?

Example: $\varphi = \text{Relu}$, $\ell = \text{logistic loss}$

- 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

$$y_i = \text{sign}\left(f_\mu(x_i)\right) \quad i=1,\dots,n$$ (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 $\ell(-z) \approx e^{-z}$ as $z \to +\infty$

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

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

Guess: for t large

$$\frac{\hat{\Pi}^t}{|\hat{\Pi}^t|} \approx \hat{\nu} \in \underset{\nu \in \mathcal{P}(S_d)}{\text{argmax}} \min_i y_i \underbrace{f_\nu(x_i)}_{\langle \Phi(x_i,\cdot), \nu\rangle}$$

Theorem (informal)

Under some mild conditions, the above guess holds true

It is then possible to prove that the classifier $\hat{h}(x) := \text{sign}\left(f_{\hat{w}}(x)\right)$ has some nice statistical properties. For example, it is able to adapt to low-dimensional structures ...

⇝ in this case implicit regularisation of G.D. + overparametrisation

⇓

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 phenomenon have to be better understood

Messages from this course :

→ in general, learning in high-dimensional spaces is hopeless

→ but feasible, when there exist some low-dimensional structures in the data

→ it is possible to adapt to low-dimensional structures, but it comes with computational challenges (convexification or iterative optimisation)

→ bias from convexification can be an issue

→ G.D. induces some regularisation beneficial in highly dimensional input spaces and/or highly overparametrised models.
