Théorie de l’apprentissage : TLDR

High-dimensional statistics : non asymptotic

- Bias-variance decomposition of risk (same approach as in classical case)
- Explicit computation impossible for typical ML problems
- Upper bounding :
    - bias : Rademacher (or Gaussian) averages
    - variance : concentration inequalities

High-dimensional statistics : asymptotic

- Different asymptotic regimes, e.g. n -> +∞ et n/p -> constante

Deep learning theory?

- Over-parametrisation + SGD -> computationally tractable minimum-norm interpolation?
- Good generalisation properties of minimum-norm interpolations?

Optimisation stochastique

Contexte : minimisation du risque empirique pour une fonction de coût “séparable par point de donnée”

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Descente de gradient stochastique

$$w_1 \in \mathbb{R}^d \text{ given}$$

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k)$$

i_k is chosen *randomly* from $\{1, \ldots, n\}$ and α_k is a positive stepsize

Optimisation stochastique

Exemple de garantie de convergence

(cf. Bottou, Curtis et Nocedal (2018) Optimisation Methods for Large-Scale Machine Learning)

Si

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$$

Assumption 4.1 (**Lipschitz-continuous objective gradients**). *The objective function* $F : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable and the gradient function of* F*, namely,* $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$*, is Lipschitz continuous with Lipschitz constant* $L > 0$*, i.e.,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leq L\|w - \overline{w}\|_2 \quad \textit{for all} \quad \{w, \overline{w}\} \subset \mathbb{R}^d.$$

plus des conditions de régularité pas très contraignantes

Alors

$$\liminf_{k \to \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0$$

Algèbre linéaire et optimisation

Normes de vecteurs

A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a *vector norm* if it has the following properties:

1. $\|\mathbf{x}\| \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar $\alpha \in \mathbb{R}$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2$$

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Si Q est une matrice orthogonale, $\|Qx\|_2 = \|x\|_2$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Normes de matrices

A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that has the following properties:

- $\|A\| \geq 0$ for any $A \in \mathbb{R}^{m\times n}$, and $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$ for any $m \times n$ matrix A and scalar α
- $\|A + B\| \leq \|A\| + \|B\|$ for any $m \times n$ matrices A and B

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

$$\|A\|_2 = \sigma_1$$

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}.$$

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$

Application de la notion de norme: lien valeurs propres, valeurs singulières

$$\sigma_r \leq |\lambda| \leq \sigma_1$$

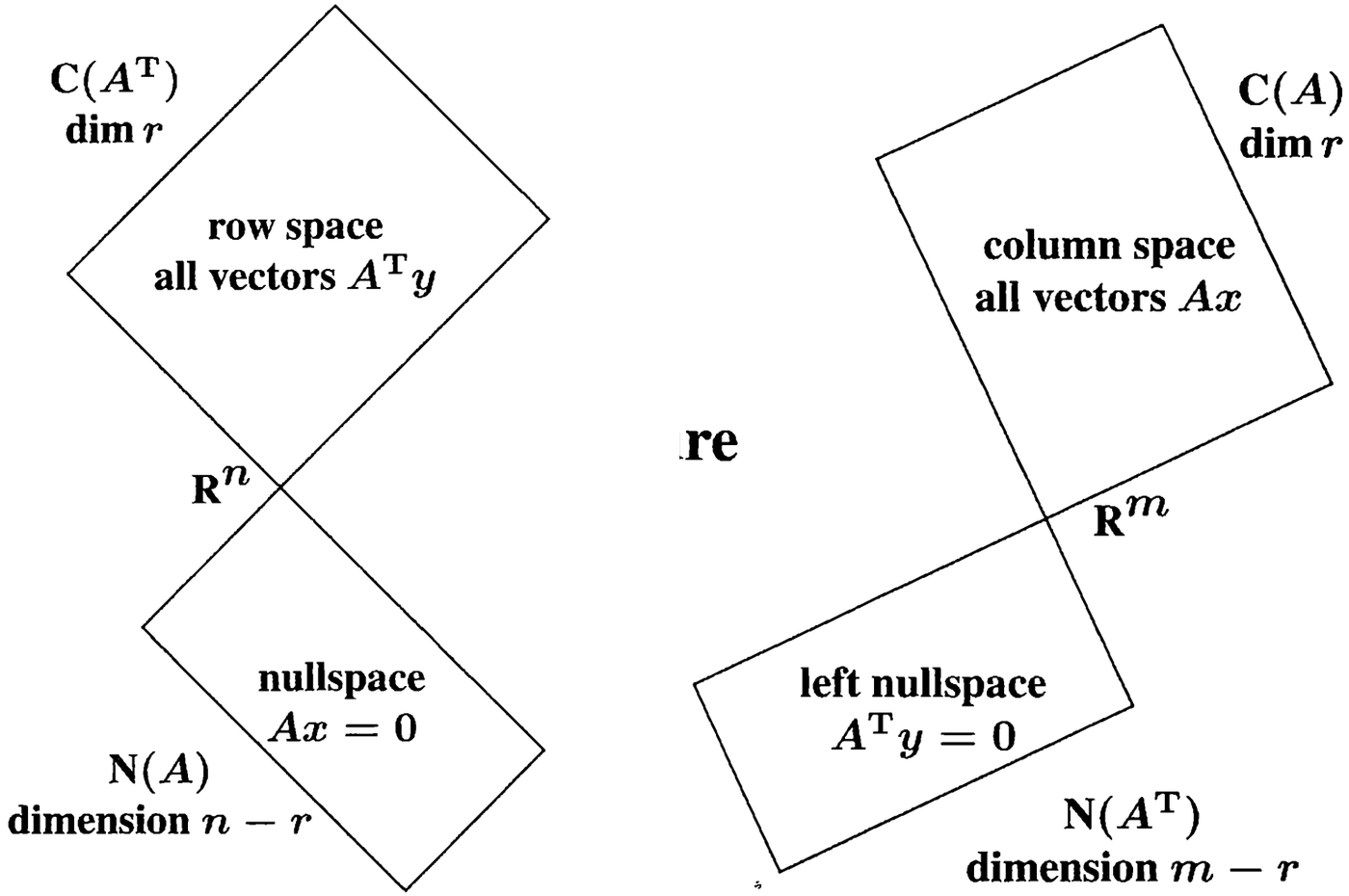
Application de la notion de norme: lien valeurs propres, valeurs singulières

$$\sigma_r \leq |\lambda| \leq \sigma_1$$

Éléments propres de $A^T A$ et $A A^T$?

Moindre carrés linéaires

$$AV = U\Sigma \qquad A\begin{bmatrix} & & \\ v_1 & .. & v_r \; .. \; v_n \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ u_1 & .. & u_r \; .. \; u_m \\ & & \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$



Algèbre linéaire numérique

Algèbre linéaire numérique

Exemple : produit matrice-matrice A: $m \times n$, B: $n \times p$.

Produit matriciel $C = AB$:

$C = 0$

Boucle $i = 1..m$, $j = 1..p$, $k = 1..n$:

$c_{ij} = c_{ij} + a_{ik} b_{kj}$

Ordre des trois boucles ?

i en premier $C = \begin{bmatrix} \tilde{c}_1^T \\ \vdots \\ \tilde{c}_m^T \end{bmatrix} = AB = \begin{bmatrix} \tilde{a}_1^T B \\ \vdots \\ \tilde{a}_m^T B \end{bmatrix}$

j en premier $C = \begin{bmatrix} c_1 \cdots c_p \end{bmatrix} = AB = \begin{bmatrix} Ab_1 \cdots Ab_p \end{bmatrix}$

k en premier $C = AB = \sum_{k=1}^{n} a_k \tilde{b}_i^T$

Algèbre linéaire numérique

Stabilité numérique

Algèbre linéaire numérique

Stabilité numérique

Supposons qu'un processus nous donne la solution estimée $\hat{x}$ d'un système linéaire $Ax = b$ en effectuant toutes les opérations matricielles de manière exacte mais qu'il y a des erreurs d'arrondi dans le stockage en nombre flottants de A et de b dans la mémoire (en pratique les opérations matricielles effectuées pour trouver x introduisent davantage d'erreurs d'arrondis, ce modèle donne donc une borne supérieure sur la qualité possible d'un algorithme numérique pour résoudre des systémes linéaires).

Alors,

$$\text{si } \mathbf{u}\kappa_\infty(A) \leq .5 \qquad \frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leq 4\mathbf{u}\kappa_\infty(A)$$

u est le unité d'arrondi, égale à la moitié de lécart entre 1 et le plus petit nombre flottant strictement supérieur à 1. Pour les nombres flottants IEEE single précision, **u** est d'environ 10^{-7}. Il est d'environ 10^{-16} pour les nombres flottants IEEE double précision.

$\kappa_\infty(A) := \|A\|_\infty \|A^{-1}\|_\infty$ est le conditionnement de A pour la norme ∞