MIA Séance 7
Théorie de l'apprentissage & Optimisation stochastique

Amal BOURAHMA
Vanessa GUERRIER
GHRISSI ALOUI Abdelaziz
KEBAB Zinedine

October 13, 2025

Contents

1 Introduction

Ce document présente les bases de l'algèbre linéaire et de l'optimisation appliquées à l'apprentissage automatique. Il décrit les normes de vecteurs et de matrices, ainsi que leur rôle dans la stabilité numérique. La notion de biais–variance y est analysée, notamment en grande dimension et en deep learning. Enfin, une introduction à l'optimisation stochastique conclut le contenu.

2 Rappel : Algèbre linéaire et optimisation

2.1 Norme de vecteur

Une fonction

$$||.|| : R^n \to R$$
$$x \to ||x||$$

est dit norme d'un vecteur si les propriete suivant sont respecter:

1. $||x|| \geq 0$ et $||x|| = 0 \iff x = 0$ pour $\forall\, x \in R^n$
2. $||\alpha x|| = |\alpha| \times ||x||$ pour $\forall\, x \in R^n$ et $\alpha \in R$
3. $||x + y|| \leq ||x|| + ||y||$ pour $\forall\, x, y \in R^n$

Pour un vecteur $x \in R^n$ la norme p de x est $||x||_p = (\sum_{i=1}^{n} x^p)^{\frac{1}{p}}$

2.1.1 Normes 1,2 et ∞

- norme 1 : $||x||_1 = |x_1| + |x_2| + ... + |x_n|$
- norme 2 : $||x||_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = \sqrt{x^T x}$
- norme ∞ : $||x||_\infty = max_{1 \leq i \leq n} |x_i|$

Si Q est une matrice orthogonales, $||Qx||_2 = ||x||_2$.

2.2 Norme de Matrice

Une fonction

$$||.|| : R^{m \times n} \to R$$
$$A \to ||A||$$

est dit norme d'une matrice si les propriete suivant sont respecter:

1. $||A|| \geq 0$ et $||A|| = 0 \iff A = 0$ pour $\forall\, A \in R^{m \times n}$
2. $||\alpha A|| = |\alpha| \times ||A||$ pour $\forall\, A \in R^{m \times n}$ et $\alpha \in R$
3. $||A + B|| \leq ||A|| + ||B||$ pour $\forall\, A, B \in R^{m \times n}$

2.3 Norme de matrice induite (ou naturelle)

La norme induite d'une matrice A est définie par :

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

Elle représente le *facteur d'agrandissement maximal* que A peut produire sur un vecteur.

2.4 Propriété sous-multiplicative

La norme induite vérifie toujours :

$$\|AB\| \leq \|A\| \, \|B\|.$$

2.5 Cas particulier : la norme 2 (norme euclidienne)

Si l'on prend la norme euclidienne $\|\cdot\|_2$, alors :

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

On cherche donc le vecteur x de norme unité qui maximise $\|Ax\|_2$.

2.6 Lien avec les valeurs propres

On définit la fonction :

$$g(x) = \|Ax\|_2^2 = x^T(A^TA)x.$$

Le maximum de $g(x)$ pour $\|x\|_2 = 1$ est atteint lorsque x est un vecteur propre de A^TA associé à la plus grande valeur propre $\lambda_{\max}$. Ainsi :

$$\|A\|_2^2 = \lambda_{\max}(A^TA), \qquad \boxed{\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}}.$$

La norme 2 d'une matrice est donc la racine carrée de la plus grande valeur propre de A^TA ,donc la plus grande valeur singuliere de A^TA.

2.7 Norme de Frobenius

La norme de Frobenius d'une matrice est définie par :

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

C'est la racine carrée de la somme des carrés de tous les éléments de A. Elle est plus facile à calculer que la norme 2, mais elle n'est pas induite par une norme vectorielle.

2.8 Stabilité numérique

Le conditionnement d'une matrice est une mesure de la sensibilité de la solution d'un système linéaire aux petites perturbations dans les données ou aux erreurs d'arrondi. Pour une matrice A $\in R^{nn}$ inversible, le nombre de conditionnement de A dans une norme p est

$$\kappa_p(A) = \|A\|_p\|A^{-1}\|_p$$

. Si on cherche a estimé $\hat{x}$ pour le systeme linéaire Ax=b

Si $\kappa_p(A) \approx 1$, la matrice est **bien conditionnée** :

les petites erreurs sur A ou b ne changent que peu la solution $\hat{x}$.

Si $\kappa_p(A) \gg 1$, la matrice est **mal conditionnée** :

une petite erreur dans les données peut produire une grande erreur dans $\hat{x}$.

3 Théorie de l'apprentissage : Biais et Variance

3.1 Biais dans les statistiques classiques et en grande dimension

Définition et rôle du biais

En apprentissage statistique, le **biais** représente l'erreur systématique d'un estimateur, c'est-à-dire la différence entre la fonction moyenne estimée $\mathbb{E}[\hat{f}(x)]$ et la fonction cible $f^*(x)$ que l'on cherche à approcher.

Formellement, pour une donnée x :

$$\text{Biais}(x) = \mathbb{E}[\hat{f}(x)] - f^*(x). \tag{1}$$

Un biais élevé traduit une incapacité du modèle à reproduire correctement la structure de la donnée (*sous-ajustement*), tandis qu'un biais faible indique que la classe de fonctions considérée est suffisamment riche pour approcher la réalité. Dans la décomposition classique du risque quadratique moyen le biais correspond donc à l'erreur de modélisation.

3.2 Le biais dans le cadre non asymptotique en grande dimension

Lorsque la dimension du problème devient grande (p comparable ou supérieur à n), la notion de biais garde le même sens conceptuel, mais son évaluation explicite devient impossible. En pratique, on ne calcule plus directement le biais : on cherche à le **borner** à l'aide d'outils probabilistes.

Dans ce cadre :

- le biais est lié à la **capacité d'approximation** de la classe de fonctions utilisée ;
- il est contrôlé par des mesures de **complexité statistique**, comme les *moyennes de Rademacher* ou les *complexités gaussiennes*.

Ces outils quantifient la richesse de la famille de modèles considérés : une classe de fonctions trop riche peut conduire à un biais très faible (le modèle peut presque tout approximer), tandis qu'une classe trop restreinte impose un biais important.

Ainsi, dans les statistiques en grande dimension, le biais devient une **mesure de la complexité d'approximation** plutôt qu'une simple différence moyenne.

3.3 Le biais dans le cadre asymptotique

Lorsque l'on considère des régimes asymptotiques où $n \to +\infty$ et $p \to +\infty$ simultanément, avec un rapport $n/p \to c$ constant, le biais dépend explicitement de ce rapport.

Dans ces régimes, l'analyse montre que le biais peut **diminuer ou augmenter** selon la manière dont la complexité du modèle croît avec le nombre d'observations. C'est notamment dans ce contexte qu'apparaît le phénomène de *double descente*, où le biais décroît "coucou vanessa":) avec la complexité du modèle jusqu'à un certain point, avant de se stabiliser ou de réaugmenter lorsque le modèle devient excessivement flexible.

3.4 Le biais dans le deep learning

Dans le cas du *deep learning*, la situation est encore plus remarquable. Les réseaux de neurones profonds sont souvent **surparamétrés** ($p \gg n$) et capables d'interpoler parfaitement les données d'apprentissage, ce qui suggérerait un biais nul. Pourtant, ces modèles généralisent bien sur des données nouvelles.

Une explication couramment admise est que la combinaison de la surparamétrisation et de la *descente de gradient stochastique (SGD)* conduit à une solution particulière parmi toutes celles qui minimisent l'erreur d'apprentissage : celle de **norme minimale**, aussi appelée *interpolation à norme minimale*.

Cette solution possède un biais faible — le modèle ajuste correctement les données — mais reste régulière grâce à une **régularisation implicite** introduite par l'algorithme d'optimisation.

Ainsi, dans le *deep learning* moderne, le biais ne résulte pas d'une contrainte explicite du modèle, mais d'un **effet émergent du processus d'apprentissage** lui-même.

3.5 Variance

La variance quantifie dans quelle mesure les prédictions d'un modèle changent lorsque celui-ci est réentraîné sur des jeux de données légèrement différents issus de la même distribution.

Un modèle à **forte variance** s'adapte excessivement aux particularités du jeu d'entraînement. Il présente donc une faible stabilité et un risque élevé de *sur-apprentissage* (overfitting). À l'inverse, un

modèle à **faible variance** produit des prédictions stables mais parfois trop rigides pour capturer les structures fines des données, conduisant à un *sous-apprentissage* (underfitting).

Sur le plan théorique, la variance intervient dans la décomposition de l'erreur quadratique moyenne :

$$\mathbb{E}_{\text{train}}\left[(\hat{f}(x) - f^*(x))^2\right] = \underbrace{(\mathbb{E}[\hat{f}(x)] - f^*(x))^2}_{\text{Biais}^2} + \underbrace{\mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Bruit irréductible}}.$$

où :

- f^* est la fonction cible réelle ;
- $\hat{f}$ est la fonction apprise par le modèle ;
- σ^2 correspond à la variance du bruit des données.

Cette décomposition montre que la variance représente la variabilité intrinsèque du modèle, due à la manière dont celui-ci s'adapte aux échantillons d'apprentissage.

3.6 Variance et Statistiques en grande dimension

Dans les **statistiques en grande dimension**, la variance devient un facteur déterminant. Lorsque le nombre de paramètres p est comparable au nombre d'observations n, la variance de l'estimateur peut croître de manière significative, rendant le modèle instable et difficile à généraliser.

- **Régime non asymptotique :** on cherche à encadrer la variance pour un nombre fini d'observations à l'aide d'outils probabilistes tels que les *inégalités de concentration* (Hoeffding, Bernstein, etc.), qui fournissent des bornes sur les fluctuations aléatoires de la performance.
- **Régime asymptotique :** on étudie le comportement lorsque $n, p \to \infty$ avec un rapport n/p tendant vers une constante. Ce cadre est souvent abordé via la *théorie des matrices aléatoires*, qui permet d'expliquer comment la variance se stabilise ou diverge selon le rapport dimension/échantillons.

Un graphique typique illustre la variance de l'erreur de test en fonction du rapport n/p : lorsque n/p est faible (modèle très paramétré, peu de données), la variance est élevée ; elle décroît ensuite lorsque le nombre d'observations augmente.

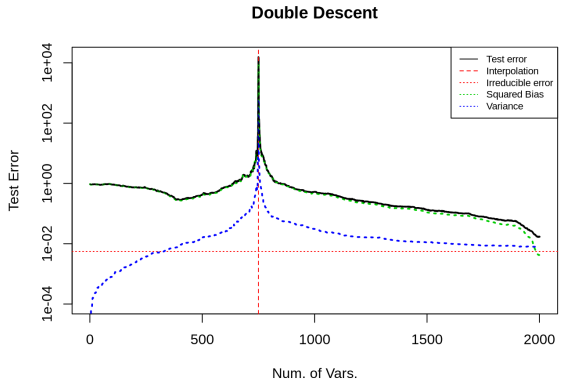

Figure 1: Courbe de *double descent* illustrant l'évolution de la variance de l'erreur de test dans un contexte de grande dimension. Lorsque le nombre de paramètres p approche le nombre d'échantillons n, la variance de la perte de test augmente fortement avant de décroître à nouveau dans le régime sur-paramétré. Source : Belkin et al. (2019), *PNAS*, 116(32), 15849–15854.

3.7 Le compromis biais–variance

La variance ne peut être étudiée isolément : elle interagit constamment avec le **biais** du modèle. On parle alors du **compromis biais–variance**, qui décrit l'équilibre entre la capacité d'un modèle à s'adapter aux données et sa stabilité.

- Un modèle **simple** (faible capacité) présente un biais élevé : il sous-estime la complexité de la relation réelle. Il commet des erreurs systématiques (biais élevé, variance faible).
- Un modèle **complexe** (haute capacité) réduit le biais mais devient très sensible au bruit du jeu d'entraînement (variance élevée, risque de sur-apprentissage).

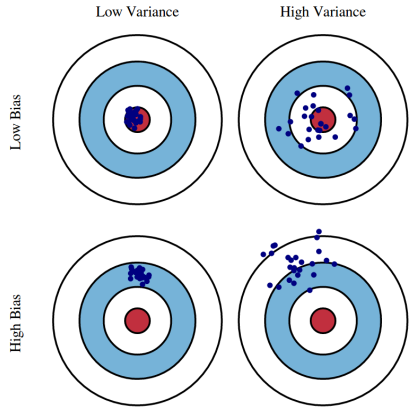

Figure 2: Illustration du compromis biais–variance, où la position des points montre le biais (écart systématique à la vérité) et leur dispersion traduit la variance (sensibilité du modèle aux données).**Source :** Cornell University, Lecture 12, "Bias–Variance Tradeoff", 2018.

L'objectif de l'apprentissage est donc de trouver un point d'équilibre où la somme du biais au carré et de la variance est minimale, ce qui correspond à la plus faible erreur de généralisation.

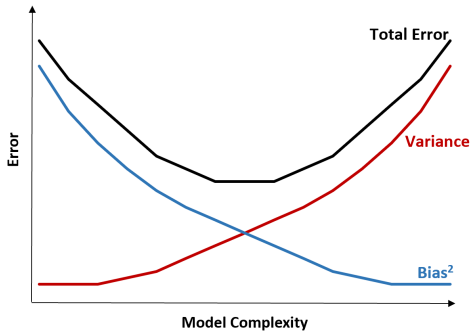

Figure 3: Illustration du compromis biais–variance : l'erreur totale (en noir) résulte de la somme du biais^2 (en bleu) et de la variance (en rouge). Le point minimal correspond à l'équilibre optimal entre les deux composantes.Source : Statology, "Bias-Variance Tradeoff".

4 Optimisation stochastique

4.1 Descente de Gradient Stochastique (SGD)

L'idée de la descente de gradient stochastique est de mettre à jour les paramètres w à partir d'un **échantillon aléatoire** du jeu de données, plutôt que d'utiliser tous les exemples à chaque itération.

$$\begin{cases} w_1 \in \mathbb{R}^d \text{ donné} \\ w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k) \end{cases}$$

où :

- i_k est choisi aléatoirement parmi $\{1, \ldots, n\}$,
- α_k est un pas d'apprentissage positif,
- $\nabla f_{i_k}(w_k)$ est le gradient du coût pour l'exemple i_k.

Cette approche réduit le coût de calcul tout en assurant une convergence vers un minimum de la fonction de coût globale.

4.2 Garantie de Convergence

Selon Bottou, Curtis et Nocedal (2018) dans *Optimisation Methods for Large-Scale Machine Learning*, la descente de gradient stochastique converge sous certaines conditions.

4.3 Hypothèses sur les pas d'apprentissage

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

Explication détaillée :
Ces deux conditions sur les pas d'apprentissage α_k sont fondamentales pour garantir la convergence :

- **Première condition :** $\sum_{k=1}^{\infty} \alpha_k = \infty$

  Cette condition assure que l'algorithme **accumule suffisamment de pas** pour pouvoir atteindre la solution optimale, même si elle est éloignée du point initial. Sans cette condition, l'algorithme pourrait s'arrêter trop tôt, avant d'avoir atteint un minimum.

- **Deuxième condition :** $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$

  Cette condition garantit que la **variance des mises à jour diminue** avec le temps. Elle permet de réduire le bruit introduit par l'échantillonnage stochastique et assure que les itérations successives deviennent de plus en plus stables.

Exemple de séquence valide : $\alpha_k = \frac{1}{k}$ satisfait ces deux conditions :

$$\sum_{k=1}^{\infty} \frac{1}{k} = \infty \quad \text{et} \quad \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} < \infty$$

4.4 Hypothèse de Lipschitz-continuité

La fonction objectif $F : \mathbb{R}^d \to \mathbb{R}$ est continûment différentiable et son gradient est Lipschitz-continu :

$$\|\nabla F(w) - \nabla F(w')\|_2 \leq L \|w - w'\|_2, \quad \forall w, w' \in \mathbb{R}^d$$

pour une constante $L > 0$.
Explication détaillée :
L'hypothèse de Lipschitz-continuité du gradient est cruciale pour plusieurs raisons :

- **Bornitude de la variation du gradient :** Elle garantit que le gradient ne peut pas varier trop brutalement entre deux points proches. Cela signifie que la fonction a une **courbure limitée**.

- **Conséquence importante :** Cette condition implique la **majoration quadratique** suivante :

$$F(w) \leq F(w') + \nabla F(w')^T (w - w') + \frac{L}{2}\|w - w'\|_2^2$$

  Cette inégalité est fondamentale pour analyser la décroissance de la fonction objectif à chaque itération.

- **Interprétation pratique :** En apprentissage automatique, cette condition est souvent vérifiée pour les fonctions de perte convexes et lipschitziennes comme la perte logistique ou le SVM à marge douce.

- **Valeur de L :** La constante L est appelée **constante de Lipschitz** et joue un rôle important dans le choix du pas d'apprentissage. Typiquement, on choisit $\alpha_k \leq \frac{1}{L}$ pour assurer la convergence.

4.5 Résultat de convergence

Sous ces hypothèses (et quelques conditions de régularité peu contraignantes), on a :

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

Explication détaillée :
Ce résultat de convergence mérite plusieurs précisions :

- **Liminf vs Lim :** On utilise $\liminf$ plutôt que lim car la suite des gradients peut osciller. Le $\liminf$ garantit qu'il existe une sous-suite qui converge vers zéro.

- **Espérance mathématique :** La convergence est en espérance car l'algorithme est stochastique. Cela signifie qu'en moyenne, le carré de la norme du gradient tend vers zéro.

- **Interprétation :**
  - $\|\nabla F(w_k)\|_2^2 \to 0$ signifie que l'algorithme atteint un **point stationnaire** (où le gradient s'annule)
  - Pour les fonctions convexes, ce point stationnaire est un **minimum global**
  - Pour les fonctions non-convexes, il peut s'agir d'un minimum local ou d'un point selle

- **Conditions supplémentaires :** Les "conditions de régularité peu contraignantes" incluent typiquement :
  - La bornitude des moments des gradients stochastiques
  - La continuité de la fonction objectif
  - L'existence de minimums

- **Vitesse de convergence :** Sous des hypothèses supplémentaires (forte convexité), on peut obtenir des vitesses de convergence plus précises, comme une convergence à taux $O(1/k)$.

5 Conclusion

Dans ce rapport, nous avons abordé les notions relatives aux normes de matrices et de vecteurs. Nous avons présenté les conditions permettant de déterminer si une fonction définit bien une norme sur une matrice ou sur un vecteur.Nous avons également traité du conditionnement d'une matrice, concept qui permet de vérifier si une matrice est stable face aux erreurs d'arrondi ou aux petites perturbations des données.Ensuite, nous avons abordé les notions de biais et de variance, en donnant leurs définitions et en les expliquant dans des cadres asymptotique et non asymptotique.Enfin, nous avons discuté de l'optimisation stochastique, de sa garantie de convergence, ainsi que des hypothèses sur le pas d'apprentissage.En somme, ce rapport établit un lien entre l'algèbre linéaire et l'apprentissage automatique, en montrant comment les concepts mathématiques fondamentaux soutiennent les méthodes modernes d'apprentissage.