Fiche de révision

Paramétrique ou non · k-NN · distances · espérance conditionnelle · biais–variance

Oulad Ali Ayoub

TD X-HEC · APM_51434_EP · École Polytechnique

Table des matières

1 Les trois familles d'apprentissage . . . 3
2 Paramétrique ou non paramétrique ? . . . 3
2.1 Les deux définitions . . . 4
2.2 Exemples en 2D . . . 4
2.3 Exemple en 3D . . . 5
2.4 Les « briques » des méthodes non paramétriques . . . 5
2.5 Pourquoi choisir l'un ou l'autre ? . . . 6
3 On se concentre sur le k-NN . . . 8
3.1 Comment on choisit les voisins . . . 8
3.2 La forme de la fonction . . . 9
3.3 Exemple de classification, à la main . . . 9
3.4 Exemple de régression, à la main . . . 10
4 Les distances . . . 10
4.1 Définition : les 4 règles . . . 11
4.2 Exemples de distances . . . 11
4.3 Distances équivalentes . . . 12
4.4 Fabriquer une distance : on part d'une norme . . . 14
4.5 Boîte à outils : les petites inégalités . . . 14
4.6 Distances entre fonctions . . . 15
5 Espérance et espérance conditionnelle : les rappels . . . 16
5.1 L'espérance et la variance . . . 16
5.2 L'espérance conditionnelle . . . 17
5.3 Les règles de calcul . . . 18
5.4 Comment reconnaître ce qui dépend de X . . . 18
6 Biais et variance d'un estimateur . . . 19
6.1 Application au k-NN : conditionnel ou inconditionnel ? . . . 21
7 Autotest : 16 questions éclair . . . 23

Le fil de la fiche

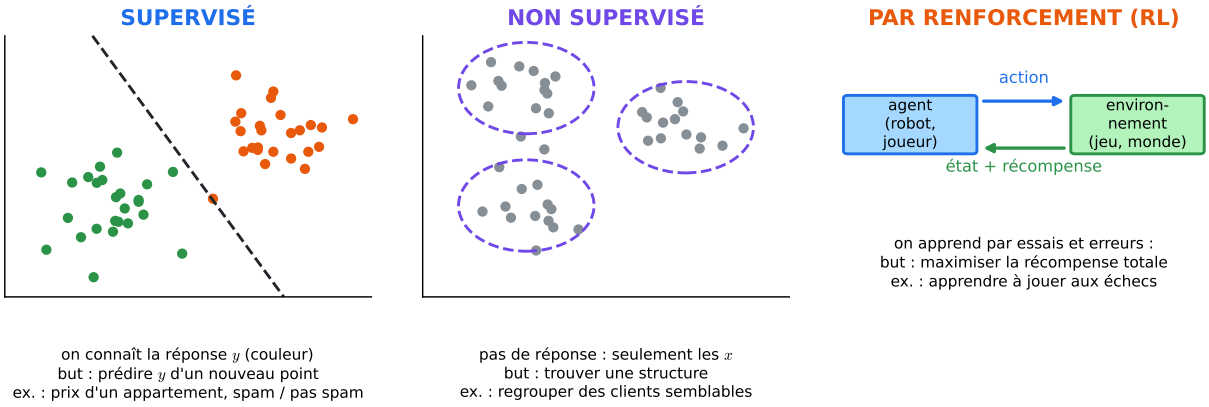
1. Les 3 familles d'apprentissage (supervisé, non supervisé, RL).
2. Paramétrique ou non paramétrique ? Définitions, exemples, quand choisir quoi.
3. On se concentre sur le **k-NN** : la forme de la fonction, un exemple pas à pas.
4. Pour le k-NN il faut une **distance** : les 4 axiomes, exemples, équivalence, comment en fabriquer une.
5. La boîte à outils d'inégalités, puis les distances entre fonctions.
6. Rappels : **espérance** et **espérance conditionnelle**, leurs règles, et comment reconnaître ce qui dépend de X (donc se comporte comme une constante quand on conditionne).
7. Biais et variance d'un estimateur $\hat{\mu}$, puis le **k-NN** : biais et variance conditionnels ou non

à X.

Comment lire

Chaque page : **un dessin + l'essentiel**. Boîte bleue = vie de tous les jours ; verte = à retenir ; rose = piège ; orange = réflexe.

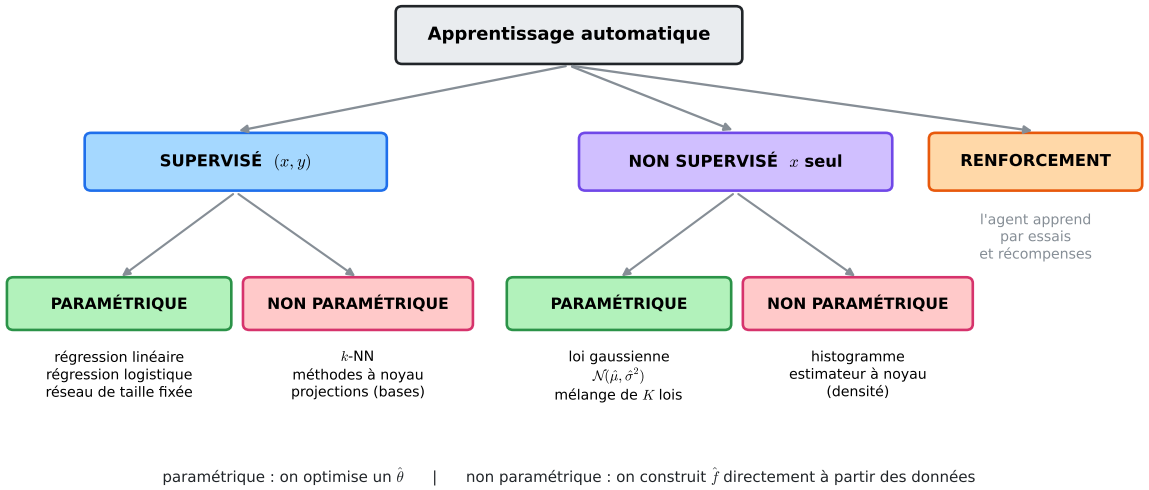
1 Les trois familles d'apprentissage



Famille	Ce qu'on a	Ce qu'on cherche	Exemple
Supervisé	des couples (x_i, y_i) : l'entrée *et* la réponse	une fonction $\hat{f}$ avec $\hat{f}(x) \approx y$	loyer selon la surface ; spam ou pas
Non supervisé	seulement des x_i	une structure cachée (groupes, densité, direction principale)	grouper des clients
Renforcement (RL)	un agent qui agit et reçoit une récompense	une stratégie qui maximise la récompense totale	apprendre à jouer aux échecs

À retenir

Dans le **supervisé** *et* dans le **non supervisé**, les méthodes se classent en deux camps : **paramétriques** et **non paramétriques**. C'est l'objet de la section suivante.



paramétrique : on optimise un $\hat{\theta}$ | non paramétrique : on construit $\hat{f}$ directement à partir des données

2 Paramétrique ou non paramétrique ?

2.1 Les deux définitions

Méthode paramétrique

On **choisit à l'avance une famille** $\{f_\theta\}$ avec $\theta \in \mathbb{R}^p$, où le nombre p de paramètres est **fixé** (il ne dépend pas de n). Apprendre = **optimiser** $\hat{\theta}$ (ou $\hat{w}$), par exemple

$$\hat{\theta} \in \arg\min_{\theta} \sum_i \ell\big(y_i, f_\theta(x_i)\big).$$

Ensuite on peut *jeter les données* : $\hat{\theta}$ résume tout. Ex. : régression linéaire $\hat{w} = (X^\top X)^{-1} X^\top y$.

Méthode non paramétrique

Il n'y a **ni forme imposée, ni vecteur θ de taille fixe**. On se donne une famille de « briques » (cases, voisinages, cloches, fonctions de base) et on **construit directement $\hat{f}$ à partir des données**. La complexité de $\hat{f}$ **grandit avec** n et on **garde les données**. Ex. : k-NN, histogramme, estimateur à noyau.

« Non paramétrique » $\neq$ « sans paramètre »

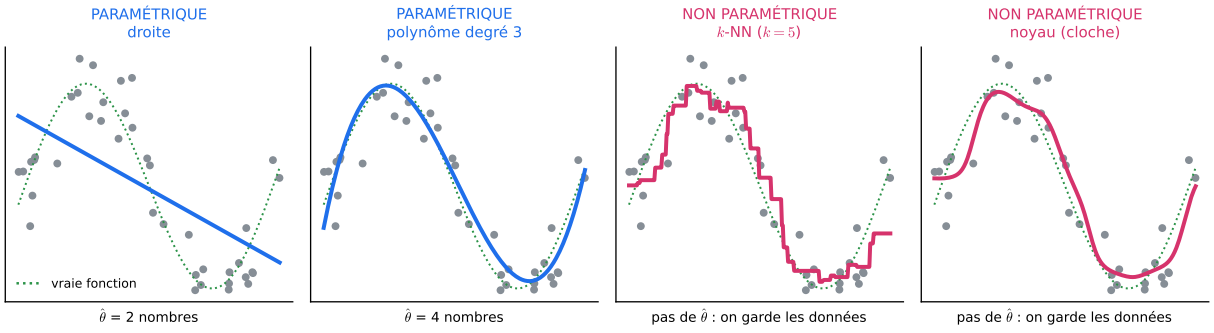
Il reste des **hyperparamètres** (k, largeur de cloche h, nombre de cases, nombre J de fonctions de base). Ils règlent la *souplesse* ; ce ne sont pas des paramètres appris par optimisation comme $\hat{w}$.

Dans la vraie vie

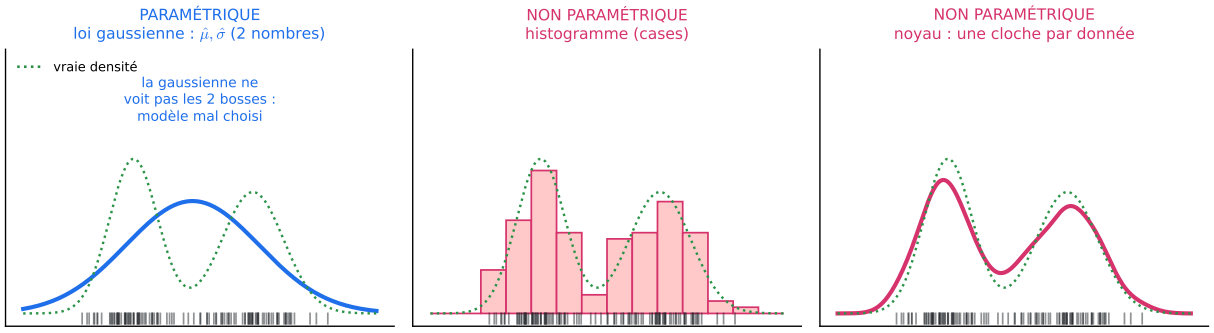
Paramétrique = une recette de cuisine : tu fixes la liste des ingrédients, tu ajustes juste les doses. **Non paramétrique** = tu demandes à tes voisins : « combien payez-vous de loyer ? » et tu fais la moyenne. Pas de recette, tu gardes l'annuaire.

	Paramétrique	**Non paramétrique**
ce qu'on choisit	une famille f_θ	une règle de construction (brique)
ce qu'on apprend	$\hat{\theta}$, de taille fixe p	rien : $\hat{f}$ est construit directement
après l'entraînement	on garde $\hat{\theta}$	on garde **toutes** les données
complexité	fixe	grandit avec n
réglages	choix du modèle	k, h, nombre de cases, J
exemples	droite, polynôme de degré fixé, logistique, gaussienne	k-NN, histogramme, noyau, projection

2.2 Exemples en 2D

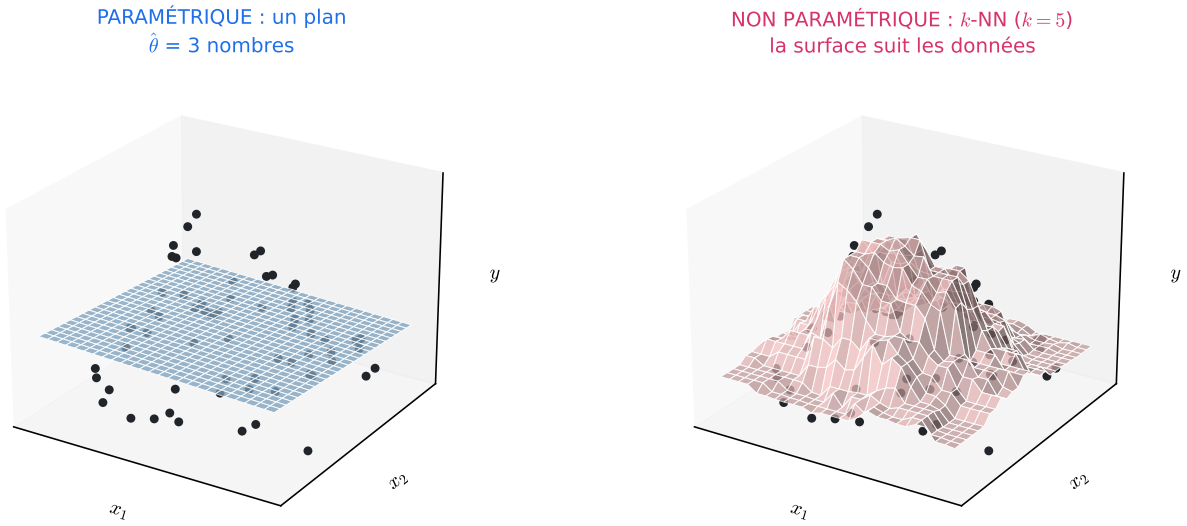


Mêmes données. À gauche, deux familles fixées : la droite se trompe (modèle trop simple), le polynôme de degré 3 suffit ici. À droite, rien n'est imposé : $\hat{f}$ suit les données.



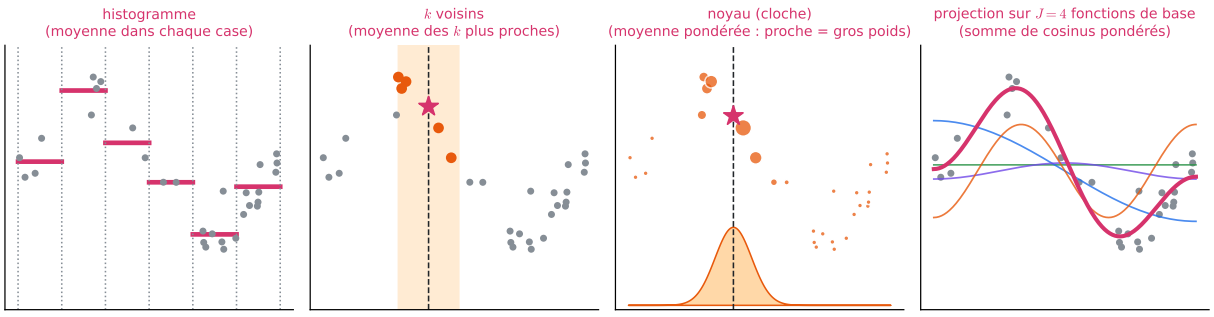
Même distinction en non supervisé (estimer une densité) : la gaussienne n'a que $\hat{\mu}, \hat{\sigma}$ et rate les deux bosses ; l'histogramme et le noyau les voient.

2.3 Exemple en 3D



Deux variables d'entrée (x_1, x_2), une sortie y (une « colline »). Paramétrique : un plan, trois nombres $\hat{\theta}$. Non paramétrique : la surface du k-NN épouse la colline.

2.4 Les « briques » des méthodes non paramétriques



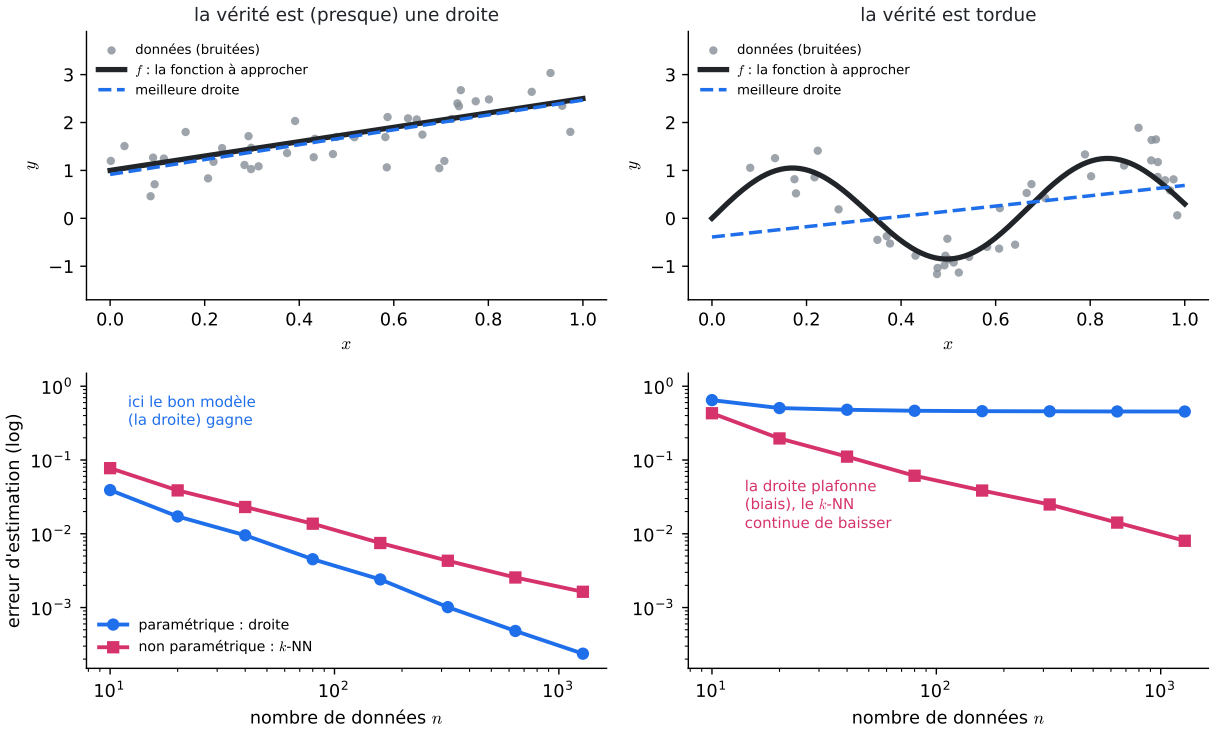
Brique	Estimateur de $f(x)$
histogramme (cases)	moyenne des y_i dont le x_i tombe dans la **même case** que x
k-NN	moyenne des y_i des k plus proches : $\hat{f}(x) = \frac{1}{k}\sum_{i\in N_k(x)} y_i$
noyau (cloche)	$\hat{f}(x) = \dfrac{\sum_i K\left(\frac{x-x_i}{h}\right) y_i}{\sum_i K\left(\frac{x-x_i}{h}\right)}$: moyenne pondérée, poids fort si x_i est proche
projection	$\hat{f} = \sum_{j=0}^{J-1} \hat{c}_j \varphi_j$ sur une base (cosinus, polynômes…) avec J qui **grandit avec** n
méthodes à noyau	$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$: autant de coefficients que de données

Piège classique

Si J est fixé une fois pour toutes (ex. polynôme de degré 3), c'est de la régression *paramétrique*. Ce qui rend la projection non paramétrique, c'est que la famille s'enrichit avec n.

2.5 Pourquoi choisir l'un ou l'autre ?

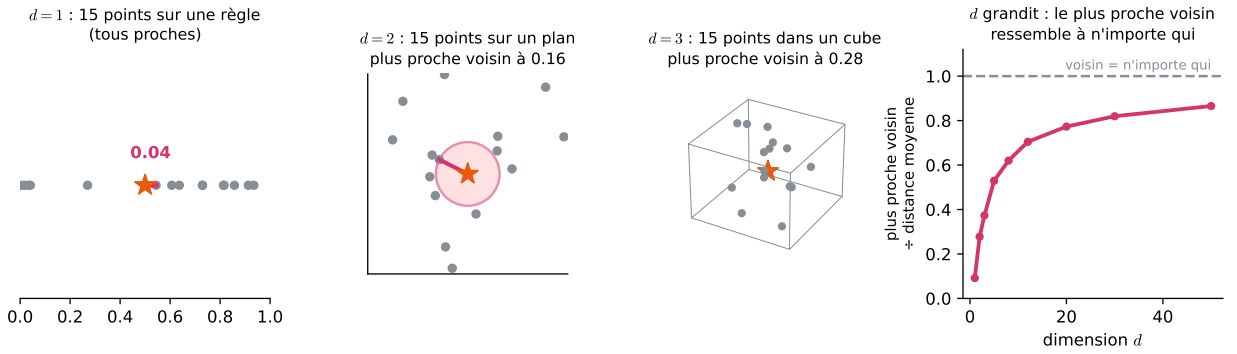
Paramétrique si…	Non paramétrique si…
peu de données	**beaucoup de données**
grande dimension d	**petite dimension** (1 à quelques variables)
tu crois à une **forme précise** (physique, expérience)	la **forme est inconnue**, tordue, locale
tu veux **interpréter** ($\hat{w}$ se lit)	tu veux **peu d'hypothèses**
prédiction **rapide**, mémoire légère	tu peux garder les données et payer le calcul
risque : modèle faux ⇒ biais qui ne baisse pas avec n	risque : fléau de la dimension, variance si n petit



En haut : la fonction f à approcher (trait noir), les données bruitées et la meilleure droite. En bas : l'erreur d'estimation quand n grandit. À gauche, f est une droite : le bon modèle paramétrique gagne. À droite, f est tordue : la droite plafonne (elle ne peut pas épouser la courbe), le k-NN continue de s'améliorer.

Le fléau de la dimension : en grande dimension, les voisins s'éloignent

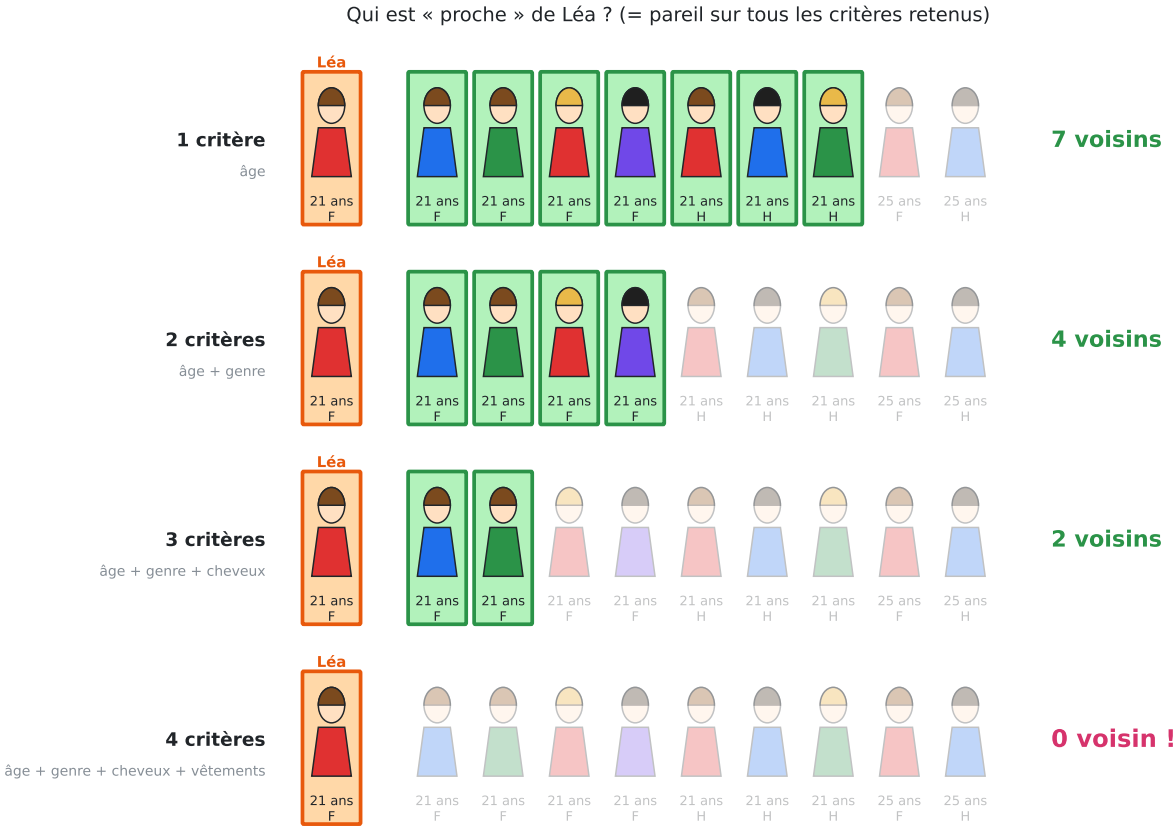
Pourquoi ? Être « proche », c'est être proche sur **toutes les coordonnées à la fois**. Chaque coordonnée en plus est une nouvelle occasion d'être loin : il y a de moins en moins de chances qu'un point soit proche de x partout en même temps. Avec n fixé, les points se « diluent » dans l'espace.



Même nombre de points (15), même point cible $\star$ au centre. Plus d grandit, plus son plus proche voisin (trait rouge) est loin. À droite : le plus proche voisin finit par être aussi loin que n'importe quel autre point (rapport $\to 1$).

Dans une classe

Tu cherches les « voisins » de Léa parmi 9 camarades. Avec l'**âge** seul, elle a beaucoup de voisins. Tu ajoutes le **genre** : moins. Puis la **couleur des cheveux** : encore moins. Puis la **couleur des vêtements** : **plus personne**. Chaque critère indépendant en plus a retiré des voisins.



Autre façon de le voir : pour que le voisinage d'un point contienne 10 % *des données, il faut un côté de* $0{,}1^{1/d}$ *:* 0,1 *en* $d = 1$, 0,32 *en* $d = 2$, 0,79 *en* $d = 10$. *En grande dimension il faut presque tout l'espace : le non paramétrique y souffre, le paramétrique tient bon.*

Choisir en deux questions

(1) Ai-je une **bonne raison de croire** à une forme précise ? → paramétrique. (2) Sinon : ai-je **beaucoup de données en petite dimension** ? → non paramétrique. Dans le doute : compare sur un jeu de **validation**.

3 On se concentre sur le k-NN

3.1 Comment on choisit les voisins

Dans la vraie vie

Tu arrives dans un nouveau quartier et tu veux estimer ton loyer : tu regardes les k appartements **les plus proches** du tien (même surface, même étage…) et tu fais la moyenne. Pour classer un fruit (pomme ou orange ?), tu regardes les k fruits les plus ressemblants et tu **votes**.

Données $(x_i, y_i)_{i=1..n}$, une **distance** d, un entier k, un nouveau point x.

1. calculer $d(x, x_i)$ pour **tous** les i ;
2. **trier** ces distances du plus petit au plus grand ;
3. garder les k premières : $N_k(x)$ = l'ensemble des indices des k plus proches ;
4. **agréger** leurs réponses y_i (moyenne ou vote).

3.2 La forme de la fonction

Régression ($y \in \mathbb{R}$) : on moyenne

$$\hat{f}_k(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i = \sum_{i=1}^{n} w_i(x)\, y_i, \qquad w_i(x) = \tfrac{1}{k}\mathbf{1}\{i \in N_k(x)\}.$$

C'est un estimateur **linéaire en** y : des poids $w_i \geq 0$, $\sum_i w_i = 1$.

Classification ($y \in \{1, \ldots, C\}$) : on vote

$$\hat{f}_k(x) = \arg\max_c \sum_{i \in N_k(x)} \mathbf{1}\{y_i = c\}$$

(la classe la plus fréquente parmi les k voisins).

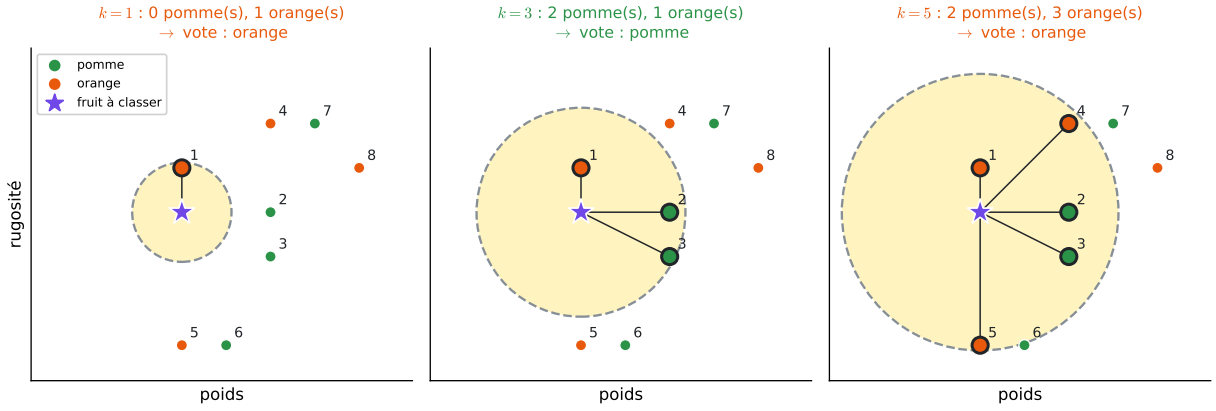
Deux classes $\{0,1\}$: $\hat{f}_k(x) = \mathbf{1}\{\frac{1}{k}\sum_{i \in N_k(x)} y_i \geq \frac{1}{2}\}$ (c'est « moyenne puis seuil à $\frac{1}{2}$ »).

Égalités

Si deux classes (ou plus) ont le même nombre de voix, pas de panique : prendre k impair aide (deux classes) mais **n'est pas obligatoire**. On fixe une **convention**, par exemple « **premier arrivé, premier servi** » : on garde la classe qui apparaît en premier (celle du voisin le plus proche parmi les ex æquo). Même chose si plusieurs points sont à la même distance : l'ordre d'apparition décide (`argsort`).

3.3 Exemple de classification, à la main

Fruits décrits par *(poids, rugosité)* ; on veut classer le fruit $x = (4,3)$.



rang	1	2	3	4	5	6	7	8
point	$(4,4)$	$(6,3)$	$(6,2)$	$(6,5)$	$(4,0)$	$(5,0)$	$(7,5)$	$(8,4)$
classe	orange	pomme	pomme	orange	orange	pomme	pomme	orange
distance à x	1	2	$\sqrt{5} \approx 2{,}24$	$\sqrt{8} \approx 2{,}83$	3	$\sqrt{10} \approx 3{,}16$	$\sqrt{13} \approx 3{,}61$	$\sqrt{17} \approx 4{,}12$

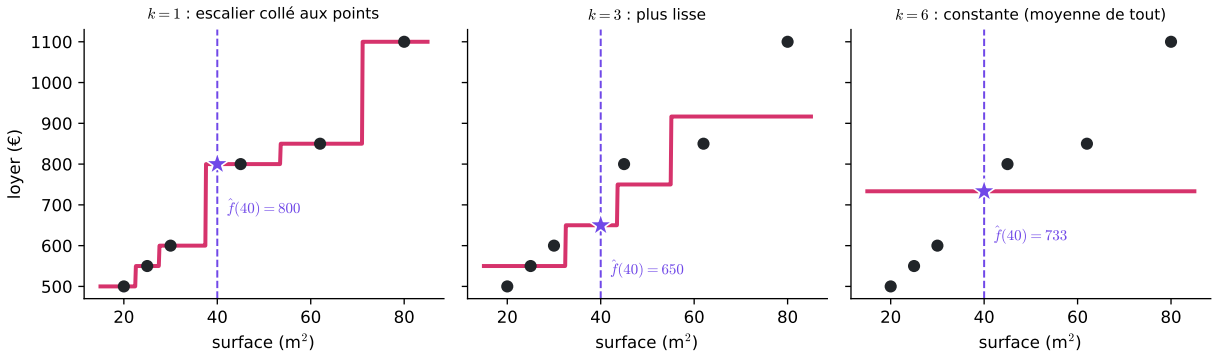
$k = 1$: **orange**. $k = 3$: orange, pomme, pomme → **pomme**. $k = 5$: orange, pomme, pomme, orange, orange → **orange**. *La prédiction change avec k !*

3.4 Exemple de régression, à la main

Loyer selon la surface ; on veut prédire le loyer d'un appartement de $x = 40$ m^2.

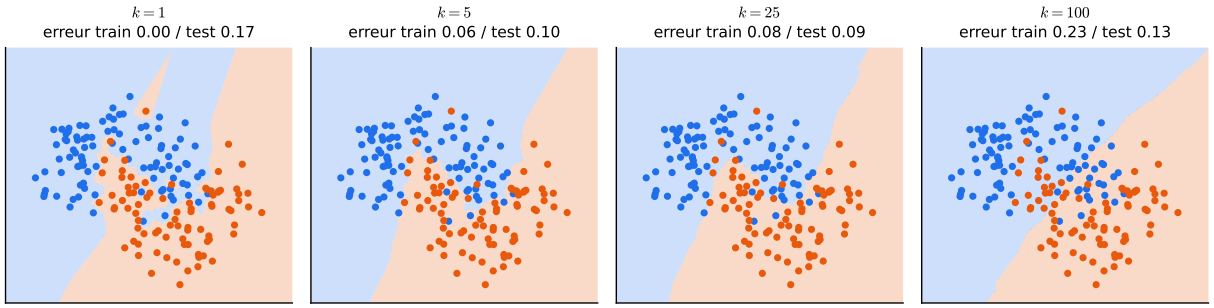
surface (m^2)	20	25	30	45	62	80
loyer (€)	500	550	600	800	850	1100
distance à 40	20	15	10	5	22	40

Classés par distance : 45 (800), 30 (600), 25 (550), 20 (500), 62 (850), 80 (1100). $\hat{f}_1(40) = 800$ $\hat{f}_2(40) = \frac{800+600}{2} = 700$ $\hat{f}_3(40) = \frac{800+600+550}{3} = 650$ $\hat{f}_6(40) = \frac{4400}{6} \approx 733$ (la moyenne de tout).



À quoi ressemble $\hat{f}_k$?

Une fonction **constante par morceaux** (un escalier). $k = 1$: l'escalier colle aux points (il suit le bruit). k plus grand : plus lisse. $k = n$: **constante** (la moyenne globale, ou la classe majoritaire globale).



Le rôle de k

k petit : **biais faible, variance forte** (sur-apprentissage ; erreur d'entraînement nulle pour $k = 1$). k grand : **biais fort, variance faible** (sous-apprentissage). On choisit k par **validation croisée**, jamais sur le test.

Piège classique

Le k-NN a besoin d'une **distance** (« proche = selon quoi ? ») et, si les variables n'ont pas la même échelle, il faut les **normaliser** (centrer-réduire). C'est le sujet de la section suivante.

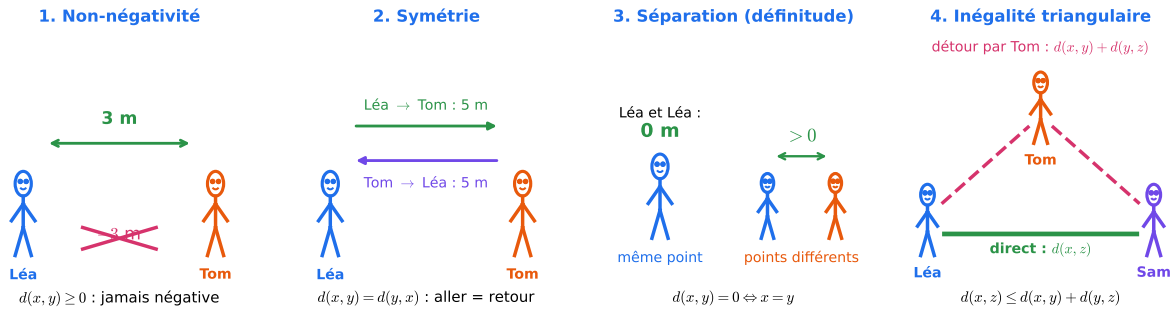
4 Les distances

4.1 Définition : les 4 règles

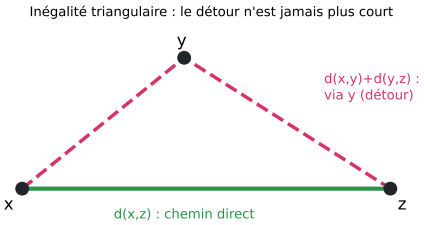
À retenir

Une **distance** sur un ensemble E est une application $d : E \times E \to \mathbb{R}_+$ telle que, pour tous $x, y, z \in E$:

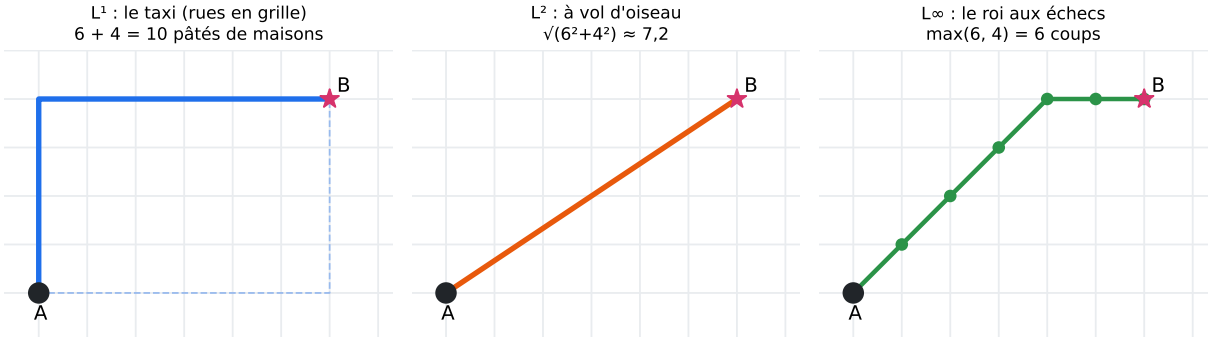
1. **non-négativité** : $d(x,y) \geq 0$;
2. **symétrie** : $d(x,y) = d(y,x)$;
3. **séparation (définitude)** : $d(x,y) = 0 \iff x = y$;
4. **inégalité triangulaire** : $d(x,z) \leq d(x,y) + d(y,z)$.



Règle	Formule	En clair
non-négativité	$d(x,y) \geq 0$	une distance mesurée au mètre ne peut pas être -3 m
symétrie	$d(x,y) = d(y,x)$	Léa→Tom = Tom→Léa
séparation	$d(x,y) = 0 \iff x = y$	distance nulle = *même* point (et un point est à 0 de lui-même)
triangle	$d(x,z) \leq d(x,y) + d(y,z)$	un détour ne raccourcit jamais

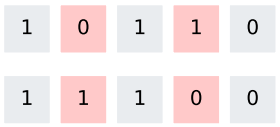


4.2 Exemples de distances



Quand	Distance	Exemple
points de $\mathbb{R}^d$	$d_1 = \sum_i \lvert x_i - y_i \rvert$ (L^1, taxi)	rues en grille
	$d_2 = \sqrt{\sum_i (x_i - y_i)^2}$ (L^2, à vol d'oiseau)	distance usuelle
	$d_\infty = \max_i \lvert x_i - y_i \rvert$ (L^∞, roi aux échecs)	plus grand écart
	$d_p = \left(\sum_i \lvert x_i - y_i \rvert^p\right)^{1/p}$, $p \geq 1$	famille L^p
	$\sqrt{(x-y)^\top \Sigma^{-1} (x-y)}$ (Mahalanobis)	tient compte des corrélations
catégories	0 si égaux, 1 sinon	pays, genre
codes, mots de même taille	Hamming : nombre de positions différentes	codes binaires
mots	Levenshtein : insertions, suppressions, substitutions	chat $\to$ chien : 3
ensembles	Jaccard : $1 - \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	paniers de courses
fonctions	d_∞, d_1, d_2 (voir 4.6)	courbes

Hamming : positions différentes



2 cases différentes : $d = 2$

(mots de même longueur, codes binaires)

Levenshtein : nb de modifications

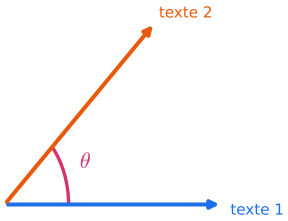
«chat» $\to$ «chien»

1. a $\to$ i (substitution)
2. t $\to$ e (substitution)
3. + n (insertion)

3 modifications : $d = 3$

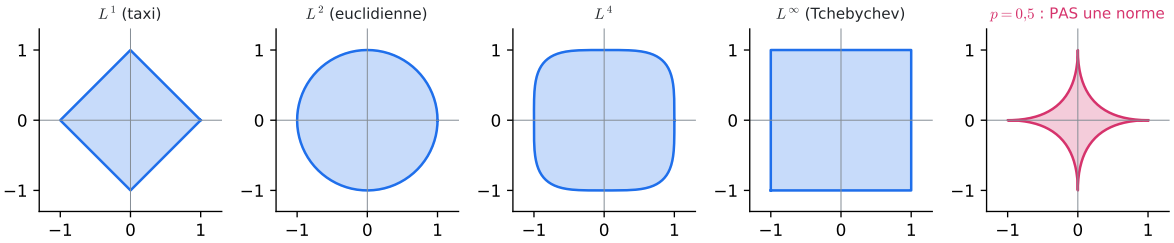
(mots de longueurs différentes)

Cosinus : l'angle entre deux directions



$d = 1 - \cos\theta$: attention, le triangle peut échouer

Boules unité $\{x : \|x\| \leq 1\}$: convexes pour une norme, non convexe pour $p < 1$



Piège classique

L^p avec $p < 1$: **pas** une norme (boule non convexe, dernière boule ci-dessus). La « distance cosinus » $1 - \cos\theta$ **n'est pas** une vraie distance (le triangle peut échouer).

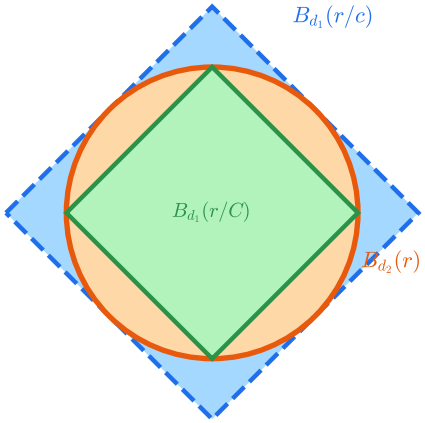
4.3 Distances équivalentes

Définition

Deux distances d_1 et d_2 sur E sont **équivalentes** s'il existe $c, C > 0$ tels que, pour tous x, y :

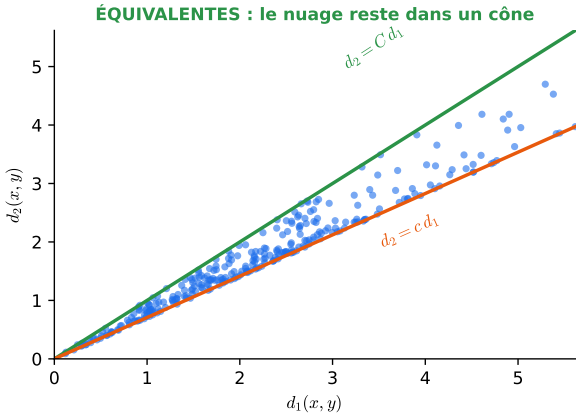
$$c\, d_1(x,y) \;\leq\; d_2(x,y) \;\leq\; C\, d_1(x,y).$$

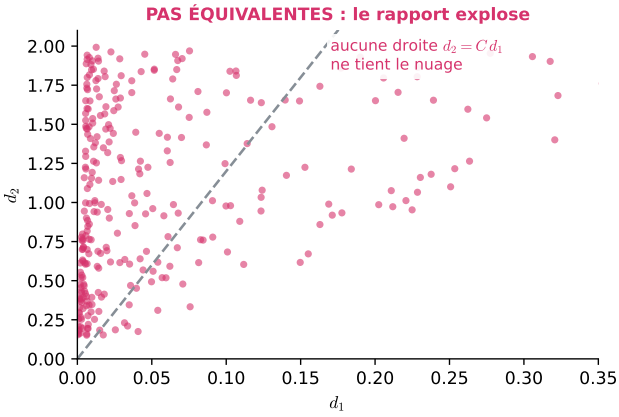
Ça veut dire : **elles se contrôlent l'une l'autre**, au facteur c ou C près. Conséquences : mêmes suites convergentes, mêmes ensembles ouverts.



une boule de d_2 est coincée entre deux boules de d_1

Ce que ça veut dire en image. Les boules de d_2 sont **coincées** entre deux boules de d_1 : si on rétrécit ou agrandit la boule de d_1 d'un facteur constant, elle englobe ou est englobée par celle de d_2. Quand d_1 est petite, d_2 aussi, et réciproquement.





Équivalentes ou pas ?

Équivalentes : le rapport d_2/d_1 reste enfermé entre deux constantes (nuage dans un cône). **Pas équivalentes** : on peut trouver des couples de points pour lesquels d_1 est minuscule alors que d_2 ne l'est pas : le rapport explose.

Montrer que deux distances ne sont PAS équivalentes

Chercher une **suite** de couples (x_n, y_n) avec $d_1(x_n, y_n) \to 0$ mais $d_2(x_n, y_n) \not\to 0$. Si $d_2 \leq C\,d_1$ était vrai, d_2 devrait aussi tendre vers 0 : contradiction.

Piège classique

Équivalentes $\neq$ mêmes voisins : deux distances équivalentes donnent les mêmes suites convergentes, mais pas forcément le même classement des plus proches voisins (le voisin le plus proche peut changer).

En dimension finie

Sur $\mathbb{R}^d$, **toutes les normes sont équivalentes**. En dimension infinie (espaces de fonctions), **non** : voir 4.6.

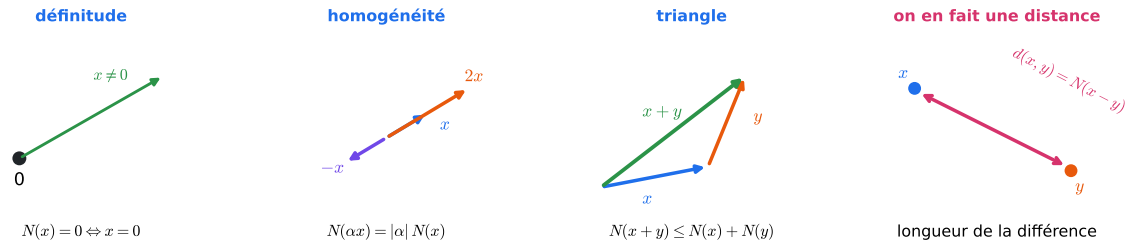
4.4 Fabriquer une distance : on part d'une norme

Norme

Une **norme** sur $\mathbb{R}^d$ est une application $N : \mathbb{R}^d \to \mathbb{R}_+$ telle que, pour tous $x, y \in \mathbb{R}^d$ et $\alpha \in \mathbb{R}$:

1. **définitude** : $N(x) = 0 \iff x = 0$;
2. **homogénéité** : $N(\alpha x) = |\alpha|\, N(x)$;
3. **inégalité triangulaire** : $N(x+y) \le N(x) + N(y)$.

Exemples : $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\|\cdot\|_p$ $(p \ge 1)$.



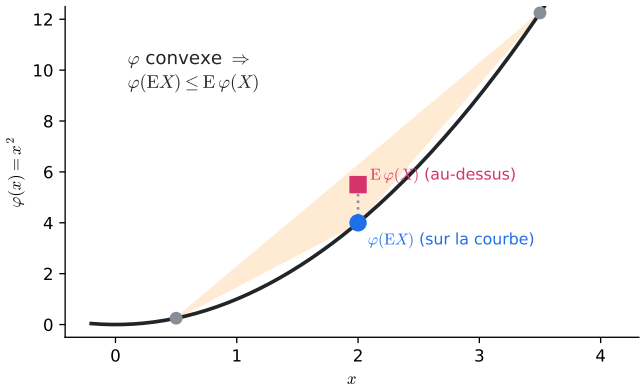
La recette

$$\boxed{d(x,y) = N(x-y)}$$

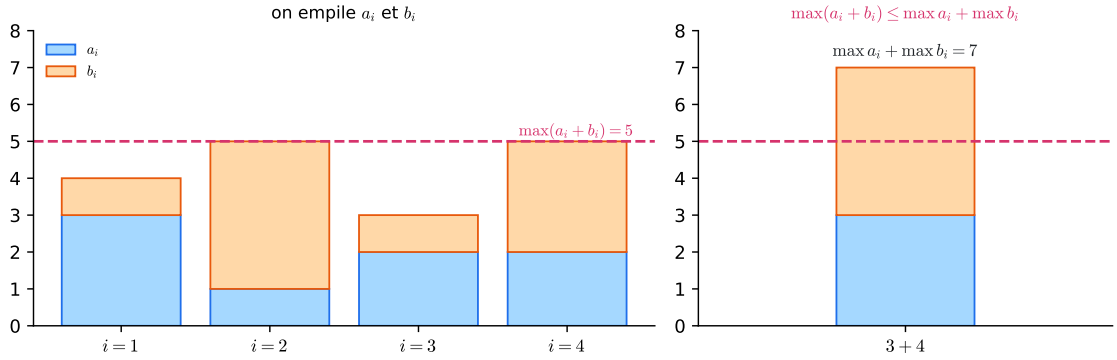
La distance entre x et y = la **longueur de leur différence**. On vérifie que les 4 règles d'une distance découlent des propriétés de N (c'est l'exercice de PC : on le fait ensemble).

4.5 Boîte à outils : les petites inégalités

Outil	Énoncé
triangle sur $\mathbb{R}$	$\lvert a+b\rvert \le \lvert a\rvert + \lvert b\rvert$
max d'une somme	$\max_i(a_i+b_i) \le \max_i a_i + \max_i b_i$ (« le max de la somme est plus petit que la somme des max »)
Jensen	φ convexe : $\varphi(\mathbb{E}X) \le \mathbb{E}\,\varphi(X)$; version finie : $\varphi\big(\sum_i \lambda_i x_i\big) \le \sum_i \lambda_i \varphi(x_i)$ pour $\lambda_i \ge 0, \sum \lambda_i = 1$
Cauchy–Schwarz	$\lvert\langle x,y\rangle\rvert \le \|x\|_2\|y\|_2$ (et $\lvert\int fg\rvert \le \sqrt{\int f^2}\sqrt{\int g^2}$)
termes positifs	chaque terme $\le$ la somme ; la somme de d termes $\le d\times$ le plus grand



Jensen en image. Pour une fonction convexe (un bol), la moyenne des valeurs est **au-dessus** de la valeur de la moyenne. Cas utiles : $(\mathbb{E}X)^2 \le \mathbb{E}[X^2]$ (c'est $\mathrm{Var}\, X \ge 0$) ; $|\mathbb{E}X| \le \mathbb{E}|X|$.



Deux équivalences pour la séparation

1. $\max_i |y_i| = 0 \iff \forall i,\ y_i = 0.$
2. Si $t_i \geq 0$ pour tout i : $\sum_i t_i = 0 \iff \forall i,\ t_i = 0.$

Elles servent à prouver la règle « $d(x,y) = 0 \iff x = y$ » pour d_∞ et pour d_1, d_2.

4.6 Distances entre fonctions

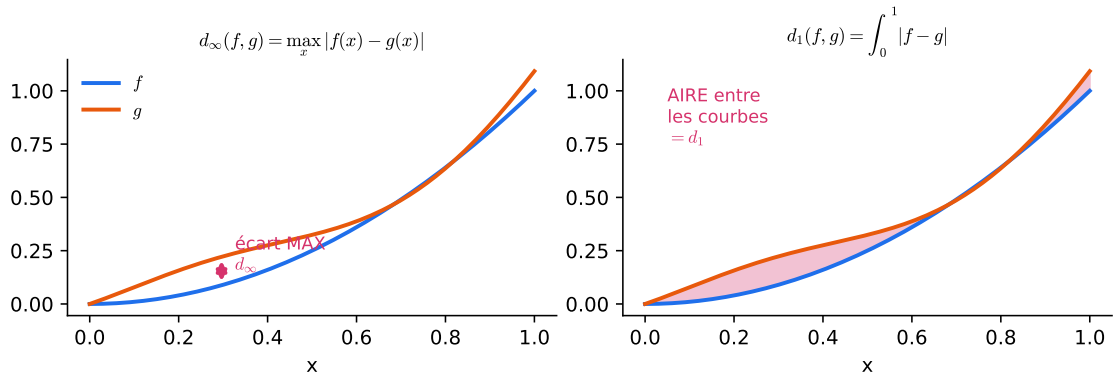
À retenir

Sur $\mathcal{C}([0,1],\mathbb{R})$ (fonctions continues), **un point est une fonction**.

$$d_\infty(f,g) = \sup_{x\in[0,1]} |f(x)-g(x)| \qquad \text{(écart maximal)},$$

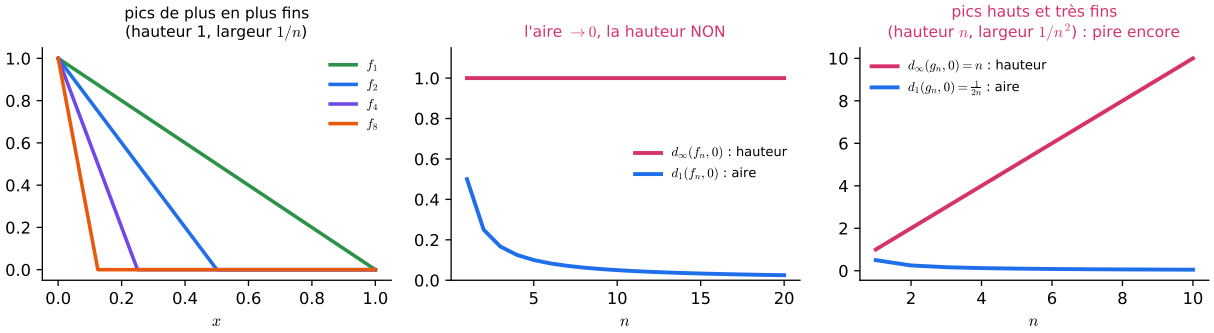
$$d_1(f,g) = \int_0^1 |f-g| \qquad \text{(aire entre les courbes)}.$$

On définit aussi $d_2(f,g) = \left(\int_0^1 (f-g)^2\right)^{1/2}$. (Sur $[0,1]$ fermé borné, une fonction continue atteint son maximum : le sup est un max.)



Propriété

Ces distances **ne sont pas toutes équivalentes**, contrairement à $\mathbb{R}^d$. Deux courbes peuvent être très proches **en aire** et très éloignées **en écart maximal**.



Suite de fonctions	**aire** $d_1(\cdot,0)$	**écart max** $d_\infty(\cdot,0)$
pic de hauteur 1, largeur $\frac{1}{n}$	$\frac{1}{2n} \to 0$	1 (ne tend pas vers 0)
pic de hauteur n, largeur $\frac{1}{n^2}$	$\frac{1}{2n} \to 0$	$n \to +\infty$

Piège classique

Ces exemples sont des *suites de pics* : « petite aire, grande hauteur ». Pour prouver que deux distances ne sont pas équivalentes, il suffit d'**une seule** suite comme celle-là.

5 Espérance et espérance conditionnelle : les rappels

5.1 L'espérance et la variance

Dans la vraie vie

L'espérance, c'est la **moyenne** qu'on obtiendrait en répétant l'expérience un très grand nombre de fois (le « centre de gravité » de la loi). La variance mesure **de combien on s'écarte de ce centre**, en moyenne quadratique.

Définitions

Pour X discrète, puis à densité p :

$$\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X=x) \quad \text{ou} \quad \mathbb{E}[X] = \int x\,p(x)\,dx, \qquad \mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

Transfert : $\mathbb{E}[g(X)] = \sum_x g(x)\,\mathbb{P}(X=x)$ ou $\int g(x)\,p(x)\,dx$ (inutile de connaître la loi de $g(X)$).

Propriété	**Énoncé**	**Condition**
linéarité	$\mathbb{E}[aX+bY+c] = a\,\mathbb{E}X + b\,\mathbb{E}Y + c$	**toujours** (sans indépendance)
constante	$\mathbb{E}[c] = c \quad \mathrm{Var}(c) = 0$	c non aléatoire
affine	$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}\,X$	décaler ne change rien, multiplier met **au carré**
somme	$\mathrm{Var}(X+Y) = \mathrm{Var}\,X + \mathrm{Var}\,Y + 2\,\mathrm{Cov}(X,Y)$	si indépendantes : Cov $= 0$
produit	$\mathbb{E}[XY] = \mathbb{E}X\,\mathbb{E}Y$	indépendance (ou Cov $= 0$)
positivité	$\mathrm{Var}\,X \geq 0$; $X \leq Y \Rightarrow \mathbb{E}X \leq \mathbb{E}Y$	toujours

Piège classique

(a) $\mathbb{E}[g(X)] \neq g(\mathbb{E}X)$ en général (Jensen : $\geq$ si g convexe). (b) $\mathbb{E}[XY] = \mathbb{E}X\,\mathbb{E}Y$ demande l'indépendance (ou Cov $= 0$) ; la réciproque est **fausse**. (c) Pour X, Y indépendantes, $\mathrm{Var}(X - Y) = \mathrm{Var}\,X + \mathrm{Var}\,Y$: **on additionne** les variances, même pour une différence. (d) $\mathrm{Var}(aX) = a^2\,\mathrm{Var}\,X$: le a sort **au carré**.

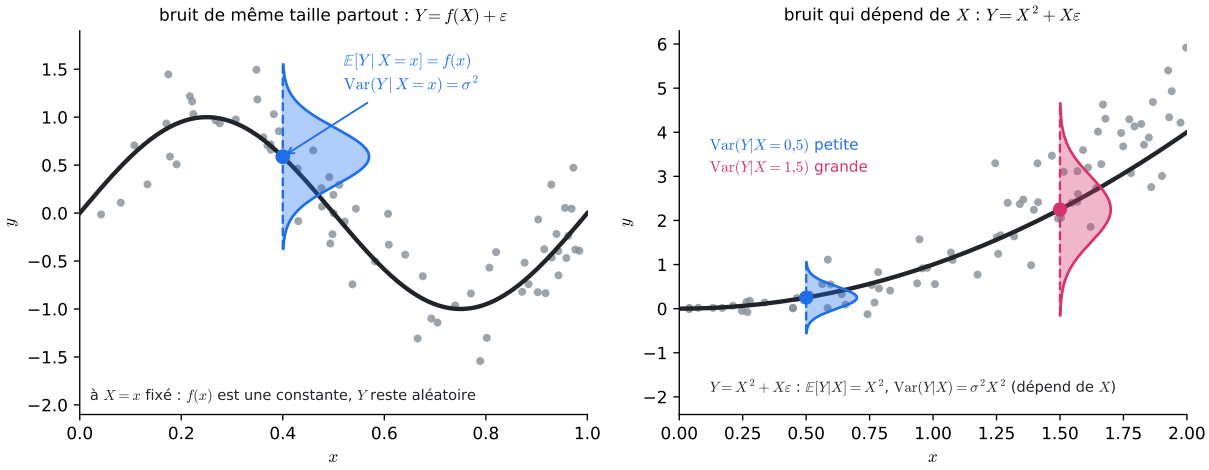
5.2 L'espérance conditionnelle

Dans une classe

Le poids Y d'un élève dépend de sa taille X. $\mathbb{E}[Y]$ est le poids moyen de **toute** la classe. $\mathbb{E}[Y \mid X = 1{,}70]$ est le poids moyen des élèves de 1,70 m **seulement** : on **connaît** X, on ne moyenne que ce qui reste aléatoire.

Définition

$g(x) = \mathbb{E}[Y \mid X = x] = \int y\, p(y \mid x)\, dy$ est une **fonction** de x. L'espérance conditionnelle est $\mathbb{E}[Y \mid X] := g(X)$: c'est une **variable aléatoire**, car c'est une fonction de X. Donc $\mathbb{E}[Y \mid X = x]$ est un **nombre**, et $\mathbb{E}[Y \mid X]$ une **variable aléatoire**.



À gauche : à $X = x$ fixé, la courbe vaut un nombre $f(x)$ (le point) et Y se disperse autour (la cloche) ; cette dispersion a la même taille partout. À droite : le bruit grandit avec x, donc $\mathrm{Var}(Y \mid X)$ ***dépend de*** X.

Pourquoi c'est la star de la régression

$\mathbb{E}[Y \mid X]$ est la **meilleure prédiction de Y à partir de X** au sens quadratique : pour toute fonction h,

$$\mathbb{E}\big[(Y - h(X))^2\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big] + \mathbb{E}\big[(\mathbb{E}[Y \mid X] - h(X))^2\big] \;\geq\; \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big].$$

La fonction à estimer, $f(x) = \mathbb{E}[Y \mid X = x]$, est la **fonction de régression** ; le k-NN l'estime par une moyenne locale.

5.3 Les règles de calcul

Espérance conditionnelle : 5 règles

Avec h une fonction (déterministe) :

1. **linéarité** : $\mathbb{E}[aY + bZ + c \mid X] = a\,\mathbb{E}[Y \mid X] + b\,\mathbb{E}[Z \mid X] + c$, y compris quand a, b, c sont de la forme $h(X)$;
2. **sortir ce qui dépend de** X : $\mathbb{E}[h(X)\,Y \mid X] = h(X)\,\mathbb{E}[Y \mid X]$, et en particulier $\mathbb{E}[h(X) \mid X] = h(X)$;
3. **indépendance** : si Y est indépendante de X, alors $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$;
4. **tour** (espérance totale) : $\mathbb{E}\big[\mathbb{E}[Y \mid X]\big] = \mathbb{E}[Y]$;
5. **emboîtement** : $\mathbb{E}\big[\mathbb{E}[Y \mid X, Z] \mid X\big] = \mathbb{E}[Y \mid X]$.

Variance conditionnelle

$$\mathrm{Var}(Y \mid X) = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2 \mid X\big] = \mathbb{E}[Y^2 \mid X] - \big(\mathbb{E}[Y \mid X]\big)^2.$$

1. $\mathrm{Var}\big(h(X)\,Y + g(X) \mid X\big) = h(X)^2\,\mathrm{Var}(Y \mid X)$: ce qui dépend de X sort **au carré**, ce qui est ajouté disparaît ;
2. $\mathrm{Var}(h(X) \mid X) = 0$ et $\mathrm{Cov}(h(X), Y \mid X) = 0$;
3. si Y est indépendante de X : $\mathrm{Var}(Y \mid X) = \mathrm{Var}(Y)$;
4. **variance totale** : $\mathrm{Var}(Y) = \mathbb{E}\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\big(\mathbb{E}[Y \mid X]\big)$.

En mots : variance totale = moyenne des variances conditionnelles + variance des moyennes conditionnelles.

5.4 Comment reconnaître ce qui dépend de X

La règle d'or

Quand on conditionne par X, **tout ce qui s'écrit** $h(X)$, avec h une fonction déterministe (connue *ou inconnue*), se comporte comme **une constante** : on le sort de l'espérance conditionnelle, et il ne compte pas dans la variance conditionnelle. Tout ce qui fait intervenir Y (ou le bruit ε) **reste aléatoire**.

Quantité	**Fonction de X seul ?**	**Sachant X**
X, X^2, $\sin X$, $\mathbf{1}\{X \geq 2\}$	oui	constante
$f(X)$, avec f *inconnue* (fonction de régression)	oui (déterministe, même inconnue)	constante : $\mathbb{E}[f(X) \mid X] = f(X)$
$\sigma^2(X)$ (bruit qui dépend de X)	oui	constante
ε, avec ε indépendant de X et $\mathbb{E}\varepsilon = 0$	non	$\mathbb{E}[\varepsilon \mid X] = 0$, $\mathrm{Var}(\varepsilon \mid X) = \sigma^2$
$Y = f(X) + \varepsilon$	non	aléatoire (à cause de ε)
voisins $N_k(x_0)$, poids $w_i(x_0)$ du k-NN	oui (ne dépendent que des X_i et de x_0)	constants (sachant $X_1, \ldots, X_n$)
$\hat{f}(x_0) = \frac{1}{k}\sum_{i \in N_k(x_0)} Y_i$	non (il contient les Y_i)	aléatoire

Méthode pour calculer $\mathbb{E}[\cdot \mid X]$ ou $\mathrm{Var}(\cdot \mid X)$

(1) **Séparer** les facteurs qui sont des fonctions de X de ceux qui contiennent Y (ou ε). (2) **Sortir** les premiers (au **carré** pour une variance). (3) **Remplacer** le reste : $\mathbb{E}[Y \mid X]$, ou $\mathbb{E}[\varepsilon \mid X] = 0$ et $\mathrm{Var}(\varepsilon \mid X) = \sigma^2$. (4) Si ε est indépendant de X, ces deux derniers ne dépendent plus de X.

Deux cas à savoir faire

Dans les deux cas ε est indépendant de X, $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}\,\varepsilon = \sigma^2$.
• $Y = f(X) + \varepsilon$: $\mathbb{E}[Y \mid X] = f(X) + \mathbb{E}\varepsilon = f(X)$ et $\mathrm{Var}(Y \mid X) = \mathrm{Var}(\varepsilon) = \sigma^2$. Au total : $\mathbb{E}[Y] = \mathbb{E}[f(X)]$ et $\mathrm{Var}(Y) = \sigma^2 + \mathrm{Var}\left(f(X)\right)$.
• $Y = X^2 + X\varepsilon$: $\mathbb{E}[Y \mid X] = X^2 + X\,\mathbb{E}[\varepsilon] = X^2$ et $\mathrm{Var}(Y \mid X) = X^2\,\mathrm{Var}(\varepsilon) = \sigma^2 X^2$. Ici la variance conditionnelle **dépend** de X ; au total $\mathrm{Var}(Y) = \sigma^2\mathbb{E}[X^2] + \mathrm{Var}(X^2)$.

Piège classique

(a) $\mathbb{E}[Y \mid X] \neq \mathbb{E}[Y]$ sauf indépendance : l'un est une fonction de X, l'autre un nombre. (b) « Constante » est **relatif au conditionnement** : $f(X)$ est constante *sachant* X, mais **aléatoire** si on ne conditionne pas ($\mathrm{Var}\,f(X) \neq 0$ en général). (c) Dans $\mathrm{Var}(h(X)\,Y \mid X) = h(X)^2\,\mathrm{Var}(Y \mid X)$, n'oublie pas le **carré**. (d) On ne sort que les **fonctions de** X, jamais Y ni $\hat{f}(x_0)$.

6 Biais et variance d'un estimateur

Dans la vraie vie

Une balance de salle de bains. **Biais** = elle marque toujours 2 kg de trop (erreur *systématique*). **Variance** = elle donne 68, puis 73, puis 65 pour la même personne (elle *tremble*). Une bonne balance n'a ni l'un ni l'autre.

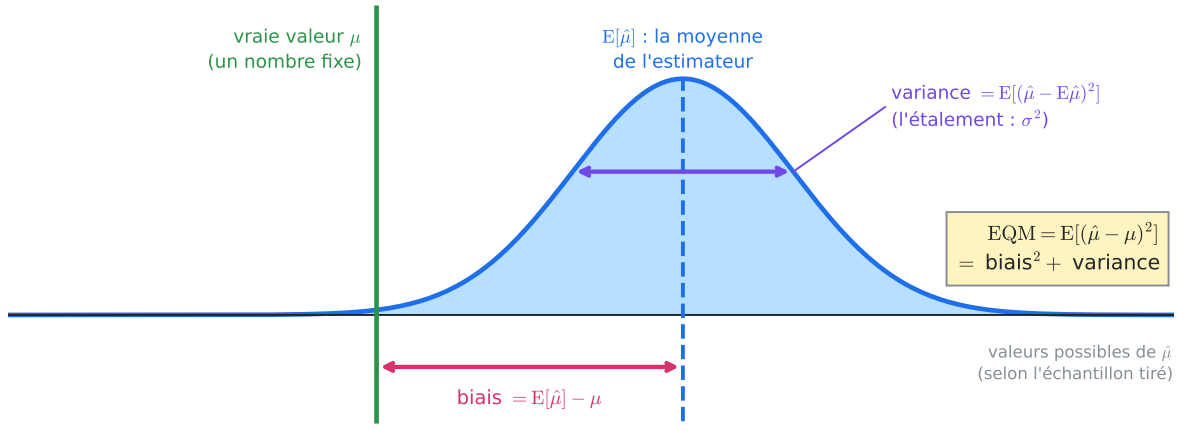
Définitions

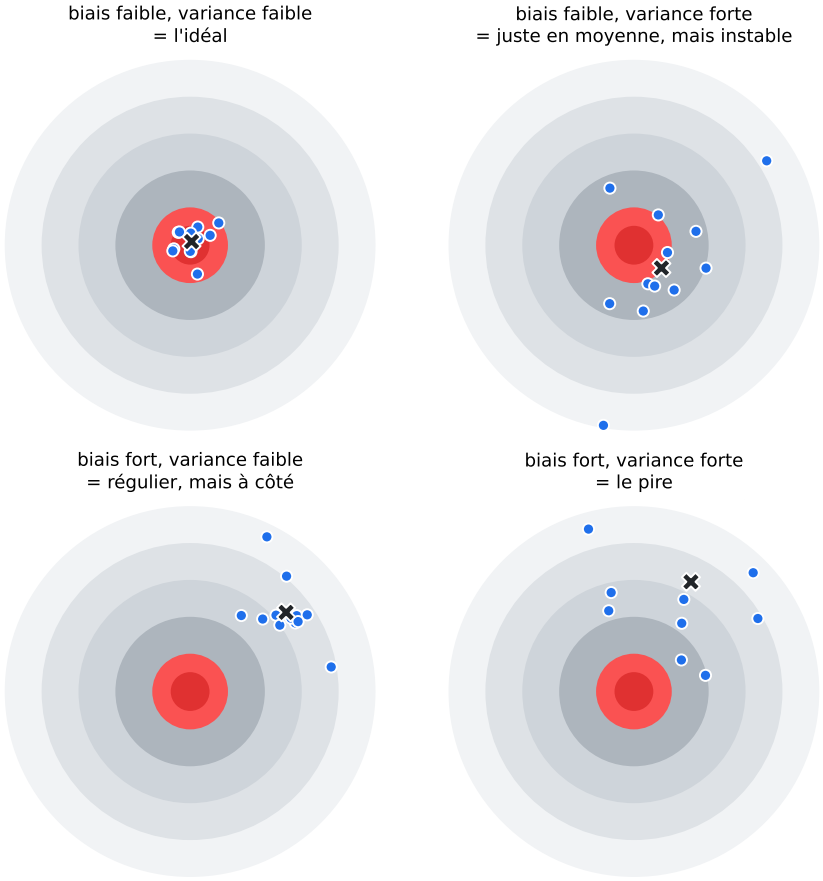
μ est un **nombre fixe** (la vraie valeur) ; $\hat{\mu}$ est un **estimateur** : il dépend des données tirées au hasard, donc il est **aléatoire**.

$$\mathrm{biais}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu, \qquad \mathrm{Var}(\hat{\mu}) = \mathbb{E}\left[(\hat{\mu} - \mathbb{E}\hat{\mu})^2\right],$$

$$\boxed{\mathrm{EQM}(\hat{\mu}) = \mathbb{E}\left[(\hat{\mu} - \mu)^2\right] = \mathrm{biais}^2 + \mathrm{Var}(\hat{\mu})} \qquad \text{(erreur quadratique moyenne).}$$

Sans biais : $\mathbb{E}\hat{\mu} = \mu$ (juste en moyenne).





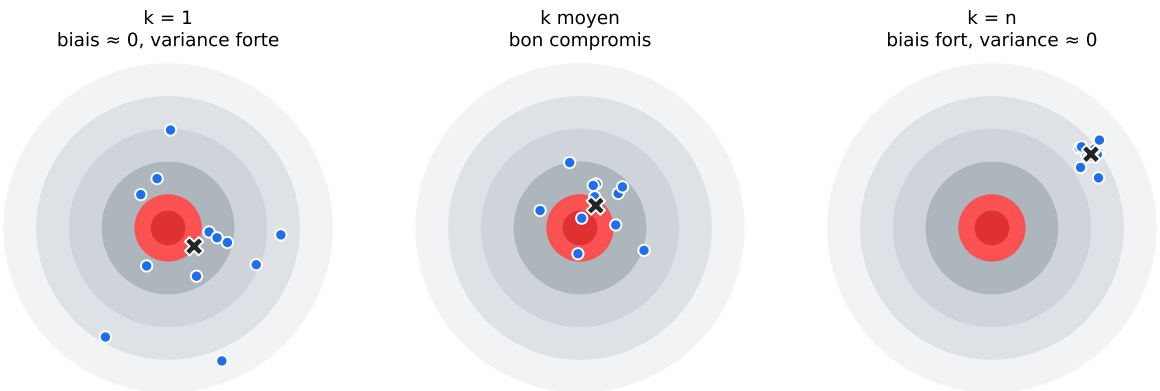
centre = la vraie valeur ; fléchettes = les estimations sur 12 échantillons différents ; × = leur moyenne

Deux estimateurs de la moyenne μ d'une loi

Données $Z_1, \dots, Z_n$ i.i.d., $\mathbb{E}Z_i = \mu$, $\operatorname{Var} Z_i = \sigma^2$.
- $\hat{\mu} = \bar{Z} = \frac{1}{n}\sum Z_i$: biais 0, variance σ^2/n.
- $\hat{\mu} = 0$ (on répond toujours 0) : variance 0, biais $-\mu$.

Le lien avec ce qu'on vient de voir

k-NN : k petit $\Rightarrow$ variance forte ; k grand $\Rightarrow$ biais fort. Modèle paramétrique trop simple $\Rightarrow$ biais fort ; méthode non paramétrique très souple $\Rightarrow$ variance forte.



kNN : plus k est grand, plus on est régulier... mais décalé

6.1 Application au k-NN : conditionnel ou inconditionnel ?

Dans la vraie vie

Deux façons de répéter l'expérience. **Design fixé** : mêmes appartements visités (mêmes x_i), on retire seulement les loyers (le bruit). **Design aléatoire** : on retire aussi les x_i. Le second ajoute de l'instabilité, car d'un tirage à l'autre **ce ne sont plus les mêmes voisins**.

Le modèle

$Y_i = f(X_i) + \varepsilon_i$, avec ε_i i.i.d., indépendants des X_j, $\mathbb{E}\varepsilon_i = 0$, $\operatorname{Var}\varepsilon_i = \sigma^2$. En x_0 : $\hat{f}(x_0) = \frac{1}{k}\sum_{i\in N_k(x_0)} Y_i$. On note

$$m(X) = \mathbb{E}\big[\hat{f}(x_0) \mid X_1, \dots, X_n\big], \qquad v(X) = \operatorname{Var}\big(\hat{f}(x_0) \mid X_1, \dots, X_n\big).$$

Ce sont des fonctions de $X = (X_1, \dots, X_n)$, donc des variables aléatoires.

Ce qui dépend de X, ici

$N_k(x_0)$ est choisi avec les distances $d(x_0, X_i)$: il ne dépend que des X_i (**jamais** des Y_i), avec une règle d'égalité qui ne regarde que les X_i. Sachant les X_i, c'est donc un **ensemble fixe**, et les poids $w_i = \frac{1}{k}\mathbf{1}\{i \in N_k(x_0)\}$ aussi. Il ne reste d'aléatoire que les Y_i : $\hat{f}(x_0)$ est une combinaison **à poids constants** des Y_i.

Le calcul sachant X

Espérance : $\mathbb{E}[Y_i \mid X] = f(X_i)$ (règle 2 : $f(X_i)$ sort ; et $\mathbb{E}\varepsilon_i = 0$), donc

$$m(X) = \frac{1}{k}\sum_{i\in N_k(x_0)} f(X_i), \qquad \text{biais conditionnel} = f(x_0) - m(X).$$

Variance : les Y_i sont indépendants sachant X, de variance σ^2, et les poids sont constants :

$$v(X) = \frac{1}{k^2}\sum_{i\in N_k(x_0)} \operatorname{Var}(Y_i \mid X) = \frac{k\,\sigma^2}{k^2} = \frac{\sigma^2}{k}.$$

Le biais conditionnel mesure l'écart entre $f(x_0)$ et la moyenne de f sur les voisins ; la variance conditionnelle, elle, **ne dépend plus de** X.

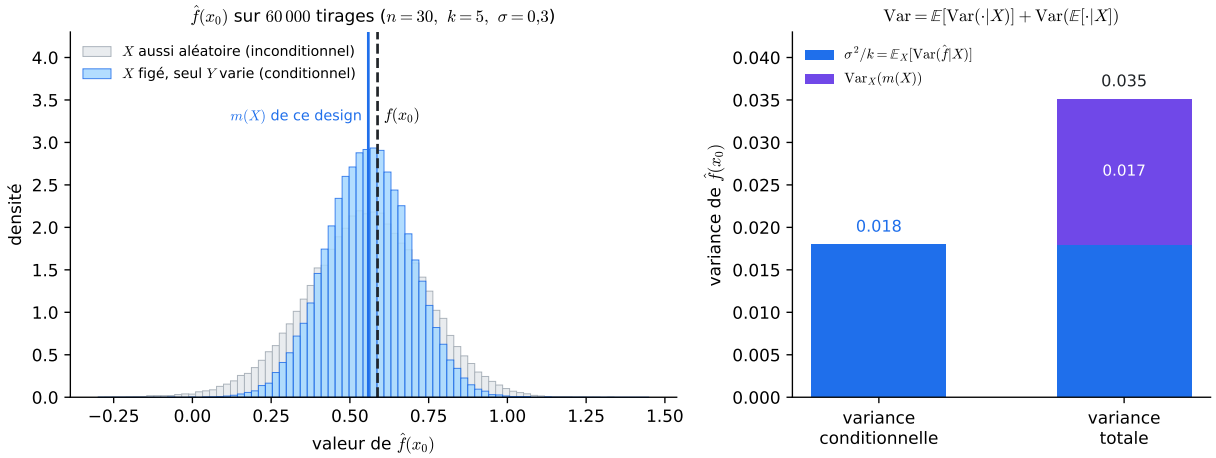
Du conditionnel à l'inconditionnel

Avec la règle du tour et la variance totale, appliquées à $Z = \hat{f}(x_0)$:

$$\operatorname{Biais}^2(x_0) = \Big(f(x_0) - \mathbb{E}_X\big[\mathbb{E}_{Y|X}[\hat{f}(x_0)]\big]\Big)^2 = \Big(f(x_0) - \mathbb{E}_X\big[m(X)\big]\Big)^2,$$

$$\operatorname{Var}\big(\hat{f}(x_0)\big) = \mathbb{E}_X\big[\operatorname{Var}_{Y|X}(\hat{f}(x_0))\big] + \operatorname{Var}_X\big(\mathbb{E}_{Y|X}[\hat{f}(x_0)]\big) = \underbrace{\mathbb{E}_X[v(X)]}_{=\sigma^2/k} + \underbrace{\operatorname{Var}_X\big(m(X)\big)}_{\text{les voisins changent avec } X}.$$

Preuve de la variance. Sachant X, $\mathbb{E}[Z^2 \mid X] = v(X) + m(X)^2$. En prenant $\mathbb{E}_X$: $\mathbb{E}[Z^2] = \mathbb{E}[v] + \mathbb{E}[m^2]$, et $\mathbb{E}[Z] = \mathbb{E}[m]$. Donc $\operatorname{Var} Z = \mathbb{E}[v] + \mathbb{E}[m^2] - (\mathbb{E}m)^2 = \mathbb{E}[v] + \operatorname{Var}(m)$.
En mots : variance totale = variance conditionnelle moyenne + variance de la moyenne conditionnelle.



Simulation ($f(x) = \sin 2\pi x$, $x_0 = 0{,}4$, $n = 30$, $k = 5$, $\sigma = 0{,}3$). Design fixé : la dispersion vaut $\sigma^2/k = 0{,}018$. Design aléatoire : s'ajoute $\mathrm{Var}_X(m) \approx 0{,}017$, d'où $0{,}035$ au total, ce que donne bien la simulation.

Moyenne des biais conditionnels $\neq$ biais inconditionnel

$\mathbb{E}_X\left[(f(x_0) - m(X))^2\right] = \mathrm{Biais}^2(x_0) + \mathrm{Var}_X\left(m(X)\right)$: si on moyenne les biais conditionnels **au carré**, on y glisse le terme de variance dû au design. Au total l'erreur quadratique est la même dans les deux écritures :

$$\mathbb{E}\left[(\hat{f}(x_0) - f(x_0))^2\right] = \mathrm{Biais}^2(x_0) + \mathbb{E}_X[v(X)] + \mathrm{Var}_X\left(m(X)\right).$$

7 Autotest : 16 questions éclair

Question	**Réponse en une ligne**
Les 3 familles d'apprentissage ?	supervisé (x, y), non supervisé x seul, renforcement (récompense)
Paramétrique : définition ?	famille f_θ choisie à l'avance, p fixe, on optimise $\hat{\theta}$
Non paramétrique : définition ?	pas de forme ni de θ fixes ; $\hat{f}$ construit directement des données, complexité qui grandit avec n
« Non paramétrique » = sans paramètre ?	non : il reste des hyperparamètres (k, h, nombre de cases, J)
Trois exemples non paramétriques ?	k-NN, histogramme, estimateur à noyau (ou projection)
Forme du k-NN, régression / classif ?	moyenne $\frac{1}{k}\sum_{N_k(x)} y_i$ / vote majoritaire des k voisins
Que vaut $\hat{f}_n$ $(k = n)$?	une constante : la moyenne (ou la classe majoritaire) de tout le jeu
Les 4 règles d'une distance ?	non-négativité, symétrie, séparation ($d = 0 \iff x = y$), triangle
Distances équivalentes ?	$c\, d_1 \leq d_2 \leq C\, d_1$; mêmes suites convergentes
Comment fabriquer une distance ?	$d(x, y) = N(x - y)$ avec N une norme
$\max(a_i + b_i)$ contre $\max a_i + \max b_i$?	$\max(a_i + b_i) \leq \max a_i + \max b_i$
Biais, variance, EQM de $\hat{\mu}$?	$\mathbb{E}\hat{\mu} - \mu$; $\mathbb{E}[(\hat{\mu} - \mathbb{E}\hat{\mu})^2]$; EQM $=$ biais^2 + variance
$\mathbb{E}[Y \mid X]$ contre $\mathbb{E}[Y]$?	une fonction de X (variable aléatoire) contre un nombre ; $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$
Un facteur $h(X)$ dans $\mathbb{E}[h(X)Y \mid X]$ et dans $\mathrm{Var}(h(X)Y \mid X)$?	il sort tel quel dans l'espérance, **au carré** dans la variance : $h(X)\mathbb{E}[Y \mid X]$ et $h(X)^2 \mathrm{Var}(Y \mid X)$
Variance totale ?	$\mathrm{Var}\, Y = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$
k-NN : $\mathbb{E}[\hat{f}(x_0) \mid X]$ et $\mathrm{Var}(\hat{f}(x_0) \mid X)$?	$\frac{1}{k}\sum_{N_k(x_0)} f(X_i)$ et σ^2/k (les voisins ne dépendent que des X_i : constants sachant X)