Mathématiques pour les économistes

Table des matières

Rappels de probabilité

En sciences humaines et en économie, une question parait centrale : Avec quel degré de certitude et quel risque d'erreur peut-on tirer une conclusion d'un jeu de donnée "imparfait" et fini ? Cette question vaste se décline en nombreuses versions : quand on ne sonde qu'une petite fraction de la population sur un sujet, quelle est la chance (ou probabilité) que l'opinion de cet *échantillon* représente de manière satisfaisante l'opinion générale ? Comment cette probabilité évolue-t-elle avec la taille de l'échantillon ? Quand on observe une *corrélation* entre deux observables, quelle est la probabilité que cette corrélation soit uniquement le fruit du hasard ? En somme que peut-on dire à partir d'un jeu de données réel ?

Nous donnons dans ce cours des pistes de réponse à cette question (aussi abordée, un peu différemment, dans le cours d'économétrie), au travers d'outils statistiques comme les tests et les estimateurs. Ces outils nécessitent de comprendre et d'utiliser une formalisation mathématique de l'aléatoire : la théorie des probabilités. Ce chapitre introduit quelques notions de probabilité indispensables pour la suite.

1 Notions de base

On appelle expérience aléatoire une expérience dont l'issue n'est pas prévisible car, répétée dans des conditions identiques, elle peut donner lieu à des résultats différents (appelés *réalisations*). On donne trois exemples canoniques : le lancer d'une pièce, le lancer d'un dé et le temps de vie d'une ampoule. On appelle *univers* et on note Ω l'ensemble des résultats possible de l'expérience. Dans le cas du lancer de pièce $\Omega = \{Pile, Face\}$, pour le lancer de dé $\Omega = \{1, 2, 3, 4, 5, 6\}$ alors que pour l'ampoule $\Omega = \mathbb{R}_+$. On appelle *événement* n'importe quel sous-ensemble de l'univers, par exemple l'événement "l'ampoule grille en moins de deux heures" ($A = [0, 2] \subset \mathbb{R}_+$), ou bien "Le dé tombe sur une face paire" ($A = \{2, 4, 6\}$).

1.1 Événements et probabilités

On associe [1] à chaque événement A une probabilité notée $\mathbb{P}(A)$, qui est la chance de réalisation de cet événement. Par exemple, pour une pièce équilibrée, $\mathbb{P}(Face) = 1/2$. $\mathbb{P}$ est donc une fonction, qui va de l'ensemble des événements possibles dans $[0, 1]$, et qui vérifie les deux propriétés suivantes :

— $\mathbb{P}(\Omega) = 1$

— Si A et B sont deux événements disjoints (c'est-à-dire que $A \cap B = 0$), alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

On peut déduire à partir de ces deux propriétés fondamentales d'autres propriétés utiles ; par exemple $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, ou bien $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Probabilité conditionnelle Pour un événement B de probabilité non-nulle, on appelle *probabilité de A sachant B* et on note $\mathbb{P}(A|B)$ la fraction $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. On mentionne au passage la formule de Bayes qui relie $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$: $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$

Indépendance On dit que deux événements sont indépendants quand le fait de savoir si le premier événement a lieu n'apporte pas d'information sur le second. Autrement dit, A et B sont indépendants si et seulement si $\mathbb{P}(A|B) = \mathbb{P}(A)$, soit $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

1. On ne s'étendra pas sur les détails techniques qui mènent à cette association.

1.2 Variable aléatoire

On appelle *variable aléatoire* (abrégé v.a.) une fonction qui à chaque résultat d'une épreuve aléatoire associe un nombre. Par exemple on peut définir, à la suite d'un lancer de pièce, une variable aléatoire X qui vaut 0 quand la pièce tombe sur pile et 1 quand la pièce tombe sur face. On a alors $\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = 1/2$. On peut aussi définir la variable aléatoire Y à la suite d'un lancer de dé, qui vaut le carré du nombre inscrit sur la face obtenue, donc $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 4) = ... = \mathbb{P}(Y = 36) = 1/6$. On peut enfin définir la variable aléatoire Z qui est égale au temps écoulé avant que l'ampoule ne grille. Contrairement aux deux premiers exemples, la variable aléatoire peut prendre une infinité de valeurs (n'importe quel réel positif)

On distingue deux sortes d'infinis. Les premiers ressemblent à $\mathbb{N}$ et sont dit dénombrables : on peut les compter, les numéroter, même s'il y en a une infinité. Par exemple $\mathbb{Z}$ est un infini dénombrable : on peut donner un ordre à ses éléments qui me permettra de tous les compter sans en oublier : $0, 1, -1, 2, -2, ...$ En revanche $\mathbb{R}$ n'est pas un infini dénombrable : on ne peut pas numéroter tous ses éléments (on peut démontrer que c'est impossible : c'est Cantor, à la fin du 19^{eme} siècle qui a trouvé la première preuve de ce résultat très important).

1.3 Variable aléatoire discrète

On dit que les variables aléatoires sont discrètes si elles peuvent prendre un nombre fini ou infini dénombrable de valeurs différentes. Ce sont les variables aléatoires les plus simples à étudier en général. On commence par donner trois exemples importants

- X suit une loi de Bernoulli de paramètre p si $\mathbb{P}(X = 1) = p$ et $\mathbb{P}(X = 0) = 1 - p$
- X suit une loi binomiale de paramètres p, n si pour $i \in \{0, 1, 2, ..., n\}$, $\mathbb{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$
- X suit une loi de Poisson de paramètre θ si, pour tout $k \in \mathbb{N}$, $\mathbb{P}(X = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}$

On voit que pour caractériser complètement une variable aléatoire discrète, il suffit de donner la probabilité de chacune des valeurs dénombrables qu'elle peut prendre. Si on note x_i ces différentes valeurs, et $p_i = \mathbb{P}(X = x_i)$, on a, d'après les deux propriétés fondamentales des probabilité, $\sum_{i\geq 0} p_i = 1$, cette somme portant éventuellement sur un nombre infini de termes. On peut définir facilement *l'espérance* de cette variable aléatoire :

$$\mathbb{E}(X) = \sum_{i\geq 0} p_i x_i$$

Cette somme porte éventuellement sur un nombre infini de termes également, et elle peut valoir l'infini ou même ne pas être calculable. L'espérance est une quantité centrale en probabilité. On peut maintenant noter deux grandes propriétés de l'espérance :

- L'espérance est *monotone*, c'est à dire que si $X \geq 0$ (c'est-à-dire que X ne peut prendre que des valeurs positives), alors $\mathbb{E}(X) \geq 0$
- L'espérance est *linéaire*, c'est-à-dire que si μ et λ sont deux réels quelconques, $\mathbb{E}(\mu X + \lambda Y) = \mu\mathbb{E}(X) + \lambda\mathbb{E}(Y)$

Si l'espérance existe, on peut définir la variance (qui peut elle aussi être impossible à calculer) comme

$$\mathrm{Var}(X) = \sum_{i\geq 0} p_i (x_i - \mathbb{E}(X))^2 = \sum_{i\geq 0} p_i x_i^2 - \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

($*$) Exercice : Prouver que les deux quantités qui définissent la variance sont égales.

Remarque : la variance, quand elle existe, est toujours positive, et elle n'est pas linéaire. On parle souvent d'écart-type d'une variable : il s'agit de la racine carrée de la variance. Pour n'importe quelle fonction f définie sur $\mathbb{R}$, $f(X)$ est une v.a. et on peut calculer $\mathbb{E}(f(X)) = \sum p_i f(x_i)$, c'est le **THEOREME DE TRANSFERT**

Malheureusement, on ne peut pas définir aussi simplement ces quantités lorsque les variables aléatoires ne sont pas discrètes. En reprenant l'exemple du temps de vie d'une ampoule, on sent bien que parler de la probabilité que Z vaille 1.7105 n'aurait pas vraiment de sens : la probabilité de tomber exactement sur ce réel, plutôt que sur 1.71051 ou 1.710501 est nulle. On est obligé de changer d'approche.

1.4 Fonction de répartition et variables aléatoires continues

Il existe un outil qui permet de manipuler et de caractériser tous les types de variables aléatoires : la fonction de répartition. On la définit comme suit : pour tout $a \in \mathbb{R}$, $F(a) = \mathbb{P}(X \leq a)$.

Dans le cas d'une v.a. discrète, F est une fonction en escalier ayant des points de discontinuité aux différents x_i. Dans tous les cas elle vérifie les conditions suivantes :

— $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$

— F est continue à droite.

— F est croissante

Réciproquement, toute fonction qui vérifie ces conditions est la fonction de répartition d'une certaine v.a. On va s'intéresser à l'étude d'un *type particulier* de v.a. très important, qui est en un sens l'opposé des v.a. discrète.

Définition (Variables aléatoires continues)**.** *On dit qu'une variable aléatoire X est continue si sa fonction de répartition F est continue et qu'il existe une fonction $f_X \geq 0$ telle que pour tout $a \in \mathbb{R}$*

$$F(a) = \int_{-\infty}^{a} f_X(t)dt$$

On appelle alors la fonction f_X la densité *de X*

La densité en a mesure la probabilité que X soit "aux alentours" de a. En fait on peut écrire que pour ϵ suffisamment petit, $\mathbb{P}(X \in [a, a+\epsilon]) \simeq f_X(a) \times \epsilon$. On a forcément $\int_{-\infty}^{\infty} f_X(t)dt = 1$.

On peut mentionner quelques densités classiques, comme la densité gaussienne (ou normale) de paramètres μ, σ^2

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

où la densité exponentielle de paramètre λ : $f_X(x) = \lambda \exp(-\lambda x)$.

Dans ce cas, on a $\mathbb{P}(X \in A) = \int_A f_x(t)dt$. Les variables aléatoires continues sont pratiques car on peut calculer facilement leur espérance. En effet, pour n'importe quelle fonction g :

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(t) f_X(t)dt$$

(version continue du théorème de transfert). Et en particulier $\mathbb{E}(X) = \int_{\mathbb{R}} t f_X(t)dt$ (intégrale qui peut éventuellement être divergente)

($*$) Exercice : Calculez la moyenne et la variance d'une loi normale de paramètre (μ, σ) et les mêmes quantités pour une loi exponentielle de paramètre λ.

Une dernière remarque important : il existe des lois qui ne sont ni discrètes, ni continues : si ces deux cas sont les plus importants (et les seuls sur lesquels nous allons travailler), ils ne sont pas les seuls possibles.

2 Vecteurs aléatoires

Les variables aléatoires de la partie précédente sont des fonctions à valeur dans $\mathbb{R}$: c'est-à-dire qu'elles peuvent prendre différentes valeurs dans $\mathbb{R}$. On s'intéresse maintenant à des vecteurs aléatoire, qui peuvent prendre des valeurs dans $\mathbb{R}^d$

2.1 Le cas discret

Le cas discret est encore une fois le plus simple : Soit Z un vecteur aléatoire discret, (c'est à dire qui ne peut prendre comme valeur qu'un nombre dénombrable de vecteurs v_i possibles). Pour simplifier les choses, on se place en dimension deux et on écrit $Z = (X, Y)$. On note x_i (resp. y_i) les différentes valeurs que peut prendre la v.a. X (resp. Y). Il y en a potentiellement un nombre infini, mais dénombrable. La probabilité jointe de Z s'écrit alors :

$$p_z(x_i, y_j) = \mathbb{P}(Z = (x_i, y_j)) = \mathbb{P}(X = x_i \cap Y = y_j)$$

Un exemple : On prend un exemple qu'on utilisera pour illustrer ce chapitre.

$$(X,Y) = \begin{cases} (0,0) \text{ avec probabilité } 0,2 \\ (0,1) \text{ a.p. } 0,3 \\ (1,0) \text{ a.p. } 0,1 \\ (1,1) \text{ a.p. } 0,4 \end{cases}$$

qu'on peut aussi représenter sous forme de tableau

X/Y	0	1
0	0,2	0,3
1	0,1	0,4

Un autre exemple : On peut par exemple prendre, pour i et j deux entiers naturels ≥ 1,

$$p_z(i,j) = 1/2^{i+j}$$

Dans ce cas, le vecteur aléatoire prend un nombre infini dénombrable de valeurs.

On se pose la question suivante : quelle est la loi suivie par la première coordonnée de Z ? On a $p_X(x_i) = \mathbb{P}(X = x_i) = \sum_j \mathbb{P}(X = x_i \cap Y = y_j) = \sum_j p_z(x_i, y_j)$. On appelle cette loi **la loi marginale** de X (parce qu'on peut l'écrire dans la marge). On définit de même la loi marginale de Y : $p_Y(y_j) = \sum_i p_z(x_i, y_j)$. Dans notre premier exemple, on peut calculer la loi de X : $\mathbb{P}(X = 0) = 0,2 + 0,3 = 1/2$, donc X suit une loi de Bernouilli de paramètre $1/2$. On peut également calculer $\mathbb{P}(Y = 0) = 0,2 + 0,1 = 0,3$.

En général, pour n'importe quelle fonction de deux variables $g(x, y)$, on définit

$$\mathbb{E}(g(X,Y)) = \sum_{i,j} g(x_i, y_j) p_Z(x_i, y_j)$$

Toujours en prenant le premier exemple, on a $\mathbb{E}((X+Y)^2) = 1 \times (0,3 + 0,1) + 4 \times 0,4 = 2$. On a $\mathbb{E}(XY) = \sum_{i,j} x_i y_j p_Z(x_i, y_j)$. On définit la covariance entre X et Y comme $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, et la corrélation comme $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$. Dans notre exemple, $\mathbb{E}(XY) = 0,4$, $\mathbb{E}(X) = 0,5$, $\mathbb{E}(Y) = 0,7$, donc $Cov(X,Y) = 0,05$

Loi conditionnelle On peut définir la loi conditionnelle de Y sachant X comme suit :

$$p_{Y|X=x_i}(y_j) = \mathbb{P}(Y = y_j | X = x_i) = \frac{\mathbb{P}(Y = y_j \cap X = x_i)}{\mathbb{P}(X = x_i)} = \frac{p_Z(x_i, y_j)}{p_X(x_i)}$$

$p_{Y|X=x_i}(y_j)$ est la probabilité que Y vaille y_j sachant que X vaut x_i : on regarde ce que la connaissance de X apporte comme connaissance sur Y. On remarque que, naturellement, $\sum_{y_i} p_{Y|X=x_i}(y_j) = 1$. Voyons par exemple la loi de Y sachant $X = 1$ dans notre premier exemple. $\mathbb{P}(Y = 0 | X = 1) = \frac{\mathbb{P}(X=1 \cap Y=0)}{\mathbb{P}(X=1)} = \frac{0,1}{0,1+0,4} = 0,2$

On peut de même définir l'espérance de Y sachant que $X = x_i$ comme $\mathbb{E}(Y|X = x_j) = \sum_j p_{Y|X=x_i}(y_j) y_j$. Dans l'exemple 1, $E(Y|X = 1) = 0,8$ et $E(Y|X = 0) = 0,6$

Définition. *On dit que deux variables aléatoires discrètes X et Y sont indépendantes quand les événements $(X = x_i)$ et $(Y = y_j)$ sont indépendants quel que soit la valeur de i et de j, c'est-à-dire si, pour tout i, j*

$$p_z(x_i, y_j) = \mathbb{P}(X = x_i \cap Y = y_j) = p_X(x_i) p_Y(y_j)$$

On remarque que si X et Y sont indépendantes, $p_{Y|X=x_i}(y_j) = p_Y(y_j)$: la connaissance de X ne change rien, elle n'apporte pas d'informations supplémentaires.

$(*)$ Exercice : Montrer que si on définit le vecteur $Z = (X, Y)$ par la loi jointe suivante $p_Z(0,0) = p_Z(1,1) = 1/2$, alors X et Y ne sont pas indépendantes.

$(*)$ Exercice : Prouvez que deux variables indépendantes ont une covariance nulle. Attention la réciproque est fausse !

2.2 Le cas continu

On traite le cas continu par analogie avec le cas discret. On note $Z = (X, Y)$ notre vecteur aléatoire, et pour $x, y \in \mathbb{R}$ on note $f_Z(x, y)$ la *densité jointe*. On a $\mathbb{P}((X \in [x, x + \epsilon]) \cap (Y \in [y, y + \eta])) \simeq f_Z(x, y)\epsilon\eta$ si ϵ et η sont suffisamment petits.

Un exemple : On se donne pour illustrer ce chapitre $f_Z(x, y) = (x + y)\mathbf{1}_{x\in[0,1]}\mathbf{1}_{y\in[0,1]}$

On peut définir la loi marginale de X comme

$$f_X(x) = \int_{y\in\mathbb{R}} f_Z(x, y)$$

Pour notre exemple, on peut ainsi calculer $f_X(x) = \int_{y\in[0,1]}(x + y)\mathbf{1}_{x\in[0,1]} = (1/2 + x)\mathbf{1}_{x\in[0,1]}$

Pour n'importe quelle fonction de deux variables $g(x, y)$, on peut toujours définir l'espérance de la variable aléatoire $g(X, Y)$, cette fois-ci par la formule

$$\mathbb{E}(g(X, Y)) = \int_{x,y} g(x, y)f_Z(x, y)dxdy$$

On peut définir comme précédemment la covariance $Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ et la corrélation. Pour notre exemple $\mathbb{E}(XY) = 1/3$, et $\mathbb{E}(X) = \mathbb{E}(Y) = 7/12$, donc $Cov(X, Y) = -1/144$. On définit également la *densité conditionnelle* :

$$f_{Y|X=x}(y) = \frac{f_Z(x, y)}{f_X(x)}$$

qui nous permet de définir l'espérance de Y sachant $X = x_i$ comme $\mathbb{E}(Y|X = x) = \int_y f_{Y|X=x}(y)ydy$.

Voyons un peu dans le cas de notre exemple, quand $x, y \in [0, 1]$: $f_{Y|X=x}(y) = \frac{x+y}{x+1/2}$, et donc $\mathbb{E}(Y|X = x) = \frac{x+2/3}{2x+1}$.

On dit que deux variables continues sont indépendantes quand, pour tout x et y réels, $f_Z(x, y) = f_X(x)f_Y(y)$. Si X et Y sont indépendante, pour tous intervalles I et J,

$$\mathbb{P}(X \in I \cap Y \in J) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J)$$

on peut montrer que si deux variables sont indépendantes, leur covariance est nulle (la réciproque est ici encore fausse). En fait, quand deux variables sont indépendantes, on a toujours $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$. On peut notamment remarquer que dans ce cas, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ (*ce qui n'est pas vrai en général*)

3 Quelques techniques utiles

3.1 Changement de variable

On motive cette partie par une simple question : si on connait la loi de X, peut on connaitre la loi de $g(X)$ pour g une fonction quelconque (par exemple, peut-on connaitre la loi de X^2 ?) . Si X est une variable aléatoire discrète, c'est facile de connaitre la loi de $g(X)$:

$$\mathbb{P}(g(X) = g(x_i)) = \sum_{j\in J} p_j$$

où J est l'ensemble des j tels que $g(x_j) = g(x_i)$

En revanche le cas continu est plus difficile à traiter. On passe souvent par la fonction de répartition : plutôt que de donner un théorème général, on montre comment résoudre des cas pratiques

1) Le cas du carré : Soit X une variable aléatoire continue de fonction de répartition F_X. On cherche la fonction de répartition de la variable $Y = X^2$: pour a réel positif,

$$F_{X^2}(a) = \mathbb{P}(X^2 \leq a) = \mathbb{P}(-\sqrt{a} \leq X \leq \sqrt{a}) = F_X(\sqrt{a}) - F_X(-\sqrt{a})$$

et pour a négatif, $\mathbb{P}(X^2 \leq a) = 0$. On peut ensuite, en dérivant, trouver la densité de la variable X^2 :

$$f_{X^2}(a) = \frac{1}{2\sqrt{a}}f(\sqrt{a}) + \frac{1}{2\sqrt{a}}f(-\sqrt{a})$$

2) Le cas de l'exponentiel : On cherche la fonction de répartition de la variable $Y = \exp(X)$.

$$F_{\exp(X)}(a) = \mathbb{P}(\exp(X) \leq a) = \mathbb{P}(X \leq \log a) = F_X(\log a)$$

3) Le cas du maximum : Soient X_1, X_2 et X_3 des v.a. indépendantes et identiquement distribuées. On veut connaitre la loi de $Y = \max(X_1, X_2, X_3)$

$$F_Y(a) = \mathbb{P}(Y \leq a) = \mathbb{P}((X_1 \leq a) \cap (X_2 \leq a) \cap (X_3 \leq a)) = \mathbb{P}(X_1 \leq a)^3 = F_X(a)^3$$

($*$) Exercice : étudiez le cas du minimum

($*$) Exercice : Si X est une variable aléatoire continue de densité f, quelle est la densité de la variable $Y = aX + b$?

3.2 Somme de variables et loi normale

On peut se demander, après avoir étudié la loi d'une fonction de X en connaissant celle de X, que peut-on dire sur la loi d'une fonction de (X, Y) (par exemple $X + Y$ ou X/Y) en connaissant la loi du couple $Z = (X, Y)$? Cette question est plus compliquée en générale, et on n'étudie dans ce cours que des cas particuliers. Le premier donne un résultat assez utiles

Proposition (Somme de deux variables)**.** *Soit $S = X + Y$.*

$$f_S(a) = \int_{t \in \mathbb{R}} f_Z(t, a - t)dt$$

Si X et Y sont indépendants on a $f_S(a) = \int_{t \in \mathbb{R}} f_X(t) f_Y(a - t)dt = f_X * f_Y(a)$ où $*$ est le symbole du produit de convolution entre deux fonction. En général, il est donc assez compliqué d'obtenir la loi d'une somme de v.a. à partir des lois des v.a. Il existe une exception notable et très importante :

Proposition (Stabilité de la famille des gaussiennes)**.** *Si $X_1, X_2, ..., X_n$ sont des gaussiennes indépendantes de moyennes $\mu_1, ..., \mu_n$ et de variances $\sigma_1^2, ..., \sigma_n^2$, alors, pour tous réels $\lambda_1, ..., \lambda_n$, la variable $\lambda_1 X_1 + ... + \lambda_n X_n$ suit une loi normale, de moyenne $\sum \lambda_i \mu_i$ et de variance $\sum \lambda_i^2 \sigma_i^2$.*

C'est la seule famille de variables aléatoires qui possède cette propriété parmi les variables qui ont une variance finie.

On finit par donner la formule pour la multiplication.

Proposition (Produit de deux variables)**.** *Soit $P = X \times Y$.*

$$f_P(a) = \int_{t \in \mathbb{R}} \frac{1}{|t|} f_Z(t, a/t)dt$$

3.3 Trois inégalités*

La première des trois inégalités qu'on va donner existe dans de nombreux domaines des mathématiques. Ici nous donnons une version appliquée à la théorie des probabilités

Proposition (Inégalité de Cauchy-Schwarz)**.** *Pour toutes variables aléatoires X et Y dont la variance existe,*

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(y^2)}$$

($*$) Exercice : Montrer, à l'aide de l'inégalité de Cauchy-Schwarz, que le coefficient de corrélation $\rho(X, Y)$ défini à la partie précédente est toujours compris entre -1 et 1. Indication : Appliquer l'inégalité aux variables centrées $X - \mathbb{E}(X)$ et $Y - \mathbb{E}(Y)$.

La seconde inégalité permet, en sachant que la variable aléatoire $X \geq 0$ admet une espérance, de majorer la probabilité que X soit grand.

Proposition (Inégalité de Markov)**.** *Pour toute variable aléatoire $X \geq 0$ et pour $a \in \mathbb{R}_+$*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Démonstration. $\mathbb{E}(X) = \mathbb{E}(X\mathbf{1}_{\{X \geq a\}}) + \mathbb{E}(X\mathbf{1}_{\{X \leq a\}}) \geq \mathbb{E}(X\mathbf{1}_{\{X \geq a\}}) \geq a\mathbb{E}(\mathbf{1}_{\{X \geq a\}}) = a\mathbb{P}(X \leq a)$ ■

Dans la preuve on se sert de deux outils utiles : la *monotonie de l'espérance* (si une variable Y est tout le temps supérieure à une variable X, comme $X\mathbf{1}_{\{X\geq a\}}$ est tout le temps supérieur à $a\mathbf{1}_{\{X\geq a\}}$, alors $\mathbb{E}(Y)\geq\mathbb{E}(X)$) et l'identité très utile $\mathbb{E}(\mathbf{1}_{\{X\geq a\}})=\mathbb{P}(X\leq a)$ (on peut vérifier que c'est vrai pour les variables discrète et continue comme exercice).

Proposition (Inégalité de Jensen)**.** *Pour toute variable aléatoire $X\geq 0$ et pour toute fonction convexe f*

$$f(\mathbb{E}(X))\leq\mathbb{E}(f(X))$$

On étudiera en détail ce qu'est une fonction convexe dans le chapitre suivant.

4 Convergence de variables aléatoires*

On peut ignorer en première lecture ce chapitre qui n'est pas nécessaire pour comprendre la suite du cours.

On considère une suite $X_1, X_2, ...$ de variables aléatoires. Il existe deux grands types de convergences possibles :

— Les convergences "locales". Elle signifient qu'avec une probabilité de plus en plus grande, la suite $X_1, ..., X_n$ converge vers une valeur, ils se rapprochent d'une valeur, qu'on note X_∞, qui peut être aléatoire. Par exemple, si $B_1, B_2, ...$ sont des v.a. de Bernouilli de paramètre $1/2$, la suite $Y_n=\sum_{i=1}^n B_i/i^2$ converge "localement", alors que la suite B_i ne converge pas localement. Il y a trois types de convergence locale : la convergence presque-sure, la convergence en probabilité ($\mathbb{P}(|X_i-X_\infty|>\epsilon)\to 0$ quel que soit $\epsilon>0$) et la convergence en moyenne ($\mathbb{E}(|X_i-X_\infty|)\to 0$)

— La convergence en loi. Les variables aléatoires ne convergent pas forcément, elles ne se rapprochent pas, mais leurs fonctions de répartitions deviennent de plus en plus proches d'une fonction de répartition limite (on dit que la suite de fonction F_{X_n} converge simplement. Par exemple, la suite B_i converge en loi.

On peut maintenant énoncer les deux grands théorèmes asymptotiques en probabilité :

Théorème 1 (Loi des grands nombres)**.** *Soit (X_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ. La moyenne empirique $\overline{X}_n=\frac{1}{n}\sum_{k=1}^n X_k$ converge en probabilité et presque surement vers l'espérance : $\overline{X}_n\to\mu$.*

La quantité $(\overline{X}_n-\mu)$ tend vers 0. On aimera savoir à quelle vitesse la convergence se fait : aussi vite que $1/n$? que $1/n^2$. Le théorème suivant donne la réponse à cette question.

Théorème 2 (Théorème central limite)**.** *Soit (X_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ et une variance σ^2 finie. Alors, quand n tend vers l'infini, $\sqrt{n}(\overline{X}_n-\mu)$ converge en loi vers une loi normale $\mathcal{N}(0,\sigma^2)$*

Ce théorème fondamental montre qu'une somme de variables aléatoires identiques, de quelque forme qu'elles soient tend vers une gaussienne quand on la divise par le bon facteur ! Cela montre le rôle central joué par les gaussiennes dans le monde des probabilités.

Optimisation

Introduction

L'optimisation est la branche des mathématiques qui étudie les problèmes de la forme

$$\min f(x), x \in K$$

où f est une fonction de E dans $\mathbb{R}$ appelée *fonction objectif*, E un ensemble quelconque ($\mathbb{Z}^d, \mathbb{R}^d$, $\mathcal{M}_n(\mathbb{R})$, $\mathbb{R}^{\mathbb{N}}$, ...) et K une partie de E (souvent appelée *contrainte*). On peut naturellement se demander :

— Un tel minimum existe-t-il ?
— Est-il unique ?
— Comment trouver (ou au moins approcher) ce minimum (s'il existe) ?

L'objectif de ce chapitre est de donner des éléments de réponse à ces questions et de montrer leur importance en économie et en statistiques. On commence par donner quelques problèmes simples pour motiver ces questions, et pour montrer que l'optimisation est un outil de base en économie.

Exemple 1 (Problème du consommateur)**.** *On considère une économie à deux biens, A et B, de prix unitaires p_A et p_B respectivement. Un consommateur a une utilité $U(x, y)$ pour la consommation de x unités du bien A et y du bien B, et il a un budget R à dépenser. Le consommateur doit donc résoudre le problème suivant*[2] *:*

$$\begin{array}{ll} \textit{maximiser} & U(x,y) \\ \textit{sous contrainte} & x \geq 0,\ y \geq 0,\ p_A x + p_b y \leq R \end{array}$$

C'est l'exemple le plus classique d'un problème d'optimisation en économie et beaucoup de problèmes s'y ramènent. On cherche les solutions dans l'espace $E = \mathbb{R}^2$, et on est contraint au domaine $K = \{(x, y) \in \mathbb{R}^2 | x \geq 0,\ y \geq 0,\ p_A x + p_b y \leq R\}$.

Exemple 2 (Optimisation à horizon infini, contrôle optimal)**.** *A chaque période t, un agent a le choix de la part de revenu qu'il consomme c_t. Avec un facteur de préférence pour le présent β, et un salaire (fixe) α il résout le problème d'optimisation suivant :*

$$\begin{array}{ll} \textit{maximiser} & \sum_{t \in \mathbb{N}} \beta^t U(c_t) \\ \textit{sous contrainte} & c_t \geq 0,\ s_{t+1} = (1+r)s_t + \alpha - c_t \geq 0 \end{array}$$

La différence majeure avec le premier exemple est que l'agent a le choix sur *une infinité* de variables $(c_t)_{t \in \mathbb{N}}$: l'espace E dans lequel on cherche est de dimension infinie, ce qui rend en général le problème plus compliqué. L'exemple très important qui suit est issu de la statistique (on le développera au prochain chapitre).

Exemple 3 (Maximum de vraisemblance)**.** *On a n données indépendantes, notées $(x_1, ..., x_n)$, tirées selon un loi normale de moyenne μ et de variance σ^2 inconnues : $f_{\mu,\sigma}(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-a^2}{2\sigma^2})$. On cherche à estimer la valeur inconnue de μ et de σ à l'aide des n tirages $(x_1, ..., x_n)$. Une des méthodes possibles est celle du* maximum de vraisemblance *: on cherche les valeurs de μ et de σ qui rendent le tirage $(x_1, ..., x_n)$ le plus vraisemblable, c'est à dire qui maximise $f_{\mu,\sigma}(x_1) \times f_{\mu,\sigma}(x_2) \times ... \times f_{\mu,\sigma}(x_n)$.*

$$\textit{maximiser} \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2} \sum_{i \leq n} (x_i - \mu)^2)$$

2. C'est un problème de maximisation, mais on le ramène facilement à un problème de minimisation en se rendant compte que maximiser la fonction $(x, y) \to U(x, y)$ revient à minimiser la fonction $(x, y) \to -U(x, y)$: on dit que les problèmes sont équivalents.

On remarque qu'on n'impose pas de contraintes sur σ. Il existe de nombreux autres domaine d'applications : apprentissage statistique (machine learning), théorie des jeux, théorie des graphes, physique, ingénierie, etc. , et de nombreux autres types de problèmes (bin-packing, MAX-CUT), qui dépassent le cadre de ce cours. On commence par répondre à la question de l'existence d'un minimum, puis on fait quelques rappels de calculs différentiels qui seront utiles. On répond ensuite a la question de l'approximation

1 Existence d'un minimum

Une fonction f peut ne pas avoir de minimum fini sur un ensemble : l'exemple le plus évident est celui de la fonction $x \to x$ considérée sur $\mathbb{R}$, où encore la fonction $x \to 1/x$ considérée sur $[-1, 0[$. Néanmoins ces cas pathologiques sont rarement intéressants en économie, et sont souvent plutôt le signe d'un problème mal posé.

On donne sans preuve quelques conditions suffisantes d'existence d'un minimum. La première est la plus importante et les autres en sont des conséquences :

Proposition (Weierstrass)**.** *Si $f : X \mapsto \mathbb{R}$ est continue sur un ensemble $X \subset \mathbb{R}^n$ fermé*[3] *et borné, alors f a un minimum (et un maximum) sur X qu'elle atteint en au moins un point $x^* \in X$.*

Remarque : cette proposition n'est vraie que pour les sous-ensemble de $\mathbb{R}^n$. Elle n'est plus vraie en dimension infinie ! Elle permet d'étudier l'exemple 1 de l'introduction : comme l'ensemble sur lequel on cherche un minimum est fermé et borné, on peut déduire que la fonction objectif admet un maximum sur cet ensemble et qu'elle l'atteint en un point. On peut donner deux autres conditions suffisantes d'existence de minimum (ou de maximum) qui découlent de la première.

Proposition. *Si $f : X \mapsto \mathbb{R}$ est continue sur un ensemble $X \subseteq \mathbb{R}^n$ et que $f(x) \to +\infty$ (resp $-\infty$) quand $||x|| \to \infty$, alors f admet un minimum (resp un maximum) qu'elle atteint en au moins un point $x^* \in X$.*

Proposition. *Si $f : X \mapsto \mathbb{R}$ est continue sur un ensemble $X \subseteq \mathbb{R}^n$ et que $f(x) \to l$ quand $||x|| \to \infty$, et qu'il existe $y \in \mathbb{R}^n$ tel que $f(y) < l$ (resp $f(y) > l$) alors f admet un minimum (resp un maximum) qu'elle atteint en au moins un point $x^* \in X$.*

Remarque : on peut prendre $X = \mathbb{R}^n$, c'est souvent le cas. La dernière proposition permet ainsi de montrer l'existence d'un maximum de vraisemblance dans l'exemple 3 de l'introduction. En effet, en prolongeant par continuité la vraisemblance en $\sigma = 0$ (expliquer le prolongement par continuité), on se rend compte que quand la norme du vecteur (μ, σ) tend vers l'infini, alors la vraisemblance tend vers 0. Hors la vraisemblance est partout strictement positive, ce qui rentre dans le cadre de la dernière proposition.

On quitte cette partie avec une "proposition" beaucoup moins rigoureuse, mais qui donne l'intuition nécessaire : Soit O un ouvert[4] de $\mathbb{R}^n$. Si une fonction f tend vers une limite l (eventuellement infinie) quand x se "rapproche des bords de l'ouvert" (par exemple pour une fonction définie sur $\mathbb{R} \times \mathbb{R}_+^*$, $x = (x_1, x_2)$ se rapproche des bords quand x_1 tend vers plus ou moins l'infini, ou quand x_2 tend vers 0 ou $+\infty$), et si à l'intérieur de l'ouvert on peut trouver un y tel que $f(y) < l$, alors f admet et atteint un minimum à l'intérieur de l'ouvert O. On voit que cette proposition généralise la précédente, et permet de traiter simplement le cas de l'exemple 3 en se restreignant à l'ensemble $\mathbb{R} \times \mathbb{R}_+^*$.

2 Éléments de calcul différentiel

2.1 Le cas uni-dimensionnel

Sur $\mathbb{R}$ on sait que, si f est une fonction dérivable en a, alors $f(a + \epsilon) \simeq f(a) + \epsilon f'(a)$ quand ϵ est "petit" (proche de 0) : on peut approximer la fonction f (à priori compliquée) par une fonction *affine* en ϵ (donc simple), mais cette approximation ne fonctionne que localement. Cette approximation, même locale permet d'avoir l'intuition d'un résultat important :

3. On dit qu'un ensemble $X \subset E$ est fermé quand toute suite convergente d'éléments de X convergente a sa limite dans X. Par exemple $]0, 1]$ n'est pas fermé dans $\mathbb{R}$ car la suite $1/n$ est une suite d'éléments de $]0, 1]$ qui tend vers $0 \notin]0, 1]$. On peut retenir qu'un ensemble est fermé quand il contient ses frontières, quand il est défini par des inégalités larges et pas strictes, comme dans l'exemple 1.

4. Par opposition avec un fermé, on dit qu'un ensemble est ouvert quand il ne contient pas ses frontières : plus rigoureusement, si un élément x est dans O, il existe une boule de centre x qui est entièrement contenue dans x. Par exemple $]0, 1]$ n'est pas ouvert, $]0, 1[$ l'est, $\mathbb{R} \times \mathbb{R}_+^*$ est ouvert, pas $\mathbb{R} \times \mathbb{R}_+$

Proposition. *Si une fonction f est dérivable sur $D \subset \mathbb{R}$ ouvert, et si elle admet un extremum en $x^* \in D$, alors $f'(x^*) = 0$.*

Preuve heuristique : supposons qu'on ait $f'(x^*) > 0$, alors, pour ϵ suffisamment petit, $f(x^* + \epsilon) \simeq f(x^*) + \epsilon f'(x^*)$, donc $f(x^* + \epsilon)$ est plus petit que $f(x^*)$ quand $\epsilon < 0$, et plus grand que $f(x^*)$ quand $\epsilon > 0$, $f(x^*)$ n'est donc pas un extremum.

La réciproque de cette proposition n'est pas vraie : une fonction dérivable peut avoir un point non-extrémal où sa dérivée s'annule (penser par exemple à la fonction $x \to x^3$ en 0).

2.2 Le cas multi-dimensionnel

On se donne une fonction linéaire $f : \mathbb{R}^d \to \mathbb{R}$ et un point a dans $\mathbb{R}^d$. On aimerait étendre l'approximation linéaire de la partie précédente c'est-à-dire trouver une fonction *linéaire* $\mathcal{L}_{f,a}$ de $\mathbb{R}^d$ dans $\mathbb{R}$, tel que, si $\epsilon \in \mathbb{R}^d$ a toutes ses coordonnées suffisamment petites, $f(a + \epsilon) \simeq f(a) + \mathcal{L}_{f,a}(\epsilon)$. On rappelle un petit lemme d'algèbre linéaire :

Lemme. *Une fonction linéaire $\mathcal{L}$ de $\mathbb{R}^d$ dans $\mathbb{R}$ peut toujours s'écrire sous la forme d'un produit scalaire : autrement dit il existe un vecteur g tel, pour tout x dans $\mathbb{R}^d$*

$$\mathcal{L}(x) = \langle x, g \rangle$$

Démonstration. C'est un cas particulier du théorème de Riesz, très facile à montrer. On note $(e_1, ..., e_d)$ la base canonique de $\mathbb{R}^d$. On peut alors écrire, pour $x \in R^d$, $x = \sum x_i e_i$. Par linéarité, on a $\mathcal{L}(x) = \mathcal{L}(\sum x_i e_i) = \sum x_i \mathcal{L}(e_i)$. En notant g le vecteur $\sum \mathcal{L}(e_i) e_i$, on a bien, pour tout x, $\mathcal{L}(x) = \langle x, g \rangle$. ■

Définition. *On dit qu'une fonction f est différentiable en a s'il existe un* vecteur*, noté ∇f_a (ou parfois $f'(a)$) et appelé gradient de f en a tel que, pour $\epsilon \in \mathbb{R}^d$ suffisamment petit ,*

$$f(a + \epsilon) \simeq f(a) + \langle \epsilon, \nabla f_a \rangle$$

Calcul du gradient Par définition du gradient, la i^{eme} coordonnée du gradient $\langle e_i, \nabla f_a \rangle = \lim_{\eta \to 0}(f(a + \eta e_i) - f(a))/\eta$, cette limite étant couramment notée $\frac{\partial f}{\partial x_i}\Big|_a$ et appelée dérivée partielle de f par rapport à la i^{eme} coordonnée. On peut donc écrire :

$$\nabla f(a) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(a) \\ \cdots \\ \frac{\partial f}{\partial x_d}(a) \end{pmatrix}$$

(∗) Exercice : Calculez le gradient en tout point de $f : (x, y) \to \sqrt{x^2 + y^2}$
(∗) Exercice (fondamental) : Soit $y \in \mathbb{R}^d$ et A une matrice carrée de taille d. Calculez le gradient de

$$f : x \to ||y - Ax||_2^2$$

On donne quelques propriétés du gradient qui permettent de le calculer facilement. On se donne f et g deux fonction définie sur $\mathbb{R}^d$

- Le gradient est linéaire : si $\lambda \in \mathbb{R}$, $\nabla(f + \lambda g) = \nabla f + \lambda \nabla g$
- On peut calculer le gradient d'un produit, $\nabla f \times g(a) = g(a)\nabla f(a) + f(a)\nabla g(a)$
- Pour la composée, si $h : \mathbb{R} \to \mathbb{R}$, alors $\nabla h \circ f(x) = h'(f(x))\nabla f(x)$

En reprenant l'heuristique de la partie précédente, on prouve le résultat fondamental suivant :

Théorème 1. *Si une fonction f est différentiable sur $\mathbb{R}^p$, et si elle admet un extremum en x^*, alors $\nabla f_{x^*} = 0_{\mathbb{R}^d}$.*

Comme dans la partie précédente, ce théorème n'a pas de réciproque, elle ne donne qu'une condition nécessaire d'extrémalité.

3 Résolution du maximum de vraisemblance

On cherche à résoudre le problème posé en introduction :

$$\underset{\mu\in\mathbb{R},\sigma\in\mathbb{R}_+^*}{\text{maximiser}} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i\leq n}(x_i-\mu)^2)$$

Astuce importante : On remarque que maximiser une fonction objective positive revient à maximiser son logarithme. Donc on peut chercher le maximum de

$$g(\mu,\sigma) = \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i\leq n}(x_i-\mu)^2)\right) = -\log\left((2\pi\sigma^2)^{n/2}\right) - \frac{1}{2\sigma^2}\sum_{i\leq n}(x_i-\mu)^2$$

On sait grace à la partie 1 que cette fonction a un maximum sur $\mathbb{R}\times\mathbb{R}_+^*$: on va donc étudier tous les "candidats possibles", c'est à dire tous les points où le gradient est nul : on calcule donc le gradient :

$$\nabla g(\mu,\sigma) = \left(-\frac{1}{\sigma^2}\sum_i(\mu-x_i), -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i(\mu-x_i)^2\right)$$

Pour qu'il soit nul, on résout le système d'équation, et on trouve qu'il n'y a qu'une seule solution : $\mu^* = \frac{1}{n}\sum x_i$ (on retrouve la moyenne empirique des différentes réalisations), et $(\sigma^*)^2 = \frac{1}{n}\sum(x_i-\mu^*)^2$ (variance empirique). On n'a qu'un seul candidat, on n'a donc pas besoin de comparer différentes Hessiennes, c'est bien notre maximum de vraisemblance.

Malheureusement ce type de résolution est très limité : il est très rare d'optimiser une fonction globalement. En effet, la plupart du temps des contraintes apparaissent naturellement dans les problèmes d'optimisation, même dans les problèmes de maximum de vraisemblance, comme on va le voir plus tard.

4 Optimisation sous contraintes

On a n données indépendantes, notées $(x_1,...,x_n)$, tirées selon, cette fois-ci, une *loi uniforme* sur le segment $[a,b]$ où a et b sont deux réels inconnus qu'on essaie d'estimer. La densité d'une loi uniforme s'écrit $f_{a,b}(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$. On cherche là encore le maximum de vraisemblance, c'est à dire les valeurs de a et de b qui rendent le tirage $(x_1,...,x_n)$ le plus vraisemblable, c'est à dire qui maximise $f_{a,b}(x_1)\times f_{a,b}(x_2)\times...\times f_{a,b}(x_n)$. Notre problème de maximisation s'écrit donc

$$\underset{a,b\in\mathbb{R},b\geq a}{\text{maximiser}} \left(\frac{1}{b-a}\right)^n \mathbf{1}_{[a,b]}(x_1)\times...\times\mathbf{1}_{[a,b]}(x_n)$$

A première vue, il pourrait s'agir d'un problème de minimisation sans contraintes : en effet, la condition qu'on impose sur a et b n'est pas très importante (on pourrait l'enlever à condition de remplacer le $b-a$ au dénominateur de la fonction objectif par $|b-a|$). Pourtant il y a une grande différence entre ce problème et le précédent. Ici, la fonction objectif n'est *pas différentiable* sur $\mathbb{R}^2$: elle n'est même pas continue ! On ne peut donc pas appliquer le théorème 1 et simplement trouver les points où le gradient s'annule. Il faut procéder différemment. En fait ici, on se rend compte que si a est plus grand que n'importe lequel des x_i, alors la fonction objectif est nulle, et n'atteint donc pas le maximum. De même si b est plus petit que n'importe lequel des x_i. On sait donc que pour trouver le maximum de la fonction objectif, on va avoir $a\leq\min_i x_i$ et $b\geq\max_i x_i$: l'astuce est de ne chercher *que* sur cet espace. Le problème se réécrit alors

$$\begin{array}{ll}\text{maximiser} & \left(\frac{1}{b-a}\right)^n \\ \text{sous contrainte} & a\leq\min_i x_i, b\geq\max_i x_i\end{array}$$

car si $a\leq\min_i x_i$ et $b\geq\max_i x_i$, alors $1_{[a,b]}(x_i)=1$. La fonction à minimiser dans ce cas, $(a,b)\to\frac{1}{b-a}$ est elle bien différentiable ! L'intuition nous dit que pour maximiser $1/(b-a)$, il faut minimiser $(b-a)$ et donc prendre b le plus petit possible, soit $\max x_i$, et a le plus grand possible, soit $\min x_i$, et en effet $a^*=\min_i x_i$ et $b^*=\max_i x_i$. L'objectif de cette partie est de justifier cette intuition afin de pouvoir traiter des cas plus compliqués (comme par

exemple l'exemple 1)

On a deux types de contraintes possibles :

— les contraintes d'égalité, où on impose qu'une fonction quelconque g du vecteur $v = (v_1, ..., v_n)$ soit nul : $g(v_1, ..., v_n) = 0$. Par exemple, en deux dimensions, on peut imposer par exemple $v_1^2 + v_2^2 = R^2$ (on doit optimiser sur un cercle), ou encore $\alpha v_1 + \beta v_2 = 1$ (on optimise sur une droite)

— les contraintes d'inégalité, où on impose qu'une fonction quelconque g du vecteur $v = (v_1, ..., v_n)$ soit positive : $g(v_1, ..., v_n) \leq 0$. On peut imposer par exemple $v_1^2 + v_2^2 \leq R^2$ (on impose cette fois sur un *disque*), ou $\alpha v_1 + \beta v_2 \leq 1$ (on optimise sur un demi-plan) .

Bien-sur, on peut coupler des contraintes d'inégalité à des contraintes d'égalité quand on a plusieurs contraintes. *Il est très important de noter que dans tout ce qui va suivre, on va supposer que les fonctions de contraintes g sont différentiables partout.* On va commencer par étudier le cas simple où il n'y a que des contraintes d'égalité

4.1 Contraintes d'égalité

On va étudier l'exemple suivant, pour fixer les idées

$$\begin{array}{ll} \text{minimiser} & f(x,y) = 3x - 2y \\ \text{sous contrainte} & g(x,y) = x^2 + y^2 - 1 = 0 \end{array}$$

D'après la partie 1, ce minimum existe et il est atteint (car on est sur un fermé borné). On se place donc en ce point optimal sous contrainte $v^* = (x^*, y^*)$. Quand on parlait de minimum global (dans la preuve du théorème 1), il fallait que, pour toute petite variation $\epsilon = (\epsilon_1, \epsilon_2)$, on ait $f(v^* + \epsilon) \geq f(v^*)$, ce qui entrainait $\nabla f(v^*) = 0$. Maintenant, il suffit que, pour tout ϵ **tel que** $g(v^* + \epsilon) = 0$ (on n'a plus besoin de tous les ϵ), $f(v^* + \epsilon) \geq f(v^*)$. Il s'agit donc de trouver les ϵ tels que $g(v^* + \epsilon) = 0$, les "bons" ϵ. Prenons l'approximation linéaire, comme g est différentiable par hypothèse : $g(v^* + \epsilon) = g(v^*) + \langle \epsilon, \nabla g(v^*) \rangle$. Comme v^* est l'optimum contraint, $g(v^*) = 0$, donc, en approximation, les ϵ tels que $g(v^* + \epsilon) = 0$ sont les ϵ tels que $\langle \epsilon, \nabla g(v^*) \rangle = 0$, c'est à dire les ϵ orthogonaux à $\nabla g(v^*)$.

On peut donc réécrire notre condition de minimum local : il faut que, pour tout ϵ orthogonal à $\nabla g(v^*)$, $f(v^* + \epsilon) \geq f(v^*)$. On réécrit notre approximation sur f : pour tout ϵ orthogonal à $\nabla g(v^*)$, $\langle \nabla f(v^*), \epsilon \rangle = 0$. Pour que cette dernière condition soit vraie, il faut et il suffit qu'il existe λ tel que $\nabla f(v^*) = \lambda \nabla g(v^*)$. On peut donc écrire notre premier lemme :

Lemme. *On se place dans le cadre suivant :*

$$\begin{array}{ll} \textit{minimiser} & f(x,y) \\ \textit{sous contrainte} & g(x,y) = 0 \end{array}$$

où f et g sont différentiables. Alors f atteint son minimum contraint en un point v^* où il existe λ tel que $\nabla f(v^*) = \lambda \nabla g(v^*)$.

Avant d'énoncer le théorème général, on va résoudre notre exemple. On peut montrer facilement que $\nabla f(x,y) = (3, -2)$ et $\nabla g(x,y) = (2x, 2y)$. En écrivant $\nabla f(v^*) = \lambda \nabla g(v^*)$, on obtient $x = \frac{3}{2\lambda}, y = \frac{-1}{\lambda}$. Pour trouver le bon λ, il suffit de se souvenir que $g(v^*) = 0$: on doit donc avoir $(\frac{3}{2\lambda})^2 + (\frac{1}{\lambda})^2 = 1$, ce qui entraine $\lambda^2 = 13/4$. On a donc deux candidats possibles pour le minimum selon que λ est positif ou négatif. Pour savoir lequel retenir il suffit de comparer la valeur de la fonction objectif en ces deux points, et on remarque que le minimum est atteint pour $\lambda = \sqrt{13/4}$ (l'autre point correspond à un maximum). On étend le lemme aux dimensions > 2, et à plusieurs contraintes.

Théorème 2. *On se place dans le cadre suivant : on considère $f, g_1, g_2, ..., g_k$ des fonctions différentiables de $\mathbb{R}^d$ dans $\mathbb{R}$. On cherche à résoudre*

$$\begin{array}{ll} \textit{minimiser} & f(x) \\ \textit{sous contrainte} & g_1(x) = 0, g_2(x) = 0, ..., g_k(x) = 0 \end{array}$$

Alors f atteint son minimum contraint en un point v^* où il existe $\lambda_1, \lambda_2, ..., \lambda_k$ tel que $\nabla f(v^*) = \sum_i \lambda_i \nabla g_i(v^*)$. Ces $\lambda_1, \lambda_2, ..., \lambda_k$ sont souvent appelés multiplicateurs de Lagrange.

Exemple : on considère le problème suivant :

$$\begin{array}{ll} \text{minimiser} & f(x,y,z) = x^2+y^2+z^2 \\ \text{sous contrainte} & x+y+z=1, xy=1 \end{array}$$

Le premier problème est de savoir si le minimum existe. L'ensemble des contraintes est fermé mais pas borné. Pour remédier à ce problème, on remarque par exemple que le point $a = (5, 1/5, -1/5)$ répond aux contraintes et $f(a) = 25 + 2/25$: on peut donc se restreindre à chercher le minimum dans l'ensemble $\{v = (x,y,z) | f(v) \leq 25+2/25, x+y+z=1, xy=1\}$: on *ajoute artificiellement une contrainte* en se servant de la fonction objectif pour se trouver dans un ensemble borné et fermé.

On calcule ensuite $\nabla f = (2x, 2y, 2z)$, $\nabla g_1 = (1,1,1)$ et $\nabla g_2 = (y, x, 0)$. En écrivant $\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2$, on obtient $z = \lambda_1/2$, et $x = y = \lambda_1/(2-\lambda_2)$. Pour trouver λ_1 et λ_2, on écrit $2\lambda_1/(2-\lambda_2) + \lambda_1/2 = 1$ et $\lambda_1^2/(2-\lambda_2)^2 = 1$: on a donc deux solutions possibles $(\lambda_1, \lambda_2) = (-2, 4)$ ou $(6, 8)$, ce qui donne deux candidats : $(1, 1, -1)$ et $(-1, -1, 3)$. On compare la valeur de la fonction objectif en ces deux points, et on remarque que le minimum est atteint pour le point $(1, 1, -1)$. On se penche maintenant sur les contraintes d'inégalités, bien plus courantes en économie.

4.2 Contraintes d'inégalité

On fixe là aussi les idées avec un exemple

$$\begin{array}{ll} \text{minimiser} & f(x,y) = (x-3)^2 + (y-2)^2 \\ \text{sous contrainte} & x \geq 0,\ y \geq 0,\ x+y \leq 2 \end{array}$$

On remarque que c'est un cas particulier proche du problème du consommateur cité en introduction. On peut réécrire les contraintes $g_1(x,y) \leq 0,\ g_2(x,y) \leq 0,\ g_3(x,y) \leq 0$ pour g_1, g_2 et g_3 bien choisis. Pas de problème pour prouver que le minimum existe et est atteint : notre ensemble contraint est bien fermé et borné (c'est un triangle). On se place donc en v^*, le point optimal. Il faut que, pour toute petite variation ϵ **tel que** $g_1(v^*+\epsilon) \leq 0$, $g_2(v^*+\epsilon) \leq 0$, **et** $g_3(v^*+\epsilon) \leq 0$ on ait $f(v^*+\epsilon) \geq f(v^*)$. Comme dans précédemment, on essaie d'abord d'identifier les "bons" ϵ. On prend n'importe quelle contrainte, la première par exemple : deux cas sont possibles :

- soit le point v^* *sature la contrainte*, c'est-à-dire vérifie $g_1(v^*) = 0$ (on peut aussi dire que la contrainte est active. Dans ce cas $g_1(v^*+\epsilon) \simeq \langle \epsilon, \nabla g_1 \rangle$ et $g_1(v^*+\epsilon) \leq 0$ si et seulement si $\langle \epsilon, \nabla g_1 \rangle \leq 0$
- soit le point v^* ne sature pas la contrainte ($g_1(v^*) < 0$), et dans ce cas, par continuité, pour *n'importe quel* ϵ assez petit, $g_1(v^*+\epsilon) \leq 0$

Donc, si on note S l'ensemble des contraintes saturées par v^*, les "bons" ϵ sont ceux tels que pour tout $i \in S$, $\langle \epsilon, \nabla g_i \rangle \leq 0$: et il faut que pour tous ces ϵ, $\langle \nabla f, \epsilon \rangle \geq 0$. Pour que cette dernière condition soit vraie, il faut et il suffit qu'il existe des $\lambda_i \geq 0$ tels que $\nabla f(v^*) = -\sum_{i \in S} \lambda_i \nabla g_i(v^*)$ [5]. En résumé, si v^* est le point minimum, nécessairement $-\nabla f(v^*)$ est une combinaison linéaire positive des différents $\nabla g_i(v^*)$ pour lesquels la contrainte est saturée. On récapitule cela dans le théorème suivant.

Théorème 3 (Conditions de Karush-Kuhn-Tucker)**.** *On se place dans le cadre suivant : on considère $f, g_1, g_2, ..., g_k$ des fonctions différentiables de $\mathbb{R}^d$ dans $\mathbb{R}$. On cherche à résoudre*

$$\begin{array}{ll} \textit{minimiser} & f(x) \\ \textit{sous contrainte} & g_1(x) \leq 0, g_2(x) \leq 0, ..., g_k(x) \leq 0 \end{array}$$

Alors f atteint son minimum contraint en un point v^* où il existe $\lambda_1, \lambda_2, ..., \lambda_k \geq 0$ tel que $\nabla f(v^*) = -\sum_i \lambda_i \nabla g_i(v^*)$. De plus, pour tout $i \leq k$, $\lambda_i \times g_i(v^*) = 0$: si la contrainte i n'est pas saturée, alors $\lambda_i = 0$.

Remarque : si on veut maximiser une fonction f, on reprend le théorème précédent. Il est valide à condition de remplacer $\nabla f(v^*) = -\sum_i \lambda_i \nabla g_i(v^*)$ par $\nabla f(v^*) = \sum_i \lambda_i \nabla g_i(v^*)$.

Reprenons notre exemple : $\nabla f = (2(x-3), 2(y-2))$, $\nabla g_1 = (-1, 0)$, $\nabla g_1 = (0, -1)$, $\nabla g_1 = (1, 1)$: on a donc $x = 3 + (\lambda_1 - \lambda_3)/2$ et $y = 2 + (\lambda_2 - \lambda_3)/2$. La condition $x + y \leq 2$ se réecrit $5 + (\lambda_1 + \lambda_2)/2 - \lambda_3 \leq 2$, donc nécessairement $\lambda_3 > 0$ et donc $x + y = 2$ (saturation de la contrainte). On a donc trois cas possibles : $x = 0$ ou $y = 0$ ou $\lambda_1 = \lambda_2 = 0$. Le premier cas donne le point $(0, 2)$, le deuxième $(2, 0)$, le troisième $(1.5, 0.5)$. On compare la valeur de la fonction objectif en ces trois points, et on voit que le minimum est atteint en $(1.5, 0.5)$

5. Montrer l'équivalence ici n'est pas évident : un des deux sens est simple, l'autre est plus avancé et nécessite quelques manipulations d'algèbres linéaire. Il s'agit en fait de montrer que le bidual de $P = (\nabla g_1, ..., \nabla g_S)$ est le plus petit cône convexe fermé contenant P : $P^{**} = \overline{\text{co}}(\mathbb{R}_+ P)$. Une simple recherche google permet de trouver des preuves élémentaires de ce résultat

4.3 Lagrangien*

Il y a une manière simple de résumer les théorèmes précédents. On introduit le *Lagrangien* du problème, qui est une fonction de $x \in \mathbb{R}^d$ et des différents λ :

$$\mathcal{L}(x, \lambda_1, ..., \lambda_k) = f(x) + \sum_i \lambda_i g_i(x)$$

S'il n'y a que des contraintes d'égalité, résoudre le problème d'optimisation revient à résoudre le problème $\nabla\mathcal{L} = 0$, ce qui va nous donner le bon minimum contraint x^* et les multiplicateurs de Lagrange (ou au moins de bons candidats). On ajoute maintenant les contraintes d'inégalité.

Soit un problème d'optimisation dans sa forme standard

$$\begin{aligned} &\text{minimiser } f(x) \\ &\quad \text{avec } g_i(x) \leq 0,\ i \in \{1, \dots, m\} \\ &\quad\quad h_j(x) = 0,\ j \in \{1, \dots, p\} \end{aligned}$$

On appelle p^* la valeur optimale de la fonction objective. Le Lagrangien $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ est définie par : $\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \nu_j h_j(x)$. On remarque qu'on peut réécrire le problème d'optimisation sous la forme

$$\min_x \sup_{\lambda_i \geq 0, \nu_i \in \mathbb{R}} f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \nu_j h_j(x) = \min_x \sup_{\lambda_i \geq 0, \nu_i \in \mathbb{R}} \mathcal{L}(x, \lambda, \nu)$$

Proposition. *Pour toute fonction u fonction de deux variables : alors* $\inf_y \sup_x u(x, y) \geq \sup_x \inf_y u(x, y)$

Il n'y a pas toujours d'égalité : par exemple supposons que $x, y \in \{0, 1\}$ et que $u(0, 0) = u(1, 1) = 1$ et $u(0, 1) = u(1, 0) = 0$. Alors $\inf_y \sup_x u(x, y) = 1$ et $\sup_x \inf_y u(x, y) = 0$

On définit la "fonction duale de Lagrange" $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$:

$$g(\lambda, \nu) = \inf_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda, \nu) = \inf_{x \in \mathbb{R}^d} \left(f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \nu_j h_j(x) \right).$$

La fonction duale est concave. On a, d'après la proposition, $\sup_{\lambda_i \geq 0, \nu_i \in \mathbb{R}} g(\lambda, \nu) \leq p^*$. La problème de maximisation de la fonction g est appelé *problème dual* du premier problème de minimisation, qui est appelé primal. Il est souvent plus simple à résoudre. Le dernier théorème nous indique qu'il est parfois suffisant de résoudre le problème dual.

Théorème 4 (Dualité forte)**.** *Si toutes les fonctions en jeu (f, g_i, h_i) sont convexes et différentiables et si l'ensemble des contraintes a un point intérieur (c'est-à-dire qui vérifie toutes les inégalités strictement) alors une solution du problème dual est solution du primal et vice-versa.*

Si on reprend l'exemple de la partie précédente, le Lagrangien s'écrit $\mathcal{L}(x, \lambda, \nu) = (x-3)^2 + (y-2)^2 - \lambda_1 x - \lambda_2 y + \lambda_3(x + y - 2)$, et donc la fonction duale s'écrit $g(\lambda_1, \lambda_2, \lambda_3) = (\lambda_1 - \lambda_3)^2/4 + (\lambda_2 - \lambda_3)^2/4 - \lambda_1(3 + (\lambda_1 - \lambda_3)/2) - \lambda_2(2 + (\lambda_2 - \lambda_3)/2) + \lambda_3(3 + (\lambda_1 + \lambda_2)/2 - \lambda_3) = -\lambda_1^2/4 - \lambda_2^2/4 - \lambda_3^2/2 - 3\lambda_1 - 2\lambda_2 + 3\lambda_3 + \lambda_1\lambda_3/2 + \lambda_2\lambda_3/2$

5 Convexité*

La convexité est une des notions les plus importantes en optimisation. Commençons par rappeler la définition d'une fonction convexe (et celle de fonction concave, très liée).

Définition. *Soit f une fonction définie sur $\mathbb{R}^p$, f est convexe (resp concave) sur D, si et seulement si, pour tout $\alpha \in [0, 1]$ et pour tout $(x, y) \in D^2$,*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y)$$

(resp $f(\alpha x + (1 - \alpha)y) \geq \alpha f(x) + (1 - \alpha) f(y)$)

Remarque : f est convexe si et seulement si $-f$ est concave.
Une fonction est donc convexe quand sa courbe entre les points x et y est en dessous du segment $[x, y]$.

Commençons par les fonctions définies sur $\mathbb{R}$ (ou sur une partie de $\mathbb{R}$). En terme économique, on peut parler de rendement croissant : pour une fonction convexe, le gain marginal augmente toujours $f(x+1)-f(x) \geq f(x)-f(x-1)$. On a envie de trouver un moyen simple de prouver qu'une fonction est convexe, c'est la proposition suivante qui nous le donne.

Proposition. *Une fonction deux fois dérivable sur $D \subset \mathbb{R}$ est convexe (resp. concave) sur D si et seulement si pour tout $x \in D$, $f''(x) \geq 0$ (resp ≤ 0)*

$(*)$ Exercice : Vérifier que les fonctions $x \to x^2$, $x \to \exp(x)$, $x \to 1/x$ sont convexes (sur $\mathbb{R}$, $\mathbb{R}$, et $\mathbb{R}_+^*$ respectivement) alors que $x \to \sqrt{x}$ ou $x \to \log(x)$ sont concaves sur leurs ensembles de définitions.

On veut ensuite étendre cette proposition aux fonctions de plusieurs variables. On sait que pour ces fonctions, le rôle de dérivée seconde est joué par la matrice Hessienne (cf éléments d'algèbre). C'est donc sans surprise que nous retrouvons cette quantité dans la proposition suivante :

Proposition. *Une fonction deux fois dérivable sur $D \subset \mathbb{R}^d$ est convexe sur D si et seulement si pour tout $x \in D$, pour tout vecteur $u \in \mathbb{R}^d$ non nul, $u^T H_f(x) u \geq 0$.*

C'est donc une condition sur la matrice Hessienne *en tout point*. C'est une condition un peut moins forte que la définie-positivité (cf éléments d'algèbre), puisqu'on ne demande pas une inégalité stricte : on l'appelle généralement *positivité* (dans la littérature anglo-saxonne on parle de semi-definite positive matrix, SDP).

- Pour M une matrice 2×2, M est positive si et seulement si $\mathrm{Tr}(M) \geq 0$ *et* $\det(M) \geq 0$
- Pour les matrices de plus grandes tailles, pour qu'une matrice symétrique $A = (a_{ij})_{1 \leqslant i,j \leqslant d}$ soit positive, il faut et suffit que toutes les matrices $A_I = \left(a_{ij}\right)_{1 \leqslant i,j \in I}$, où I est n'importe quelle partie de $[\![1, n]\!]$, aient leur déterminant positif ou nul

Remarque très importante : La plupart du temps, on ne montre pas qu'une fonction est convexe en calculant sa Hessienne est en montrant qu'elle est positive : on utilise des règles d'association à partir de "briques de bases" qu'on sait déjà convexes :

- Le produit d'une fonction convexe par un réel positif est convexe
- La somme de fonctions convexes est convexe
- Toutes les fonctions linéaires sont convexes ET concaves (ce sont les seules)
- Si $g : \mathbb{R} \to \mathbb{R}$ est convexe et croissante et que f est convexe alors $g \circ f$ est convexe
- Les fonctions du type $x \to x^T M x$ où M est définie positive sont convexes.

On énonce ensuite le théorème principal de cette partie,

Théorème 5. *Tout minimum local x^* d'une fonction convexe sur $\mathbb{R}^d$ f est un minimum global de cette fonction. Si de plus $H_f(x^*)$ est définie positive, x^* est le seul minimum de f.*

Remarque : on déduit que pour les fonctions concaves, tout *maximum* local est un maximum global.

Démonstration. Supposons que x ne soit pas un minimum global. Comme x n'est pas un minimum global, il existe y tel que $f(y) < f(x)$. Mais alors, pour tout $\epsilon > 0$, $f(x + \epsilon(y - x)) = f((1 - \epsilon)x + \epsilon y) \leq (1 - \epsilon)f(x) + \epsilon f(y) < f(x)$. Donc, aussi petit que soit ϵ, le minimum de f sur la boule $\mathcal{B}(x, \epsilon)$ n'est pas en x : x n'est donc pas un minimum local. C'est ce qu'il fallait démontrer.

■

On ajoute à ce constat un lemme très lié et bien pratique :

Lemme. *Si f est une fonction convexe sur $\mathbb{R}^d$, f a un minimum global en x^* si et seulement si $\nabla f(x^*) = 0$*

On a donc une réciproque du théorème 1, dans le cas où la fonction objectif est convexe.

6 Un peu d'algorithmique

6.1 Descente de gradient et gradient projeté

La méthode la plus courante et la plus basique est celle de la descente de gradient, qui a de nombreuses variantes.

On s'intéresse pour commencer à un problème non contraint : on cherche à minimiser la fonction f sur $\mathbb{R}^d$. Une méthode classique est celle de la descente de gradient : on se donne un point de départ x_0, et l'algorithme calcule l'itération x_{i+1} à partir de x_i comme suit : $x_{i+1} = x_i - \nabla f(x_i) \times \eta_i$ où η_i est un pas qui peut prendre différentes valeurs (qui constituent différentes méthodes de descente possibles). L'exemple le plus simple est celui où η_i est constant égal à a.

A noter que ces méthodes nécessitent presque toujours une fonction f convexe.

Proposition. *Supposons que la fonction f est $\mathcal{C}^2$ et que, en tout x, $cId \preceq H_f(x) \preceq KId$ et qu'on prenne un pas constant $\leq 2/K$. Alors l'algorithme de descente de gradient à pas constant converge vers le minimum x^**

On n'abordera pas ici les questions essentielles de vitesse de convergence, qui nécessitent de faire plus d'hypothèses et d'utiliser des outils plus avancés. On peut faire mieux en optimisant à chaque étape le pas η_k. Pour les problèmes d'optimisation contraints à un domaine K, une méthode populaire est celle du gradient projeté : à chaque étape, on calcule $y_i = x_i - \nabla f(x_i) \times \eta_i$. Ce résultat intermédiaire peut éventuellement se trouver hors de K. Dans ce cas on le projette sur l'ensemble K : $x_{i+1} = \mathbf{proj}_K(y_i)$.

6.2 Méthodes des barrières et points intérieurs

Une autre méthode pour les problèmes contraintes, autre que le gradient projeté, est ce qu'on appelle la méthode des barrières. On optimise globalement (via une descente de gradient simple par exemple) mais on remplace la fonction objectif f par une fonction à laquelle on ajoute des "pénalités" ou des "barrières". Par exemple le problème

$$\begin{aligned} &\text{minimiser } f(x) \\ &h_j(x) \leq 0,\ j \in \{1, \dots, p\} \end{aligned}$$

se transformerai en $\min f(x) - \lambda \sum_j \log(-h_j(x))$: on ne contraint plus le minimum, mais il est naturellement contraint par les fonctions barrières qui tendent vers l'infini quand on arrive au bord de la contrainte. Le problème, c'est bien sur qu'on ne minimise plus la bonne quantité : pour pallier ce problème, on va faire progressivement tendre λ vers 0. On résout $\min f(x) - \lambda_n \sum_j \log(-h_j(x))$ pour différents λ_n de plus en plus proches de 0. On prend le point obtenu à la fin du n^{eme} algorithme comme point de départ du $n + 1^{eme}$ algorithme. Les différents points obtenus sont appelés point intérieurs ou, dans la littérature anglo-saxonne, central-path.

6.3 Méthode du simplexe

Dans de nombreux cas, la fonction à minimiser ainsi que les contraintes sont linéaires. Ce type de problème est appelé *programme linéaire* (PL). Il peut s'écrire

$$\begin{aligned} &\text{minimiser } \langle c, x \rangle \\ &\text{sous contrainte } Ax \leq b \end{aligned}$$

avec A une matrices, b un vecteur (qui a autant d'éléments qu'il y a de contraintes). L'ensemble contraint peut éventuellement être, de plus, borné : dans ce cas il s'agit d'un *polytope*, aussi appelé parfois *simplexe*. On appelle point extreme d'un convexe un point qui ne peut pas s'écrire comme combinaison convexe d'autres points. Les points extrêmes d'un simplexe sont appelés sommets.

Proposition. *On note $\mathcal{C}$ l'ensemble $\{x \in \mathbb{R}^d \mid Ax \leq b\}$. On suppose que $\mathcal{C}$ est non vide et compact. Dans ce cas, au moins un des sommets de l'ensemble des contraintes $\mathcal{C}$ est solution du problème*

La méthode du simplexe propose juste d'explorer successivement les différents sommets jusqu'à trouver celui qui minimise la fonction objectif. Deux problèmes avec cette méthode : elle ne marche que dans le cas de la programmation linéaire, et elle peut être longue (il peut y avoir beaucoup de sommets). On trouve les sommets en cherchant les points qui saturent certaines contraintes d'inégalités (n'importe quel d d'entre elles).

7 En dimension infinie ?*

On reprend l'exemple de l'introduction. A chaque période t, un agent a le choix de la part de revenu qu'il consomme c_t. Avec un facteur de préférence pour le présent β, et un salaire (fixe) α il résout le problème d'optimisation suivant :

$$\begin{array}{ll} \text{maximiser} & \sum_{t\in\mathbb{N}} \beta^t U(c_t) \\ \text{sous contrainte} & c_t \geq 0,\ s_{t+1} = (1+r)s_t + \alpha - c_t \geq 0 \end{array}$$

La différence majeure avec tout ce qui précède est que l'agent a le choix sur *une infinité* de variables $(c_t)_{t\in\mathbb{N}}$: l'espace E dans lequel on cherche est de dimension infinie. On voit que dans ce cas, il faut introduire une contrainte supplémentaire, pour limiter les dépenses, d'une forme qu'on n'a pas vue jusqu'ici : $\lim_{t\to\infty} s_t \geq 0$ (une contrainte de "solvabilité").

On dit que les variables sur lesquelles on peut agir (ici c_t) sont des variables de "contrôle" et que les variables sur lesquelles on ne peut pas jouer directement (ici s_t) sont des variables d'état. La branche de l'optimisation qui étudie les problèmes où il y a une infinité (voire une infinité continue) de variables de contrôle est appelée théorie du contrôle optimal.

Dans ce cas, on a deux techniques équivalentes pour résoudre le problème. Pour fixer les idées, on prend $U(x) = \log(x)$ pour la suite de cette partie.

7.1 Méthode du Hamiltonien

La solution naturelle est d'essayer d'adapter ce qu'on a vu dans le cas où il y avait un nombre fini de variables à choisir. On écrit donc $s_t = \sum_0^{t-1} \alpha(1+r)^i - \sum_0^{t-1} c_i(1+r)^{t-1-i}$. On n'incorpore pas la contrainte $c_t \geq 0$ car elle est forcément non saturée (sinon l'utilité tend vers moins l'infini). On a donc

$$\beta^t U'(c_t) = +\sum_{j\geq t} \lambda_j (1+r)^{j-t},$$

avec des multiplicateurs λ_i positifs. On voit que $U'(c_{t+1}) = (U'(c_t) - \lambda_t/\beta^t)/((1+r)\beta)$. Supposons que $(1+r)\beta > 1$, alors $U'(c_{t+1}) > U'(c_t)$, donc $c_{t+1} > c_t$ (on suppose la fonction U à rendement décroissant, ou concave). La contrainte $s_{t+1} \geq 0$ n'est donc pas saturée, donc $\lambda_t = 0$ pour tout t. Subtilité : on pourrait croire que $\beta^t U'(c_t) = 0$, ce qui n'aurait pas de sens (ça implique $c_t = \infty$), en fait, on fait "comme si" λ_∞ était non-nul, c'est la contrainte en l'infini qui est saturée. Mais on peut donc écrire $((1+r)\beta)U'(c_{t+1}) = U'(c_t)$, soit avec notre exemple $c_{t+1} = c_t\beta(1+r)$, et donc $c_t = c_0\beta^t(1+r)^t$. Pour trouver c_0, on sature la dernière contrainte, en l'infini, $s_N = \alpha((1+r)^N - 1)/r - (1+r)^{N-1}c_0/(1-\beta)$, donc $c_0 = (1-\beta)\alpha(1+r)/r < \alpha$.

Supposons maintenant que $(1+r)\beta < 1$. Soit $\lambda_t = 0$, soit $s_{t+1} = 0$. Si $\lambda_t = 0$, alors $((1+r)\beta)U'(c_{t+1}) = U'(c_t)$, et $c_{t+1} = c_t\beta(1+r)$, sinon, $c_t = (1+r)s_t + \alpha$. Dans tous les cas, $c_{t+1} \leq c_t$. Une fois que $c_t = (1+r)s_t + \alpha$, pour tout $t' > t$, $c_{t'} = \alpha$. Notons t_1 le premier temps ou $c_t = (1+r)s_t + \alpha$, on a donc $c_{t-1} = c_t/\beta(1+r)$, et par récurrence $c_0 > \alpha$, absurde, donc $t_1 = 0$, et on a tout le temps $c_t = \alpha$.

(∗) Exercice : Que se passe-t-il si on ajoute $s_0 > 0$?
(∗) Exercice : Refaire l'exercice avec $U(x) = \sqrt{x}$

7.2 Principe de Bellman

L'équation de Bellman est une équation dite de "programmation dynamique". On peut s'en servir même pour un nombre fini de contrôles, particulièrement dans les cas où l'utilisation des techniques classiques d'optimisation différentiable vues plus haut est impossible (par exemple le "problème du sac à dos"). Cette technique est basée sur la séparation du problème de manière récursive.

Notons $V(s_0)$ l'utilité optimale que l'on peut obtenir en posant comme condition $s(0) = s_0$. Alors en séparant la première périodes des autres, on obtient l'équation suivante.

$$V(s_0) = \max_{c_0} \log(c_0) + \beta V(s_0(1+r) + \alpha - c_0)$$

Résoudre cette équation fonctionnelle (l'inconnue de cette équation est une fonction, la fonction V) permet de déduire la statégie c_t optimale, mais peut être très compliqué.

7.3 Introduction à la théorie du controle optimale

Supposons maintenant qu'on essaye de trouver le maximum de

$$\int_0^\infty U(c_t, s_t, t)dt$$

sous contrainte $c_t > 0$, $\dot{s_t} = G(c_t, s_t, t)$, et $s_t \geq 0$ où U et G sont des fonctions d'utilités et d'évolution du capital. On voit que c'est une généralisation du problème précédent à des fonctions quelconques et au cas continu (on ne choisit plus une infinité discrète de variable c_t, mais une infinité continue).

On peut réécrire comme précédemment $s_t = \int^t G(c_u, s_u, u)du$. Question : si on bouge un peu un certain c_u, comment est modifié s_t ? En fait c'est une question difficile. On voit assez facilement que $\frac{\partial s_t}{\partial c_t} = \frac{\partial G(c_t,s_t,t)}{\partial c_t}dt$. En "propageant" le long de l'équation différentielle cette différence (on peut le voir en écrivant les différences infinitesimales, le faire en exercice ?) on a $\frac{\partial s_t}{\partial c_u} = \frac{\partial G(c_u,s_u,u)}{\partial c_u}du \exp(\int_u^t \frac{\partial G(c_v,s_v,v)}{\partial s_v}dv)$.
On va associer à la contrainte $s_t \geq 0$ la covariable $\lambda_t \geq 0$, on a donc l'équation suivante (en "dérivant" en c_t).

$$\frac{\partial U}{\partial c}(c_t, s_t, t) + \int_{v\geq t} \frac{\partial U}{\partial s}(c_v, s_v, v)\frac{\partial s_v}{\partial c_t} = -\int_{v\geq t} \lambda_v \frac{\partial s_v}{\partial c_t}$$

En fait comme dans le cas précédent, tous les λ_t où $s_t > 0$ sont nuls. Donc en réalité, à moins d'être à la saturation de la contrainte, on a $\lambda_t = 0$, donc pouvoir dériver le membre de gauche par rapport à t sans se soucier de $v \geq t$ (ce qui va nous permettre de nous "débarasser" des λ). On obtient, en passant au log puis en dérivant par rapport à t :

$$\frac{\partial G}{\partial c}\frac{d}{dt}\frac{\partial U}{\partial c}(c_t, s_t, t) - \frac{\partial U}{\partial s}\frac{\partial G}{\partial c}^2 = \frac{d}{dt}\frac{\partial G}{\partial c}\frac{\partial U}{\partial c} - \frac{\partial G}{\partial s}\frac{\partial U}{\partial c}\frac{\partial G}{\partial c}(c_t, s_t, t)$$

Il est très important (crucial !) de noter que $\frac{d}{dt}\ln(\frac{\partial s_v}{\partial c_t})$ *ne dépend pas de* v, c'est ce qui permet que tout ça marche !

On donne un exemple pour fixer les idées : $s'(t) = \alpha + (1 + r)s(t) - c(t)$, et on cherche à optimiser pour $U = \ln(c(t))e^{-\rho t}$, avec $r > \rho$. On obtient

$$-1 \times \frac{d}{dt}\frac{1}{c(t)}e^{-\rho t} = (1 + r)\frac{1}{c(t)}e^{-\rho t}.$$

D'ou $\frac{1}{c(t)}e^{-\rho t} = Ae^{-(1+r)t}$, et donc $c(t) = Ae^{(1+r-\rho)t}$. Pour trouver A, condition en T temps final, qu'on fait tendre vers l'infini, on veut que $s(T) = 0$. On trouve $A = \alpha/(1 + r)$.

Pour se souvenir de cette équation, un bon moyen mnémotechnique est d'utiliser le Hamiltonien : $H = U + \lambda_t G$, on obtient alors, $\lambda_t \geq 0$, $\frac{\partial H}{\partial c} = 0$ et $\frac{\partial H}{\partial s} = -\frac{d}{dt}\lambda_t$, auquel on peut rajouter $\lim s_t\lambda_t = 0$ (condition de transversalité).

Estimations et tests statistiques

1 Définitions

On commence par donner les définitions des concepts de ce chapitre :

Définition (Modèle statistique)**.** *Un modèle statistique ($\mathcal{X}$, $\mathcal{P}$) est la donnée d'un espace des observations $\mathcal{X}$, et d'une famille de mesures de probabilité sur cet espace.*

Par exemple, pour un lancer de pièce, on aurait $\mathcal{X} = \{pile, face\}$ et on peut se donner comme famille de probabilité toutes les probabilités $\mathbb{P}(face) = p = 1 - \mathbb{P}(pile)$ avec $p \in [0, 1]$. Les observations sont les réalisations d'une variable aléatoire X qui suit l'une des lois de probabilités de $\mathcal{P}$ mais cette loi est inconnue. L'un des objectif est de déterminer cette loi de probabilité à partir des observations.

Définition (Echantillon)**.** *Un échantillon est une collection de n observations.*

Remarque : Pour toute la suite, on suppose que les n observations sont indépendantes et identiquement distribuées ; ce sont n réalisations indépendantes de la même expérience. On peut faire des statistiques même hors de ce cadre : penser par exemple à l'auto-régression $X_i = \theta X_{i-1} + \xi_i$, ou aux données de réseaux. Il reste tout de même le plus simple et le plus élémentaire et on est souvent obligé de le supposer.

Définition (Statistique)**.** *Une statistique est une fonction mesurable des n observations d'un échantillon : $T : \mathcal{X}^n \to \mathcal{Y}$*

Une statistique est donc une variable aléatoire. Par exemple dans l'exemple de lancer de pièce, on enregistre $x_i = 1$ si le i^{eme} lancer donne pile et $x_i = 0$ sinon. Une statistique possible est $S(x_1, ..., x_n) = \sum_1^n x_i$, la somme de toutes les observations. On distingue deux grands types de modèles statistiques : les modèles paramétriques et non-paramétriques.

1.1 Modèles paramétriques

Une famille importante de modèles, ceux avec lesquels on va majoritairement travailler sont dits "paramétriques".

Définition (Modèle paramétriques)**.** *Un modèle est dit paramétrique si les éléments de $\mathcal{P}$ peuvent être décrits par un nombre fini de paramètres réels.*

Par exemple, si $\mathcal{P}$ est l'ensemble des lois normales, $\mathcal{P} = \{\mathcal{N}(\mu, \sigma), \mu \in \mathbb{R},\ \sigma \in \mathbb{R}_+^*\}$, les lois de $\mathcal{P}$ sont toutes décrites par deux paramètres réels. L'ensemble des lois exponentielles est décrit par un seul paramètre. Par contre, si $\mathcal{P}$ est l'ensemble de toutes les lois à densité, P n'est pas paramétrable, et donc un modèle avec cet ensemble de loi est dit non-paramétrique. Le modèle de régression $Y_i = aX_i + \xi_i$, $X, \xi \sim \mathcal{N}(0, 1)$ est paramétrique. A noter qu'un modèle de régression où on ne précise pas la forme de l'erreur est souvent appelé "semi-paramétrique". La dernière définition dont on aura besoin est celle de la vraisemblance :

Définition (Vraisemblance)**.** *On considère un modèle statistique paramétrique, paramétré par $\theta \in \mathbb{R}^d$: $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \mathbb{R}^d\}$. Alors la vraisemblance du modèle est une fonction de θ, $\theta \to L(x_1, ..., x_n, \theta)$ définie par*

— *Si les lois de notre modèles sont toutes discrètes, on définit la vraisemblance comme*

$$L(x, \theta) = \mathbb{P}_\theta(x_1)\mathbb{P}_\theta(x_2)...\mathbb{P}_\theta(x_n)$$

— *Si les lois de notre modèles sont toutes continues, on définit la vraisemblance comme*

$$L(x, \theta) = f_\theta(x_1)f_\theta(x_2)...f_\theta(x_n)$$

2 Estimation ponctuelle

On peut estimer deux types de quantités différentes : dans le cas des modèles paramétriques, on cherche à estimer le paramètre θ du modèle, qui nous permettra d'avoir accès à la distribution. Dans le cas de modèles non-paramétriques, on peut estimer des quantités liées à la distribution : sa moyenne, sa variance, ses différents quantiles. On commence par l'estimation de θ. On appelle biais d'un estimateur T la quantité $b(\theta) = \mathbb{E}_\theta(T) - \theta$ et erreur quadratique moyenne d'un estimateur $R(\theta) = \mathbb{E}_\theta[(T-\theta)^2] = \mathrm{Var}_\theta(T) + b(\theta)^2$. Ici $\mathbb{E}_\theta$ et Var_θ signifie qu'on calcule l'espérance et la variance avec la loi paramétrée par θ.

On dit qu'un estimateur $\hat{\theta}$ est *consistant* quand, pour tout $\theta \in \mathbb{R}^d$, $\hat{\theta}_n \to \theta$ quand le nombre de données tend vers l'infini

2.1 Estimateur des moments

La méthode des moments a été proposée par Karl Pearson en 1894. On suppose que $\Theta \subset \mathbb{R}^p$, et que pour tout $\theta \in \Theta$ les moments d'ordre p de X_1 existent. On note, pour tout $\theta \in \Theta$,

$$\mu_r(\theta) = \mathbb{E}_\theta(X^r) = \int x^r f_\theta(x) dx$$

On appelle estimateur des moments $\hat{\theta}^{MM} \in \Theta \subset \mathbb{R}^p$ la solution du système à p équations suivant, d'inconnue θ

$$\mu_r(\theta) = 1/n \sum x_i^k \text{ pour } k \in \{1, ..., p\}$$

Exemple : on se donne un modèle statistique où les variables $X_1, ..., X_n$ suivent la loi suivante

$$f_{a,b}(x) = \frac{1}{a} \exp(-(x-b)/a) \mathbf{1}_{x \geq b}$$

On cherche à déterminer deux paramètres a et b, donc on utilise une méthode des moments d'ordre 2. On vérifie que $\mu_1(a,b) = b + a$ et que $\mu_2(a,b) = (a+b)^2 + a^2$. On résout le système suivant

$$\begin{cases} \hat{a} + \hat{b} = \frac{1}{n} \sum X_i \\ (\hat{a} + \hat{b})^2 + \hat{a}^2 = \frac{1}{n} \sum X_i^2 \end{cases}$$

Ce qui donne $\hat{a} = \sqrt{\frac{1}{n} \sum X_i^2 - (\frac{1}{n} \sum X_i)^2}$ et $\hat{b} = \frac{1}{n} \sum X_i - \hat{a}$

$(*)$ Exercice : Déterminez l'estimateur des moments d'un n échantillon de loi uniforme sur $[0, \theta]$ où θ est le paramètre inconnu.

Estimateur des moments généralisé

Une démarche similaire à la méthode des moments peut être effectuée avec des fonctions générales $\phi_r(x)$ plutôt que les fonctions puissances x^r. Par exemple dans le cas du modèle de Cauchy, $f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$, même le moment d'ordre 1 n'existe pas, on ne peut donc pas utiliser la méthode des moments classique. Dans ce cas, on peut par exemple utiliser une fonction bornée, comme la fonction $\phi(x) = \mathrm{sgn}(x)$. On a alors $\mu_1(\theta) = \int \mathrm{sgn}(x) f_\theta(x) dx = \frac{2}{\pi} \arctan(\theta)$.

On résout donc $\frac{2}{\pi} \arctan(\hat{\theta}) = 1/n \sum \mathrm{sgn}(X_i)$ soit

$$\hat{\theta} = \tan(\frac{\pi}{2n} \sum \mathrm{sgn}(X_i))$$

2.2 Estimateur du maximum de vraisemblance et information de Fisher

On s'intéresse toujours à l'estimation paramétrique. Un estimateur du maximum de vraisemblance, est, sous réserve d'existence, une solution de l'équation

$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}}\, L(x, \theta)$$

Remarque : l'estimateur du maximum de vraisemblance peut ne pas exister ou ne pas être unique.

On essaie le plus souvent de maximiser non pas la vraisemblance mais son logarithme, ce qui est équivalent. On note $l(X, \theta)$ la log-vraisemblance, et $S(X, \theta)$ le score du modèle $S(X, \theta) = \nabla l(X, \theta)$.

Exemple : Considérons le modèle statistique de Laplace, $f_\theta(x) = \frac{1}{2}\exp(-|x-\sigma|)$

$$L(X,\theta) = \frac{1}{2^n}\prod_i \exp(-|X_i - \sigma|),$$

donc $l(X,\theta) = -\sum_i |X_i - \sigma|$. La médiane empirique des X_i est donc un des maximiseurs de la vraisemblance. Quand n est pair, il y a plusieurs maximiseurs.

Exemple : Considérons le modèle statistique de Poisson, $\mathbb{P}_\theta(X = i) = \frac{\theta^i}{i!}\exp(-\theta)$

$$L(X,\theta) = \exp(-n\theta)\prod_i \frac{\theta^{X_i}}{X_i!},$$

donc $l(X,\theta) = -n\theta + \log(\theta)\sum_i X_i - \sum \log(X_i!)$. En dérivant par rapport à θ, on conclut que seul maximiseur de la vraisemblance est donc $\hat{\theta} = \sum X_i/n$, la moyenne empirique. Remarque : dans ce cas l'EMV et l'EMM coincident.

On a nécessairement $\mathbb{E}_\theta(S(X,\theta)) = 0$, en effet $\mathbb{E}_\theta(S(X,\theta)) = \int \nabla L(x,\theta)dx = \nabla 1 = 0$ L'information de Fisher est la variance du score :

$$\mathbb{I}(\theta) = \underset{\theta}{\mathrm{Var}}(S(X,\theta))$$

Ici Var_θ est la matrice de variance-covariance de $S(X,\theta)$ sous θ.

Proposition.

$$\mathbb{I}(\theta) = -\mathbb{E}_\theta[H_l(X,\theta)]$$

où $H_l(X,\theta)$ est la hessienne de la log-vraisemblance.

Démonstration : $\int \nabla L(x,\theta)/L(x,\theta) \times L(x,\theta)dx = 0$. En dérivant on obtient

$$\int \left[\frac{\nabla L(x,\theta)}{L(x,\theta)} \times \frac{\nabla L(x,\theta)}{L(x,\theta)} + H_l(x,\theta)\right] L(x,\theta)dx = 0$$

L'information de Fisher nous permet d'avoir une borne inférieure très intéressante sur le risque quadratique des estimateurs.

Théorème 1 (Borne de Fréchet, Darmois, Cramer, Rao)**.** *Soit T un estimateur sans biais de θ. Alors*

$$\underset{\theta}{\mathrm{Var}}(T) \geq \mathbb{I}(\theta)^{-1}$$

Démonstration. Soit T un estimateur sans biais de θ. On pose $Z = T - \theta - \mathbb{I}(\theta)^{-1}S(X,\theta)$. On sait que le score est d'espérance nulle donc

$$\mathrm{Var}(Z) = \mathbb{E}(T - \theta - \mathbb{I}(\theta)^{-1}S(X,\theta))^2 = \mathrm{Var}(T) + \mathbb{I}(\theta)^{-2}\mathrm{Var}(S) - 2\mathbb{I}(\theta)^{-1}\mathbb{E}(TS(X,\theta))$$

On sait que $\mathbb{E}(TS(X,\theta)) = \int T(x)\nabla L(x,\theta)/L(x,\theta) \times L(x,\theta)dx = \int \nabla T(x)L(x,\theta) = \nabla \int T(x)L(x,\theta) = 1$, donc, comme $\mathrm{Var}_\theta(S(X,\theta)) = \mathbb{I}(\theta)$, $\mathrm{Var}(Z) = \mathrm{Var}(T) - \mathbb{I}(\theta)^{-1}$, comme une variance est toujours positive, on a le résultat. ■

Un estimateur est efficace si il atteint la borne FDCR.

Exemple : Considérons, comme ci-dessus, le modèle statistique de Poisson. On avait trouvé $l(X,\theta) = -n\theta + \log(\theta)\sum_i X_i - \sum \log(X_i!)$, donc $S(X,\theta) = -n + \sum_i X_i/\theta$. On calcule l'information de Fisher $I(\theta) = n/\theta$. Dans cet exemple, l'estimateur du maximum de vraisemblance décrit plus haut atteint la borne FDCR, et est donc efficace.

Théorème 2. *L'estimateur du maximum de vraisemblance est asymptotiquement efficace.*

2.3 Estimateur plug-in

Cette technique est surtout utilisé pour l'estimation non-paramétrique. On définit la mesure empirique des observations $x_1, ..., x_n$ comme la mesure de fonction de répartition $F_n(x) = 1/n \sum_1^n \mathbf{1}_{\{x_i \leq x\}}$. Cette mesure, qui est *aléatoire* (toutes les autres mesures qu'on a vues étaient déterministes : même si X est aléatoire, F_X était une fonction fixe, deterministe. Ici F_n dépend des X_i qui sont aléatoire, c'est donc une mesure aléatoire) se rapproche, quand la taille n de l'échantillon grandit, de la "vraie" fonction de répartition de la variable aléatoire dont nos observations sont les réalisations. C'est ce qu'explique le théorème suivant, parfois appelé théorème fondamental de la statistique.

Théorème 3 (Glivenko-Cantelli)**.**

$$||F_n - F||_\infty = \sup_x |F_n(x) - F(x)| \to 0$$

et la convergence a lieu presque-surement.

On cherche à connaître n'importe quelle fonctionnelle $\mathbb{E}_f(a(X))$ de notre loi f inconnue. On ne connait pas f, on ne peut donc bien-sur pas calculer directement l'intégrale. L'estimation plug-in consiste à estimer $E_n(a(X))$ où l'espérance est prise par rapport à la mesure empirique :

$$\mathbb{E}_n(a(X)) = 1/n \sum_1^n a(X_i)$$

Encore une fois, cette espérance empirique dépend des observations (comme la mesure empirique), il s'agit donc d'une variable aléatoire et non d'un réel. Grace à la loi des grands nombre, on sait que si l'espérance $\mathbb{E}_f(a(X))$, alors $E_n(a(X))$ converge bien presque surement vers $\mathbb{E}_f(a(X))$ quand n tend vers l'infini (on dit que l'estimateur est *consistant*)

3 Tests statistiques

Définition. *Un test est une fonction ϕ des observations $(x_1, ..., x_n)$ à valeur dans $\{0, 1\}$. $R = \{(x_1, ..., x_n) | \phi(x_1, ..., x_n) = 1\}$ est appelé la zone de rejet du test. $\bar{R}$ est appelé la zone d'acceptation.*

On se place dans le cadre de l'estimation paramétrique pour commencer. On note Θ l'ensemble dans lequel le paramètre θ prend ses valeurs.

Définition. *Faire une hypothèse nulle sur θ consiste à se donner un sous ensemble $\Theta_0 \subset \Theta$. L'hypothèse nulle s'écrit*

$$\mathcal{H}_0 : \theta^* \in \Theta_0$$

L'hypothèse alternative s'écrit $\mathcal{H}_1 : \theta^* \in \Theta_1$, où $\Theta_0 \cap \Theta_1 = \varnothing$

Exemple Détection de missile. Une des premières applications de la théorie des tests statistiques était liée au problème militaire de détection de la présence d'un missile à l'aide de radar. L'écho de radar est "grand" si un missile est présent et il est "petit" dans le cas contraire. Supposons que l'on observe une suite de valeurs $X_1, ..., X_n$ de l'écho de radar aux instants $1, ..., n$. On peut supposer que les X_i sont des variables aléatoires (à cause des effets de bruit de propagation d'ondes, erreurs de mesures, etc.), qu'elles sont i.i.d. et, plus particulièrement, on se place dans le cadre d'un modèle paramétrique.

Supposons donc que l'on connaît la famille paramétrique de fonctions de répartition $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$ de sorte que la fonction de répartition F des X_i appartient à F, i.e. $F = F_{\theta *}$ pour une valeur inconnue $\theta^* \in \Theta$ (θ^* est la vraie valeur du paramètre). Supposons aussi que l'ensemble Θ peut être décomposé en deux sous-ensembles disjoints Θ_0 et Θ_1 : de sorte que alors que $\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \varnothing$, $\theta^* \in \Theta_0$ si et seulement si un missile est présent, $\theta^* \in \Theta_1$ si et seulement si il n'y a pas de missile. Notre objectif est le suivant : à partir des observations $X_1, ..., X_n$, décider si le missile est présent (i.e. $\theta^* \in \Theta_0$) ou non (i.e. $\theta^* \in \Theta_1$).

Une hypothèse nulle ou alternative est dite *simple* si Θ_0 ou Θ_1 ne contient qu'un seul élément. Elle est dit *composite* dans le cas contraire. Il faut faire attention : dans ce formalisme, les deux hypothèses ne sont pas symétriques. L'hypothèse nulle peut souvent s'avérer plus "dangereuse" que l'alternative (comme dans notre exemple). En fait, on part de l'idée que $\mathcal{H}_0$ est vrai *à priori*, et on la considère vraie jusqu'à preuve du contraire. En économie, si on veut "prouver" que le nombre d'année d'étude a un impact sur le salaire, on écrit $Y_i = \theta X_i + \xi_i$, et on teste $\theta = 0$ (hypothèse nulle) contre $\theta \neq 0$ (hypothèse alternative). On suppose qu'il n'y a pas de lien de causalité jusqu'à preuve du contraire. Dans ce cas, l'hypothèse nulle est simple, tandis que l'hypothèse alternative est composite. De même, si l'on teste un médicament, on prend comme H_0 l'hypothèse que le médicament n'est pas efficace

3.1 Test de deux hypothèses simples

On étudie le cas le plus basique : on teste l'hypothèse $\theta = \theta_0$ contre l'hypothèse $\theta = \theta_1$.

Définition. *L'erreur de première espèce consiste à rejeter $\mathcal{H}_0$ à tort.*
L'erreur de seconde espèce consiste à ne pas rejeter (ou accepter) $\mathcal{H}_0$ à tort

On part de l'idée que $\mathcal{H}_0$ est vrai *à priori*, et on la considère vraie jusqu'à preuve du contraire. Le risque de première espèce s'écrit $\alpha(R) = \mathbb{P}_{\theta_0}(R) = \mathbb{P}_{\theta_0}(\phi(X_1, ..., X_n) = 1)$. Le risque de seconde espèce s'écrit $\beta(R) = \mathbb{P}_{\theta_1}(\bar{R})$.

Comment choisir la région critique R de manière optimale ? Il est clair qu'on veut minimiser les deux risques, on veut dans l'idéal qu'ils soient tous les deux très faible. Cependant, on ne peut pas minimiser en R les deux risques simultanément. En effet, pour minimiser le risque de première espèce il faut choisir R aussi petit que possible. Pour minimiser le risque de deuxième espèce, au contraire, il faut choisir R aussi grand que possible.

Donc, il faut chercher une méthode de choix de R permettant d' établir un compromis entre les deux risques. L'approche la plus courante est celle de Neyman – Pearson. Elle est fondée sur l'idée de dissymétrie entre H_0 et H_1. L'erreur de première espèce coûte beaucoup plus cher au statisticien : donc on commence par fixer une borne, ou un seuil pour le risque de première espèce. C'est celui dont on s'occupe en premier. On *impose* la contrainte $\mathbb{P}_{\theta_0}(R) \leq \alpha$ pour α "petit" (typiquement, $\alpha = 0.01$ ou 0.001, selon ce qu'on étudie et la "gravité" d'une "fausse découverte" etc.)

Définition. *La valeur α est appelée niveau du test : $\alpha = \mathbb{P}_{\theta_0}(R)$*

On cherche ensuite à minimiser, parmi tous les tests de niveau α, le risque de seconde espèce : c'est donc un problème d'optimisation contraint (de dimension infinie car on cherche une région de l'espace) !

Définition (Paradigme de Neyman-Pearson)**.** *On déclare optimal tout test R^* de niveau α qui atteint le minimum du risque de deuxième espèce parmi tous les tests de niveau α.*

On note parfois $\pi(R) = 1 - \beta(R)$ la puissance du test, et on cherche donc le test de niveau α le plus puissant. Ce qui est remarquable est que non seulement il existe des tests optimaux, mais en plus, dans de nombreux cas, on peut les expliciter ! C'est l'objet du théorème suivant.

Théorème 4 (Neyman-Pearson)**.** *On appelle test du rapport de vraisemblance un test dont la région de rejet a pour forme $R(c) = \{x_1, ..., x_n | LR(x_1, ..., x_n) > c\}$, où $LR(x_1, ..., x_n) = L(x_1, ..., x_n, \theta_1)/L(x_1, ..., x_n, \theta_0)$ est le* rapport de vraisemblance*. S'il existe une valeur c_α tel que $\mathbb{P}_{\theta_0}(L(x_1, ..., x_n, \theta_1) \geq c_\alpha L(x_1, ..., x_n, \theta_0)) = \mathbb{P}_{\theta_0}(R(c_\alpha)) = \alpha$, alors le test du rapport de vraisemblance à région critique $R(c_\alpha)$ est le plus puissant parmi les tests de niveau α, c'est à dire qu'il minimise le risque de deuxième espèce parmi tous les tests de niveau α.*

Les tests du rapport de vraisemblance sont donc toujours les optimaux dans le cas le plus basique où les deux hypothèses sont simples.

Exemple Considérons le modèle statistique $\{\mathcal{N}(\theta, 1)\}$. Supposons que l'on souhaite tester l'hypothèse $H_0 : \theta = 0$, contre l'alternative $H_1 : \theta = 1$ (i.e. $\theta_0 = 0,\ \theta_1 = 1$). Dans ce cas le rapport de vraisemblance vaut

$$LR(x_1, ..., x_n) = \frac{L(x_1, ..., x_n, \theta_1)}{L(x_1, ..., x_n, \theta_0)} = \exp(\frac{1}{2}[2(X_1 + ... + X_n) - 1]) = \exp(\frac{n}{2}[2\bar{X} - 1]),$$

où $\bar{X} = \frac{X_1 + ... + X_n}{n}$ est la moyenne empirique des différentes observations. Le test de Neyman-Pearson a donc pour région critique

$$R(c) = \{\exp(\frac{n}{2}[2\bar{X} - 1]) > c\}.$$

On peut aussi l'écrire sous la forme $R(c') = \{\bar{X} \geq c'\}$. On choisit maintenant la constante c' de sorte que notre test soit de niveau α, c'est-à-dire que $\mathbb{P}_0(R) = \alpha$. Comme on a $\mathbb{P}_0(R) = 1 - \phi(\sqrt{n}c')$ (ou ϕ est la fonction de répartition de la normale centrée réduite $\mathcal{N}(0, 1)$), on a l'équation $1 - \phi(\sqrt{n}c') = \alpha$, dont la solution est $c' = q^{\mathcal{N}}_{1-\alpha}/\sqrt{n}$, $q^{\mathcal{N}}_{1-\alpha}$ étant le quantile d'ordre $1 - \alpha$ de la loi normale centrée réduite. Finalement le test le plus puissant de niveau α a pour région critique $R = \{\bar{X} \geq q^{\mathcal{N}}_{1-\alpha}/\sqrt{n}\}$. On peut ensuite calculer la puissance de ce test.

Remarques :

1. On ne peut pas simultanément diminuer le risque de première espèce α et augmenter la puissance π (cf. Fig. 1).
2. Quand $n \to \infty$ le test devient de plus en plus puissant : $\pi \to 1$.

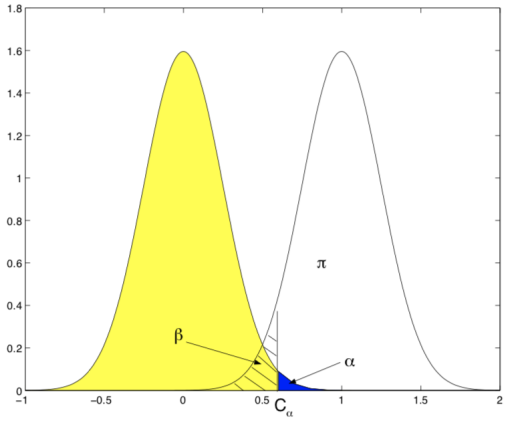

FIGURE 1 – Densité de $\bar{X}$ sous $\mathbb{P}_0$ et $\mathbb{P}_1$

Supposons pour simplifier qu'il existe une statistique à valeur réelle $T(X)$ telle que $R = \{T(X) \geq s\}$. On remarque que plus s est grand, plus le niveau du test est petit, et moins le test est puissant.

Définition. *La p-value (ou valeur critique) d'un test ϕ est, pour toute observation $x = x_1, ..., x_n$, $p(x) = \mathbb{P}_{\theta_0}(T(X) \geq T(x))$*

Intuitivement, il s'agit de dire quel serait le niveau du test si on prenait $T(x)$ comme "seuil". Pour tout $\alpha > p(x)$, rejette $\mathcal{H}_0$, pour $\alpha < p(x)$, on ne rejette pas l'hypothèse nulle.

Une p-value très faible s'interprète comme une indication contre l'hypothèse H_0, alors qu'une grande p-value indique qu'on ne peut pas rejetter avec beaucoup de certitude l'hypothèse nulle.

Points fixes et équations différentielles

Dans de nombreuses applications en macro-économie, on étudie l'évolution de quantités comme le stock de capital, la consommation, le PIB au cours du temps. Pour décrire leur évolution, on fait appel à des modèles (comme le modèle de Ramsey-Cass-Koopmans, ou le modèle de Solow) qui décrivent la variations d'une quantité au cours d'une unité de temps en fonction de cette quantité (et éventuellement d'autres quantités). Par exemple, dans le modèle de Solow, le stock de capital par tête évolue selon l'équation suivante : pour tout temps t,

$$k(t+\epsilon) - k(t) = \epsilon(sy(t) - \delta k(t))$$

En réalité il ne s'agit pas d'une équation, mais d'une infinité d'équations. En rendant ϵ proche de zéro et en prenant une production par tête courante $y(t) = k(t)^\alpha$, on tombe sur l'équation d'évolution suivante

$$\dot{k}(t) = sk(t)^\alpha - \delta k(t)$$

Si on note $k_0 = k(0)$, on peut vérifier que $k(t)^{1-\alpha} = s/\delta - (s/\delta - k_0^{1-\alpha})\exp(-(1-\alpha)\delta t)$ est bien une solution de cette équation différentielle. On peut en fait prouver que c'est la seule possible ! L'objectif de ce chapitre est de permettre aux étudiants de résoudre certaines équations différentielles simples.

1 Equation du premier ordre

Définition. *Une équation différentielle est* ***d'ordre*** k *si k est l'ordre maximal des dérivées qu'on rencontre dans l'équation* $F(x, y^{(k)}, \dots, y) = 0$

Exemple :

— $k'(t) = sk(t)^\alpha - \delta k(t)$ est une équation différentielle d'ordre 1

— $\theta^{(2)}(t) = -\omega \sin(\theta(t))$ est une équation différentielle d'ordre 2

On commence par se concentrer sur les équations différentielles d'ordre 1. On commence par se pencher sur l'existence de solutions, et on rappelle le théorème le plus fondamental sur les équations différentielles.

Théorème 1 (Cauchy-Lipsthitz)**.** *On considère l'équation sur I, pour F n'importe quelle fonction differentiable :*

$$(E)\ y' = F(y(t), t).$$

Pour tout couple (y_0, t_0) de ***conditions initiales****, il* ***existe une unique solution*** *à l'équation différentielle.*

Même si les équations différentielles d'ordre 1 sont les plus élémentaires, leur résolution peut-être très compliqué, et dans certains cas impossibles. On se limite à deux sous-cas simples :

— Les équations différentielles à variables séparables

— Les équations différentielles linéaires.

1.1 Les équations différentielles à variables séparables

Une équation est dite "à variable séparable" quand elle s'écrit sous la forme $y'(t) = f(y(t))g(t)$ avec f et g des fonctions quelconque : les dépendances de y' en y et en t sont "séparées". Dans le cas où $y'(t) = f(y(t))$, l'équation est dite "autonome".

Exemple :

— $k'(t) = \exp(-t)k(t)^\alpha$ est une équation différentielle à variable séparable

— $k'(t) = k(t)^2 - tk(t)$ n'est pas à variable séparable

Cette forme particulière permet d'obtenir des solutions. La méthode consiste à séparer les différentielles en écrivant $\frac{1}{g(y)}\frac{dy}{dx} = f(x)$ sous la forme :

$$\frac{1}{g(y)}dy = f(x)dx$$

En intégrant séparément chaque membre : :$H(y) = F(x) + K$

où H représente une primitive de $1/g$ et F représente une primitive de f et K une constante arbitraire. En outre, la fonction H est continûment dérivable, strictement monotone, donc admet une fonction réciproque de sorte que la solution s'exprime comme : $y = H^{-1}(F(x) + K)$

Exemple :

On résout l'équation différentielle suivante : $y' = -y^2 \times \exp(t)$ avec la méthode précédente : $\int y'/y^2 = -\exp(t) + K$ donc $-1/y = -\exp(t) + K$ soit $y(t) = \frac{1}{\exp(t) - K}$

1.2 Les équations linéaires d'ordre 1

Définition. *Un équation linéaire d'ordre 1 surI* $\subset \mathbb{R}$ *est une équation de la forme :*

$$(E)\ y' - a(t)y = b(t),$$

ou a et b sont continues de I vers E. b est appelé second membre, et l'équation homogène associée est :

$$(H)\ y' - a(t)y = 0.$$

On remarque que quand b est non nul, les équations linéaires ne sont pas des équations à variable séparables. On appelle ensemble des solutions toutes les fonctions vérifiants

$$\forall t \in I\ ,\ y'(t) - a(t)y(t) = b(t)$$

Théorème 2 (Principe de superposition)**.** *Considérons l'équation d'ordre* (E) *Les solutions forment un espace affine de* $C^1(I)$ *de même dimension que E. Les solutions générales du problème sont de la forme d'une somme entre une solution particulière et n'importe laquelle des solutions générales de l'équation homogène.*

1- Résoudre l'équation homogène On peut montrer grâce au théorème de Cauchy que n'importe quelle solution non-nulle d'une équation homogène ne s'annule jamais. L'équation homogène est une équation à variable séparable, donc on sait la résoudre facilement : si w est une solution de l'équation homogène, alors

$$\frac{w'}{w} = a(t)$$

c'est à dire $\ln(w(t)) = \int a(u)du + c$, où c est n'importe quel réel. On en déduit

$$w(t) = C\exp\left(\int a(u)du\right)$$

où C est une constante quelconque (qu'on ne peut fixer que grâce aux conditions initiales).

2- Méthode de variation de la constante C'est la partie plus complexe de la résolution. On cherche une solution de la forme

$$v(t) = C(t)\exp A(t);\ \text{où}\ A(t) = \int a(u)du.$$

On calcule

$$\begin{aligned} v'(t) - a(t)v(t) =& C'(t)\exp A(t) + C(t)A'(t)\exp A(t) - C(t)a(t)\exp A(t) \\ =& C'(t)\exp(A(t)). \end{aligned}$$

Donc v est une solution si $C'(t) = b(t)\exp[-A(t)]$ et il suffit donc de calculer une primitive à la fonction $b(t)\exp[-A(t)]$.

La forme générale de la solution est donc la suivante

$$y(t) = C\exp(A(t)) + \exp(A(t))\int_0^t b(u)\exp[-A(u)]du$$

2 Equations du second ordre

Ces équations sont beaucoup plus compliquées à traiter, on ne fait qu'effleurer le sujet : on ne traite les équations du second ordre que dans le cas de coefficients constants.

$$(E)\ ay" - by' - cy = g(x)$$

où a, b, et c sont des réels (et non des fonctions comme dans la section précédente).

On considère l'équation homogène

$$(H)\ ay" - by' - cy = 0$$

et on introduit le polynôme : $P = aX^2 + bX + c$ et on note Δ son discriminant.

1. Si $\Delta > 0$, l'équation caractéristique a deux racines distinctes λ et μ, et

   $$(t \mapsto \exp(\lambda t), t \mapsto \exp(\mu t))$$

   est le système fondamental de solutions.
2. Si $\Delta = 0$, l'équation a une unique racine λ, et

   $$(t \mapsto \exp(\lambda t), t \mapsto t\exp(\lambda t))$$

   est le système fondamental de solutions.
3. Si $\Delta < 0$, l'équation caractéristique a deux racines imaginaires $k \pm i\omega$, et

   $$(t \mapsto \exp(kt)\cos(\omega t), t \mapsto \exp(kt)\sin(\omega t))$$

   est le système fondamental de solutions.

Méthode de variation de la constante Considérons l'équation

$$(E)\ ay" - by' - cy = g(x)$$

On connait un système de solutions non proportionnelles w_1 et w_2. On cherche les solutions sous la forme :

$$y(t) = c_1(t)w_1(t) + c_2(t)w_2(t);\ y'(t) = c_1(t)w_1'(t) + c_2(t)w_2'(t)$$

Alors c_1 et c_2 vérifient :

$$\begin{aligned} c_1'w_1 + c_2'w_2 &= 0 \\ c_1'w_1' + c_2'w_2' &= g. \end{aligned}$$

$(*)$ Exercice : Donnez la solution v de l'équation :

$$y''(t) - 2y'(t) + y(t) = 2\exp(t)$$

telle que $v(0) = v'(0) = 1$.

Apprentissage statistique

Introduction et exemples

Avec l'avènement de l'informatique, il est devenu possible de génerer et de stocker de grandes quantités de données dans de nombreux domaines, en biologie, en économie, en robotiques. Le travail du statisticien consiste à leur donner un sens : extraire les modèles et les tendances importants, et comprendre "ce que disent les données". C'est ce que nous appelons l'apprentissage à partir des données.

Les défis posés par l'apprentissage à partir des données ont conduit à une révolution dans les sciences statistiques. Le calcul informatique jouant un rôle essentiel, il n'est pas surprenant qu'une grande partie de ces nouveaux développements ait été réalisée par des chercheurs dans d'autres domaines tels que l'informatique et l'ingénierie. Les problèmes d'apprentissage que nous considérons peuvent être grossièrement classés en deux catégories : supervisé et non supervisé. Dans l'apprentissage supervisé, l'objectif est de prédire la valeur d'un résultat Y (*outcome*, *label*, réponses) sur la base d'un certain nombre de mesures d'entrée X (*features*, imput, *predictors*, covariables, covariates, régresseurs, design, descripteurs) ; dans l'apprentissage non supervisé, il n'y a pas de mesure de résultat Y, et l'objectif est de décrire les associations et les modèles parmi un ensemble de mesures d'entrée.

On donne pour fixer les idées quelques problèmes typiques d'apprentissage :

Exemple 1 (Prédiction d'une espèce de fleur)**.** *Un des jeux de données les plus célèbre en machine learning est le jeu de données Iris de Fisher. Il a été pour la première fois utilisé en 1936 par Ronald Fisher dans son papier* The use of multiple measurements in taxonomic problems*. Le jeu de données comprends des échantillons (environ une cinquantaine dans sa version originelle, beaucoup plus dans les versions récentes). Pour chaque échantillon ont été mesurés quatre "features" : la longueur et la largeur des sépales et des pétales, exprimées en centimètres, et on a également accès à l'outcome qui est l'espèce d'Iris, qui compte parmi trois espèces distinctes (Iris setosa, Iris virginica et Iris versicolor). A partir de ces données on se pose la question suivante : comment trouver un mécanisme qui permettra de classifier efficacement une nouvelle donnée où l'on me donne les features mais pas l'étiquette ?*

On a donc affaire à un problème d'apprentissage supervisé, puisqu'on cherche à prédire une étiquette, ou un label auquel on a accès dans la base d'entraînement. Dans cet exemple, l'outcome est une variable qualitative, descriptive, une étiquette et pas un nombre, on ne peut pas ordonner les classes les unes par rapport aux autres. Quand la variable de sortie ou outcome est qualitative (ou catégorique, ou discrète), on parle d'un problème de classification.

Exemple 2 (Reconnaissance de chiffres écrits à la main)**.** *Les données de cet exemple proviennent de codes ZIP manuscrits sur les enveloppes du courrier postal américain. Chaque image est un segment d'un code ZIP à cinq chiffres, isolant un seul chiffre. Les images sont des cartes en niveaux de gris* 16×16 *chaque pixel ayant une intensité comprise entre 0 et 100.*

Les 15000 *images à notre disposition ont été normalisées pour avoir approximativement la même taille et la même orientation. La tâche consiste à prédire, à partir de la matrice* 16×16 *des intensités des pixels, l'identité de nouvelles images* $\{0, 1, ..., 9\}$ *de manière rapide et précise. S'il est suffisamment précis, l'algorithme résultant sera utilisé dans le cadre d'une procédure de tri automatique des enveloppes. Il s'agit encore d'un problème de classification, avec cette fois* 10 *catégories à la place de* 3*, et* 256 *features.*

Exemple 3 (Prédiction du salaire)**.** *On a un échantillon où ont été relevées des features (âge, niveau d'étude, domaine d'étude, taille, etc) et l'outcome qui est le salaire mensuel de notre échantillon. On veut prédire l'outcome de manière efficace pour de nouveaux individus pour lesquels on connaît les features. L'outcome n'est plus une variable qualitative mais une variable quantitative*

Noter que les features peuvent être quantitatives ou éventuellement catégorielles, c'est en tout cas bien la nature de l'outcome et pas des features qui définit la nature du problème (classification vs regression). Parler d'apprentissage non-supervisé (clustering + réduction de dimension, ACP) et d'apprentissage par renforcement/on-line learning.

Pour le moment, nous pouvons formuler la tâche d'apprentissage comme suit : étant donné un vecteur de features X, faire une bonne prédiction de la sortie Y, dénotée par $\hat{Y}$ (prononcé "y-hat"). Si Y prend des valeurs réelles, alors $\hat{Y}$ aussi ; de même pour les outcomes catégorielles.

Nous avons besoin de données pour construire des règles de prédiction, bien souvent de beaucoup de données. Nous supposons donc que nous disposons d'un *sample* $\mathcal{D}_N$ de données $(x_i, y_i)_{i\leq N}$ que l'on appelle les données d'apprentissage, avec lesquelles nous allons construire notre règle de prédiction. *Remarque : on suppose que les nouvelles données qui arrivent pour lesquelles on souhaite faire une prédiction sont distribuées selon la même loi que les données du sample.* De récents travaux de recherche proposent de sortir de cette hypothèse (transfert learning) mais c'est largement hors du cadre de ce cours. Une procédure prend en argument les données $\mathcal{D}_N$ et sort une règle de prédiction (donc une fonction !).

Qu'est ce qui distingue l'apprentissage statistique de l'économétrie ? Il y a de nombreux points communs entre les deux domaines : l'utilisation de la théorie statistique pour obtenir des informations sur un phénomène à partir de données. Les buts ne sont pas exactement les mêmes dans les deux cas : dans le premier on veut optimiser la qualité de la prédiction, et l'on s'intéresse moins que dans le second aux questions de causalité, les mesures de performance d'une procédure ne sont donc pas les mêmes (même si il y a des évolutions dans les deux domaines, qui s'entre-nourrissent et interagissent de plus en plus, cf cours de Machine Learning for Econometrics à l'ENSAE...).

Ou trouve-t-on des applications au machine learning ? Partout ! Reconnaissance d'images, Recherche sur le web, Recommandation, Publicité, Traduction automatique, Reconnaissance vocale, Voitures à conduite autonome, Santé, etc.

1 Mesure de l'efficacité d'une procédure

Pour mesurer l'efficacité d'une procédure, on doit d'abord avoir une fonction de perte (ou *loss* en anglais), qui mesure le cout de prédire $\hat{Y} = 1$ alors qu'en réalité $Y = 0$ et vice-versa, notée $\ell(Y, \hat{Y})$. Cette fonction de perte est souvent imposées par le problème, par l'industrie, mais elle peut être sujette à débat et à interprétation.

Pour la classification Une perte naturelle est la perte 0/1 suivante : $\ell(Y, \hat{Y}) = \mathbf{1}_{Y\neq\hat{Y}}$. Mais ce n'est pas la seule ! Elle présuppose que les deux types d'erreurs vaillent autant, ce qui n'est pas forcément le cas ! En général $\ell(Y, \hat{Y}) = c_0\mathbf{1}_{Y=0,\hat{Y}=1} + c_1\mathbf{1}_{Y=1,\hat{Y}=0}$.

On verra plus tard que dans le cas de la classification, on peut avoir envie de prédire $\hat{Y} \in [0, 1]$, ou même dans $\mathbb{R}$ pour plusieurs raisons (efficacité informatique, possibilité d'ajuster, augmentation de l'information, etc...) et dans ce cas, on a besoin de nouvelles pertes, la plus utilisée étant celle de la régressions logistique $\ell(Y, \hat{Y}) = \log(1+\exp(-Y\hat{Y}))$

Pour la régression La perte la plus classique est la perte L_2 qui est la plus proche de la théorie économétrique "classique" $\ell(Y, \hat{Y}) = (Y - \hat{Y})^2$, mais on peut imaginer beaucoup d'autres pertes, plus "robustes", par exemple $\ell(Y, \hat{Y}) = |Y - \hat{Y}|$, ou la perte Huber.

Prédicteur de Bayes Une fois ces fonctions de perte définies, on aimerait minimiser la quantité suivante appelée risque de la fonction f. $\mathcal{R}(f) = \mathbb{E}_{X,Y}(\ell(f(X), Y))$. On note, parmi *toutes les fonctions possibles*, f^* celle qui minimise $\mathcal{R}(f)$: $f^* = \text{argmin}_f \mathcal{R}(f)$. Cette fonction est appelée fonction oracle, ou *prédicteur de Bayes*. Le prédicteur de Bayes dépend de la loss qu'on choisit. Il atteint un risque noté $\mathcal{R}^*$.

Exercice : pour les différentes pertes, trouvez le prédicteur de Bayes. Par exemple pour la perte quadratique $\eta(x) = \mathbb{E}(Y|X = x)$ est le prédicteur de Bayes. Bien noter que le prédicteur de Bayes est parfaitement défini localement, point par point, que le comportement en un point n'influe pas celui en un autre point : c'est parce qu'on cherche le minimum parmi *toutes* les fonctions. Si on cherchait le minimum parmi les fonctions linéaires (cas de la

régression linéaire), ce ne serait plus le cas, on serait obligé de réfléchir globalement et non plus point par point.

Underfitting/Overfitting On ne peut pas atteindre le prédicteur de Bayes, parce qu'on ne connait pas la loi de (X, Y), ni la loi de $Y|X$. Plutôt que de minimiser le risque on pourrait minimiser le risque "empirique" ? $\mathcal{R}_n(f) = 1/N \sum^N \ell(f(X_i), Y_i) \simeq \mathcal{R}(f)$ (l'un tend vers l'autre par la loi des grands nombres). Voir les limites de cette approches, premières idées sur l'overfitting et l'underfitting grace à l'exemple $Y = 1.5X^3 - X^2 - .75X + 1 + \epsilon$ avec ϵ Gaussien, une perte L_2, et des classes polynomiales de plus en plus grandes. Dans le premier cas (si on prend parmi trop peu de fonctions), $\mathcal{R}_n(f) \simeq \mathcal{R}(f)$ pour toutes les fonctions de la classes mais on n'arrive pas à rendre ce risque empirique petit. Dans le deuxième cas (si on agrandit trop la classe de fonction), on arrive à rendre le risque empirique très petit, mais en général $\mathcal{R}_n(f) \neq \mathcal{R}(f)$, ce qui est plus grave et correspond à l'overfitting. Comment choisir la classe de fonctions de manière correcte ? (choisir assez de fonctions mais pas trop ?)

2 Minimisation du risque empirique.

On essaie de répondre à la question posées au dessus. Comment maîtriser la différence entre $\mathcal{R}_n(f) - \mathcal{R}(f)$?

On note $f_{erm,n,\mathcal{F}} = \text{argmin}_{f\in\mathcal{F}} \mathcal{R}_n(f)$, souvent notée simplement f_{erm} quand il n'y a pas de confusion possible. On a deux termes d'erreurs :

$$\mathcal{R}(f_{erm,n,\mathcal{F}}) - \mathcal{R}^* = \mathcal{R}(f_{erm,n,\mathcal{F}}) - \min_{f\in\mathcal{F}} \mathcal{R}(f) + \min_{f\in\mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*,$$

erreur d'estimation (premier terme) qui s'agrandit quand on agrandit la classe de fonction (cause de l'overfitting), et erreur d'approximation qui réduit quand on agrandit la classe de fonction (cause de l'underfitting). On voudrait borner le premier terme (c'est souvent ce qu'on fait en pratique, on prend une classe de fonction aussi grande que possible sans que l'erreur d'estimation explose).

Lemme.

$$\mathcal{R}(f_{erm,n,\mathcal{F}}) - \min_{f\in\mathcal{F}} \mathcal{R}(f) \leq 2 \sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_n(f)|$$

Exercice : Démonstration.

Donc il nous reste a borner la différence entre risque et risque empirique pour toutes les fonctions de la classe $\mathcal{F}$ (et on voit que ce supremum va augmenter avec $\mathcal{F}$). Comment faire ? On se place dans le cas de la classification (qui est largement majoritaire), et on commence avec le cas d'une classe $\mathcal{F}$ comportant un nombre fini de fonctions et on utilise le théorème suivant (Hoeffding), qu'on admet.

Theorem 1 (Hoeffding). *Si $U_1, ..., U_N$ sont N variables i.i.d. dans $[0, 1]$ et que $S = \sum_1^N U_k$, alors*

$$\mathbb{P}(|S - \mathbb{E}(S)| > t) < 2\exp(-2t^2/N)$$

Donc, pour *une fonction f donnée*, $\mathbb{P}(|\mathcal{R}(f) - \mathcal{R}_n(f)| > t) < 2\exp(-2t^2N^2/N) = 2\exp(-2t^2N)$. Donc en utilisant une borne d'union, $\mathbb{P}(\exists f, |\mathcal{R}(f) - \mathcal{R}_n(f)| > t) \leq 2M\exp(-2t^2N)$. Donc on peut choisir M de l'ordre de $\exp(N)$: si on veut une accuracy de 95%, on a t de l'ordre de $\sqrt{\frac{\log(40M)}{N}} \to$ forme très classique ! Une complexité en haut, une décroissance en $\sqrt{N}^{-1}$ (on peut parfois observer, dans des situations bien précises des "fast rates" de l'ordre de $1/n$).

Comment faire maintenant si on a un nombre infini de fonction dans notre classe, comme dans le cas des regressions linéaires et logistiques par exemple ? $\to$ on peut en fait remplacer la complexité $\log(40M)$ par la VC-dimension de $\mathcal{F}$ (non démontré ici) !

Définition (VC-dimension). *On dit qu'une classe de fonction $\mathcal{F}$ "pulvérise" un ensemble de données $(x_1, x_2, \ldots, x_n)$ si, pour tout étiquetage de cet ensemble de données, il existe un $f \in \mathcal{F}$ tel que la fonction f ne fasse aucune erreur dans l'évaluation de cet ensemble de données. On appellera alors dimension VC d'une classe $\mathcal{F}$ le cardinal du plus grand ensemble pulvérisé par $\mathcal{F}$.*

En notant $VC(\mathcal{F})$ la dimension VC du modèle $\mathcal{F}$, on a donc : $VC(\mathcal{F}) = \max\{k \mid \text{Card}(S) = k \text{ et } \mathcal{F} \text{ pulverise } S\}$

Exercice : Trouver la VC-dimension des classificateurs suivants :

1. $S = \{(-\infty, b] : b \in \mathbb{R}\}$.
2. $S = \{(a, b] : a, b \in \mathbb{R}\}$.

3. $S = \{(-\infty, b_1] \times (0, b_2] : b_1, b_2 \in \mathbb{R}\}$
4. $S = \{[a, b] \times [c, d] : a, b, c, d \in \mathbb{R}\}$

Souvent, la VC-dimension se confond avec la dimension. Une bonne règle est donc de prendre un espace de dimension petite devant le nombre de données.

3 Quelques méthodes globales et locales

Regression linéaire Très utilisée en économie et en économétrie. On se donne une loss quadratique, et on cherche le minimiseur du risque empirique parmi les fonctions linéaires en les features, donc le minimum de $\sum(Y_i - \beta X_i)^2$. Deux points positifs majeurs :

— On a une forme close pour β en fonction des données : $\beta_{erm} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$ (quand $N > d$ et que la matrice $(\mathbb{X}^T\mathbb{X})$ est inversible). Très pratique, facile à calculer.

— Facilement interprétable : une augmentation d'une unité de telle feature conduit à telle variation sur la prédiction.

Points négatifs : très sensible aux outliers et aux fausses données (peu robuste), ce qui conduit à de l'erreur de généralisation. Montrer un exemple $Y = X$, et des X entre 0 et 1 sauf un qui vaut 100 (alors qu'il devrait valoir 1), on obtient $\beta \sim 0$. Pas adapté à la classification (qui représente la majorité des problèmes de learning).

Regression logistique Relativement utilisée en économie. On se donne une log-loss

$$\ell(a, Y) = \log(1 + \exp(-a(2Y - 1))),$$

et on cherche à trouver un minimum du risque parmi les fonctions linéaires $\text{argmin}_\beta \sum \ell(\beta X_i, Y_i)$.

A noter que c'est une loss un peu bizarre : on a bien un Y en qui vaut 0 ou 1, mais la loss le "compare" à une valeur réelle. Cette loss en fait est une fonction convexe et dérivable qui reproduit la loss $0 - 1$.

Trois points positifs :

— Comme cette loss est convexe, même s'il n'y a pas de forme close pour β, il est solution d'un problème d'optimisation convexe, donc assez facile à trouver informatiquement.

— Assez facilement interprétable encore une fois.

— Marche très bien en pratique : la plupart de l'AI est en fait de la regression logistique.

Support Vector Machine (SVM) Supposons qu'on ait des données *séparables*. On peut donc obtenir une erreur empirique (en utilisant la loss $0-1$) de 0 en ne prenant que les séparateurs linéaires (donc une classe assez petite), et en fait on peut même obtenir plusieurs séparateurs linéaires (une infinité, faire un dessin). Comment faire pour choisir un "bon" séparateur linéaire ? On pourrait utiliser la loss logistique à la place et se ramener au cas précédent. En fait on peut faire autrement : on cherche le séparateur qui a la plus forte *marge* (en anglais margin) $\gamma = \min_i \tilde{Y}_i(\omega X_i + b)$

$$\max \gamma \ \textbf{ sous contrainte } \tilde{Y}_i(\omega X_i + b) \geq \gamma \textbf{ et } ||\omega|| = 1$$

équivalent à

$$\min ||\omega|| \textbf{ sous contrainte } \tilde{Y}_i(\omega X_i + b) \geq 1$$

On résoud, et on obtient (en utilisant le Lagrangien par exemple...) et on obtient $\omega^* = \sum \lambda_i \tilde{Y}_i X_i$ avec $\lambda_i > 0$ si et seulement si la contrainte est saturée (sur les "support vectors" à marge minimum).

Pourquoi les utiliser ? En fait ils gèrent bien l'extension à de grandes dimensions et l'ajout de features grace au Kernel Trick. On pourrait prouver (hors cadre du cours) que les λ_i s'expriment comme des fonctions des Y_i et des différents *produits scalaire* $\langle X_i, X_j \rangle$ sans jamais avoir recours aux X_j en question, contrairement par exemple à la régression logistique. L'astuce du Kernel (ou Kernel Trick) consiste à remplacer le produit scalaire $\langle X_i, X_j \rangle$ par quelque chose de plus compliqué partout qu'il apparaît (par exemple $\langle X_i, X_j \rangle^2$ ou $\exp(||X_i - X_j||^2)$). On peut montrer (cours d'advanced ML) que remplacer les produits scalaires de la sorte est équivalent à rajouter un grand nombre de features, et donc à se placer dans un nouvel espace ou les données seront séparables. Les méthodes qui ne font apparaitre que le produit scalaire permettent donc de "gerer" l'ajout de nouvelles features *de manière implicite*, en modifiant juste dans le calcul un produit scalaire par une autre fonction, sans avoir a les gérer de manière explicite.

Point positif :

— Faible complexité grace au Kernel Trick... mais encore faut-il choisir le bon kernel... .

Arbre de décision Classification and regression tree (CART)

Pratique car il peut gérer des features catégoriques (ce qui était difficile avec les trois méthodes présentées précédemment, ou alors à implémenter à la main). Arbre binaire : Chaque noeud = une condition (par exemple $x_1 < 2.3$ ou alors $x_2 =$"Iris" ou "Coquelicot"), chaque feuille = un output Y, catégorique ou quantitatif (arbre de régression dans ce cas).

Comment construire un arbre ? On choisit une feature x^j et un seuil s de sorte à ce que les ensembles $\{Y_i | X_i^j > s\}$ et $\{Y_i | X_i^j < s\}$ soient "le plus différent possible" (pas la même composition et pas qu'un point dans l'un ou l'autre, ça poserait en plus d'être trop long des problèmes de généralisations). Comment mesurer cette différence ? Quand arreter de séparer ?

Mesure de la différence : on regarde $D_> = \{(X_i, Y_i) | X_i^j > s\}$ et $D_<$, on note $p_> = \text{Card}\{(X_i, Y_i) | X_i^j > s \textbf{ et } Y_i = 1\} / \text{Card}\, D_>$ et $p_<$. On essaie ensuite de minimiser $\text{Card}\, D_> I(p_>) + \text{Card}\, D_< I(p_<)$ où I est une mesure de l'incertitude, nulle en 0 et en 1 et maximum en 1/2, par exemple $I(p) = p(1-p)$ ou $I(p) = p\log(p)$. On choisit le split avec la plus faible incertitude, et on recommence avec les deux branches, jusqu'à ne plus avoir de nouvelles données.

→ Problème pour calculer : il faut explorer à chaque fois toutes les features et tous les seuils possibles... Souvent en réalité tirés au hasard quand il y a un grand nombre de features.

Random forest et bagging ? Peu de théorie, fait empirique. On tire à partir du sample $\mathcal{D}_n$ avec remise des nouveaux sample $\mathcal{D}_n^1$, $\mathcal{D}_n^2$, ..., on construit à partir de chacun un prédicteur (donc un arbre dans notre cas, mais en fait on peut faire du bagging avec n'importe quelle classe), et ensuite on fait voter ces prédicteurs (ou en prend la moyenne dans le cas de la regression). Ca marche très bien en pratique, gros gain d'erreur de généralisation, mais plus difficile à interpréter (alors qu'un des gros avantages des arbres est leur interprétabilité) .

Neural Network Les neural network permettent d'introduire de la non-linéarité. Ce sont les ERM d'une classe spécifique. Dessiner NN. $f_j^{(k)}(x) = \sigma(\sum_i \omega_{ij}^{(k)} f_j^{(k-1)}(x))$ pour la k^{th} couche, où σ est une fonction d'activation non-linéaire (par exemple $\max(0, x)$, $1/(1+\exp(x))$, ou d'autres).

Expliquer la backpropagation.

Problème : cette classe a un grand nombre de paramètre, une énorme VC-dimension, et pourtant l'erreur de généralisation est faible. Les Reseaux de neurones semble aller au delà de l'overfitting, et bénéficier de régularisation implicite dans la manière de les optimiser (très avancé et au delà du cadre de ce cours).

Deuxième problème : Optimisation non-convexe donc même avec une descente de gradient, on n'est pas sur de trouver l'optimum global... et ça marche quand même. Encore une fois de nombreuses théories pour expliquer ça, pas de consensus.

Nearest Neighbours et Plug-in methods Pas une méthode globale, mais une méthode locale, aussi appelée "plug-in". En fait on essaie d'estimer $\eta(x) = \mathbb{E}(Y|X = x)$ au point où on a la nouvelle donnée. Regressogramme le plus simple pour expliquer le trade-off biais/ variance underfitting/overfitting. Plus on agrandit les cases, plus on introduit de biais, mais plus on a de données par case, en fait on a moins de paramètre à estimer.

Problème :

— empty bins,

— la partition choisie ne dépend pas des données, on a envie d'avoir plus de bin là ou il y a plus de données.

→ KNN. On prend les k plus proches voisins de la nouvelle donnée. Marche bien en pratique : k trop petit, overfitting, k trop grand underfitting. Problème : la frontière de décision est compliquée.

Mentionner les Nadaraya-Watson Kernel.

4 Choisir un modèle

La question reste de savoir comment choisir en pratique quels paramètres choisir quand la VC-dimension est difficile à calculer, où même qu'elle n'est pas pertinente (exemple des neural network). On veut savoir comment sélectionner les bon "hyper-paramètres" d'un modèle, par exemple nombre de neurones à mettre par couche, bon Kernel à choisir pour un SVM, quelle profondeur choisir pour un arbre, comment choisir le k dans les k nearest

neighbours. On n'essaie pas de trouver les paramètres mais de choisir combien de paramètre on va estimer : → hyperparamètre.

Souvent on a un split train/test, on peut donc regarder l'erreur de généralisation d'un modèle sur le test set. **Mais on ne peut pas choisir le meilleur modèle en se basant sur une faible test error !** En fait on veut une faible erreur de generalisation $\mathcal{R}(f)$, et $\min_f \sum_{test} \ell(Y_i, f(X_i))$ n'est pas $\min_f \mathcal{R}(f)$: on risque un overfitting aux données de test !

Cross-validation k-fold cross validation :

— Découper l'échantillon $\mathcal{D}_n$ en k sous-ensembles de données.

— Apprendre chaque modèle sur $k-1$ sous ensemble, tester le modèle sur le sous-ensemble restant.

— Choisir le modèle avec la plus petite erreur moyenne (moyenne parmi K).

Pénalisation On veut éviter l'overfitting dans une regression ou dans une regression logistique avec un grand nombre de features (par exemple $d > N$, catastrophique). On peut sélectionner "à la main" des variables pertinentes, mais ça peut prendre du temps, de l'expertise, biais dans la sélection, etc. On peut aussi forcer le modèle de regression à être simple automatiquement, et à avoir la plupart de ses coefficients nuls ou faibles.

Least Absolute Shrinkage and Selection Operator (LASSO)

Idée : on "pénalise" le fait d'avoir beaucoup de coefficient non-nuls :

$$\min_\beta \sum_i \ell(Y_i, \beta X_i) + \lambda ||\beta||_0,$$

où $||\beta||_0$ est le nombre de coefficient non-nuls de β et λ un paramètre (ou plutôt un hyperparamètre, à choisir par cross-validation) de trade-off entre précision du modèle et simplicité du modèle. λ très faible = overfitting, (presque) pas de pénalisation, λ trop fort au contraire pénalise trop les coefficients non nuls et on choisit toujours $\beta = 0$, underfitting.

Problème, on ne peut pas résoudre ce problème d'optimisation non-différentiable, non convex avec des techniques usuelles. Solution : on minimise plutôt

$$\min_\beta \sum_i \ell(Y_i, \beta X_i) + \lambda ||\beta||_1,$$

c'est le LASSO. ATTENTION, TRES IMPORTANT de NORMALISER les données avant d'utiliser le LASSO, sinon on met à 0 des coefficients très importants mais faibles parce que les données étaient énormes (écrites en unité petites, cm vs m). Sélection automatique de variables.

Ridge Regression

Même idée en remplaçant $||\beta||_1$ par $||\beta||_2$: ici on ne veut plus beaucoup de coeffs nuls mais plutot tous les coefs faibles, pas de trop gros coef.

Méthode du coude (heuristique)
Considérer des modèles de plus en plus complexes, tant que l'erreur d'apprentissage diminue beaucoup. Très utile pour l'apprentissage non-supervisé, surtout pour le clustering. (within cluster SoS vs number of cluster) !

Annexes

1 Éléments d'intégration

1.1 Integration simple

On rappelle ici comment intégrer certaines fonctions usuelles.

On se donne une fonction f, continue par morceaux. On dit que F est une primitive de f si, pour tout x, $F'(x) = f(x)$. Toute fonction f a une infinité de primitives, séparées les une des autres par des constantes : si F_1 et F_2 sont des primitives de f alors $F_1 = F_2 + C$ où C est une constante. On joint un tableau des primitives usuelles :

Fonction f	Fonction primitive F	Intervalle
x^n, $n \in \mathbb{N}$	$\dfrac{x^{n+1}}{n+1}$	$\mathbb{R}$
$\dfrac{1}{x}$	$\ln \lvert x \rvert$	$]-\infty, 0[$ ou $]0, +\infty[$
$\dfrac{1}{x^n} = x^{-n}$, $n \in \mathbb{N} \setminus \{1\}$	$\dfrac{x^{-n+1}}{1-n} = \dfrac{1}{(1-n)x^{n-1}}$	$]-\infty, 0[$ ou $]0, +\infty[$
x^α, $\alpha > 0$	$\dfrac{x^{\alpha+1}}{\alpha+1}$	$]0, +\infty[$
e^x	e^x	$\mathbb{R}$
$\cos x$	$\sin x$	$\mathbb{R}$
$\sin x$	$-\cos x$	$\mathbb{R}$
$\cosh x$	$\sinh x$	$\mathbb{R}$
$\sinh x$	$\cosh x$	$\mathbb{R}$
$\dfrac{1}{\sqrt{1-x^2}}$	$\arcsin x$	$]-1, 1[$
$\dfrac{1}{1+x^2}$	$\arctan x$	$\mathbb{R}$

La relation fondamentale de l'intégration est la suivante : si F est une primitive de f, alors, pour $b > a$

$$\int_a^b f(t)dt = F(b) - F(a) = [F]_a^b$$

On donne quelques propriétés usuelles de l'intégration :

— Pour deux fonctions f et g, $\int_c^d f + g = \int_c^d f + \int_c^d g$.

— Pour f une fonction et a un nombre réel, $\int_c^d af = a\int_c^d f$.

— Si f est une fonction positive, alors $\int_c^d f \geq 0$

Changement de variable Parfois pour calculer une intégrale, on a besoin de faire un changement de variable. La formule est la suivante : si v est une fonction dérivable, n'importe laquelle, alors,

$$\int_a^b f(v(x))v'(x)dx = \int_{v(a)}^{v(b)} f(u)du$$

Par exemple

$$\int_1^2 \frac{1}{3t-1}3dt = \int_2^5 \frac{1}{t}dt = \log(5) - \log(2)$$

1.2 Intégration de plusieurs variables

On a parfois besoin de calculer des intégrales multiples, de la forme : $\int_a^b (\int_c^d f(x,y)dy)dx$
Pour cela on rappelle un théorème un théorème utile d'interversion :

Théorème 1 (Fubini)**.** *Si f est une fonction de deux variables intégrable, alors,*

$$\int_a^b (\int_c^d f(x,y)dy)dx = \int_c^d (\int_a^b f(x,y)dx)dy.$$

Par exemple, supposons qu'on veuille calculer l'intégrale suivante : $\int_0^1 \int_0^1 x^y dydx$, on utilise le théorème de Fubini pour intervertir les intégrales et on obtient $\int_0^1 1/(y+1)dy = \log(2)$

2 Éléments d'algèbre linéaire

2.1 Hessienne

On rappelle d'abord le principe de la formule de Taylor pour une dimension :

Proposition (Formule de Taylor)**.** *Soit f une fonction deux fois dérivable en $a \in \mathbb{R}$, on a*

$$f(a+\epsilon) = f(a) + f'(a)\epsilon + \frac{f''(a)}{2}\epsilon^2 + o(\epsilon^2)$$

Ici on approxime notre fonction f par une fonction quadratique, plus compliquée qu'une fonction affine, mais qui approxime mieux la fonction f. En effet, le "reste" de notre approximation linéaire, $f(a+\epsilon) - (f(a) + \epsilon f'(a))$, était comparable à ϵ^2, alors que le reste de notre approximation quadratique est $o(\epsilon^2)$ (négligeable par rapport à ϵ^2). Maintenant, on voudrait étendre ces résultats aux cas de fonctions de plusieurs variables. On est donc là aussi obligé d'affiner notre approximation linéaire en approximation quadratique.

Autrement dit, si on note notre perturbation $\epsilon = (\epsilon_1, ..., \epsilon_n)$, on aimerait passer d'une approximation de la forme

$$f(a+\epsilon) = f(a) + \epsilon_1 \frac{\partial f}{\partial x_1}(a) + ... + \epsilon_n \frac{\partial f}{\partial x_d}(a) + o(||\epsilon||)$$

à quelque chose de la forme

$$f(a+\epsilon) = f(a) + \epsilon_1 \frac{\partial f}{\partial x_1}(a) + ... + \epsilon_n \frac{\partial f}{\partial x_d}(a) + \sum_{i,j} \alpha_{i,j} \times \epsilon_i \epsilon_j + o(||\epsilon||^2)$$

avec des $\alpha_{i,j}$ à déterminer. La formule de Taylor multidimensionnelle nous permet d'avoir un résultat de ce type. On commence par définir l'équivalent multidimensionnel d'une dérivée seconde

Définition. *Soit f une fonction définie sur une partie D de $\mathbb{R}^p$, deux fois différentiable en a*[6]*. On appelle* ***Hessienne*** *de f en a la matrice $H_f(a)$ définie par :*

$$H_f(a) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\right]$$

où $\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial}{\partial x_i}(\frac{\partial f}{\partial x_j}(a))$ est la dérivée partielle par rapport à x_i de la dérivée partielle de f par rapport à x_j.

6. On ne rentre pas dans le détail de ce que cette hypothèse signifie. Il suffit de savoir que si les d^2 dérivées doubles existent toutes et sont toutes en plus continues, alors la fonction est deux fois différentiable. La plupart des fonctions usuelles qu'on manipulera vérifieront cette hypothèse.

En fait l'ordre dans lequel on dérive importe peu, comme l'affirme la proposition suivante parfois appelée théorème de Schwartz.

Proposition. *Si la fonction f est deux fois différentiable en a, alors la Hessienne est symétrique, c'est-à-dire :*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

($*$) Exercice : Calculer la Hessienne de $f(x_1, x_2) = \sqrt{x_1 x_2}$.

On peut maintenant donner une version multi-dimensionnelle de la formule de Taylor :

Proposition. *Soit f une fonction définie sur une partie D de $\mathbb{R}^p$, deux fois différentiable en $a \in R^d$, et $\epsilon \in \mathbb{R}^d$ une perturbation.*

$$f(a+\epsilon) = f(a) + \sum_{i=1}^{p} \epsilon_i \frac{\partial f}{\partial x_i}(a) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \epsilon_i \epsilon_j \frac{\partial^2 f}{\partial x_j \partial x_i}(a) + o\left(\sum_{i=1}^{p} \epsilon_i^2\right)$$

ou bien en notation matricielle

$$f(a+\epsilon) = f(a) + \langle \nabla f(a) | \epsilon \rangle + \frac{1}{2} \epsilon^T H_f(a) \epsilon + o\left(\|h\|^2\right)$$

2.2 Positivité d'une matrice

Pour montrer qu'une fonction d'une variable réelle est convexe, il suffit de montrer que sa dérivée double est toujours positive. Pour les fonctions de plusieurs variables, on aimerait avoir une caractérisation semblable. Le rôle de la dérivée seconde étant jouée par la Hessienne, se pose la question suivante : quand peut-on dire qu'une matrice est positive ? La définition suivante donne une réponse.

Définition. *On dit qu'une matrice carrée symétrique M de taille d est* définie positive *quand pour tout vecteur $u \in \mathbb{R}^d$ non nul, $u^T M u > 0$*

Exemple : Montrez que la matrice $M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ est définie positive. Qu'en est-il de la matrice $M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$?

La deuxième condition demande donc que la Hessienne en x^* soit définie positive, ce qui amène la question naturelle : comment vérifier qu'une matrice est définie positive ? Réponse : une matrice est définie positive si et seulement si toutes ses valeurs propres sont strictement positives. Les valeurs propres d'une matrice sont en général difficiles à trouver, donc on peut utiliser des caractérisations plus simple :

- Pour M une matrice 2×2, M est définie positive si et seulement si $\mathrm{Tr}(M) > 0$ *et* $\det(M) > 0$
- Pour les matrices de plus grandes tailles, le *critère de Sylvester* peut être utile : Pour qu'une matrice symétrique $A = (a_{ij})_{1 \leqslant i,j \leqslant d}$ soit définie positive, il faut et suffit que les d matrices $A_p = \left(a_{ij}\right)_{1 \leqslant i,j \leqslant p}$ pour p allant de 1 à d, aient leur déterminant strictement positif