Mathématiques générales 2

Objectif de préparation :

Décortiquer le cours et arriver en séance avec :

1. **les points incompris et des questions précises associées : faites un tableau avec l'endroit approximatif du problème dans une première colonne et la question associée dans une seconde colonne.**
2. **des réponses aux questions de fin de paragraphe**
3. **des réponses aux qcm de la séance.**

Les parties avec $\star$ sont des parties hors programme motivant le cours et sont donc facultatives. En séance, vous poserez vos questions à vos camarades et vous répondrez aux leurs. Le mot d'ordre pour ce travail est FAIRE L'EFFORT DE COMPRENDRE EN DETAIL.

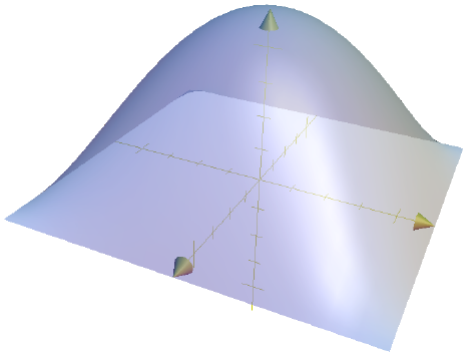
Le cours est construit autour d'un problème fil rouge : ce problème sera traité au fur et à mesure du chapitre en fonction de l'arrivée de nouveau concept. L'objectif de ce problème est de vous faire voir, partant d'un problème concret de départ, comment sont utilisés les mathématiques et particulièrement les concepts de l'UE. Il est l'illustration d'une branche des mathématiques appliquées très prisée actuellement. Ce problème est inévitablement compliqué car il mêle beaucoup de concepts mathématiques. Pour cette raison, il ne figurera pas à l'évaluation de l'UE. S'il s'avère trop compliqué voire décourageant, il vous est possible de ne pas le considérer et d'aller directement aux objectifs du cours : les sections le traitant ont dans leur titre le symbole $\star$.

1 $\star$ Problème fil rouge du cours $\star$

1.1 $\star$ Problème de membranes élastiques $\star$

1.1.1 $\star$ Des problèmes concrets $\star$

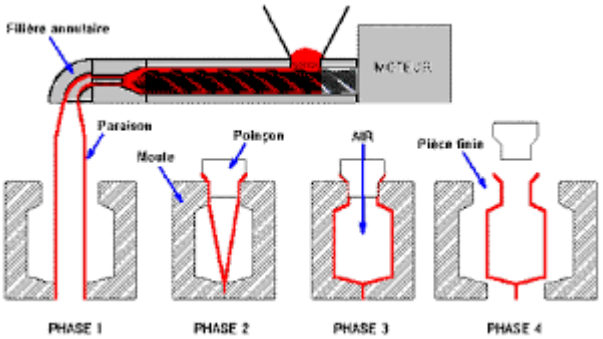
Nous nous intéressons au problème suivant : une membrane carrée plane a ses bords fixés à une altitude 0, on applique une force verticale en un point de la membrane. Quelle est l'altitude $u(x)$ de la membrane en chaque point x du domaine ?



Pourquoi se poser ce genre de questions ?

Il existe de nombreux domaines pour lesquels ce genre de problème a de l'importance.

Voici un exemple industriel qui met en oeuvre la déformation de membranes élastiques : la fabrication de bouteilles en plastique par extrusion soufflage. Des tubes en plastique préalablement fabriquées sont introduites dans un moule, de l'air sous pression est ensuite injecté dans la bouteille pour déformer le plastique afin qu'il épouse les contours du moule. Pour un descriptif plus poussé, consulter la vidéo : `https://www.youtube.com/watch?v=wCWG58e1dC4`



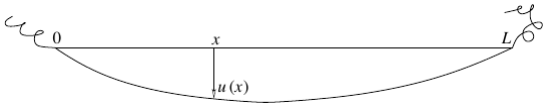
La biomécanique contient également des problématiques de membrane élastique. Cette discipline a pour objectif la compréhension de la mécanique du corps humain afin de détecter d'éventuelles pathologies chez des patients. Le corps humain est rempli de membrane élastiques. Le coeur pompe et rejette du sang dans le corps humain via les artères : ces artères subissent des variations de pression brutales entraînant de légères déformations de leur membrane. Pour implanter des microvalves cardiaques palliant un dysfonctionnement du coeur, il faut étudier finement quelle pression imposer afin de préserver les artères du patient. Les globules rouges sont des disques remplis d'hémoglobine également constituée d'une membrane élastique. Cette membrane leur confère une grande élasticité et la capacité à adopter des formes permettant une viscosité faible du sang. Les veines, l'arbre bronchique sont d'autres d'organes faisant intervenir des membranes élastiques soumises à des forces intérieures.

Pourquoi le faire sur une membrane carrée alors que les applications concernent des membranes avec des géométries beaucoup plus sophistiquées ?

Les mathématiciens ont toujours pour principe de partir de modèles très simples, de les décortiquer en profondeur avant de passer à plus compliqué. En l'occurrence avant de s'intéresser à une géométrie complexe, il est plus que raisonnable de voir si on arrive à s'en sortir avec des géométries simples (qui peuvent mener malgré tout à des problèmes mathématiques compliquées). Par ailleurs, il ne faut pas se tromper de métier : le matheux va élaborer une théorie, la tester, la valider sur un certains nombres de cas tests (benchmarks) vus comme référence par la communauté. Si sa théorie est validée, elle pourra être proposée à l'entreprise, au médical afin que des ingénieurs mathématiciens et des programmeurs l'adaptent au cas particulier qui les intéresse.

1.1.2 ⋆ Modélisation du problème de la membrane ⋆

On considère le cas unidimensionnel d'un fil élastique attaché aux bords du domaine à l'altitude $u(0) = 0 = u(L)$. On fixe L à 1 pour simplifier la situation. Une force F est exercée verticalement sur le fil : la force en chaque point x est donnée par $F(x)$. On note u la fonction donnant l'altitude $u(x)$ de la corde en chaque x. Ce qui nous intéresse, c'est cette altitude à l'équilibre, une fois la force appliquée.



L'équation différentielle vérifiée par u, pour T la tension du fil, est donnée par

$$\begin{cases} -u''(x) = \dfrac{F(x)}{T}, & \forall x \in]0,1[\\ u(0) = 0 \\ u(1) = 0 \end{cases}$$

Si vous êtes curieux de la physique menant à cette équation, l'explication est donnée dans la partie 1.1 du cours suivant : `https://www.ljll.math.upmc.fr/ledret/M1/ComplementsM1ApproxEDP.pdf`

Pour la suite du cours, nous compliquons légèrement le problème. Nous traitons ainsi une classe de problèmes englobant celui de la membrane (qui correspond à $c = 0$). Le problème étudié est le suivant pour f et c deux fonctions définies sur $[0, 1]$:

$$(P) : \begin{cases} -u''(x) + c(x)u(x) = f(x), & \forall x \in]0,1[\\ u(0) = 0 \\ u(1) = 0 \end{cases}$$

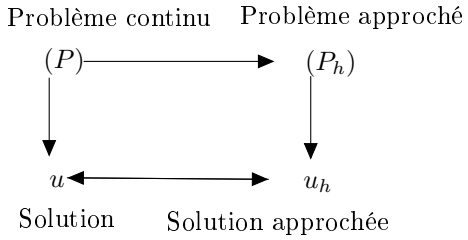
1.1.3 ⋆ Comment résoudre cette équation différentielle ? ⋆

Là vous commencez à toucher les limites des équations différentielles dont on peut obtenir une solution explicite. Si c est constante, il est possible de trouver la solution de ce système. Pour cela, vous déterminez les solutions de l'équation homogène puis on peut trouver une solution particulière en utilisant la méthode de la variation de la constante. Cependant si c se met à varier, l'équa diff change peu mais suffisamment pour qu'on ne puisse plus la résoudre. On est alors bloqué.

Le mathématicien a alors deux réflexes. Le premier est de s'assurer que le modèle a bien une solution. Vous allez me dire comment faire pour assurer l'existence d'une solution si on n'a pas son expression. En réalité, les mathématiciens ont développé des outils permettant d'assurer l'existence et l'unicité d'une solution sans savoir l'écrire. Un résultat fondamental pour cela est **le théorème de Cauchy-Lipschitz**. Il est hors programme, j'en parle juste pour la culture.

Le second réflexe qui motivera l'UE est le suivant : certes on ne peut pas résoudre le problème analytiquement mais n'y a-t-il pas moyen "d'approcher l'équation différentielle par une équation qui elle serait résolvable" ? On entre alors dans le domaine de **l'analyse numérique**, domaine en plein expansion depuis l'avènement de l'ordinateur.

Plus précisément, on part de notre problème (P) dont on veut trouver la solution u. L'idée est de créer un problème (P_h) qui approche (P) (quand le paramètre h est proche de 0, (P_h) est proche de (P)) et dont la solution u_h est facile à trouver. On peut alors espérer (vérifier) que lorsque h tend vers 0, u_h est proche de u et on a une estimation proche de la vraie solution u.



Pour ce cours, la méthode d'approximation que nous utiliserons s'appelle la méthode des éléments finis.

1.2 ⋆ La méthode des éléments finis : formulation variationnelle ⋆

L'objectif de cette partie est de vous donner les fondements de la méthode des éléments finis sur un exemple. Cette méthode est hors programme de l'UE donc ne paniquez pas si c'est trop compliqué. J'en parle dans ce cours car c'est un très bel exemple d'application de beaucoup de concepts que vous verrez dans ce chapitre et de beaucoup de concepts déjà vus (espaces vectoriels, applications linéaires, équations différentielles ...). Par ailleurs, si vous continuez les maths dans vos formations futures, vous serez certainement amenés à l'utiliser.

La méthode d'approximation du problème se base sur une réécriture du problème (P) sous une forme différente : la formulation variationnelle. C'est cette formulation variationnelle qui sera approximée (on dit **discrétisée**) dans une section ultérieure.

Obtenir la formulation variationnelle :

Soit $v \in C^1([a,b])$ telle que $v(0) = v(1) = 0$, intégrons l'équation contre v. Comme $-u''(x)+c(x)u(x) = f(x)$, on a

$$\int_0^1 -u''(x)v(x) + c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

et donc par linéarité de l'intégrale

$$\int_0^1 -u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

Comme $v \in C^1([a,b])$, on peut intégrer par parties la première intégrale. On obtient

$$-[u'(x)v(x)]_0^1 + \int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

et donc

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

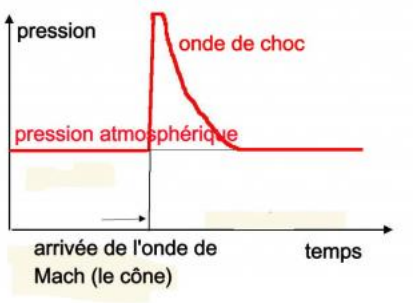
C'est ce qu'on appelle la formulation variationnelle de (P). Il se trouve qu'on peut relaxer les hypothèses qu'on a faites sur u et v tout en gardant cette égalité vraie. On verra dans la suite que pour que les intégrales précédentes aient un sens, il suffit que u, v, u' et v' soient de carré intégrables ($\int_0^1 f(t)^2 dt$ converge). Nous noterons

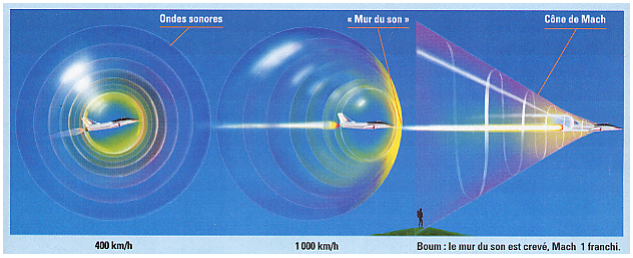
$$H_0^1([0,1]) = \left\{ f \in \mathcal{F}([0,1],\mathbb{R}) | \int_0^1 f(t)^2 dt < +\infty, \int_0^1 f'(t)^2 dt < +\infty, f(0) = f(1) = 0 \right\}$$

Mais pourquoi fait-on cela ? ?

Il est à ce stade naturel de se demander quel est l'intérêt de transformer le problème initial en une égalité portant sur des intégrales faisant intervenir une fonction v inconnue.

Jusqu'ici, on vous a toujours dit "Si tu veux que la dérivée existe, il faut que la fonction soit dérivable". De même si vous vouliez résoudre une équation différentielle d'ordre 1 représentant un phénomène physique donné, il fallait que la fonction soit au moins dérivable. Tout ceci part du postulat que dans la vraie vie, les phénomènes physiques sont bravement continus, voire dérivables. Or ceci est faux. Dans la nature il existe de nombreux phénomènes discontinus notamment ce qu'on appelle communément les ondes de chocs qui interviennent par exemple en mécanique des fluides (exemple : explosions nucléaires, traffic routier, mur du son).





Or une fonction dérivable ne peut pas représenter ce type onde qui sont discontinues. Il a donc fallu inventer des outils mathématiques permettant de représenter des phénomènes physiques non réguliers (non continus). C'est ce qu'a fait un mathématicien nommé Laurent Schwartz en créant la notion de dérivée au sens faible appelée distribution (dérivée de fonction non classiquement dérivable). Dès lors, les mathématiciens ont pu écrire des modèles mathématiques intégrant des données physiques discontinues. Alors finalement à quoi correspond la formulation variationnelle ? La formulation variationnelle est construite pour être l'équivalent au sens faible de $-u''(x) + c(x)u(x) = f(x)$.

Résumons en quelques mots, les modèles mathématiques (équations différentielles) fabriqués jusqu'ici et exigeant des hypothèses assez fortes sur la solution ne permettent pas de décrire des phénomènes physiques discontinus qui existent dans la nature. On écrit donc des modèles "au sens faible" qui contiennent le modèle initial et qui autorisent des solutions bien plus biscornues.

Revenons à des choses plus terre à terre et observons la structure de la formulation variationnelle :

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

Celle-ci se récrit sous la forme

$$a(u,v) = l(v)$$

pour

$$a: \begin{array}{l} H_0^1([0,1])^2 \longrightarrow \mathbb{R} \\ (u,v) \mapsto \displaystyle\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx \end{array}, \qquad l: \begin{array}{l} H_0^1([0,1]) \longrightarrow \mathbb{R} \\ v \mapsto \displaystyle\int_0^1 f(x)v(x)dx \end{array}$$

Ce sont deux applications qui sont définies sur des ev de fonctions et qui sont à valeurs dans $\mathbb{R}$. L'application l est linéaire et l'application a est bilinéaire cad linéaire par rapport à la première et la seconde variable séparément. Voici donc deux applications linéaires ou bilinéaires apparaissant dans un problème très concret.

▸ **Exercice 1**. Démontrer que l est bien une application linéaire et que a est bien bilinéaire.
Si on y regarde de plus près à vrai dire, on se rend compte que les trois intégrales sont de la forme $\displaystyle\int_0^1 f(x)g(x)dx$ pour f et g deux fonctions. L'application

$$(f,g) \mapsto \int_0^1 f(x)g(x)dx$$

est une application bilinéaire spéciale. Il s'agit en réalité de ce qu'on appelle un produit scalaire. Pourquoi est-ce important ? Car cette propriété à la base d'un théorème qui permet d'assurer l'existence d'une solution u à la formulation variationnelle. Ainsi on s'assure que notre modèle a un intérêt : on l'a construit pour représenter un phénomène physique (la déformation d'une membrane), s'il n'avait pas de solution, il pourrait difficilement reproduire la solution physique qui elle existe bien ! Il est donc indispensable de se pencher dès maintenant sur ce qu'on appelle un **produit scalaire**. Nous reviendrons régulièrement sur le problème de la membrane.

2 Produit scalaire et espaces préhilbertiens/euclidiens

2.1 Définitions propriétés

2.1.1 Le produit scalaire que vous connaissez

Plaçons-nous sur $\mathbb{R}^2$ et donnons-nous deux vecteurs $u = (u_1, u_2)$, $v = (v_1, v_2)$. Depuis le lycée, vous avez-vu sur $\mathbb{R}^2$ ces deux formules que vous avez appelé "produit scalaire" :

$$\langle u, v\rangle = ||u||.||v|| \cos(u,v) = u_1v_1 + u_2v_2.$$

On note que cette quantité s'annule soit si un des vecteurs est nul, soit si le cosinus s'annule cad si u et v sont orthogonaux. Ce produit scalaire est donc un bon outil pour mesurer l'orthogonalité de deux vecteurs : il suffit de regarder s'il s'annule.

On souhaiterait généraliser cette notion de produit scalaire à des espaces vectoriels plus exotiques. Après tout la notion d'orthogonalité existe aussi dans $\mathbb{R}^3$. L'extension qui semblerait naturelle du produit scalaire dans $\mathbb{R}^3$ serait la suivante :

$$\forall u = (u_1,u_2,u_3) \in \mathbb{R}^3, v = (v_1,v_2,v_3) \in \mathbb{R}^3, \quad \langle u, v\rangle = u_1v_1 + u_2v_2 + u_3v_3.$$

De même, si on voulait étendre à $\mathbb{R}^n$ cette expression, il semblerait naturel de poser :

$$\forall u = (u_1,\dots,u_n) \in \mathbb{R}^n, v = (v_1,\dots,v_n) \in \mathbb{R}^n, \langle u, v\rangle = \sum_{i=1}^n u_iv_i$$

Allons plus loin : vous avez découvert que $\mathbb{R}^n$ n'est qu'un cas particulier d'une notion plus vaste appelé espace vectoriel : on vous a présenté des espaces vectoriels de polynômes comme $\mathbb{R}[X]$, de matrices comme $\mathcal{M}_n(\mathbb{R})$, de fonctions comme $\mathcal{F}(\mathbb{R},\mathbb{R})$... En mathématiques, on aime généraliser le cas particulier pour mieux le comprendre. En ce sens serait-il possible de définir une notion de produit scalaire générale pour un ev donné ? Pour cela, il est nécessaire de comprendre quelles propriétés satisfont les produits scalaires ci-dessus.

Considérons le produit scalaire défini sur $\mathbb{R}^2$: il est **symétrique** cad que le sens du produit scalaire n'a pas d'importance

$$\langle u, v\rangle = u_1v_1 + u_2v_2 = v_1u_1 + v_2u_2 = \langle v, u\rangle.$$

Par ailleurs, l'application $u \in \mathbb{R}^2 \mapsto \langle u, v\rangle \in \mathbb{R}$ est linéaire. On dit alors que $\langle \cdot, \cdot \rangle$ est **linéaire à gauche**. En effet soient $\lambda, \mu \in \mathbb{R}, u, v, w \in \mathbb{R}^2$, alors

$$\langle \lambda u + \mu v, w\rangle = (\lambda u_1 + \mu v_1)w_1 + (\lambda u_2 + \mu v_2)w_2 = \lambda(u_1w_1 + u_2w_2) + \mu(v_1w_1 + v_2w_2) = \lambda\langle u, w\rangle + \mu\langle v, w\rangle$$

Le fait que le produit scalaire soit symétrique implique que l'application $v \in \mathbb{R}^2 \mapsto \langle u, v\rangle \in \mathbb{R}$ est linéaire c'est à dire que $\langle \cdot, \cdot \rangle$ est **linéaire à droite**. On dit alors que l'application est **bilinéaire**. Par ailleurs, le produit scalaire d'un vecteur $u = (u_1, u_2)$ contre lui-même donne un scalaire positif : $\langle u, u\rangle = u_1^2 + u_2^2 \geqslant 0$. On dit alors que $\langle \cdot, \cdot \rangle$ est une application **positive**. Enfin $\langle u, u\rangle = u_1^2 + u_2^2$ s'annule seulement si $u_1 = u_2 = 0$ cad si $u = (0, 0)$. On parle d'application **définie**. Ces 4 propriétés sont les axiomes de définition du produit scalaire.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quel objet est le produit scalaire ?
- A votre avis, combien pourra-t-on définir de produits scalaires ?
- Faire les questions QCM du paragraphe.

2.1.2 Généralisation de la notion de produit scalaire

Définition 1:

On appelle produit scalaire sur E une application

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle : & E \times E & \longrightarrow & \mathbb{R} \\ & (u, v) & \mapsto & \langle u, v\rangle \end{array}$$

qui soit une forme

- bilinéaire (linéaire par rapport à chacune de ses variables).
  - $\forall (u, v, w) \in E^3, \forall \lambda \in \mathbb{R}, \langle \lambda u + v, w\rangle = \lambda\langle u, w\rangle + \langle v, w\rangle$.
  - $\forall (u, v, w) \in E^3, \forall \lambda \in \mathbb{R}, \langle w, \lambda u + v\rangle = \lambda\langle w, u\rangle + \langle w, v\rangle$.
- symétrique : $\forall (u, v) \in E^2, \langle u, v\rangle = \langle v, u\rangle$.
- définie : si $\langle u, u\rangle = 0$ alors $u = 0_E$.
- positive : $\forall u \in E, \langle u, u\rangle \geqslant 0$.

On appelle alors un espace $(E, \langle \cdot, \cdot \rangle)$ muni d'un produit scalaire **espace préhilbertien**. Si E est de dimension finie, il est appelé **espace euclidien**.

Remarque 1: étymologie

• **Pourquoi préhilbertien ?** Parce qu'il y a des espaces dits "de Hilbert" qui sont préhilbertiens mais qui ont une propriété théorique en plus des ev préhilbertiens.
• **Pourquoi euclidien ?** Parce qu'on peut munir ces ev d'un repère euclidien (ils sont de dimension finie). Tout cela se rapporte à la géométrie euclidienne que vous connaissez depuis le collège.
• **Pourquoi produit scalaire ?** Parce que cette application "multiplie" deux vecteurs pour donner un scalaire.

Exemple 1: Comment démontrer que quelque chose est un produit scalaire ?

Démontrons que $\langle \cdot, \cdot \rangle$ définit un produit scalaire sur $\mathbb{R}^n$ si

$$\forall u = (u_1, \ldots, u_n) \in \mathbb{R}^n, v = (v_1, \ldots, v_n) \in \mathbb{R}^n, \langle u, v\rangle = \sum_{i=1}^{n} u_i v_i$$

• **Symétrie** : Soient $u = (u_1, \dots, u_n) \in \mathbb{R}^n, v = (v_1, \dots, v_n) \in \mathbb{R}^n$ alors

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i = \sum_{i=1}^{n} v_i u_i = \langle v, u \rangle$$

• **Linéarité à gauche** : Soient $\lambda, \mu \in \mathbb{R}, u, v, w \in \mathbb{R}^n$, alors

$$\langle \lambda u + \mu v, w \rangle = \sum_{i=1}^{n} (\lambda u_i + \mu v_i) w_i = \sum_{i=1}^{n} (\lambda u_i w_i + \mu v_i w_i) = \lambda \sum_{i=1}^{n} u_i w_i + \mu \sum_{i=1}^{n} v_i w_i = \lambda \langle u, w \rangle + \mu \langle v, w \rangle$$

La symétrie du produit scalaire entraîne alors la linéarité à droite.

• **Positivité** : Soit $u \in \mathbb{R}^n, \langle u, u \rangle = \sum_{i=1}^{n} u_i^2 \geqslant 0$.

• **Définie** : Si $\langle u, u \rangle = 0$ alors $\sum_{i=1}^{n} u_i^2 = 0$. Comme on somme des quantités positives, alors pour tout i dans $\{1, \dots, n\}$, $u_i = 0$ et donc $u = (0, \dots, 0)$.

Donc il s'agit bien d'un produit scalaire sur $\mathbb{R}^n$ qui étant de dimension finie est donc un espace euclidien.

Avec cette définition générale, on peut définir toute une panoplie de produits scalaires sur des ev divers et variés :

- On peut généraliser le produit scalaire de $\mathbb{R}^n$ à $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées de taille n qui est donc un ev euclidien.

$$\forall M = [m_{ij}] \in \mathcal{M}_n(\mathbb{R}), N = [n_{ij}] \in \mathcal{M}_n(\mathbb{R}), \quad \langle M, N \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} n_{ij}$$

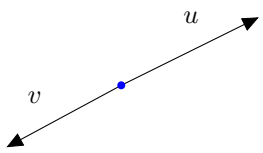
- Sur $C^0([a, b], \mathbb{R})$, l'ev des fonctions continues sur le segment $[a, b]$, on peut définir le produit scalaire

$$\langle f, g \rangle = \int_a^b f(t) g(t) dt$$

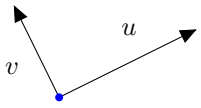
  Ainsi on munit $C^0([a, b], \mathbb{R})$ d'une structure d'ev préhilbertien puisqu'il est de dimension infinie. Ce produit scalaire est celui rencontré dans la formulation variationnelle du problème de la membrane élastique. Nous y reviendrons ultérieurement.

Que représente géométriquement le produit scalaire de deux vecteurs ?
On peut voir le produit scalaire de deux vecteurs u, v comme le degré de corrélation entre ces deux vecteurs. Pour s'en convaincre, reprenons le produit scalaire sur $\mathbb{R}^2$:

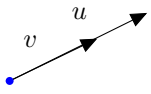
$$\langle u, v \rangle = ||u||.||v|| \cos(u, v)$$



$cos(u, v) = -1$
$|\langle u, v \rangle|$ maximal



$cos(u, v) = 0$
$|\langle u, v \rangle|$ minimal



$cos(u, v) = 1$
$|\langle u, v \rangle|$ maximal

Si les deux vecteurs sont colinéaires alors $\cos(u, v)$ vaut soit 1 soit -1. Le produit scalaire vaut alors en valeur absolue $||u||.||v||$. Si on fait tourner v de sorte que les deux vecteurs soient de moins en moins lié, alors le cosinus s'approche de 0 jusqu'à l'atteindre lorsque les deux vecteurs sont orthogonaux. Ainsi le produit scalaire en valeur absolue est d'autant plus grand que les vecteurs sont "proches de la colinéarité". La borne maximale est alors donnée par ce qu'on appelle l'inégalité de Cauchy-Schwarz.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quels sont les ensembles de départ et d'arrivée d'un produit scalaire ?
- Comment montre-t-on que quelque chose est un produit scalaire ?
- Citer de tête deux produits scalaires définis sur des ev autres que $\mathbb{R}^n$.
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

2.1.3 Inégalité de Cauchy-Schwarz

Proposition 1: Inégalité de Cauchy Schwarz

Soit E un espace préhilbertien muni de $\langle \cdot, \cdot \rangle$, alors

$$\forall (u,v) \in E^2, |\langle u,v\rangle| \leq \sqrt{\langle u,u\rangle}\sqrt{\langle v,v\rangle}.$$

Cette inégalité devient une égalité si et seulement si (u,v) est une famille liée.

Passons à la preuve de cette inégalité fondamentale :

Preuve :

On évacue tout de suite le cas $v = 0_E$ car alors l'inégalité est triviale : on a bien

$$|\langle u, 0_E\rangle| = 0 \leq 0 = \sqrt{\langle u,u\rangle}\sqrt{\langle 0_E, 0_E\rangle}$$

Soient $u, v \in E$, $\lambda \in \mathbb{R}$. La preuve consiste à calculer la quantité $\langle u+\lambda v, u+\lambda v\rangle$ en utilisant la bilinéarité du produit scalaire. On obtient

$$\begin{array}{lll} \forall \lambda \in \mathbb{R}, \quad \langle u+\lambda v, u+\lambda v\rangle & = \langle u, u+\lambda v\rangle + \lambda\langle v, u+\lambda v\rangle & \text{linéarité à gauche du produit scalaire} \\ & = \langle u,u\rangle + \lambda\langle u,v\rangle + \lambda\langle v,u\rangle + \lambda^2\langle v,v\rangle & \text{linéarité à droite du produit scalaire} \\ & = \langle u,u\rangle + 2\lambda\langle u,v\rangle + \lambda^2\langle v,v\rangle & \text{par symétrie du produit scalaire} \end{array}$$

On a donc montré que $\langle u+\lambda v, u+\lambda v\rangle$ est un polynôme de degré 2 en λ. Par ailleurs, le polynôme $\langle u+\lambda v, u+\lambda v\rangle \geqslant 0$ pour tout λ par positivité du produit scalaire. Ainsi son discriminant est négatif ou nul. Calculons son discriminant : il vaut

$$\Delta = 4\langle u,v\rangle^2 - 4\langle u,u\rangle\langle v,v\rangle$$

Ainsi

$$\langle u,v\rangle^2 \leq \langle u,u\rangle\langle v,v\rangle$$

et on obtient l'inégalité de Cauchy-Schwarz en prenant la racine carrée de cette inégalité.

Il nous reste à traiter le cas d'égalité. Tout d'abord, si (u,v) est une famille liée alors $\exists \lambda \in \mathbb{R}, v = \lambda u$. Donc par bilinéarité du produit scalaire $|\langle u,v\rangle| = |\lambda\langle u,u\rangle| = \sqrt{\langle u,u\rangle}\sqrt{\lambda^2\langle u,u\rangle} = \sqrt{\langle u,u\rangle}\sqrt{\langle v,v\rangle}$.

Réciproquement si cette égalité est vraie, ceci signifie que le discriminant précédent est nul. Donc le polynôme $\langle u+\lambda v, u+\lambda v\rangle$ admet une racine réelle double. Il existe donc $\lambda \in \mathbb{R}$ telle que $\langle u+\lambda v, u+\lambda v\rangle = 0$. Comme le produit scalaire est une forme définie alors $u+\lambda v = 0_E$. Donc (u,v) est une famille liée.

▸ **Exercice 2**. ⋆ Démontrer que

$$\langle f,g\rangle = \int_a^b f(t)g(t)dt$$

définit bien un produit scalaire sur $C^0([a,b],\mathbb{R})$. Appliquez alors l'inégalité de Cauchy-Schwarz à ce produit scalaire. En déduire que si $u \in H^1([a,b])$ et $v \in H^1([a,b])$ où

$$H^1([a,b]) = \left\{ f \in \mathcal{F}([a,b],\mathbb{R}) | \int_a^b f(t)^2 dt < +\infty, \int_a^b f'(t)^2 dt < +\infty \right\}$$

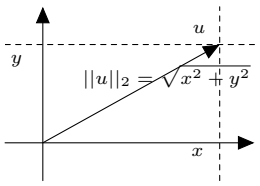
alors $\int_a^b u(t)v(t)dt$ et $\int_a^b u'(t)v'(t)dt$ sont des intégrales bien définies. Ceci permet alors de justifier l'existence de la forme bilinéaire définissant la formulation variationnelle du problème de la membrane.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Si deux vecteurs sont sur une même droite, le produit scalaire de ces vecteurs est-il proche de 0 ou loin de 0 ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

2.1.4 Norme euclidienne

Voyons maintenant une application mathématique importante de l'inégalité de Cauchy-Schwarz. Dans notre vie, la mesure est une notion cruciale. Par exemple votre cerveau instinctivement vous dit qu'une personne est plus grande qu'une autre : il a mesuré la différence de taille entre ces personnes. En mathématiques, la notion de mesure est également omniprésente : vous mesurez des distances dans le plan $\mathbb{R}^2$ depuis le collège avec la norme euclidienne (dessin ci-contre), vous mesurez la taille d'un réel à l'aide de la valeur absolue, vous mesurez la taille d'un complexe à l'aide du module. On peut alors se demander s'il est possible de construire des instruments de mesures (appelés normes) sur des espaces vectoriels exotiques (de même qu'on s'était posé la question pour le produit scalaire). L'enjeu est alors de faire ressortir les propriétés communes de la valeur absolue, du module et de la norme euclidienne afin de deviner la bonne définition.



• Premièrement, ces trois applications sont à valeurs positives : un module, une valeur absolue et la norme euclidienne d'un vecteur sont positives.

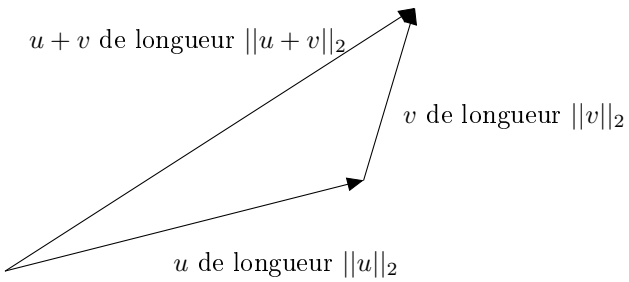
• Deuxièmement la valeur absolue vérifie $|\lambda x| = |\lambda||x|$ tout comme le module. Et $||\lambda(x,y)||_2 = ||(\lambda x, \lambda y)||_2 = \sqrt{\lambda^2 x^2 + \lambda^2 y^2} = \sqrt{\lambda^2}\sqrt{x^2+y^2} = |\lambda|||(x,y)||_2$. Cette propriété commune s'appelle l'homogénéité.

• Troisièmement : $|x| = 0 \iff x = 0$, $|z| = 0 \iff z = 0$ et $||(x,y)||_2 = 0 \iff \sqrt{x^2+y^2} = 0 \iff x^2 + y^2 = 0 \iff x = y = 0$. (Un objet de taille nulle est nécessairement le vecteur nul) : on dit que ces applications sont définies.

• Ces trois applications vérifient l'inégalité triangulaire : $|x+y| \leq |x| + |y|$, $|z_1 + z_2| \leq |z_1| + |z_2|$ et

$$||u+v||_2 \leq ||u||_2 + ||v||_2.$$

Pour la norme euclidienne dans $\mathbb{R}^2$ (dessin ci-dessous), cette inégalité signifie que dans un triangle, la longueur d'un côté est inférieure à la somme des longueurs des deux autres côtés : d'où le nom inégalité triangulaire.



Ces 4 points communs forment les 4 axiomes de la définition de norme sur un ev.

Définition 2: Norme

Soit E un ev. On appelle norme toute application définie sur E à valeurs dans $\mathbb{R}$ qui est :

- **à valeurs positives :** $\forall u \in E, ||u|| \geq 0$.
- **homogène** : $\forall \lambda \in \mathbb{R}, u \in E, ||\lambda u|| = |\lambda| \cdot ||u||$.
- **définie** : $\forall u \in E$, si $||u|| = 0$ alors $u = 0_E$.
- **vérifie l'inégalité triangulaire** : $\forall u, v \in E, ||u+v|| \leq ||u|| + ||v||$.

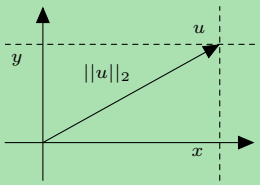
Il existe de nombreuses normes. L'inégalité de Cauchy-Schwarz permet ici de construire un tel objet sur n'importe quel espace préhilbertien.

Proposition 2: Norme euclidienne généralisée

Tout espace préhilbertien associé à $\langle \cdot, \cdot \rangle$ peut être muni d'une norme définie par $\forall u \in E, ||u||_2 = \sqrt{\langle u, u \rangle}$ et appelée norme euclidienne.

Exemple 2:

Vous travaillez depuis tout petit sur l'ev euclidien $\mathbb{R}^2$ dont la norme euclidienne est donnée par $||u||_2 = \sqrt{x^2+y^2}$. C'est en réalité la norme associée au produit scalaire $\langle u, v \rangle = u_1v_1 + u_2v_2$. Ainsi si $u = (x, y)$ alors $||u||_2 = \sqrt{\langle u, u \rangle} = \sqrt{x^2+y^2}$



Exemple 3:

Sur $C^0([a, b], \mathbb{R})$, l'ev des fonctions continues sur le segment $[a, b]$, on peut définir le produit scalaire

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt$$

La norme euclidienne associée vérifie

$$||f||_2 = \left(\int_a^b f(t)^2 dt \right)^{\frac{1}{2}}.$$

Preuve :

C'est un exercice : Démontrez que $|| \cdot ||_2$ est définie, homogène et positive à l'aide de la définition du

produit scalaire. On considère $u, v \in E$ et on veut prouver l'inégalité triangulaire. Démontrez que

$$||u+v||_2^2 = ||u||_2^2 + 2\langle u, v\rangle + ||v||_2^2.$$

Déduire grâce à l'inégalité de Cauchy-Schwarz que

$$||u+v||_2^2 \leq (||u||_2 + ||v||_2)^2.$$

Conclure.

Remarque 2: Cauchy Schwarz

Notez qu'avec les notations de norme l'inégalité de Cauchy-Schwarz s'écrit $|\langle u, v\rangle| \leq ||u||_2||v||_2$.

Revenons maintenant à la formulation variationnelle du problème de la membrane maintenant que nous avons introduit les notions de produit scalaire et de normes. Ces notions vont s'avérer cruciales dans la résolution du problème.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Prendre deux produits scalaires dont au moins un n'est pas sur $\mathbb{R}^n$ et donner l'expression des normes euclidiennes associées.
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

2.2 ⋆ Discrétisation de la formulation variationnelle ⋆

Revenons à la formulation variationnelle qui se déclinait ainsi

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

et se récrit sous la forme

$$a(u, v) = l(v)$$

pour

$$a : \begin{array}{l} H_0^1([0,1])^2 \longrightarrow \mathbb{R} \\ (u,v) \mapsto \int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx \end{array}, \qquad l : \begin{array}{l} H_0^1([0,1]) \longrightarrow \mathbb{R} \\ v \mapsto \int_0^1 f(x)v(x)dx \end{array}$$

On pourrait également récrire cette formulation à l'aide du produit scalaire $\langle f, g\rangle = \int_0^1 f(x)g(x)dx$ sous la forme

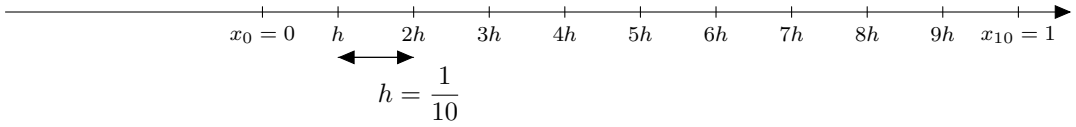
$$a(u, v) = \langle f, v\rangle, \quad a(u, v) = \langle u', v'\rangle + \langle cu, v\rangle.$$

Nous ne nous servirons pas explicitement de cette écriture sous forme de produits scalaires mais ceci a rappelons-le un intérêt certain pour prouver qu'il existe une solution u à ce problème.

Revenons à notre problèmatique de départ : nous ne sommes pas capables de déterminer l'expression de la solution u du problème des membranes, l'idée est maintenant d'approcher la formulation variationnelle par un problème qui lui sera résolvable.

Comment approximer la formulation variationnelle ?

L'idée de toutes les méthodes d'approximation d'analyse numérique est de calculer l'approximation de la solution u en un nombre fini de points et non pas en tous les points de l'intervalle $[0, 1]$. Pour cela, on **discrétise** l'intervalle d'étude (ici $[0, 1]$) : il s'agit de découper l'intervalle en plein de sous intervalles d'une longueur notée h.



Ainsi lorsqu'on diminue h, on calcule la solution approchée en de plus en plus de points et on espère que la solution approchée tende vers la bonne solution.
La première chose à faire est donc de se doter d'un maillage régulier (dessin de maillage à 11 points ci-dessus) :

$$a = x_0 < x_1 < \ldots < x_N, \quad \forall i \in \{0, \ldots, N\}, x_i = ih, \quad h = \frac{1}{N}$$

Ensuite, l'idée est de chercher la solution sous une forme simplifiée afin de pouvoir mener les calculs assez aisément. Pour cela on cherche non pas la solution u dans $H_0^1([0,1])$ mais dans l'espace suivant

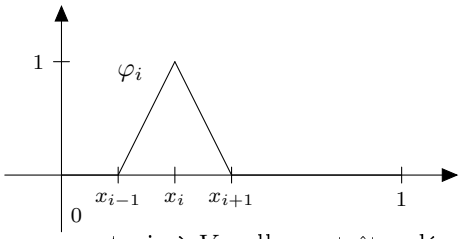
$$V_h = \{v_h \in C^0([0,1]) | v_h \text{ affine par morceaux sur chaque } [x_j, x_{j+1}], v_h(a) = v_h(b) = 0\}$$

Le formulation variationnelle approchée est alors : trouver $u_h \in V_h$ telle que

$$(P_h) : a(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h.$$

En quoi a-t-on avancé par rapport à avant ?

Il se trouve qu'on peut montrer que V_h est un sev de $C^0([0,1])$ de dimension finie dont on peut déterminer une base. Celle-ci est donnée par $(\varphi_1, \ldots, \varphi_{N-1})$ où les fonctions φ_j sont des fonctions affines par morceaux définies par $\phi_j(x_i) = 1$ si $i = j$, 0 sinon.



L'intérêt est alors que la solution approchée u_h étant choisie pour appartenir à V_h, elle peut être décomposée sur la base de V_h : autrement dit, il existe des constantes μ_i telles que $u_h = \sum_{i=1}^{N-1} \mu_i \varphi_i$. Alors cette décomposition en base accompagnée du fait que a soit bilinéaire et l linéaire va permettre de transformer le problème (P_h) en système linéaire à résoudre ! Vous avez ainsi une application directe de la théorie des espaces vectoriels et des applications linéaires.

En effet, si on remplace u_h dans (P_h) par linéarité à gauche de a, on a

$$\sum_{i=1}^{N-1} \mu_i a(\varphi_i, v_h) = l(v_h), \quad \forall v_h \in V_h$$

Cette égalité est équivalente (vous pourrez le démontrer dans l'exercice de fin de partie) à

$$\sum_{i=1}^{N-1} \mu_i a(\varphi_i, \varphi_j) = l(\varphi_j), \quad \forall j \in \{1, \ldots, N-1\}$$

Et ceci est exactement équivalent au système linéaire

$$\begin{pmatrix} a(\varphi_1, \varphi_1) & \ldots & a(\varphi_{N-1}, \varphi_1) \\ \vdots & & \vdots \\ a(\varphi_1, \varphi_{N-1}) & \ldots & a(\varphi_{N-1}, \varphi_{N-1}) \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{N-1} \end{pmatrix} = \begin{pmatrix} l(\varphi_1) \\ \vdots \\ (\varphi_{N-1}) \end{pmatrix}$$

Ce système est de la forme $A\mu = b$: A et b sont calculables puisque a, l et les φ_i sont connues. Le vecteur inconnue est μ.

Qui est μ déjà ?
Si vous trouvez μ alors vous avez tous les μ_i et donc vous avez l'expression de u_h la solution de (P_h) : il est alors possible de tracer en python par exemple u_h puis de faire diminuer h pour se rapprocher de la "vraie" solution u. Ainsi on peut voir si la solution approchée a un comportement pertinent, cohérent avec la situation physique qu'elle décrit.

Comment calculer A et b ?

Il est facile de calculer ces deux matrices : il s'agit simplement de calculer les intégrales de fonctions qui sont connues : la première étape consiste à déterminer les expressions des φ_i. Je vous le laisse en exo (voir ci-dessous). En revanche, un certain nombre de coefficients de A sont automatiquement nuls : en effet, rappelons que

$$a(\varphi_i, \varphi_j) = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx$$

et que φ_i est nulle partout sauf sur $[x_{i-1}, x_{i+1}]$. Ainsi si φ_i et φ_j sont telles que $[x_{i-1}, x_{i+1}]$ et $[x_{j-1}, x_{j+1}]$ ne s'intersectent pas, alors toutes ces intégrales sont nulles et $a(\varphi_i, \varphi_j) = 0$.

Ceci nous permet de faire la transition avec la section suivante : le fait que par exemple $\int_0^1 \varphi_i'(x)\varphi_j'(x)dx = 0$ a un sens algébrique profond si on réfléchit à l'aide du produit scalaire $\langle f, g\rangle = \int_0^1 f(x)g(x)dx$. Cela signifie que φ_i' et φ_j' sont **orthogonales** pour ce produit scalaire. Vous connaissiez l'orthogonalité de vecteurs dans $\mathbb{R}^2$ ou $\mathbb{R}^3$, il est temps pour vous de découvrir cette notion pour des objets beaucoup plus exotiques tels que les fonctions.

▸ **Exercice 3**. Cet exercice a pour objectif, si vous le souhaitez, de vous plonger plus dans la méthode des éléments finis en vue de le résoudre ultérieurement et de tracer avec python la solution approchée.

1. Démontrer que V_h est bien un ev, démontrer que les φ_i en forme bien une base et en déduire la dimension. (Vous pouvez vous aider des révisions sur les bases ci-dessous).
2. Démontrer que
$$\sum_{i=1}^{N} \mu_i a(\varphi_i, v_h) = l(v_h), \quad \forall v_h \in V_h$$
est bien équivalente à
$$\sum_{i=1}^{N} \mu_i a(\varphi_i, \varphi_j) = l(\varphi_j), \quad \forall j \in \{1, \dots, N\}$$
3. Calculer les matrices A et b entièrement lorsque $c = 1$.

2.3 Révision sur les bases

L'objectif de cette partie est de faire quelques rappels sur les bases des ev.

Qu'est-ce qu'une base "avec les mains" ?

Une base d'un ev E est une famille optimale de vecteurs de E avec lesquels on peut construire tout vecteur de E en faisant une combinaison linéaire de ces vecteurs. L'adjectif optimal vient du fait qu'enlever un vecteur ne permet plus de construire l'ev entier.

Qu'est-ce qu'une base rigoureusement ?

C'est une famille libre et génératrice. Si une famille de vecteurs est génératrice, tout vecteur de E admet **au moins** une décomposition comme combinaison linéaire des vecteurs de la famille : autrement dit si $(u_1, \dots, u_n)$ est génératrice alors

$$\forall u \in E, \exists (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n, u = \sum_{i=1}^{n} \lambda_i u_i.$$

Si une famille de vecteurs est libre, tout vecteur de E admet **au plus** une décomposition comme combinaison linéaire des vecteurs de la famille : autrement dit si $(u_1, \dots, u_n)$ est libre alors

$$\sum_{i=1}^{n} \lambda_i u_i = \sum_{i=1}^{n} \gamma_i u_i \Longrightarrow \forall i, \lambda_i = \gamma_i.$$

Le point d'équilibre est la base : tout vecteur de E admet **exactement une** décomposition comme combinaison linéaire des vecteurs de la famille : autrement dit si $(u_1, \dots, u_n)$ est une base alors

$$\forall u \in E, \exists!(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n, u = \sum_{i=1}^{n} \lambda_i u_i.$$

Qu'est-ce que la dimension d'un espace vectoriel ?

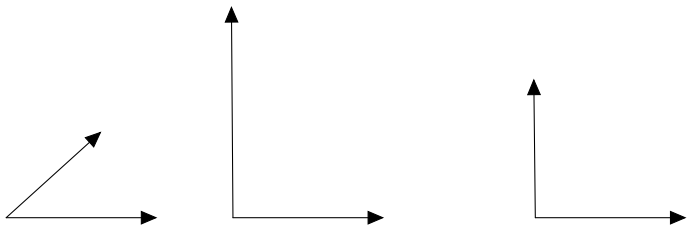
Il existe un théorème assurant que **toutes les bases d'un ev ont le même nombre de vecteurs.** Ce nombre est alors appelé **dimension de l'ev.** Par exemple, en tant que $\mathbb{R} - ev$, $\mathbb{R}^n$ est de dimension n, $\mathcal{M}_n(\mathbb{R})$ est de dimension n^2, $\mathcal{F}(\mathbb{R}, \mathbb{R})$ est de dimension infinie.

Comment démontrer qu'une famille de vecteurs est une base ?

Il y a plusieurs solutions. La première consiste à utiliser la définition : il suffit de montrer que la famille est libre et génératrice. Cependant le théorème sur la dimension apporte des informations supplémentaires sur les bases : elles ont nécessairement le même nombre de vecteurs que la dimension. Ainsi si vous connaissez la dimension (n), pour montrer qu'une famille de vecteurs est une base, il suffit de montrer qu'elle est libre et qu'elle contient n vecteurs. Ou alors il suffit de montrer qu'elle est génératrice et qu'elle contient n vecteurs. Par exemple, on peut montrer que $((1,1,1),(1,1,0),(1,-1,0))$ est une base de $\mathbb{R}^3$ et pour cela on se contente de montrer qu'elle est libre et possède 3 vecteurs dans un ev de dimension 3.

Quelle est la tête d'une base de $\mathbb{R}^2$?

Voici trois bases de $\mathbb{R}^2$: il s'agit de trois familles libres (car les vecteurs ne sont pas colinéaires) qui contiennent deux vecteurs dans un ev de dimension 2. Elles ne seraient pas des bases si les deux vecteurs étaient colinéaires.



Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Faire les questions QCM du paragraphe.

2.4 Familles orthogonales

2.4.1 Définitions et propriétés

Comment vérifiez-vous l'orthogonalité de deux vecteurs de $\mathbb{R}^2$?

Vous regardiez si le produit scalaire de ces deux vecteurs était nul. Par exemple, le produit scalaire de $(1,1)$ par $(1,-1)$ vaut $1-1=0$. Ceci est aussi valable dans $\mathbb{R}^3$ où par exemple le produit scalaire de $(1,1,1)$ par $(0,1,-1)$ vaut 0. Ainsi pour généraliser la notion d'orthogonalité à un ev préhilbertien quelconque, on regarde quand le produit scalaire associé est nul.

Définition 3:

Soit $(E, \langle \cdot, \cdot \rangle)$ un ev préhilbertien.

• On dit que deux **vecteurs** $(u, v) \in E^2$ sont **orthogonaux** si et seulement si $\langle u, v \rangle = 0$.

• On dit qu'une **famille** finie $(u_i)_{i \in \{1,\dots,p\}}$ de vecteurs de E est **orthogonale** si et seulement si ses vecteurs sont deux à deux orthogonaux :

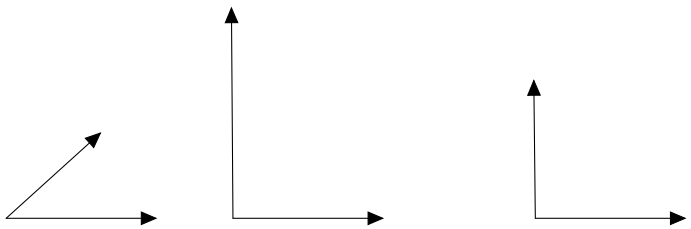
$$\forall 1 \leq i \neq j \leq p, \quad \langle u_i, u_j \rangle = 0.$$

• On dit qu'une **famille** finie $(u_i)_{i \in \{1,\dots,p\}}$ de vecteurs de E est **orthonormale** si et seulement si c'est une famille **orthogonale** dont tous les vecteurs sont **unitaires** :

$$\forall 1 \leq i \leq p, \quad ||u_i||_2 = 1.$$

Visualisons sur $\mathbb{R}^2$ ces notions :

Le couple de vecteurs de gauche n'est pas orthogonal, celui du milieu est orthogonal mais pas orthonormal, celui de droite est orthonormal (les vecteurs sont de norme 1).



Exemple 4: Comment montrer qu'une famille de vecteurs est orthogonale ?

Il suffit de calculer le produit scalaire des vecteurs deux à deux : montrons par exemple que $((1,1,0),(-1,1,0),(0,0,1))$ est une famille orthogonale de $\mathbb{R}^3$.

• $\langle (1,1,0),(-1,1,0) \rangle = 1 \times (-1) + 1 \times 1 + 0 \times 0 = 0$.
• $\langle (1,1,0),(0,0,1) \rangle = 1 \times 0 + 1 \times 0 + 0 \times 1 = 0$.
• $\langle (-1,1,0),(0,0,1) \rangle = -1 \times 0 + 1 \times 0 + 0 \times 1 = 0$.
Donc cette famille est bien orthogonale.

Exemple 5:

La notion d'orthogonalité dépend évidemment du produit scalaire de l'espace considéré. Voici deux exemples :

• On se place sur $\mathcal{M}_2(\mathbb{R})$ qu'on munit du produit scalaire $\langle M, N \rangle = M_{11}N_{11} + M_{12}N_{12} + M_{21}N_{21} + M_{22}N_{22}$. Pour ce produit scalaire, vous pouvez montrer que les matrices

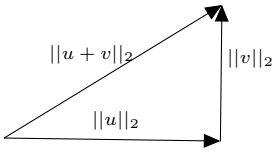
$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

sont orthogonales.

• On se place sur $C^0([-\pi, \pi])$ muni du produit scalaire $\langle f, g \rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$. Vous pouvez montrer que les fonctions cos et sin sont orthogonales pour ce produit scalaire.

Quelles propriétés possèdent les familles orthogonales ?

L'un des plus célèbres théorème lié à l'orthogonalité est le théorème de Pythagore. Evidemment vous ne le connaissez que dans l'ev $\mathbb{R}^2$:



Il nous dit que le carré de la longueur de l'hypoténuse est la somme des carrés des côtés. Ici les trois longueurs sont $||u||_2, ||v||_2, ||u+v||_2$. Ainsi le théorème de Pythagore s'écrit

$$||u+v||_2^2 = ||u||_2^2 + ||v||_2^2.$$

Le cadre des espaces préhilbertiens nous permet de voir ce théorème historique comme un cas particulier d'un théorème beaucoup plus général. La généralisation porte sur deux points : premièrement, il est valable dans tout ev préhilbertien (pas seulement pour $\mathbb{R}^2$) et deuxièmement il est vrai pour un nombre arbitraire de vecteurs. L'égalité précédente généralisée à une famille orthogonale de p vecteurs s'écrirait naturellement ainsi :

$$||u_1 + \ldots + u_p||_2^2 = ||u_1||_2^2 + \ldots + ||u_p||_2^2.$$

Ceci se récrit de manière plus propre ainsi :

Théorème 1: de Pythagore

Soit $(u_i)_{1\leq i\leq p}$ une famille **orthogonale**, alors $||\sum_{i=1}^{p} u_i||_2^2 = \sum_{i=1}^{p} ||u_i||_2^2$.

Ainsi ce théorème est valable en toute généralité sur n'importe quel ev préhilbertien (de fonctions, de matrices...) !

Exemple 6:

Comment s'écrit ce théorème dans le cas $p = 2$ sur un ev exotique ?

Prenons $M_2(\mathbb{R})$ muni de

$$\langle M, N\rangle = M_{11}N_{11} + M_{12}N_{12} + M_{21}N_{21} + M_{22}N_{22}.$$

Si M et N sont orthogonales, le théorème de Pythagore nous dit que $||M+N||_2^2 = ||M||_2^2 + ||N||_2^2$. Ceci s'écrit

$$(M_{11}+N_{11})^2+(M_{12}+N_{12})^2+(M_{21}+N_{21})^2+(M_{22}+N_{22})^2 = M_{11}^2+M_{12}^2+M_{21}^2+M_{22}^2+N_{11}^2+N_{12}^2+N_{21}^2+N_{22}^2$$

Le caractère orthogonal se traduit par l'absence de termes croisés. Normalement $(a+b)^2 = a^2+2ab+b^2$, ici les termes en $2ab$ sont annulés par l'orthogonalité.

Preuve :

On a que

$$||\sum_{i=1}^{p} u_i||_2^2 = \langle \sum_{i=1}^{p} u_i, \sum_{i=1}^{p} u_i\rangle = \langle \sum_{i=1}^{p} u_i, \sum_{j=1}^{p} u_j\rangle$$

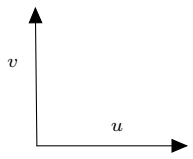
la dernière égalité étant due au caractère muet de l'indice de sommation.
En utilisant d'abord la linéarité à gauche puis la linéarité à droite,

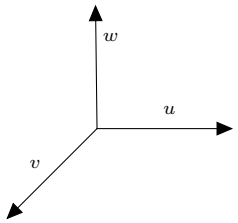
$$\langle\sum_{i=1}^{p} u_i, \sum_{j=1}^{p} u_j\rangle = \sum_{i=1}^{p}\langle u_i, \sum_{j=1}^{p} u_j\rangle = \sum_{i=1}^{p}\sum_{j=1}^{p}\langle u_i, u_j\rangle.$$

Comme $(u_i)_{1\leq i\leq p}$ est une famille orthogonale alors $\forall i \neq j, \langle u_i, u_j\rangle = 0$. Donc

$$\langle\sum_{i=1}^{p} u_i, \sum_{j=1}^{p} u_j\rangle = \sum_{i=1}^{p}\langle u_i, u_i\rangle = \sum_{i=1}^{p}||u_i||_2^2.$$

Passons maintenant à une propriété plus algébrique des familles orthogonales. Depuis le collège, vous manipulez des familles orthogonales sans le dire. Quand vous écrivez un vecteur (x, y), x correspond à la coordonnée selon le vecteur $(1, 0)$ et y selon $(0, 1)$. Vous écrivez le vecteur dans le repère euclidien formé des vecteurs $((1, 0), (0, 1))$. Cette famille de vecteurs est en réalité une famille orthogonale (même orthonormale) de même que $((1, 0, 0), (0, 1, 0), (0, 0, 1))$ l'est sur $\mathbb{R}^3$. Par ailleurs ce sont des bases respectivement de $\mathbb{R}^2$ et de $\mathbb{R}^3$. Il s'avère qu'une famille orthogonale est un bon point de départ pour construire une base de vecteurs car une famille orthogonale de vecteurs non nuls est libre. Prenons par exemple dans $\mathbb{R}^2$, le repère orthogonal euclidien : il apparaît clair que la famille de vecteurs (u, v) définissant ce repère est une famille libre car les deux vecteurs ne sont pas colinéaires. On peut faire le même constat de liberté sur le repère orthogonal de $\mathbb{R}^3$ représenté ici par (u, v, w).





Cette propriété est en réalité beaucoup plus générale :

Proposition 3:

Toute famille orthogonale de vecteurs non nuls de E est une famille libre de E.

Preuve :

Donnons-nous une famille orthogonale $(u_1, \ldots, u_p)$ de E.
Si $\sum_{i=1}^{p} \lambda_i u_i = 0_E$ alors $\forall j \in \{1, \ldots, p\}, \langle u_j, \sum_{i=1}^{p} \lambda_i u_i\rangle = 0$. Donc par linéarité à droite du produit scalaire,

$$\forall j \in \{1, \ldots, p\}, \sum_{i=1}^{p} \lambda_i \langle u_j, u_i\rangle = 0.$$

Comme $(u_i)_{1\leq i\leq p}$ est une famille orthogonale alors $\forall i \neq j, \langle u_j, u_i\rangle = 0$. Donc

$$\sum_{i=1}^{p} \lambda_i \langle u_j, u_i\rangle = \lambda_j \langle u_j, u_j\rangle = \lambda_j ||u_j||^2.$$

On en déduit que pour tous les j $\lambda_j ||u_j||^2 = 0$. Comme les u_j sont supposés non nuls, leur norme est non nulle et pour tout j, $\lambda_j = 0$.

Ainsi toute famille orthogonale de vecteurs non nuls est libre. Il est alors naturel de se dire qu'une famille orthogonale est un bon point de départ pour construire une base d'un ev donné : en effet, une base de E est une famille libre et génératrice de E. Cependant il s'agira de bases "spéciales" ayant une propriété

d'orthogonalité.

Questions :

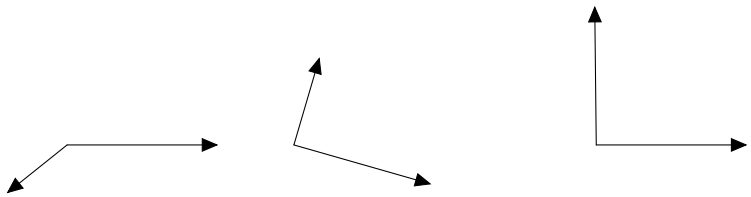
- Quels sont les points essentiels de ce paragraphe ?
- Ecrivez le théorème de Pythagore pour $p = 2$. Ecrivez l'égalité donnée par le théorème de Pythagore sur deux fonctions f et g orthogonales pour le produit scalaire $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$.
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

2.4.2 Bases orthonormales

Définition 4:

On appelle base orthonormale (bon) d'un ev euclidien toute base de E qui est une famille orthonormale.

Visualisons tout de suite à quoi ressemble une bon sur $\mathbb{R}^2$. La figure la plus à gauche est une base de $\mathbb{R}^2$ car elle est libre et constitué de deux vecteurs et $\dim(\mathbb{R}^2) = 2$. En revanche, la famille représentée n'est clairement pas orthogonale. Les deux dernières figures sont orthogonales mais seule la dernière a une chance d'être orthonormale car les deux vecteurs sont de même longueur.



Exemple 7: Comment montrer qu'une famille est une bon d'un espace euclidien ?

Prenons l'exemple de la famille $(e_1, e_2, e_3) = ((1, 0, 0), (0, 1, 0), (0, 0, 1))$.
• On montre que c'est une famille orthogonale : $\langle e_1, e_2\rangle = 1 \times 0 + 0 \times 1 + 0 \times 0 = 0$. De même $\langle e_1, e_3\rangle = \langle e_2, e_3\rangle = 0$. C'est donc une famille orthogonale : on en déduit d'après ce qui précède que c'est une famille libre. Comme elle a 3 vecteurs et qu'on travaille dans un ev de dimension 3, c'est donc une base de $\mathbb{R}^3$.

• Reste à montrer qu'elle est orthonormale : $||e_1||_2 = \sqrt{\langle e_1, e_1\rangle} = \sqrt{1 \times 1 + 0 \times 0 + 0 \times 0} = 1$. De même $||e_2||_2 = ||e_3||_2 = 1$. Donc c'est une bon de $\mathbb{R}^3$.

Quelles sont les coordonnées d'un vecteur en bon ?

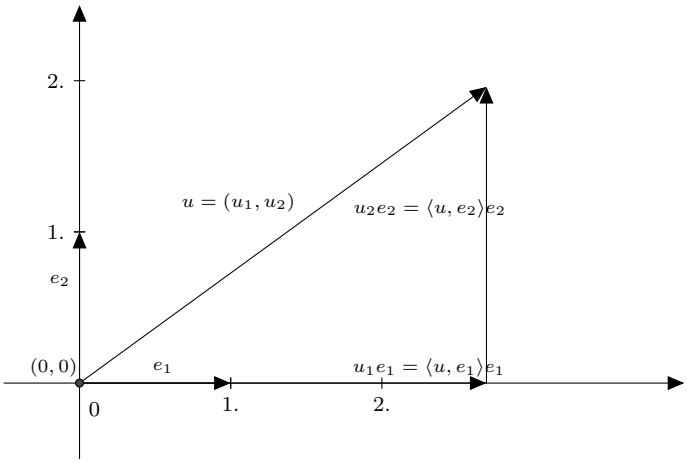
Il s'avère que dans une bon les coordonnées d'un vecteur donné sont aisées à trouver : elles s'expriment grâce au produit scalaire. Pour le comprendre, revenons à $\mathbb{R}^2$ où tout vecteur $u = (u_1, u_2)$ de $\mathbb{R}^2$ se construit à partir de $e_1 = (1, 0)$ et $e_2 = (0, 1)$ car

$$u = (u_1, u_2) = u_1(1, 0) + u_2(0, 1) = u_1 e_1 + u_2 e_2.$$

Les constantes u_1, u_2 sont les coordonnées du vecteur u dans la base (e_1, e_2). Si de plus (e_1, e_2) est une bon alors par linéarité à gauche du produit scalaire

$$\langle u, e_1\rangle = \langle u_1 e_1, e_1\rangle + \langle u_2 e_2, e_1\rangle = u_1\langle e_1, e_1\rangle = u_1$$

et on a trouvé que $u_1 = \langle u, e_1\rangle$. De même on peut trouver la deuxième coordonnée $u_2 = \langle u, e_2\rangle$.



Preuve :

de la proposition qui suit : La notion de coordonnée est valable pour tout ev de dimension finie. Si E est un tel ev ayant pour base $(e_1, \ldots, e_n)$, tout vecteur u de E s'exprime dans cette base

$$u = \sum_{i=1}^{n} u_i e_i = u_1 e_1 + \ldots + u_n e_n,$$

où les u_i sont les coordonnées de u dans la base $(e_1, \ldots, e_n)$. Si $(e_1, \ldots, e_n)$ est une bon alors le raisonnement fait pour deux vecteurs tient toujours :
Comme $u = \sum_{i=1}^{n} u_i e_i$ alors par linéarité à gauche du produit scalaire,

$$\forall j \in \{1, \ldots, n\}, \langle u, e_j\rangle = \sum_{i=1}^{n}\langle u_i e_i, e_j\rangle = \sum_{i=1}^{n} u_i\langle e_i, e_j\rangle.$$

Comme $(e_1, \ldots, e_n)$ est une bon alors

$$\forall j \in \{1, \ldots, n\}, \langle u, e_j\rangle = u_j\langle e_j, e_j\rangle = u_j||e_j||^2 = u_j.$$

Nous venons de démontrer la proposition suivante :

Proposition 4:

Soit $\mathcal{B} = (e_1, \ldots, e_n)$ une **bon** d'un espace euclidien E de dimension n, $u = \sum_{j=1}^{n} u_j e_j \in E$ alors

$$\forall i \in \{1, \ldots, n\}, \quad u_i = \langle u, e_i\rangle.$$

▸ **Exercice 4**. Soit $\mathcal{B} = (e_1, \ldots, e_n)$ une **bon** d'un espace euclidien E de dimension n, $u = \sum_{j=1}^{n} u_j e_j \in E$, $v = \sum_{j=1}^{n} v_j e_j \in E$, démontrez que :

1. $\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i = {}^tUV$ en notant $U = \begin{pmatrix} u_1 \\ . \\ u_n \end{pmatrix}$, $V = \begin{pmatrix} v_1 \\ . \\ v_n \end{pmatrix}$.
2. $||u||^2 = \sum_{j=1}^{n} u_j^2$.

Ces expressions ne vous rappellent-ils pas quelque chose ? Où les aviez-vous vu ? Pourquoi selon vous étaient-elles valables dans le cadre dans lequel vous les utilisiez ?

Remarque 3:

Attention ces expressions sont valables uniquement en **bon**. Souvenez-vous au lycée, vous écriviez que le produit scalaire de deux vecteurs (u_1, u_2) et (v_1, v_2) vaut $u_1v_1 + u_2v_2$. Cette écriture n'était valable que parce que u_1, u_2, v_1, v_2 sont des coordonnées dans la base canonique $((1,0),(0,1))$ qui est une bon !

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quelle est la différence entre une bon et une base ? une base orthogonale ?
- Comment déterminer les coordonnées d'un vecteur en bon ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

2.5 Borne inférieure d'un ensemble

Dans cette partie, nous introduisons la notion de borne inférieure d'un sous-ensemble de $\mathbb{R}$. Pour ceux qui ont fait analyse 1, c'est l'analogue de la notion de borne supérieure. Afin de comprendre de quoi il s'agit : considérons l'ensemble $]0,1[$.
Cet ensemble possède une infinité de minorants c'est-à-dire de réels inférieurs à tout point de l'ensemble. Par exemple, -2 et -12 sont inférieur à tout réel de $]0,1[$. Lorsqu'on parcourt la droite des réels de $-\infty$ en $+\infty$, il y a un réel critique qui marque la différence entre les minorants de $]0,1[$ et les non minorants. Ici ce réel est 0. Tout point inférieur ou égal à 0 minore tous les points de $]0,1[$.

En revanche, dès qu'on passe au dessus de 0, même légèrement comme 0.1 sur le dessin, on ne minore plus l'ensemble : en effet, il y a des éléments de $]0,1[$ entre 0 et 0.1 comme 0.05 par exemple ! Ce réel limite est appelé borne inférieure et se définit comme "le plus grand des minorants de l'ensemble". Evidemment, pour espérer définir "le plus grand des minorants" il faut qu'il y ait des minorants, ce qui explique l'hypothèse "partie minorée" de la définition suivante :

Définition 5: Borne inférieure

Soit A une partie de $\mathbb{R}$ minorée et non vide, on appelle borne inférieure de A et on note $\inf(A)$ le réel tel que :
• tout réel de A est supérieur à $\inf(A)$ **(c'est un minorant)**.
• si un réel x vérifie $x > \inf(A)$, on peut trouver un élément a de A vérifiant $\inf(A) \leq a \leq x$ **(c'est le plus grand des minorants)**.

Notez que cette borne inférieure peut autant appartenir à l'ensemble que ne pas y appartenir. A titre d'exemple, $0 = \inf(]0,1[)$ mais $0 \notin]0,1[$ alors que $0 = \inf([0,1[)$ et $0 \in [0,1[$. Lorsqu'il appartient à l'ensemble, la borne inférieure est alors le minimum de l'ensemble. Ainsi 0 est le minimum de $[0,1[$ tandis que $]0,1[$ n'admet aucun minimum !

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Donner un ensemble tel que sa borne inférieure y appartienne. Donnez-en un autre tel qu'elle n'y appartienne pas.
- Faire les questions QCM du paragraphe.

2.6 ⋆ Distance à un sev et résolution de système linéaire ⋆

Revenons au problème de la membrane et à sa résolution par la méthode des éléments finis. Nous avons discrétisé le problème de départ et nous avons abouti à la nécessité de résoudre un système linéaire du type $A\mu = b$ où μ est le vecteur inconnu à trouver. Nous avons vu que la matrice A contient beaucoup de 0 en raison de l'orthogonalité entre elles de nombreuses fonctions φ_i pour le produit scalaire

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx.$$

Cependant, la matrice A est de grande taille : je rappelle que la taille de la matrice est de l'ordre du nombre de points choisis dans le maillage. Ce nombre est amené à grandir puisque nous voulons mettre beaucoup de points afin d'être très précis. Ceci fait que résoudre le système linéaire en inversant la matrice A grâce à un pivot de Gauss est très lourd et fastidieux bien qu'envisageable. L'idée maintenant est de vous présenter deux autres stratégies. La première que nous allons voir maintenant se base sur le concept de distance. L'idée est la suivante : puisque $A\mu$ et b sont deux vecteurs de $\mathbb{R}^{N-1}$, on peut calculer la norme euclidienne suivante $||A\mu - b||_2$ pour

$$||u||_2 = \sqrt{\sum_{i=1}^{N-1} u_i^2}.$$

Alors résoudre $A\mu = b$ revient à annuler la norme $||A\mu - b||_2$. Cette norme étant toujours positive, ceci revient à minimiser $||Av - b||_2$ et donc à trouver la quantité suivante

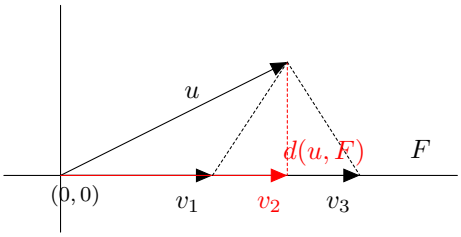
$$\inf\{||Av - b||_2 | v \in \mathbb{R}^{N-1}\}.$$

Il est temps d'introduire le concept qui va nous permettre de résoudre ce problème.

3 Distance et projection orthogonale

Plaçons-nous dans un ev préhilbertien et introduisons le concept de distance d'un vecteur à un sev de F. Pour cela, devinons la définition dans $\mathbb{R}^2$ à l'aide du dessin ci-dessous en devinant la distance de $(2, 1)$ à F (le sev donné par l'axe des abscisses).

Par distance, on entend plus petit écart entre la pointe du vecteur $(2, 1)$ et F. Ainsi en traçant un vecteur v de F et en le déplaçant, la distance sera le plus petit écart entre la pointe du vecteur $(2, 1)$ et la pointe du vecteur v. Il s'agit donc de trouver le vecteur v minimisant la longueur de $u - v$ cad minimisant $||u - v||_2$. Sur le dessin, les trois traits pointillés représentent les longueurs des vecteurs $u - v_1$, $u - v_2$ et $u - v_3$. Il apparait clair que la distance est donnée par la longueur de $u - v_2$ (en rouge).



Définition 6:

Soit F un sev de dimension finie d'un espace préhilbertien E, soit $u \in E$, on appelle distance de u à F : $d(u, F) = \inf\{||u - v||_2, v \in F\}$.

3.1 ⋆ Réécriture sous forme de distance ⋆

Le problème de la membrane élastique a abouti à minimiser

$$\inf\{||A\mu - b||_2 | \mu \in \mathbb{R}^{N-1}\}.$$

Quel est le lien avec la notion de distance ?

Lorsque μ parcourt $\mathbb{R}^{N-1}$, $A\mu$ parcourt l'image de A vue comme application linéaire. En effet $\mathrm{Im}(A)$ est l'ensemble des vecteurs qui s'écrivent sous la forme $A\mu$

$$\mathrm{Im}(A) = \{v \in \mathbb{R}^{N-1} | \exists \mu \in \mathbb{R}^{N-1}, v = A\mu\}$$

Ainsi on a

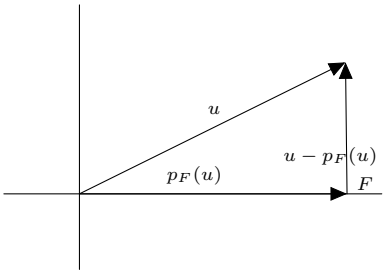
$$\inf\{||A\mu - b||_2 | \mu \in \mathbb{R}^{N-1}\} = \inf\{||v - b||_2 | v \in \mathrm{Im}(A)\} = d(b, \mathrm{Im}(A)).$$

Donc il s'agit exactement de trouver la distance de b au sev $\mathrm{Im}(A)$.

3.2 Lien entre distance et projection orthogonale

Au paragraphe précédent, on a vu qu'on pouvait ramener la résolution du système linéaire $Ax = b$ du problème de la membrane et au problème de minimisation suivant : trouver la distance de b au sev $\mathrm{Im}(A)$. Il est donc nécessaire de se poser la question suivante : **comment calculer, dans un ev préhilbertien E, la distance d'un vecteur donné à un sev de E ?** Si on répond à cette question alors on peut résoudre entièrement le problème de la membrane élastique.

Pour comprendre cela, rien ne vaut un dessin dans $\mathbb{R}^2$: choisissons $F = \mathrm{Vect}(1,0)$ le sev de $\mathbb{R}^2$ correspondant à l'axe des abscisses et essayons de deviner la distance du vecteur u dessiné. Cette distance correspond à l'écart minimal entre la pointe du vecteur u et l'axe des abscisses. Le plus petit écart est obtenu avec la perpendiculaire à l'axe des abscisses et vaut $||u - p_F(u)||_2$ où $p_F(u)$ est la projection orthogonale sur F du vecteur u.



Ce qu'on constate dans le cas particulier de $\mathbb{R}^2$ ($d(u, F) = ||u - p_F(u)||_2$) va être vraie pour tout ev euclidien en vérité : c'est un théorème fondamental qu'on verra au moment opportun. Ainsi comme u est connu, si on est capable de déterminer la projection orthogonale $p_F(u)$, on connaîtra la distance et on pourra résoudre totalement le problème de la membrane. La clé consiste dès lors à comprendre ce qu'est la projection orthogonale d'un vecteur et comment la calculer. Pour cela, il est nécessaire de se remémorer ce qu'est une projection vectorielle (pas nécessairement orthogonale). C'est l'objet du paragraphe suivant.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Faire les questions QCM du paragraphe.

3.3 Comment définir une projection ?

3.3.1 Révision : Qu'est-ce que deux sev supplémentaires

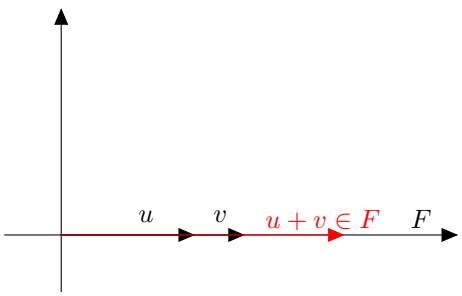
Qu'est ce que la somme de deux sev ?

Si F et G sont deux sev d'un ev E alors leur somme est défini par

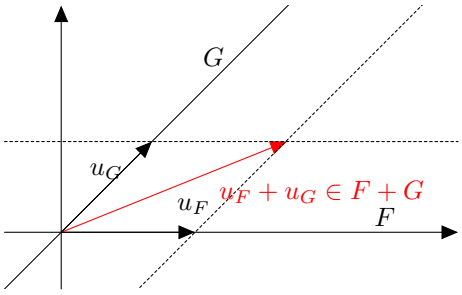
$$F + G = \{u_F + u_G | u_F \in F, u_G \in G\}.$$

Autrement dit c'est l'ensemble des vecteurs de E qui s'écrivent comme somme d'un élément de F et d'un élément de G. Vous avez vu que cet ensemble est lui-même un sev de E.

Prenons des exemples : on se place dans l'ev $E = \mathbb{R}^2$ choisissons $F = \mathrm{Vect}(1,0)$ et $G = \mathrm{Vect}(1,1)$ Commençons par sommer F avec lui-même. Que vaut $F + F$? C'est l'ensemble des vecteurs qui sont somme de deux vecteurs de F. Une somme de deux vecteurs de F a pour ordonnée 0 ne peut qu'appartenir à F. Graphiquement, les seuls vecteurs qui peuvent s'écrire comme somme de deux vecteurs de F sont sur F. Donc $F + F = F$.



Déterminons maintenant $F+G$. La question est quels vecteurs s'écrivent comme somme d'un vecteur de F et d'un vecteur de G ? Autrement dit quels vecteurs sont combinaisons linéaires de $(1,0)$ et $(1,1)$. En réalité c'est le cas pour tout vecteur de $\mathbb{R}^2$ (notamment pour le vecteur en rouge du dessin) car $((1,0),(1,1))$ est une base de $\mathbb{R}^2$. Ainsi $F+G=\mathbb{R}^2$. Ceci montre que $F+G$ n'a rien à voir avec $F\cup G$ qui n'est que la réunion de deux droites.



Alors qu'est-ce que deux sev supplémentaires ?

Deux sev sont supplémentaires "s'ils permettent de construire l'ev entier de manière optimale". Autrement dit si on les somme, on obtient l'ev entier et si ces deux ev n'ont que l'élément neutre en commun. Par exemple, $F+G$ dans l'exemple précédent sont supplémentaires dans $\mathbb{R}^2$ car $F+G=\mathbb{R}^2$ et leur intersection est réduite à l'élément neutre. En revanche F et $\mathbb{R}^2$ ne sont pas supplémentaires dans $\mathbb{R}^2$ car même si $F+\mathbb{R}^2=\mathbb{R}^2$, l'intersection de F et $\mathbb{R}^2$ est F : autrement dit, l'information donnée par F est utilisée deux fois, dans F et dans $\mathbb{R}^2$. C'est le sens de l'expression "de manière optimale".

Définition 7:

Soit E un $\mathbb{R}-$ev, on dit que F et G sont supplémentaires dans E si et seulement si

- $E=F+G$.
- $F\cap G=\{0_E\}$

On note alors $E=F\oplus G$.

Comment montrer que deux sev sont supplémentaires ?

Proposition 5:

Soit E un $\mathbb{R}-$ev, on dit que F et G sont supplémentaires dans E si et seulement **deux des trois** conditions suivantes sont satisfaites :

- $E=F+G$.
- $F\cap G=\{0_E\}$.
- $\dim(E)=\dim(F)+\dim(G)$.

Alors quel est l'intérêt de cette proposition ? En pratique, il est souvent plus facile de montrer les deux dernières conditions que la première. On démontre toujours la seconde (l'intersection), la première et la troisième étant laissée à l'appréciation de l'étudiant.

Pourquoi j'ai besoin de parler de sev supplémentaires dans ce cours ?

Ceci est indispensable pour définir la notion de projection vectorielle, et donc a fortiori de projection orthogonale.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Construire deux couples de sev supplémentaires de $\mathbb{R}^2$. Construire deux sev qui ne le sont pas. Même question dans $\mathbb{R}^3$.
- Faire les questions QCM du paragraphe.

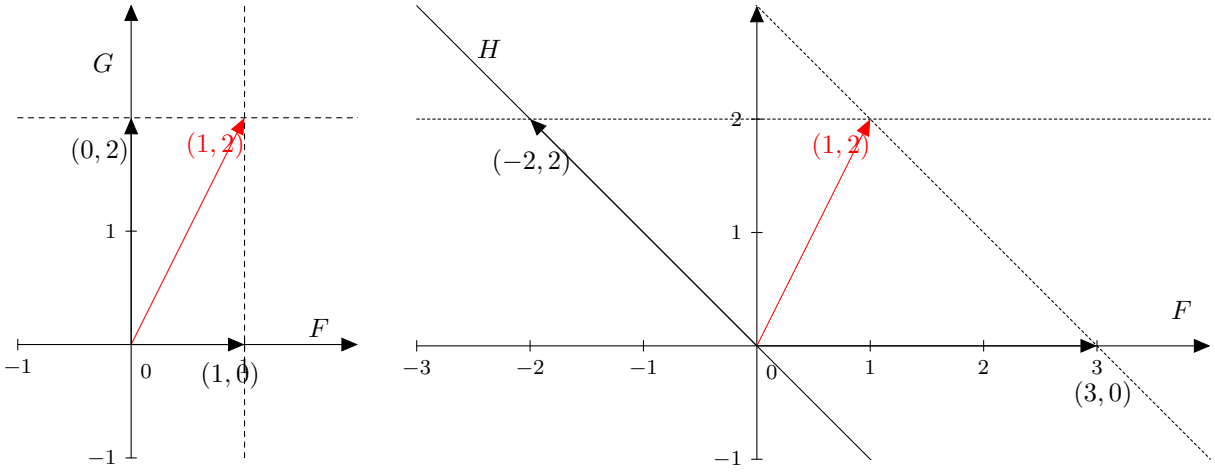
3.3.2 Révision : Projection vectorielle

Cette partie n'est pas obligatoire : elle explique comment on définit la notion de projection vectorielle. Si

vous ne vous sentez pas à l'aise, vous pouvez passer cette partie.

La première chose que nous allons essayer de comprendre, c'est en quoi la notion de sev supplémentaires est indispensable pour définir la notion de projection.

Plaçons-nous dans $\mathbb{R}^2$ et observons les dessins ci-dessous. Comment définir une projection ? Voici celle que vous connaissez depuis toujours : la projection sur l'axe des abscisses. Si on projette le vecteur $(1,2)$, on obtient le vecteur $(1,0)$. De même sur l'axe des ordonnées la projection est $(0,2)$. Seulement, la projection qu'on a utilisé n'est pas n'importe laquelle : on a projeté orthogonalement sur l'axe des abscisses c'est-à-dire parallèlement à l'axe des ordonnées. On pourrait très bien choisir de projeter autrement comme par exemple parallèlement à la droite dirigée par $(-1,1)$ auquel cas la projection de $(1,2)$ est le vecteur $(3,0)$.



Alors comment définir ces projections ?

Pour trouver la projection, on a décomposé sans le dire le vecteur $(1,2)$ en $(1,0)+(0,2)$ (voir dessin) quand on projetait parallèlement à l'axe des ordonnées et en $(3,0)+(-2,2)$ (voir dessin) quand on projetait parallèlement à $\mathrm{Vect}(-1,1)$. Notons $F=\mathrm{Vect}(1,0)$, $G=\mathrm{Vect}(0,1)$ et $H=\mathrm{Vect}(-1,1)$. Dans le premier cas, on a écrit la décomposition de $(1,2)$ sur $F+G$ et dans le second cas dans $F+H$. Là des liens commencent à se tisser avec le paragraphe sur les supplémentaires.

Prenons encore un autre exemple sur $\mathbb{R}^3$

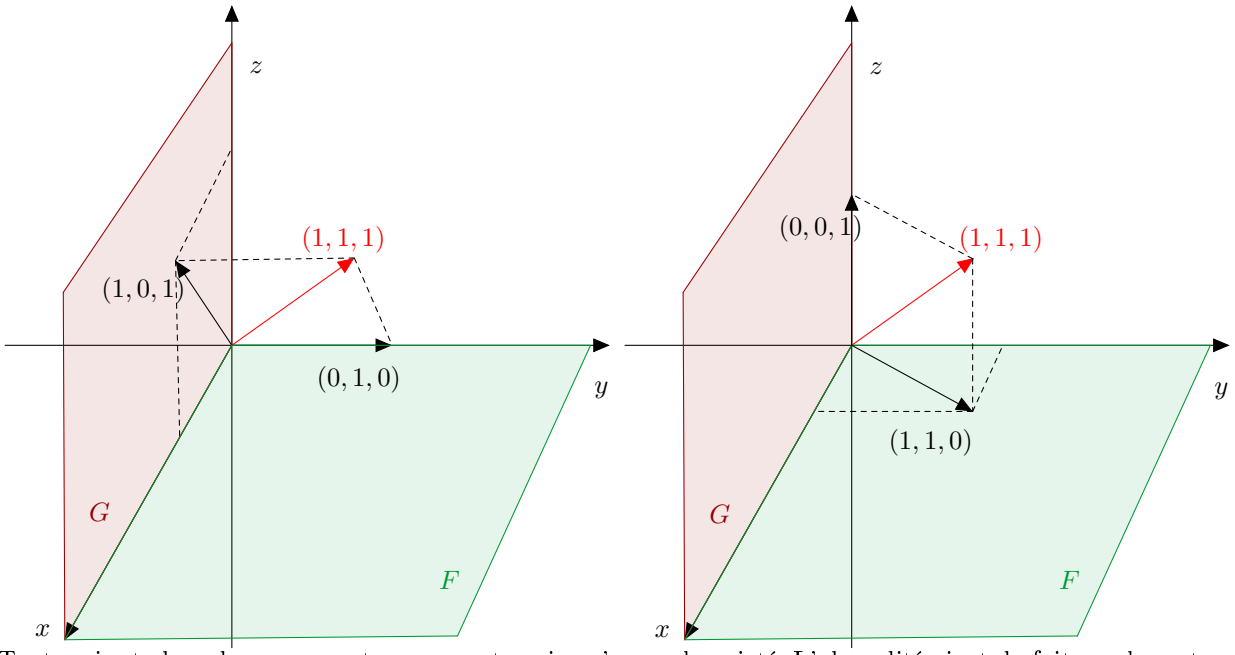
Plaçons nous sur $\mathbb{R}^3$ et considérons $F=\mathrm{Vect}((1,0,0),(0,1,0))$ et $G=\mathrm{Vect}((1,0,0),(0,0,1))$: F est le plan contenant les axes des x et des y et G est celui contenant les axes des x et des z. Un quart de chacun de ces plans est représenté sur le dessin ci-dessous. Cherchons sur le même principe la projection de $(1,1,1)$ sur F parallèlement à G. Pour cela décomposons $(1,1,1)$ en un élément de F plus un élément de G. On a

$$(1,1,1)=(1,1,0)+(0,0,1), \qquad \text{avec } (1,1,0)\in F \text{ et } (0,0,1)\in G.$$

Mais on a également

$$(1,1,1)=(0,1,0)+(1,0,1), \qquad \text{avec } (0,1,0)\in F \text{ et } (1,0,1)\in G.$$

Alors quelle est la projection de $(1,1,1)$ sur F parallèlement à G ? Est-ce $(1,1,0)$ ou $(0,1,0)$?



Tout ceci est absurde car un vecteur ne peut avoir qu'un seul projeté. L'absurdité vient du fait que le vecteur se décompose de plusieurs façons dans $F+G$. La projection sur F parallèlement à G ne peut tout simplement pas être définie. Pour espérer la définir, il faudrait que **tout vecteur ait une unique décomposition dans** $F+G$. Et ce qui va assurer cela, c'est le fait que F et G soient **supplémentaires**.

Pour l'exemple donné sur $\mathbb{R}^2$, F et G étaient supplémentaires et on avait bien unicité de la décomposition. En revanche dans l'exemple donné sur $\mathbb{R}^3$, F et G ont pour intersection l'axe des abscisses donc ne sont pas supplémentaires. Et comme par hasard, il n'y a pas unicité de la décomposition.

Proposition 6: et définition

Soit F et G deux sev supplémentaires dans E, alors tout vecteur u de E se décompose **uniquement** en $u = u_F + u_G$ où $u_F \in F, u_G \in G$. On appelle alors projection sur F parallèlement à G l'application

$$\begin{array}{rccc} p_F & : & E = F \oplus G & \to & F \\ & & u = u_F + u_G & \mapsto & u_F \end{array}$$

Revenons à notre objectif : nous souhaitons définir la notion de projection orthogonale. La projection orthogonale sur F serait une projection parallèlement à un sev noté $F^\perp$ qui serait orthogonal à F. La première chose à faire est alors de donner un sens à ce sev ? Existe-il vraiment ? Est-ce vraiment un sev ? Et est-ce que F et $F^\perp$ sont bien supplémentaires (indispensable pour définir la projection) ?

Questions :

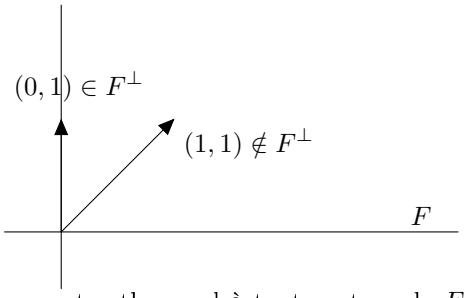
- Quels sont les points essentiels de ce paragraphe ?
- Pourquoi faut-il que F et G soient supplémentaires pour définir la projection sur F parallèlement à G ?
- Faire les questions QCM du paragraphe.

3.4 Orthogonal d'un sev et projection orthogonale

3.4.1 Définition

Quel objet est l'orthogonal d'un sev F **?** C'est un ensemble, plus précisément c'est un sous-ensemble de vecteurs de l'espace préhilbertien dans lequel on travaille.

Comment le définir ? C'est assez simple à décrire : il s'agit d'un ensemble dans lequel on met tous les vecteurs qui sont orthogonaux à tous les vecteurs de F. Par exemple le vecteur $(0,1)$ est orthogonal à tous les vecteurs de l'axe des abscisses comme on le voit sur le dessin. Donc il appartient à l'orthogonal de l'axe des abscisses. En revanche le vecteur $(1,1)$ lui n'y appartient pas.



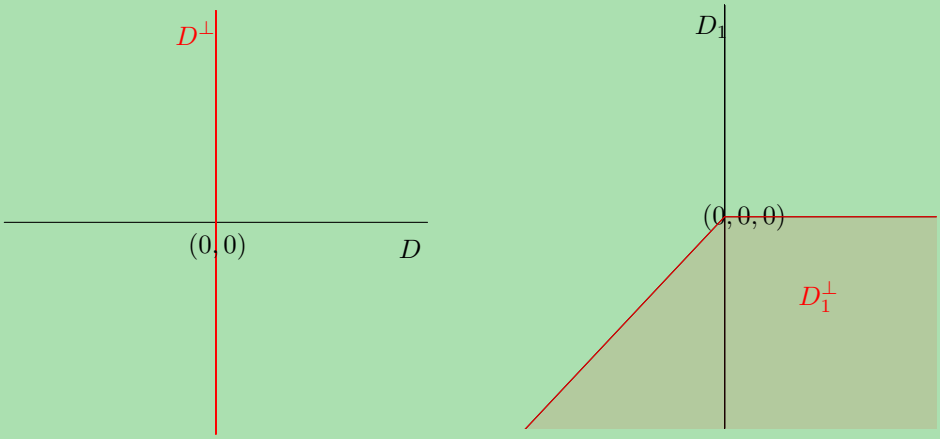
Comment traduire cela mathématiquement ? Un vecteur est orthogonal à tout vecteur de F si son produit scalaire avec tout vecteur de F est nul. C'est l'origine de la définition suivante :

Définition 8:

Soit F un sev de E, on appelle orthogonal de F l'ensemble $F^\perp = \{u \in E | \forall v \in F, \langle u, v\rangle = 0\}$.

Exemple 8:

On se place dans $\mathbb{R}^2$. Soit la droite vectorielle $D = \text{Vect}(1,0)$ de $\mathbb{R}^2$. Alors l'orthogonal de D est l'ensemble des vecteurs qui sont orthogonaux à tout vecteur de D : c'est une autre droite notée $D^\perp$ sur le dessin ci-dessous. Elle est donnée par les vecteurs (x,y) vérifiant $\langle (x,y),(1,0)\rangle = 0$ cad $x = 0$: il s'agit de l'axe des ordonnées. Plaçons nous maintenant dans $\mathbb{R}^3$, l'orthogonal de la droite D_1 est le plan $D_1^\perp$. A l'inverse l'orthogonal d'un plan est une droite passant par $(0,0)$: sur le dessin par exemple D_1 est l'orthogonal de $D_1^\perp$.



Exemple 9: Orthogonal de 0_E et E

On se place dans E un ev préhilbertien. Quel est l'orthogonal de $\{0_E\}$?

Tout vecteur de E est orthogonal au vecteur nul car $\forall u \in E, \langle u, 0_E\rangle = 0$. Donc $\{0_E\}^\perp = E$.

Quel est l'orthogonal de E ? Il s'agit de trouver l'ensemble des vecteurs orthogonaux à tout vecteur de E. Soit u un tel vecteur, étant orthogonal à tout vecteur de E alors il est orthogonal à lui-même ! Donc $\langle u, u\rangle = 0$ et comme le produit scalaire est défini, alors $u = 0_E$. Donc $E^\perp = \{0_E\}$.

En résumé, tout le monde est orthogonal au vecteur nul et si on est orthogonal à tout le monde, c'est qu'on est le vecteur nul.

Si on regarde de plus près les exemples précédents, on note que tous les espaces orthogonaux sont des espaces vectoriels : ils contiennent le vecteur nul 0_E et sont soit une droite vectorielle soit un plan soit l'ev entier. La proposition suivante prouve que ceci n'est pas un hasard.

Proposition 7:

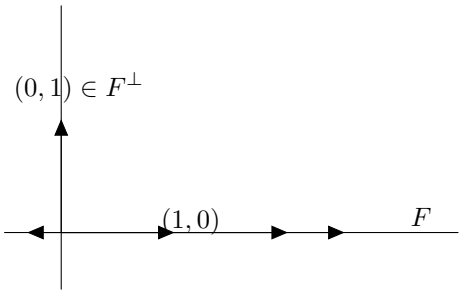
Soit F un sev de E, alors $F^\perp$ est un sev de E.

Preuve :

En exercice c'est un des problèmes de la fin de partie.

Comment déterminer simplement l'orthogonal d'un sev ?

A priori cela semble fastidieux : il s'agirait de voir si un vecteur donné est orthogonal à tous les vecteurs de l'ev considéré. En somme, il faudrait calculer autant de produits scalaires que de vecteurs dans l'ev soit éventuellement une infinité ! En réalité c'est beaucoup plus simple. Considérons par exemple le vecteur $(1,0)$: alors $(0,1)$ lui est orthogonal. Il apparait évident que $(0,1)$ est aussi orthogonal à tout vecteur lié à $(1,0)$ comme $(2,0),(-3,0)$... Ainsi $(0,1) \in \text{Vect}(1,0)^\perp$. On voit ici que pour être orthogonal à l'ev $\text{Vect}(1,0)$, il suffit d'être orthogonal à $(1,0)$ c'est-à-dire à une base de cet ev. La proposition suivante le confirme.



Proposition 8:

Soit E un ev euclidien, F un sev de E, un vecteur est dans l'orthogonal de F si et seulement si il est orthogonal à tout vecteur d'une base de F.

Preuve :

Soit un tel vecteur $u \in E$ et une base $(e_1, \dots, e_p)$ de F.
• $\Longrightarrow$ Si u est orthogonal à F alors par définition il est orthogonal à tout vecteur de F donc en particulier aux e_i.

• $\Longleftarrow$ Supposons u orthogonal aux e_i. Soit $v \in F$, alors il existe des constantes $\lambda_1, \dots, \lambda_p$ telles que

$$v = \sum_{i=1}^{p} \lambda_i e_i.$$

Donc par linéarité à droite

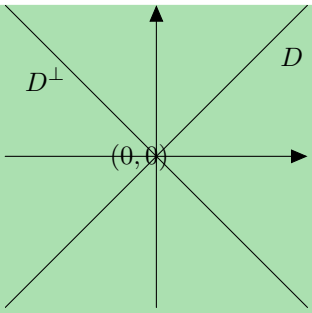
$$\langle u, v\rangle = \langle u, \sum_{i=1}^{p} \lambda_i e_i\rangle = \sum_{i=1}^{p} \lambda_i \langle u, e_i\rangle = 0$$

Donc u est orthogonal à v et ceci vrai pour tout v dans F. Donc u est orthogonal à F.

Exemple 10: Pourquoi est-ce utile en pratique ?

Déterminons l'orthogonal dans $\mathbb{R}^2$ de la droite vectorielle, $D = \text{Vect}(u)$ où $u = (1,1)$. La proposition précédente simplifie le problème puisque pour chercher les vecteurs qui sont orthogonaux à D, il suffit de chercher ceux qui sont orthogonaux à u qui est une base de D par définition.

Ainsi $D^\perp = \{v = (x,y) \in \mathbb{R}^2 | \langle v, u\rangle = 0\} = \{v = (x,y) \in \mathbb{R}^2 | x + y = 0\}$ est une droite d'équation $x + y = 0$. C'est celle représentée sur le dessin ci contre :



Exemple 11: Déterminer l'orthogonal d'un sev en pratique

Déterminer l'orthogonal de $F = \{(x,y,z) \in \mathbb{R}^3, x+2y+5z=0, x+y=0\}$.

Etape 1 : trouver une base de F :

$$\begin{cases} x+2y+5z=0 \\ x+y \quad\quad =0 \end{cases} \underset{L_1 \leftarrow L_1 - L_2}{\iff} \begin{cases} \quad y+5z=0 \\ x+y \quad\quad =0 \end{cases} \iff \begin{cases} z=-y/5 \\ x=\ -y \end{cases}$$

Donc $(x,y,z) \in F \iff (x,y,z) = (-y,y,-y/5) = y(-1,1,-1/5)$.

Notons $v = (-1,1,-1/5)$. La famille (v) est génératrice de F d'après ce qui précède. Elle est libre puisqu'il s'agit d'un vecteur non nul. Donc (v) est une base de F.

Etape 2 : En déduire $F^\perp$:
$F^\perp$ est l'ensemble des vecteurs qui sont orthogonaux à la base (v) :

$$F^\perp = \{u = (x,y,z) \in \mathbb{R}^3 | \langle u, v\rangle = 0\} = \{u = (x,y,z) \in \mathbb{R}^3 | -x+y-z/5 = 0\}.$$

C'est un plan de $\mathbb{R}^3$.

▸ **Exercice 5**. Montrer que si F est un sev de E alors $F^\perp$ l'est aussi.

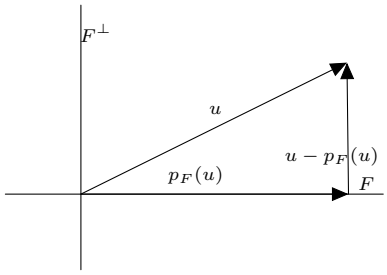
Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Dessiner l'orthogonal d'une droite passant par $(0,0,0)$ dans $\mathbb{R}^3$.
- Comment déterminer en pratique l'orthogonal d'un sev ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

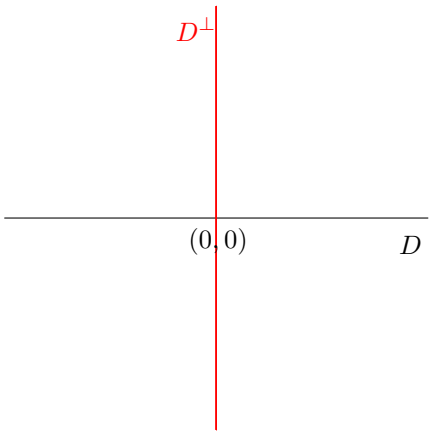
3.4.2 Décomposition fondamentale de l'espace

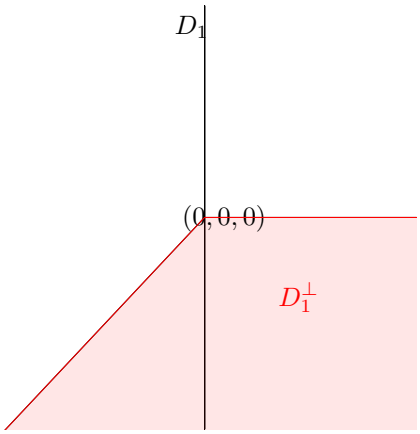
Ne perdons pas de vue notre objectif : définir la notion de projection orthogonale en vue de calculer la distance à un sev donné et de pouvoir résoudre le problème de la membrane. On a vu auparavant que définir une projection vectorielle sur un sev F parallèlement à G nécessite une décomposition de l'ev en somme directe : $E = F \oplus G$.

La façon la plus naturelle de définir la projection orthogonale sur F serait comme projection sur F parallèlement à $F^\perp$. Par ailleurs, sur un dessin on voit bien qu'on projette parallèlement à $F^\perp$. Pour que cette définition soit valide, il faudrait que $E = F \oplus F^\perp$.



Alors partons de l'exemple :





Il apparaît clair que l'intersection entre un espace et son orthogonal est réduit à $(0,0)$ et que $\dim(E) = \dim(F)+\dim(F^\perp)$. En effet, pour le dessin de gauche, $E = \mathbb{R}^2$ est de dimension 2 et $\dim(D) = \dim(D^\perp) = 1$. Pour le dessin de droite, $E = \mathbb{R}^3$ est de dimension 3, $\dim(D_1) = 1$ et $\dim(D_1^\perp) = 2$. Sur ces exemples, les ev F et $F^\perp$ sont supplémentaires. Il y a un théorème général qui généralise cela.

Théorème 2: fondamental

Soit E un espace préhilbertien et F un sev de E **de dimension finie** alors $E = F \oplus F^\perp$.
$F^\perp$ est le supplémentaire orthogonal de F.

Preuve :

On se contente ici de démontrer que les espaces sont en somme directe c'est-à-dire que $F \cap F^\perp = \{0_E\}$. Il faut cependant remarquer que pour montrer que ces espaces sont supplémentaires dans E, il faut prouver que $E = F + F^\perp$.

Si un vecteur u est dans $F \cap F^\perp$ alors il est orthogonal à lui-même : $\langle u, u\rangle = 0$ et donc $u = 0_E$ puisque le produit scalaire est une forme définie. Ceci montre que $F \cap F^\perp \subset \{0_E\}$.

On a par ailleurs $\{0_E\} \subset F \cap F^\perp$ car F et $F^\perp$ sont deux sev de E. D'où $F \cap F^\perp = \{0_E\}$.

▸ **Exercice 6**. Si on prend l'orthogonal de l'orthogonal d'un ev F retombe-t-on sur F ? Démontrer la proposition suivante : **Soit E un ev euclidien et F un sev de E, alors** $(F^\perp)^\perp = F$**.** Vous procéderez par inclusion + raisonnement sur la dimension.

Avec ce théorème fondamental, on est paré pour définir la notion de projection orthogonale.

Questions :

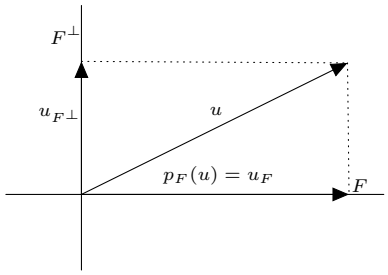
- Quels sont les points essentiels de ce paragraphe ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

3.4.3 Définition de la projection orthogonale

Si F est un sev de dimension finie d'un ev préhilbertien alors $E = F \oplus F^\perp$. Ainsi comme le montre le dessin sur $\mathbb{R}^2$ ci-contre tout vecteur u peut se décomposer de manière unique de la façon suivante

$$u = u_F + u_{F^\perp}.$$

Ainsi un vecteur a un unique projeté sur F parallèlement à $F^\perp$. Ceci permet de définir la projection orthogonale sur F.



Définition 9:

Soit F un sev de dimension finie d'un espace préhilbertien : on appelle projecteur orthogonal p_F de E sur F le projecteur sur F parallèlement à $F^\perp$. Il est défini par

$$\begin{array}{rccc} p_F & : & E = F \oplus F^\perp & \to & F \\ & & u = u_F + u_{F^\perp} & \mapsto & u_F \end{array}$$

Questions :

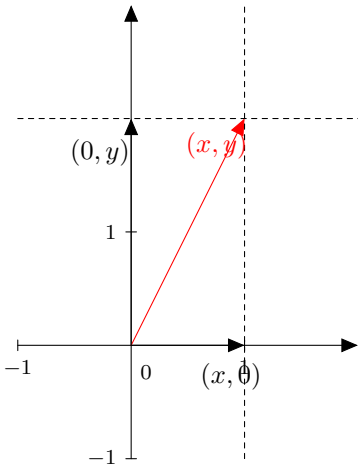
- Quels sont les points essentiels de ce paragraphe ?
- Quelle est la condition essentielle permettant de définir la projection orthogonale.
- Faire les questions QCM du paragraphe.

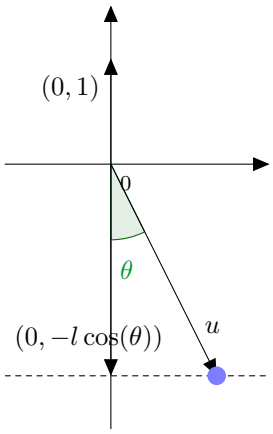
3.5 Comment calculer la projection orthogonale ?

3.5.1 Propriétés fondamentales de la projection orthogonale

Comment trouver l'expression de la projection orthogonale sur un sev de dimension finie donné ?

Le plus raisonnable est d'essayer de commencer sur un exemple des plus simples : on se place dans $\mathbb{R}^2$ et déterminons la projection orthogonale du vecteur $u = (x, y)$ sur l'axe des ordonnées cad sur l'ev $\text{Vect}(0, 1)$. La projection orthogonale a pour coordonnées $(0, y) = y(0, 1)$.



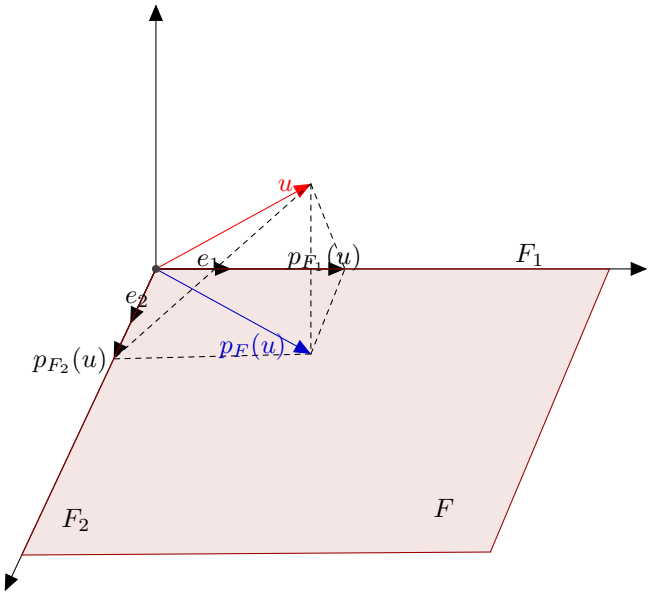


De même en physique, vous avez eu à déterminer la projection sur l'axe des ordonnées d'un pendule. Celle-ci est donnée par $(0, -l\cos(\theta)) = -l\cos(\theta)(0, 1)$ (cosinus=côté adjacent sur hypoténuse). Ces deux expressions sont très semblables : elles témoignent du fait que la projection du vecteur est un vecteur dirigé par le vecteur unitaire $(0, 1)$ pondéré par un coefficient y dans le premier cas, $-l\cos(\theta)$ dans le second. Qu'ont ces deux coefficients en commun ?

Le second nous met la puce à l'oreille : $-l\cos(\theta)$ est exactement le produit scalaire $\langle (u, (0, 1)\rangle = ||u||_2||(0, 1)||_2 \cos(\pi - \theta) = -l\cos(\theta)$. On remarque par ailleurs que pour le premier exemple, on a $y = \langle (x, y), (0, 1)\rangle$. C'est trop pour être un hasard : si vous faites la projection d'un vecteur u sur un sev 1d $F = \text{Vect}(e_1)$ où e_1 est un vecteur **unitaire**, vous avez toujours

$$p_F(u) = \langle u, e_1\rangle e_1.$$

Compliquons maintenant les choses : déterminons la projection orthogonale de u sur un sev de dimension 2 $F = \text{Vect}(e_1, e_2)$ où (e_1, e_2) est une **bon** de F.



Il apparaît clair sur le dessin que $p_F(u)$ est obtenu comme somme de la projection orthogonale de u sur $F_1 = \text{Vect}(e_1)$ et de la projection orthogonale de u sur $F_2 = \text{Vect}(e_2)$:

$$p_F(u) = p_{F_1}(u) + p_{F_2}(u).$$

Maintenant, on sait déterminer $p_{F_1}(u)$ et $p_{F_2}(u)$ grâce au cas 1d ! On a

$$p_F(u) = \langle u, e_1\rangle e_1 + \langle u, e_2\rangle e_2.$$

Et si on augmente encore la dimension de F ? En dimension 3, pour (e_1, e_2, e_3) une **bon** de F, on a envie de dire que

$$p_F(u) = \langle u, e_1\rangle e_1 + \langle u, e_2\rangle e_2 + \langle u, e_3\rangle e_3.$$

Et pour F de dimension p quelconque ? Pour $(e_1, \ldots, e_p)$ une **bon** de F, on a envie de dire que

$$p_F(u) = \langle u, e_1\rangle e_1 + \ldots + \langle u, e_p\rangle e_p.$$

Proposition 9:

Soit F un sev de dimension finie d'un espace préhilbertien E et $(e_1, \ldots, e_p)$ **une bon** de F, alors

1. $\forall u \in E, \quad p_F(u) = \sum_{j=1}^{p} \langle u, e_j\rangle e_j.$
2. $u - p_F(u) \in F^{\perp}$.

Remarque 4:

En résumé la projection orthogonale sur un sev de dimension p est la somme des projections orthogonales sur chacun des sev 1D engendrés par les vecteurs de la bon.

Preuve :

On sait que $E = F \oplus F^{\perp}$ d'après le théorème fondamental. Donc

$$\forall u \in E, \exists!(v, w) \in F \times F^{\perp}, u = v + w$$

et $v = p_F(u), w = p_{F^{\perp}}(u)$. L'idée est de démontrer qu'une autre décomposition est donnée par

$$u = \sum_{j=0}^{p} \langle u, e_j\rangle e_j + u - \sum_{j=0}^{p} \langle u, e_j\rangle e_j.$$

Il suffit alors de montrer que $\sum_{j=0}^{p} \langle u, e_j\rangle e_j \in F$ et que $u - \sum_{j=0}^{p} \langle u, e_j\rangle e_j \in F^{\perp}$. Alors par unicité de la décomposition, on aurait,

$$p_F(u) = \sum_{j=0}^{p} \langle u, e_j\rangle e_j \quad \text{et} \quad u - p_F(u) = u - \sum_{j=0}^{p} \langle u, e_j\rangle e_j \in F^{\perp}$$

ce qui démontre la proposition.

Pourquoi $\sum_{j=0}^{p} \langle u, e_j\rangle e_j \in F$ **?** : car c'est une combinaison linéaire des e_j appartenant à F et que F est un sev de E.

Pourquoi $v = u - \sum_{j=0}^{p} \langle u, e_j\rangle e_j \in F^{\perp}$ **?** : on montre pour cela que ce vecteur est orthogonal à la base $(e_1, \ldots, e_p)$ de F. Soit $i \in \{1, \ldots, p\}$, alors par linéarité à gauche du produit scalaire :

$$\langle v, e_i\rangle = \langle u, e_i\rangle - \sum_{j=0}^{p} \langle \langle u, e_j\rangle e_j, e_i\rangle = \langle u, e_i\rangle - \sum_{j=0}^{p} \langle u, e_j\rangle \langle e_j, e_i\rangle.$$

Comme $(e_1, \dots, e_p)$ est une bon de F alors le seul terme non nul de la somme est celui pour $j = i$. Donc

$$\langle v, e_i \rangle = \langle u, e_i \rangle - \langle u, e_i \rangle \langle e_i, e_i \rangle = 0.$$

Cette proposition est très utile en pratique : elle offre deux méthodes pour déterminer l'expression de la projection orthogonale sur un sev de dimension finie. Nous étudions les deux façons de faire dans les deux paragraphes suivants.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Pourquoi dans la preuve du théorème, l'argument bon est-il crucial ? Cela marcherait-il pour une base ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

3.5.2 La calculer à l'aide de l'orthogonal

Le principe : On utilise le fait que pour tout u dans E, $p_F(u) \in F$ et $u - p_F(u) \in F^\perp$.

Exemple 12: en pratique

Par exemple plaçons-nous dans $\mathbb{R}^2$ et déterminons l'expression de la projection orthogonale sur $F = \text{Vect}(v)$ où $v = (1, 1)$.

a) Trouver une base de F : une base de F n'est autre que (v). Ceci est important car sachant que $p_F(u) \in F$, on peut la décomposer dans cette base.

b) Décomposer la projection dans cette base :

$$\exists! a \in \mathbb{R}, \quad p_F(u) = av = (a, a).$$

Ne reste plus qu'à trouver a et c'est gagné.

c) Utiliser $u - p_F(u) \in F^\perp$: On sait que $u - p_F(u)$ est orthogonal à toute base de F. On en déduit que $\langle u - p_F(u), v \rangle = 0$. Donc $\langle (x - a, y - a), (1, 1) \rangle = 0$. Donc $a = \dfrac{x + y}{2}$. Donc

$$\forall u = (x, y) \in \mathbb{R}^2, \quad p_F(u) = (a, a) = \left(\frac{x + y}{2}, \frac{x + y}{2} \right).$$

Remarque 5:

Le calcul est d'autant plus fastidieux que F est de dimension élevée car alors on a à l'étape c) un système linéaire avec autant d'équations que la dimension. L'intérêt de cette méthode est qu'on n'a jamais besoin d'une bon de F : on peut se contenter d'une base de F.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Faire les questions QCM du paragraphe.

3.5.3 La calculer à l'aide d'une bon

Le principe : Utiliser l'expression en bon de p_F .

Exemple 13: en pratique

Par exemple plaçons-nous dans $\mathbb{R}^2$ et déterminons l'expression de la projection orthogonale sur $F = \text{Vect}(v)$ où $v = (1,1)$.

a) Trouver une base de F : une base de F n'est autre que (v).

b) Trouver une bon de F : Pour fabriquer une bon, comme il n'y a qu'un vecteur, il suffit de normaliser le vecteur (diviser par sa norme). Ainsi $e_1 = v/||v||_2 = \frac{1}{\sqrt{2}}(1,1)$ est une bon de F.

c) Connaître l'expression de la projection orthogonale sur F : Dès lors, d'après le cours,

$$\forall u = (x,y) \in \mathbb{R}^2, p_F(u) = \langle u, e_1\rangle e_1 = \frac{\langle u, v\rangle}{||v||^2} v = \frac{x+y}{2}(1,1) = \left(\frac{x+y}{2}, \frac{x+y}{2}\right).$$

Faisons un autre exemple, déterminons l'expression de la projection orthogonale sur $F = \{(x,y,z,t) \in \mathbb{R}^4 | x+y=0, x+z+t=0\}$. Reprenons la méthode :

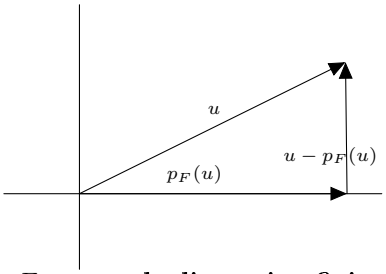
a) Trouver une base de F : $(x,y,z,t) \in F \iff (x,y,z,t) = (x,-x,z,-x-z) = x(1,-1,0,-1) + z(0,0,1,-1)$.

Notons $v = (1,-1,0,-1)$, $w = (0,0,1,-1)$. La famille (v,w) est génératrice de F d'après ce qui précède. Elle est libre puisque ces deux vecteurs ne sont clairement pas colinéaires. Donc (v,w) est une base de F.

b) Trouver une bon de F : Cette étape est essentielle pour terminer l'exercice. Lorsqu'on n'avait qu'un vecteur de base, il suffisait de le normaliser. Mais si on n'a plus d'un vecteur de base, comment construire une bon ? La réponse est donnée à la section suivante.

▸ **Exercice 7**.

Plaçons-nous dans le cas particulier de $\mathbb{R}^2$ et représentons la projection orthogonale sur le sev $F = \text{Vect}((1,0))$. Il apparaît clair que la longueur du vecteur $p_F(u)$ est inférieure à celle de u par une simple application du théorème de Pythagore. Ce résultat se généralise à un ev préhilbertien quelconque.



L'objectif de cet exercice est de démontrer la proposition suivante : **Soit F un sev de dimension finie d'un espace préhilbertien E, soit p_F la projection orthogonale sur F alors $\forall u \in E, ||p_F(u)||_2 \leq ||u||_2$.**

Indication : Partez de $||u||_2^2 = ||p_F(u) + u - p_F(u)||_2^2$.

▸ **Exercice 8**. Si on prend l'orthogonal de l'orthogonal d'un ev F retombe-t-on sur F ? Démontrer la proposition suivante : **Soit E un ev préhilbertien et F un sev de E de dimension finie, alors $(F^\perp)^\perp = F$.**

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Faire les questions QCM du paragraphe.

3.6 Orthonormalisation de Gram Schmidt

L'objectif de cette section est à partir d'une base donnée d'un ev E de construire une bon de E. Il existe une procédure algorithmique permettant d'y parvenir. Nous la présentons dans un premier temps sans la

justifier. Nous expliquons ensuite d'où viennent ces formules "magiques".

Remarque 6: Le procédé d'orthonormalisation de Gram-Schmidt

Donnons la méthode générale pour fabriquer une bon d'un ev de dimension n (cette bon a donc n vecteurs).

- On normalise le premier vecteur : $\epsilon_1 = \dfrac{e_1}{||e_1||}$.
- On crée un vecteur orthogonal à ϵ_1 : $f_2 = e_2 - \langle e_2, \epsilon_1\rangle \epsilon_1$. On le normalise pour obtenir ϵ_2.
- On crée un vecteur orthogonal à ϵ_1 et ϵ_2 : $f_3 = e_3 - \langle e_3, \epsilon_1\rangle \epsilon_1 - \langle e_3, \epsilon_2\rangle \epsilon_2$. On le normalise pour obtenir ϵ_3.
- ...
- Supposons qu'on ait construit au rang p une famille orthonormée $(\epsilon_1, \dots, \epsilon_p)$. On crée un vecteur orthogonal à ces p vecteurs qu'on normalise pour obtenir ϵ_{p+1}.

Pour créer un vecteur orthogonal à $(\epsilon_1, \dots, \epsilon_p)$, voici la formule qui généralise les précédentes (à connaître)

$$f_{p+1} = e_{p+1} - \sum_{i=1}^{p} \langle e_{p+1}, \epsilon_i\rangle \epsilon_i.$$

Notez que dans cette formule comme dans les précédentes, le vecteur choisi est combinaison linéaire de tous les vecteurs précédemment construits (au rang $p+1$, p ont été construits). L'algorithme s'arrête lorsque vous avez construit n vecteurs.

Exemple 14: Un exemple concret ?

Construire une bon en appliquant le procédé d'orthonomalisation de Gram Schmidt à la base de $\mathbb{R}^3$: $(e_1, e_2, e_3) = ((1,1,0),(1,0,1),(0,1,1))$.

Il s'agit de calquer la démonstration précédente :
On normalise le premier vecteur pour obtenir

$$\epsilon_1 = \frac{e_1}{||e_1||_2} = \frac{1}{\sqrt{2}}(1,1,0).$$

On construit ensuite un vecteur orthogonal à ϵ_1

$$f_2 = e_2 - \langle e_2, \epsilon_1\rangle \epsilon_1 = (1,0,1) - \frac{1}{2}(1,1,0) = \frac{1}{2}(1,-1,2).$$

On a alors

$$\epsilon_2 = \frac{f_2}{||f_2||_2} = \frac{1}{\sqrt{6}}(1,-1,2).$$

On construit enfin un vecteur orthogonal à ϵ_1 et ϵ_2

$$f_3 = e_3 - \langle e_3, \epsilon_1\rangle \epsilon_1 - \langle e_3, \epsilon_2\rangle \epsilon_2 = (0,1,1) - \frac{1}{2}(1,1,0) - \frac{1}{6}(1,-1,2) = \frac{1}{3}(-2,2,2).$$

On a alors

$$\epsilon_3 = \frac{f_3}{||f_3||_2} = \frac{1}{\sqrt{12}}(-2,2,2).$$

$(\epsilon_1, \epsilon_2, \epsilon_3)$ est une bon de $\mathbb{R}^3$.

D'où sortent les formules des vecteurs formant la bon ?

L'idée est la suivante : exploiter le fait que pour tout u, $p_{F^\perp}(u) = u - p_F(u) \in F^\perp$.

Reprenons la construction pour une base de départ $(e_1, \dots, e_n)$ qu'on souhaite transformer en bon $(\epsilon_1, \dots, \epsilon_n)$.

• Le premier vecteur e_1 est normalisé pour obtenir ϵ_1.

• Notons $F = \text{Vect}(\epsilon_1)$. On construit à partir de e_2 un vecteur orthogonal à ϵ_1 via

$$p_{F^\perp}(e_2) = e_2 - p_F(e_2) \in F^\perp.$$

Il est orthogonal à ϵ_1 puisqu'il appartient à $\text{Vect}(\epsilon_1)^\perp$! Or $p_F(e_2) = \langle e_2, \epsilon_1\rangle \epsilon_1$ (expression de la projection orthogonale en bon). On retrouve bien la formule voulue. Le vecteur est enfin normalisé pour donner ϵ_2.

• Notons $G = \text{Vect}(\epsilon_1, \epsilon_2)$. On recommence : le vecteur orthogonal est

$$p_{G^\perp}(e_3) = e_3 - p_G(e_3) \in G^\perp.$$

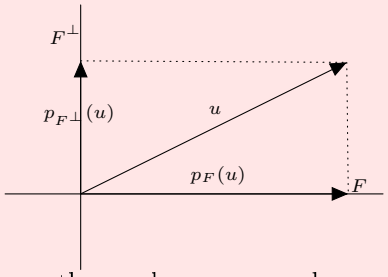
Il est orthogonal à ϵ_1 et ϵ_2 puisqu'il appartient à $\text{Vect}(\epsilon_1, \epsilon_2)^\perp$! Or $p_G(e_3) = \langle e_3, \epsilon_1\rangle \epsilon_1 + \langle e_3, \epsilon_2\rangle \epsilon_2$ (expression de la projection orthogonale en bon). On retrouve à nouveau la formule voulue. Le vecteur est enfin normalisé pour donner ϵ_3.

• On continue avec $H = \text{Vect}(\epsilon_1, \epsilon_2, \epsilon_3)$ et ainsi de suite.

Il est temps maintenant de revenir au calcul de la distance car nous avons construit tous les outils nécessaires.

Remarque 7: importante

On sait que pour p_F (resp $p_{F^\perp}$) projection orthogonale sur F (resp. $F^\perp$) parallélement à $F^\perp$ (resp F) : $u = u_F + u_{F^\perp} = p_F(u) + p_{F^\perp}(u)$.
Donc $p_F = id_E - p_{F^\perp}$.



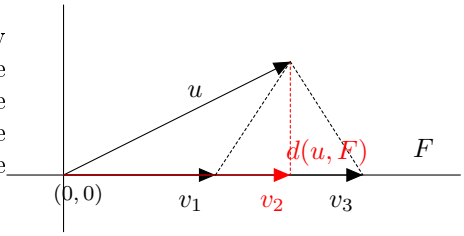
Ce constat s'avère utile en pratique lorsqu'on demande la projection orthogonale sur un sev de grande dimension. Déterminer une bon d'un ev de grande dimension est fastidieux. La dimension de $F^\perp$ vaut $n - \dim(F)$: elle est donc petite si celle de H est grande. Ainsi déterminer $p_{F^\perp}$ peut s'avérer pratique pour obtenir p_F : vous en verrez un exemple en TD.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Qu'est-ce qui fait que la méthode d'orthonormalisation de Gram-Schmidt permette de construire une bon ? D'où sortent les formules ?
- Faire les questions QCM du paragraphe.

3.7 Calculer la distance à l'aide de la projection orthogonale

Nous avons vu comment calculer la projection orthogonale sur un sev donné. Il est temps de revenir au problème initial pour achever le problème de la membrane. Nous devions pouvoir calculer la distance d'un vecteur à un sev. Nous avions intuité sur un dessin que cette distance était déterminée par la projection orthogonale sur le sev. Le paragraphe suivant concrétise cela.



3.7.1 Théorème

Théorème 3:

Soit F un sev de dimension finie d'un espace préhilbertien E, soit $u \in E$, alors

$$d(u, F) = ||u - p_F(u)||_2$$

où p_F est la projection orthogonale sur F.

Preuve :

On procède par double inégalité cad qu'on montre tour à tour que $d(u, F) \leq ||u - p_F(u)||_2$ et $d(u, F) \geqslant ||u - p_F(u)||_2$.

• $d(u, F) \leq ||u - p_F(u)||_2$ est évidente car $p_F(u) \in F$ et $d(u, F) = \inf\{||u - v||_2, v \in F\}$.
• Soit $v \in F$, on part de $||u-v||_2 = ||u-p_F(u)+p_F(u)-v||_2$. Comme $p_F(u)-v \in F$ et $u-p_F(u) \in F^\perp$ alors en appliquant le théorème de Pythagore :

$$||u - v||_2^2 = ||u - p_F(u) + p_F(u) - v||_2^2 = ||u - p_F(u)||_2^2 + ||p_F(u) - v||_2^2 \geqslant ||u - p_F(u)||_2^2.$$

Ceci étant vrai pour tout $v \in V$ alors par définition de l'inf, $d(u, F)^2 \geqslant ||u - p_F(u)||_2^2$. ce qui donne l'autre inégalité.

Remarque 8: Méthode pour trouver la distance d'un vecteur à un sev

Cet exercice est un mélange des différentes notions vues précédemment. Soit $u \in E$ et F un sev de E. Le principe est le suivant : pour la distance de u à F, le théorème précédent impose de connaître l'expression de $p_F(u)$. Les étapes sont donc les suivantes :
• Déterminer $p_F(u)$: souvenez-vous, vous avez pour cela deux méthodes.
• Calculer la distance en utilisant le théorème : il s'agit juste d'un calcul de norme, le plus dur est la première étape.

Exemple 15:

Déterminez la distance de $\begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$ au sev $\mathcal{D}_2(\mathbb{R})$ des matrices diagonales de $\mathcal{M}_2(\mathbb{R})$ pour le produit scalaire canonique de $\mathcal{M}_2(\mathbb{R})$

Notons $M_\epsilon = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$. D'après le théorème précédent, la distance de M_ϵ à $\mathcal{D}_2(\mathbb{R})$ est donnée par $||M_\epsilon - p_{\mathcal{D}_2(\mathbb{R})}(M_\epsilon)||_2$. L'exercice consiste donc à trouver la projection orthogonale de M_ϵ sur $\mathcal{D}_2(\mathbb{R})$. Pour cela, on détermine une bon de ce sev. Elle est donnée par

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

C'est clairement une base car elle est libre et toute matrice diagonale s'écrit comme combinaison linéaire de ces deux matrices : en effet,

$$\forall (a,b) \in \mathbb{R}^2, \quad \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} = a \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + b \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

C'est une famille clairement orthonormale pour le produit scalaire canonique donc c'est bien une bon de $\mathcal{D}_2(\mathbb{R})$ (ici pas besoin d'orthonormalisation !).

On connait l'expression de la projection orthogonale **en bon** :

$$\forall M \in \mathcal{M}_2(\mathbb{R}), p_{\mathcal{D}_2(\mathbb{R})}(M) = \langle M, E_{11} \rangle E_{11} + \langle M, E_{22} \rangle E_{22}.$$

On en déduit que

$$p_{\mathcal{D}_2(\mathbb{R})}(M_\epsilon) = \langle M_\epsilon, E_{11} \rangle E_{11} + \langle M_\epsilon, E_{22} \rangle E_{22} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Ainsi la distance vaut

$$||M_\epsilon - p_{\mathcal{D}_2(\mathbb{R})}(M_\epsilon)||_2 = \left|\left| \begin{pmatrix} 0 & \epsilon \\ \epsilon & 0 \end{pmatrix} \right|\right|_2 = \sqrt{2}\epsilon.$$

Le résultat est cohérent car à mesure que ϵ tend vers 0, la matrice "se rapproche" de la matrice diagonale identité donc la distance au sev des matrices diagonales doit tendre vers 0 ce qui est le cas.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Expliquer sur un dessin la formule donnant la distance à un sev.
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Faire les questions QCM du paragraphe.

3.8 Synthèse de la démarche

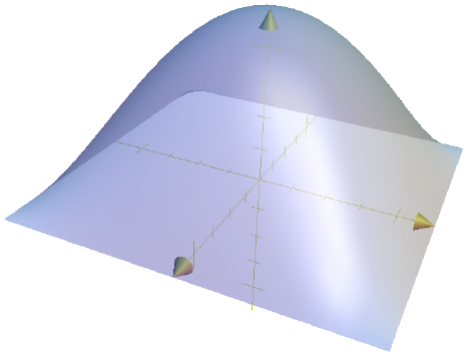
Voici les étapes pour déterminer la distance d'un vecteur à un sev F.

- Déterminer une base du sev F.
- Déterminer la projection orthogonale p_F sur le sev F.
  - En en trouvant une bon de F par orthonormalisation de Gram-Schmidt.
  - OU en utilisant le fait que $u - p_F(u) \in F^{\perp}$.
- Déterminer la distance grâce à la projection orthogonale.

3.9 ⋆ Transition : le problème des membranes bidimensionnel ⋆

Nous avons, grâce au premier chapitre, posé les fondements de la résolution du problème de la membrane élastique en 1D. Il est clairement indispensable pour des applications concrètes de généraliser cela à plusieurs dimensions. La question est alors : comment généraliser en 2D le modèle

$$(P) : \begin{cases} -u''(x) + c(x)u(x) = f(x), & \forall x \in]0,1[\\ u(0) = 0 \\ u(1) = 0 \end{cases}$$



En 2D, l'altitude u dépend maintenant de x et de y : u est donc une fonction à deux variables $u : (x, y) \mapsto u(x, y)$. Le problème 1D fait intervenir la dérivée de u par rapport à x, on imagine que la généralisation en 2D le fera aussi. Sauf qu'on ne sait pas ce qu'est la dérivée d'une fonction à deux variables ni même si ça a un sens quelconque ! Le bon sens impose alors de s'intéresser aux fonctions à plusieurs variables et de construire la notion de "dérivation" pour de telles fonctions.

4 Fonctions à plusieurs variables

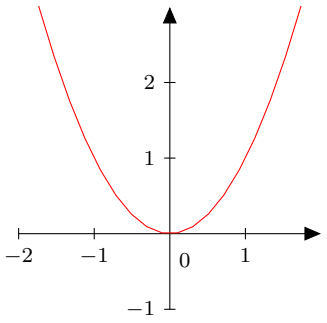
4.1 Continuité

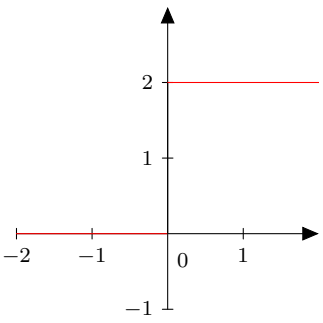
Afin d'étudier l'équation aux dérivées partielles relatives au problème de la membrane, il est indispensable, de définir la notion de dérivée partielle. Lorsqu'on calcule la dérivée d'une fonction à variable réelle, la première chose à faire est de garantir son existence, cad que la fonction est dérivable. En d'autres termes il est nécessaire d'étudier la "régularité" de cette fonction : sa continuité, sa dérivabilité, sa classe. Pour les fonctions à plusieurs variables, c'est pareil. On commence donc par définir ce qu'est la continuité d'une fonction à plusieurs variables : pour cela, il est indispensable de se remémorer de la définition de la continuité pour une fonction d'une variable.

4.1.1 Fonctions à variables réelles

Qu'est-ce que la continuité d'une fonction d'une variable ?

Voici à gauche une fonction continue à gauche et à droite une fonction discontinue en 0.





Voici une définition avec les mains de la continuité : une fonction est continue sur un intervalle I si "on peut la dessiner sur I sans lever le stylo".

Comment traduire cette définition mathématiquement ?

L'idée est la suivante : prenons un stylo dans la main gauche et un autre stylo dans la main droite, plaçons le stylo de gauche en $a-\nu$ sur la courbe et le stylo de droite en $a+\nu$. Parcourons alors la courbe en rapprochant les deux stylos de a. Si la fonction est continue au point a, alors les deux stylos se toucheront au point a.

Derrière cette métaphore se cache les concepts mathématiques permettant de définir la continuité. Le stylo de gauche "s'approchant de a" représente le fait qu'on étudie la limite à gauche de la fonction en a, celui de droite la limite à droite. Enfin les deux stylos se touchent si la limite à gauche est égale à la limite à droite.

Définition 10:

Soit I un intervalle de $\mathbb{R}$, on dit que :

- f est continue en $a \in I$ si et seulement si $\lim\limits_{x\to a} f(x) = f(a)$. Cette condition s'écrit aussi $\lim\limits_{x\to a^+} f(x) = \lim\limits_{x\to a^-} f(x) = f(a)$.
- f est continue sur I si et seulement si f est continue en tout point de I. On note alors $f \in C^0(I)$.

Comment mesure-t-on dans $\mathbb{R}$ que $f(x) \longrightarrow f(a)$ quand $x \to a$?

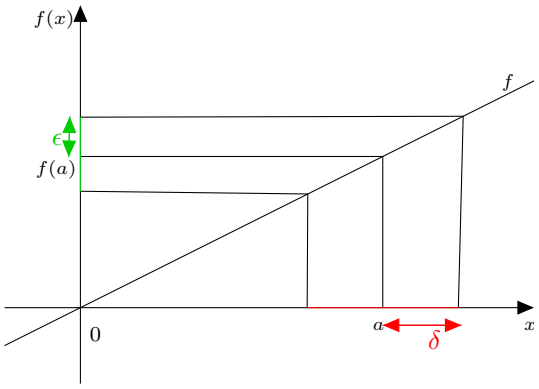
L'outil qui sert à mesurer des distances sur $\mathbb{R}$ est la valeur absolue. La variable x tend vers a quand la distance entre x et a notée $|x-a|$ s'approche de 0. La définition en quantificateurs de la limite

$$\lim_{x\to a} f(x) = f(a)$$

s'écrit d'ailleurs

$$\forall \epsilon > 0, \exists \delta > 0, \forall x \in I, |x-a| \leq \delta \Longrightarrow |f(x)-f(a)| \leq \epsilon.$$

Ceci signifie que quel que soit le voisinage V de $f(a)$, il existe un petit intervalle autour de a où l'image de tout x est dans V (voir dessin).



Comment alors généraliser cela aux fonctions à plusieurs variables ?

Evidemment on va s'inspirer fortement de la définition ci-dessus. Une difficulté apparait cependant immédiatement. Si on considère une fonction de $f : (x,y) \in \mathbb{R}^2 \mapsto f(x,y) \in \mathbb{R}$ et qu'on veut définir la continuité en $a = (a_1, a_2)$, comment évaluer la distance de (x,y) à (a_1,a_2) ? La valeur absolue n'étant valable que sur $\mathbb{R}$, il faut en trouver un analogue sur $\mathbb{R}^2$ et plus généralement sur $\mathbb{R}^n$. C'est la notion de norme qui répond à cela.

Questions :

- Quels sont les points essentiels de ce paragraphe ?

- Pourquoi la continuité en un point a se définit naturellement par l'égalité des limites à gauche et à droite en a ?
- Qcm associé.

4.1.2 Normes

Rappelons la définition de norme.

Définition 11: Norme

Soit E un ev. On appelle norme toute application définie sur E à valeurs dans $\mathbb{R}$ qui est :

— **à valeurs positives :** $\forall u \in E, ||u|| \geq 0$.

— **homogène** : $\forall \lambda \in \mathbb{R}, \forall u \in E, ||\lambda u|| = |\lambda| \cdot ||u||$.

— **définie** : $\forall u \in E, ||u|| = 0 \Longrightarrow u = 0_E$.

— **vérifie l'inégalité triangulaire** : $\forall u, v \in E, ||u + v|| \leq ||u|| + ||v||$.

Donnons-en quelques exemples :

Exemple 16:

• Sur $\mathbb{R}$, la valeur absolue $|\cdot|$ est une norme : elle est bien définie, homogène et vérifie l'inégalité triangulaire.

• Sur $\mathbb{R}^n$, le produit scalaire canonique $\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i$ permet de définir la norme euclidienne $||\cdot||_2$: $u \in \mathbb{R}^n \mapsto \sqrt{\sum_{i=1}^{n} |u_i|^2}$.

• Sur $\mathbb{R}^n$, $||\cdot||_1 : u \in \mathbb{R}^n \mapsto \sum_{i=1}^{n} |u_i|$ et $||\cdot||_\infty : u \in \mathbb{R}^n \mapsto \max |u_i|$ sont également des normes.

• Sur $C^0([a, b])$, on peut par exemple définir les normes $||\cdot||_1 : f \mapsto \int_a^b |f(t)|dt$, $||\cdot||_\infty : f \mapsto \max_{t\in[a,b]} |f(t)|dt$, $||\cdot||_2 : f \mapsto \sqrt{\int_a^b |f(t)|^2 dt}$ (cette dernière est la norme euclidienne issue du produit scalaire $\langle f, g\rangle = \int_a^b f(t)g(t)dt$.

Exemple 17: Montrer que quelque chose est une norme

- L'application $||\cdot||_1$ est à valeurs dans $\mathbb{R}_+$.
- Si $||u||_1 = 0$, alors $\forall i, |u_i| = 0$ puisque c'est une somme de termes positfs donc $\forall i, u_i = 0$.
- $\forall \lambda \in \mathbb{R}, \forall u \in \mathbb{R}^n, ||\lambda u||_1 = \sum_{i=1}^{n} |\lambda u_i| = \sum_{i=1}^{n} |\lambda||u_i| = |\lambda| \sum_{i=1}^{n} |u_i| = |\lambda| ||u||_1$.
- $\forall u \in \mathbb{R}^n, \forall v \in \mathbb{R}^n, ||u + v||_1 = \sum_{i=1}^{n} |u_i + v_i| \leq \sum_{i=1}^{n} |u_i| + |v_i| \leq \sum_{i=1}^{n} |u_i| + \sum_{i=1}^{n} |v_i| \leq ||u||_1 + ||v||_1$.
  Notez que pour les deux derniers axiomes, nous avons utilisé l'homogénéité de la valeur absolue (qui est une norme) et l'inégalité triangulaire sur la valeur absolue.

▸ **Exercice 9**. Démontrez que toutes ces normes sont effectivement des normes.

Essayons de nous représenter géométriquement sur $\mathbb{R}^2$ les normes $||\cdot||_1, ||\cdot||_2, ||\cdot||_\infty$. Pour cela regardons ce que signifie "être éloigné d'une distance au plus 1 de $(0,0)$" cad être de norme inférieure à 1. Cela revient à déterminer les ensembles suivants appelés "boules de centre $(0,0)$ et de rayon 1". $B_1((0,0),1) = \{u \in \mathbb{R}^2, ||u||_1 < 1\}, B_2((0,0),1) = \{u \in \mathbb{R}^2, ||u||_2 < 1\}, B_\infty((0,0),1) = \{u \in \mathbb{R}^2, ||u||_\infty < 1\}$.

Commençons par $B_2((0,0),1) = \{u \in \mathbb{R}^2, ||u||_2 < 1\} = \{u = (x,y) \in \mathbb{R}^2, x^2 + y^2 < 1\}$. On sait que $x^2 + y^2 = 1$ est un cercle de rayon 1. Ainsi être de norme $||\cdot||_2$ inférieure à 1 signifie être dans le disque de rayon 1.

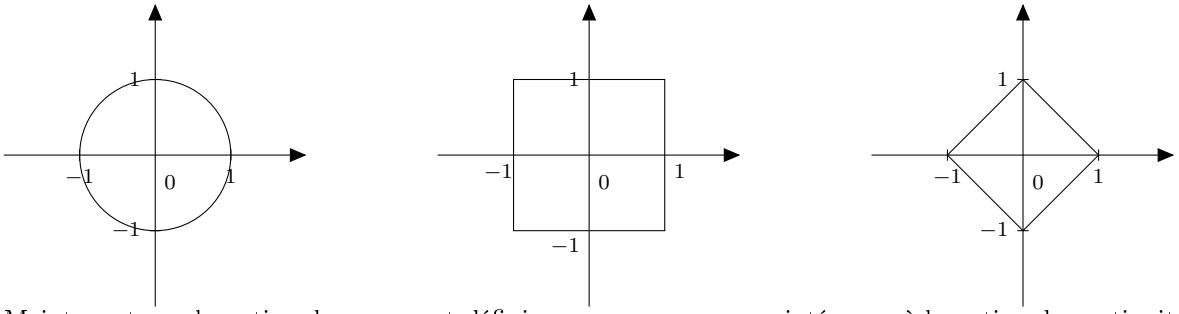
$B_\infty((0,0),1) = \{u \in \mathbb{R}^2, ||u||_\infty < 1\} = \{u = (x,y) \in \mathbb{R}^2, \max(|x|,|y|) < 1\}$.
Donc un vecteur dans cette boule vérifie $|x| < 1$ **et** $|y| < 1$ c'est-à-dire $-1 < x < 1$ et $-1 < y < 1$. Il s'agit donc du carré $[-1,1]^2$ (dessin du centre). Ainsi être de norme ∞ inférieure à 1 signifie être dans un carré centré en $(0,0)$ de côté de longueur 1

$B_1((0,0),1) = \{u \in \mathbb{R}^2, ||u||_1 < 1\} = \{u = (x,y) \in \mathbb{R}^2, |x| + |y| < 1\}$.
Un vecteur dans cette boule vérifie $|x| + |y| < 1$. On a alors 4 cas possibles. (rappelant que $|x| = x$ ou $-x$ suivant qu'il soit positif ou non) :

- soit $x \geqslant 0, y \geqslant 0$: alors $x + y < 1$ et donc $y < 1 - x$.
- soit $x \leq 0, y \geqslant 0$: alors $-x + y < 1$ et donc $y < 1 + x$.
- soit $x \geqslant 0, y \leq 0$: alors $x - y < 1$ et donc $y > x - 1$.
- soit $x \leq 0, y \leq 0$: alors $-x - y < 1$ et donc $y > -1 - x$.

De ces 4 inégalités on constate que y doit être sous les droites d'équations $y = 1 - x$ et $y = 1 + x$ et au dessus des droites d'équations $y = x - 1$ et $y = -1 - x$.

Ainsi chaque norme permet de mesurer à sa façon la distance d'un vecteur à l'origine du repère. Il n'y a pas une unique façon de mesurer cela bien au contraire.



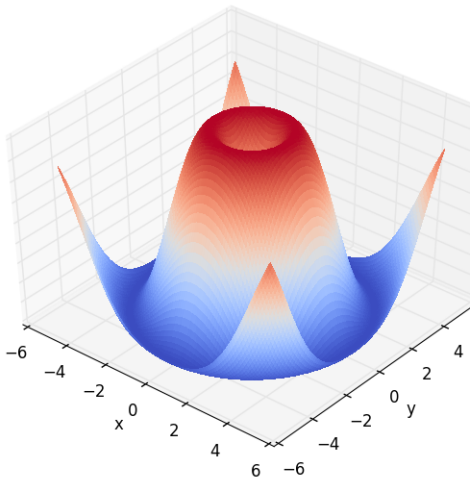
Maintenant que la notion de norme est définie, nous pouvons nous intéresser à la notion de continuité.

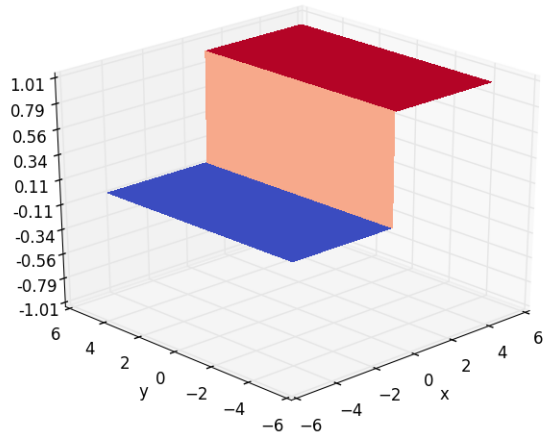
Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quel est l'argument clé dans la démonstration que $||\cdot||_1$ vérifie l'inégalité triangulaire ?
- Qcm associé.

4.1.3 Continuité des fonctions à plusieurs variables

Voici deux exemples de fonctions de $\mathbb{R}^2$ dans $\mathbb{R}$: celle de gauche est continue partout, celle de droite ne l'est pas en un point.





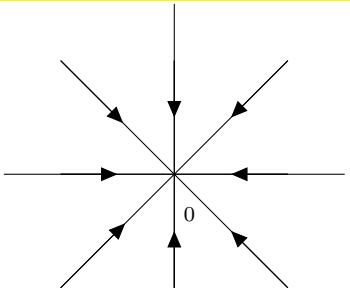
Nous avions précédemment dit qu'une fonction est continue en a si sa limite en a est $f(a)$. Cette définition se généralise à $\mathbb{R}^n$

Définition 12:

Soient A un sous-ensemble de $\mathbb{R}^n$, $f : A \to \mathbb{R}$ et $a \in A$. On se donne une norme quelconque sur $\mathbb{R}^n$, $||\cdot||$:
on dit que f est continue en a si et seulement si $\lim_{x\to a} f(x) = f(a)$ c'est-à-dire si et seulement si

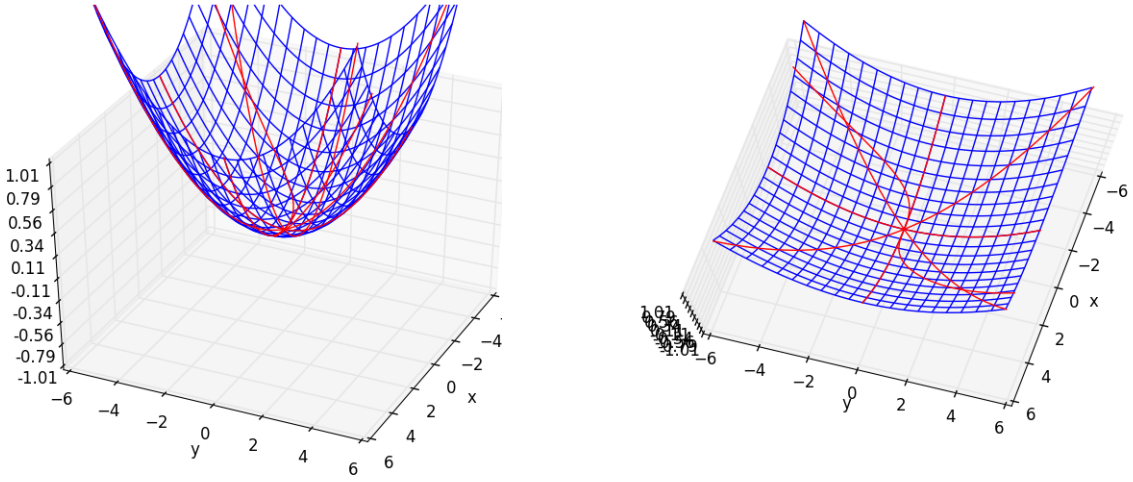
$$\forall \epsilon > 0, \exists \delta > 0, \forall x \in A, ||x - a|| \leq \delta \Longrightarrow |f(x) - f(a)| \leq \epsilon$$

Il est important de comprendre ce que signifie $\lim_{x\to a} f(x) = f(a)$. Pour des fonctions d'une variable réelle, il n'y a que deux façons de tendre vers a : par la gauche ou par la droite. La fonction était alors continue quand limite à gauche et à droite étaient égales. Pour une fonction définie sur $\mathbb{R}^2$, il y a une infinité de façons de tendre (directions) vers un point $a = (a_1, a_2)$ comme on peut le voir sur le dessin ci-contre.



Alors que signifie $\lim_{x\to a} f(x) = f(a)$?

Cela signifie que quelle que soit la façon de tendre vers a, la limite est toujours $f(a)$. Pour la continuité, il faut donc que dans toutes les directions on tende vers $f(a)$. Par exemple, la fonction $(x, y) \mapsto x^2+y^2$ tracée ci-dessous est continue en $(0, 0)$: les lignes rouges sont les images des directions $(x, 0), (0, y), (x, x), (x, -x), (x, 0.5x^3)$ contenant toutes $(0, 0)$, on constate que les images s'intersectent en un point qui est $f(0, 0)$ (pas de saut de continuité).



Notez que pour montrer la non continuité en un point, il suffit de trouver deux directions dont la limite est différente ! On reverra cela tout à l'heure.

Comment en pratique montrer qu'une fonction est continue ?

De la même manière que pour les fonctions d'une variable réelle, on utilise des fonctions de référence continues comme par exemple sur $\mathbb{R}^2$ $p_1 : (x, y) \mapsto x$ et $p_2 : (x, y) \mapsto y$.

Exemple 18:

Les projections canoniques sont continues : $\forall i \in \{1, \dots, n\}, p_i : (x_1, \dots x_n) \mapsto x_i$ est continue sur $\mathbb{R}^n$.

L'idée est ensuite de dire que les fonctions qu'on étudie sont des sommes, produits ou quotients de fonctions continues.

Proposition 10:

Soient A un sous-ensemble de $\mathbb{R}^n$ et $a \in A$, f et g deux fonctions continues en a alors :

- $\forall \lambda \in \mathbb{R}$, $\lambda f + g$ est continue en a.
- $f \times g$ est continue en a.
- si $f(a) \neq 0$, $\dfrac{1}{f}$ est continue en a.

Soit I un intervalle de $\mathbb{R}$ vérifiant $g(A) \subset I$ et $h : I \to \mathbb{R}$.
Si g est continue en a et h est continue en $g(a)$ alors $h \circ g$ est continue en a.

Exemple 19:

La fonction polynomiale suivante est continue sur $\mathbb{R}^3$:

$$f : (x, y, z) \mapsto xy + yz + 2x + 1.$$

En effet, il s'agit de produits et de sommes de fonctions continues sur $\mathbb{R}^3$. De manière générale toute fonction polynomiale de $\mathbb{R}^n$ est continue sur $\mathbb{R}^n$.

Cependant, tout ne se passe pas toujours pour le mieux. Revenons sur $\mathbb{R}$ et considérons la fonction

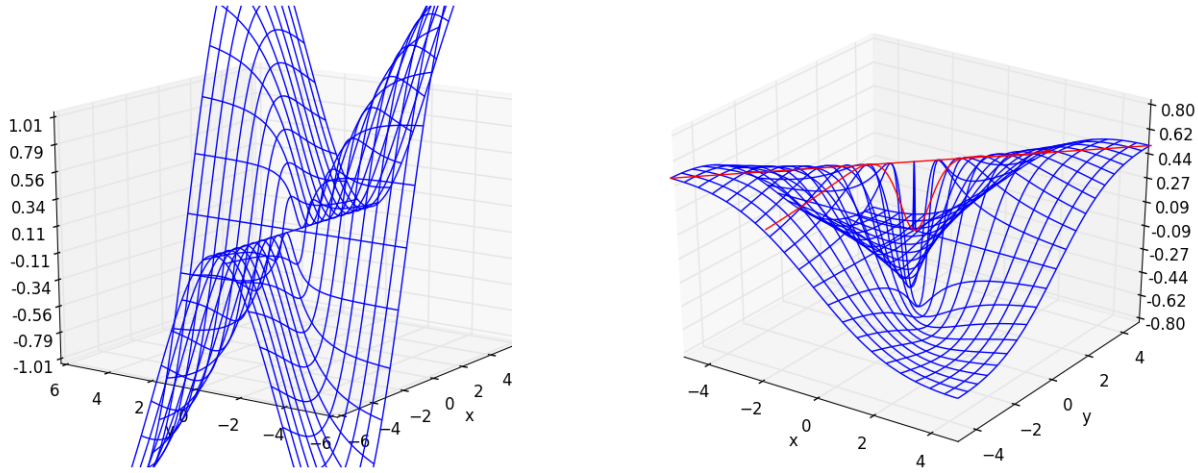
$$f : x \mapsto \begin{cases} \dfrac{\sin(x)}{x} & \text{si } \ x \neq 0 \\ 1 & \text{si } \ x = 0 \end{cases}$$

Cette fonction est continue sur $\mathbb{R}^\star$ comme quotient de fonctions continues dont le dénominateur ne s'annule pas. En revanche en 0, un traitement particulier s'impose : il faut étudier la limite en 0 de $\dfrac{\sin(x)}{x}$. A l'aide d'un DL, on obtient 1 ce qui garantit la continuité.

On peut trouver le même genre de problèmes pour les fonctions à plusieurs variables. Considérons les fonctions

$$f_1 : (x,y) \mapsto \begin{cases} \dfrac{xy^2}{x^2+y^2} & \text{si} \quad (x,y) \neq (0,0) \\ 0 & \text{si} \quad (x,y) = (0,0) \end{cases} \quad , \quad f_2 : (x,y) \mapsto \begin{cases} \dfrac{xy}{x^2+y^2} & \text{si} \quad (x,y) \neq (0,0) \\ 0 & \text{si} \quad (x,y) = (0,0) \end{cases}$$

qui sont continues sur $\mathbb{R}^2 - \{(0,0)\}$ comme quotient de fonctions continues dont le dénominateur ne s'annule pas. Le dessin de gauche représente f_1, celui de droite f_2.



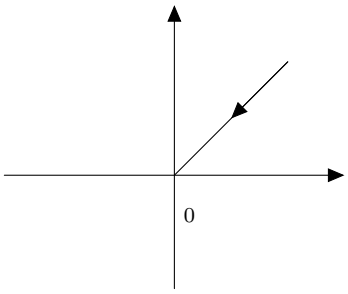
Il est alors légitime de se demander si f_1 et f_2 sont continues en $(0,0)$ c'est-à-dire, d'après la définition de la continuité, si $f_1(x,y) \underset{(x,y)\to(0,0)}{\longrightarrow} f_1(0,0) = 0$ et si $f_2(x,y) \underset{(x,y)\to(0,0)}{\longrightarrow} f_2(0,0) = 0$. La fonction de gauche semble bien être continue en $(0,0)$. En revanche, celle de droite (f_2) présente un comportement singulier en $(0,0)$. Les deux courbes en rouge correspondent aux images de $f(x,x)$ et $f(x,x^3)$ pour x compris entre -5 et 5 : (x,x) et (x,x^3) passent toutes deux par $(0,0)$ et pourtant l'image de $f(0,0)$ ne semble pas être au même endroit.

Comment deviner si les fonctions sont continues ou non ?

Considérons f_1. Son numérateur est xy^2 : quand x et y vont tendre vers 0, xy^2 va tendre avec une vitesse "d'ordre 3" c'est-à-dire qu'on a un produit de 3 quantités tendant vers 0. Au dénominateur, c'est de l'ordre 2. On peut donc intuiter qu'à la fin ne restera de l'ordre 3 sur 2 soit de l'ordre 1. On s'attend donc à ce que $f_1(x,y)$ tende bien vers 0. A l'inverse pour f_2, le numérateur est d'ordre 2. On s'attend donc à ce que f_2 soit d'ordre 0 en (x,y) et donc à ce que la fonction ne soit pas continue en $(0,0)$.

Comment prouver rigoureusement la non continuité en un point problématique ?

Il est plus aisé d'infirmer la continuité d'une fonction en un point que de la montrer. Rappelons que pour montrer qu'une fonction f est continue en un point $a = (a_1, a_2)$, il faut que $f(x,y)$ tende vers $f(a)$ quelle que soit la direction avec laquelle on tend. En revanche pour infirmer la continuité, il suffit de trouver une direction pour laquelle on ne tend pas vers $f(a)$. Pour f_2 en $(0,0)$, on veut trouver une direction dans laquelle $f_2(x,y)$ ne tend pas vers $f_2(0,0) = 0$. En général, on commence par les directions les plus simples : $(x,0)$ et $(0,y)$. Ici dans les deux cas, la limite est 0. Essayons avec (x,x).



On a $\forall x \in \mathbb{R}^\star, f(x,x) = \dfrac{xx}{x^2+x^2} = \dfrac{1}{2}$. Quand (x,x) tend vers $(0,0)$, $f(x,x)$ ne tend pas vers $f(0,0) = 0$. Donc f n'est pas continue en $(0,0)$.

Comment prouver rigoureusement la continuité en un point problématique ?

L'idée est la suivante : pour montrer que $f(x,y)$ tend vers $f(a_1,a_2)$, on essaie de majorer $|f(x,y)-f(a_1,a_2)|$ par quelque chose qui tend vers 0 quand (x,y) tend vers (a_1,a_2). Ainsi par encadrement on aurait le résultat voulu. Il s'agit donc ici d'encadrer $|f_1(x,y) - f_1(0,0)| = |f_1(x,y)|$ (dans ce cas précis $f_1(0,0) = 0$!).

$$\forall (x,y) \neq (0,0), \quad |f_1(x,y)| = \left|\frac{xy^2}{x^2+y^2}\right| = \frac{|x|y^2}{x^2+y^2} = \frac{|x|y^2}{||(x,y)||_2^2}.$$

Donc par définition de la norme infinie qui est le maximum de $|x|$ et $|y|$, on a

$$\forall (x,y) \neq (0,0), \quad |f_1(x,y)| \leq \frac{||(x,y)||_\infty^3}{||(x,y)||_2^2}.$$

Pourquoi faire cela ? Imaginez qu'on ait la même norme au numérateur et au dénominateur alors, on aurait $|f_1(x,y)| \leq ||(x,y)||_2$ et donc $f_1(x,y)$ tendrait bien vers 0 quand (x,y) tend vers $(0,0)$. Seulement on n'a pas la même norme donc cette opération est proscrite. La question qui se pose alors : peut-on estimer une norme en fonction d'une autre ? Peut-on majorer une norme avec une autre ?

Essayons avec les deux normes en présence : on sait que $||(x,y)||_\infty = |x|$ ou $|y|$. Supposons que ce soit $|x|$ (par symétrie le même raisonnement marche pour $|y|$). On a $|x|^2 \leq x^2 + y^2$. On en déduit en prenant la racine carrée que $||(x,y)||_\infty \leq ||(x,y)||_2$.

Ainsi

$$\forall (x,y) \neq (0,0), \quad 0 \leq |f_1(x,y)| \leq \frac{||(x,y)||_2^3}{||(x,y)||_2^2} \leq ||(x,y)||_2.$$

Donc si (x,y) tend vers $(0,0)$ alors par encadrement $f_1(x,y)$ tend vers 0 qui est égal à $f(0,0)$. Donc f est continue en $(0,0)$.

Remarque 9:

La question d'estimer les normes entre elles est une question beaucoup plus générale à laquelle il existe une réponse donné par le **théorème d'équivalence des normes**. Celui-ci assure l'existence de constantes $A > 0, B > 0$ telles que pour deux normes quelconques $||\cdot||$ et $|||\cdot|||$ sur un ev E de dimension finie, on ait

$$\forall u \in E, A||u|| \leq |||u||| \leq B||u||.$$

Nous ne ferons pas la démonstration générale mais en td, nous déterminerons sur des exemples précis de normes les constantes A et B.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Que signifie, en termes de direction, le fait qu'une fonction à plusieurs variables soit continue en un point a ? Est-ce différent pour une fonction d'une variable réelle ?
- Quelle stratégie permet de montrer qu'une fonction de plusieurs variables n'est pas continue en un point ?
- Quelle stratégie permet de montrer qu'une fonction de plusieurs variables est continue en un point non problématique ? en un point problématique ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Qcm associé.

4.2 Dérivées partielles

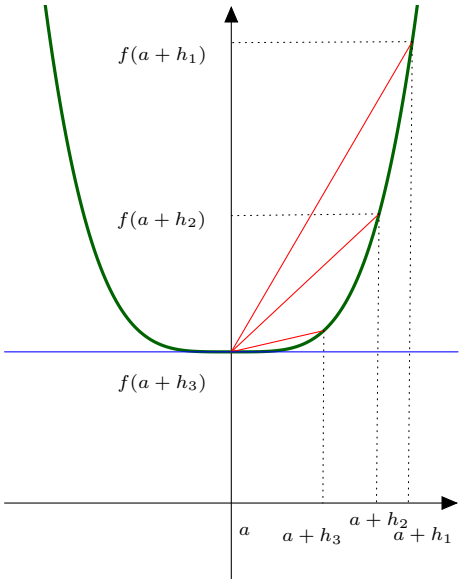
4.2.1 Qu'est-ce qu'une dérivée sur $\mathbb{R}$?

Afin de comprendre, la notion de "dérivée" pour une fonction à plusieurs variables, il est indispensable d'avoir compris le concept pour des fonctions $f : \mathbb{R} \to \mathbb{R}$. Rappelons la définition de f dérivable en un point a de $\mathbb{R}$:

$$\exists l \in \mathbb{R}, \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} = l.$$

Le réel l est alors appelé dérivée de f en a et est noté $f'(a)$.

Géométriquement le taux d'accroissement représente les pentes des cordes dessinées ci-contre en rouge. Lorsque le paramètre tend vers 0, les cordes s'aplatissent et tendent vers la pente de la tangente représentée en bleu. Cette pente est la dérivée $f'(a)$. En d'autres termes une fonction est dérivable en a si la pente des cordes quand $x \to a$ tend vers un réel. Ce réel est la dérivée au point a.

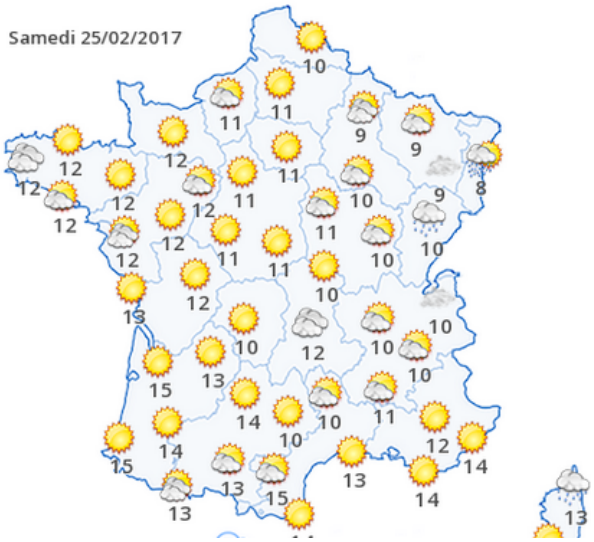


Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Que représente géométriquement le fait que le taux d'accroissement ait une limite ?
- Qcm associé.

4.2.2 Dérivées partielles

Dans la vraie vie, on ne se contente pas de fonction à une variable. Par exemple, les météorologistes vous fournissent au jour le jour l'évolution de la température en différentes villes de France. La température varie de jour en jour, elle dépend donc du temps t. Elle varie également en fonction de la position (x, y, z) où vous vous trouvez. Ainsi il s'agit d'une fonction de quatre variables $T : (t, x, y, z) \mapsto T(t, x, y, z)$. C'est la même chose pour le problème de la membrane : l'altitude u de la membrane dépend de x et y.



Comme nous étudions les variations des fonctions d'une variable réelle à l'aide de la dérivée, il faut donc se munir des outils pour faire de même avec des fonctions à plusieurs variables : les dérivées partielles.

Commençons par le cas de $\mathbb{R}^2$ avant de généraliser à $\mathbb{R}^n$. En maths, pour introduire un nouveau concept, on aime partir de ce qu'on connaît : ici les dérivées de fonctions d'une variable réelle. Alors essayons de définir une notion de dérivée pour $f(x, y) = x + y^2$ par exemple. On ne sait pas dériver par rapport à deux variables en même temps donc choisissons en une et fixons l'autre : si on dérive par rapport x en fixant y, y est indépendante de x donc la dérivée est $\partial_x f(x, y) = 1$. De même en y en fixant x, on a $\partial_y f(x, y) = 2y$. Voilà pour l'aspect calculatoire.

Maintenant ce calcul a ses limites, si on choisit $g(x, y) = x \ln(y)$ par exemple, on ne peut pas calculer la dérivée en tout point car ln n'est pas dérivable sur les négatifs. Il est donc indispensable tout comme sur $\mathbb{R}$ d'introduire la notion de dérivabilité.

Comment définir la dérivabilité par rapport à une des variables ?

On recycle l'idée précédente : on sait définir la dérivabilité pour une fonction d'une variable donc on va fixer une des deux variable et appliquer cette définition. Par exemple pour étudier en quels points (x_0, y_0) la fonction est dérivable par rapport à y, on fixe $x = x_0$ et on regarde quand le taux d'accroissement $\dfrac{x_0 \ln(y_0 + h) - x_0 \ln(y_0)}{h}$.

Définition 13: Dérivées partielles dans $\mathbb{R}^2$

Soit A un sous ensemble de $\mathbb{R}^2$, $f : A \to \mathbb{R}$, $(x_0, y_0) \in A$, on dit que f admet une dérivée partielle par rapport à x en (x_0, y_0) si et seulement si la limite suivante existe et est finie

$$\lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

On définit alors la dérivée partielle par rapport à x en (x_0, y_0) par

$$\partial_x f(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

. De même, si elle existe et est finie, on définit alors la dérivée partielle par rapport à y en (x_0, y_0) par

$$\partial_y f(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}$$

.

Comment montrer qu'une fonction a des dérivées partielles sur un ensemble ?

Une façon de répondre à cette question est de se remémorer ce qu'on faisait pour les fonctions d'une variable réelle. A vrai dire, on n'utilisait le taux d'accroissement qu'en cas de nécessite absolue. Prenons l'exemple de la valeur absolue : cette fonction est dérivable sur $\mathbb{R}^\star$ au moins puisque $x \mapsto x$ et $x \mapsto -x$ sont dérivables sur $\mathbb{R}$. C'est uniquement pour $x = 0$ qu'on utilise le taux d'accroissement pour montrer que

$$\lim_{h\to 0^+} \frac{|0+h|-|0|}{h} = 1 \neq -1 = \lim_{h\to 0^-} \frac{|0+h|-|0|}{h}$$

et donc que la valeur absolue n'y est pas dérivable. En résumé, en majorité, on dit que la fonction est dérivable par somme, produit de quotients dérivables, sauf en les quelques éventuels points problématiques.

Pour les fonctions à plusieurs variables, c'est pareil. Plaçons nous sur $\mathbb{R}^2$ et reconsidérons la fonction à deux variables (déjà étudiée auparavant)

$$f : (x,y) \mapsto \begin{cases} \dfrac{xy}{x^2+y^2} & \text{si} \quad (x,y) \neq (0,0) \\ 0 & \text{si} \quad (x,y) = (0,0) \end{cases}$$

Hors de $(0,0)$, cette fonction admet des dérivées partielles car c'est le quotient de deux fonctions en admettant et dont le dénominateur ne s'annule pas. En revanche en $(0,0)$, l'utilisation du taux d'accroissement est indispensable. On a

$$\forall h \in \mathbb{R}, \quad \frac{f(h,0)-f(0,0)}{h} = 0$$

qui tend vers 0. Donc $\partial_x f(0,0) = 0$. De même

$$\forall h \in \mathbb{R}, \quad \frac{f(0,h)-f(0,0)}{h} = 0$$

qui tend vers 0. Donc $\partial_y f(0,0) = 0$.

L'existence de dérivées partielles implique-t-elle la continuité ?

Pour les fonctions d'une variable réelle, la dérivabilité implique la continuité. Il semblerait logique que la propriété se prolonge pour les fonctions à plusieurs variables. Seulement la fonction f admet des dérivées partielles en $(0,0)$ (on vient de le montrer) mais elle n'est pas continue en $(0,0)$: on l'a montré dans la section continuité.

Alors pourquoi est-ce que ça marche dans $\mathbb{R}$ et pas dans $\mathbb{R}^2$?

Une fonction d'une variable réelle est dérivable en un point dès lors que le taux d'accroissement admet une limite finie en ce point c'est-à-dire que les limites à gauche et à droite de ce taux sont les mêmes (les limites à gauche et droite sont les deux seules possibles dans $\mathbb{R}$). Par exemple, la valeur absolue n'est pas dérivable en 0 car la limite du taux d'accroissement à gauche est -1 alors qu'à droite elle vaut 1. En revanche pour $x \mapsto x$, le taux d'accroissement tend vers 1 en 0 ce qui implique sa dérivabilité en 0.

Pour une fonction à plusieurs variables, l'équivalent de cette notion de dérivabilité serait : il faudrait que le taux d'accroissement quelle que soit la **courbe** (pas forcément une direction vectorielle) avec laquelle on tend vers le point en question ait une limite finie.
Or l'existence de dérivées partielles ne donne l'existence de cette limite que dans deux directions : $(1,0)$ et $(0,1)$. On perd ainsi une quantité d'information importante : il est intuitif que l'existence de ces deux dérivées partielles cantonées aux axes des abscisses et ordonnées ne permettent pas d'assurer la continuité dans les autres directions. En l'occurence pour f, c'est la direction (x,x) qui coince.

Et si on définissait les dérivées partielles pour les fonctions de $\mathbb{R}^n$?
Généralisons encore : ces notions de dérivées partielles sont valables pour des fonctions à n variables. L'idée est la même que pour deux variables : on fixe toutes les variables sauf une en laquelle on prend le taux d'accroissement. En voici la définition mathématique.

Définition 14: Dérivées partielles

Soient A un sous ensemble de $\mathbb{R}^n$ $f : A \to \mathbb{R}$, $a = (a_1, \ldots, a_n) \in A$, on dit que f admet une dérivée partielle par rapport à $x_j, 1 \leq j \leq n$ en a si et seulement si la limite suivante existe et est finie

$$\lim_{h \to 0} \frac{f(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - f(a_1, \ldots, a_{j-1}, a_j, a_{j+1}, \ldots, a_n)}{h}$$

On définit alors la dérivée partielle par rapport à x_j par

$$\partial_{x_j} f(a) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - f(a_1, \ldots, a_{j-1}, a_j, a_{j+1}, \ldots, a_n)}{h}$$

.

Une dérivée partielle n'est rien d'autre que la dérivée d'une fonction d'une variable en fixant toutes les autres. Il n'est alors pas étonnant de voir se prolonger toutes les propriétés de dérivation entrevues sur $\mathbb{R}$ (dérivée d'un produit, d'un quotient, ...). Ainsi pour calculer une dérivée partielle, vous faites comme si vous dériviez une fonction d'une variable.

Questions :

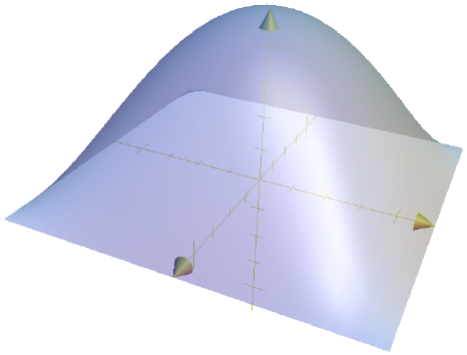
- Quels sont les points essentiels de ce paragraphe ?
- En quoi la définition des dérivées partielles des fonctions à plusieurs variables découle-t-elle de celle des fonctions d'une variable réelle ?
- Pour démontrer l'existence de dérivées partielles en un point, utilise-t-on toujours le taux d'accroissement ? Pourquoi ?
- A une variable, dérivabilité implique continuité. A plusieurs variables, l'existence de dérivées partielle n'implique pas la continuité. A quoi est due cette différence ?
- Qcm associé.

4.2.3 ⋆ Membrane élastique 2d ⋆

Voici ce que devient le problème de la membrane élastique en 2d

$$(P) : \begin{cases} (-\Delta u + cu)(x,y) = f(x,y), & \forall (x,y) \in \Omega =]0,1[^2 \\ u(x,y) = 0 & \forall (x,y) \in \partial\Omega \end{cases}$$

où $\Delta = \partial^2_{xx} + \partial^2_{yy}$ est l'opérateur laplacien (utilisant des dérivées partielles secondes : voir la section "Classe C^2") et $\partial\Omega$ est le bord de Ω. Cela signifie que la membrane est fixe au bord du domaine.



4.2.4 Classe C^1

Ayant défini les dérivées partielles, on peut définir le caractère C^1 d'une fonction. A nouveau on s'inspire de la définition pour une fonction de la variable réelle. On rappelle qu'une fonction de I un intervalle de $\mathbb{R}$ dans $\mathbb{R}$ est de classe $C^1(I)$ si et seulement si elle est dérivable sur I et sa dérivée est continue sur I. Pour les fonctions à plusieurs variables, la notion de dérivée n'existe pas. En revanche, celle de dérivée partielle oui.

Définition 15: Fonctions de classe C^1

Soit A un sous-ensemble de $\mathbb{R}^n$. On dit que $f : A \to \mathbb{R}$ est de classe $C^1(A)$ si seulement si

1. $\forall a \in A, \forall k \in \{1, \dots, n\}, \partial_{x_k} f(a)$ existent.
2. $\forall k \in \{1, \dots, n\}, \partial_{x_k} f$ est continue sur A.

Cette définition donne lieu aux habituelles propriétés garantissant la transmission du caractère C^1 par la somme, le produit et la composition.

Proposition 11: Somme, produit, composée de fonctions de classe C^1

Soit A un sous-ensemble de $\mathbb{R}^n$. Soient $f, g : A \subset \mathbb{R}^n \to \mathbb{R}$, si f et g sont $C^1(A)$ alors

- $f + g$ et $\forall \lambda \in \mathbb{K}, \lambda f, fg \in C^1(A)$.
- Si g ne s'annule pas sur A, alors $\dfrac{f}{g} \in C^1(A)$.

Soient $f : A \to I \subset \mathbb{R}$, $g : I \to \mathbb{R}$ deux applications,

$$\left.\begin{array}{l} f \in C^1(A) \\ g \in C^1(I) \end{array}\right\} \Longrightarrow g \circ f \in C^1(A).$$

Des exemples de fonctions de classe C^1 :

Plaçons nous dans $\mathbb{R}^2$, $p_1 : (x, y) \mapsto x$ est de classe $C^1(\mathbb{R}^2)$ car ses dérivées partielles en tout point valent 0 ou 1 et sont donc continues. De même $p_1 : (x, y) \mapsto y$ l'est aussi. En sommant, multipliant ces deux fonctions (appelées projections canoniques), on construit les fonctions polynomiales comme par exemple $f : (x, y) \mapsto x^3 - xy + 12$. Ces fonctions sont aussi de classe C^1 d'après la proposition précédente.

De manière plus générale, les projections canoniques de $\mathbb{R}^n$ $\forall i \in \{1, \dots, n\}, p_i : (x_1, \dots x_n) \mapsto x_i$ sont de classe $C^1(\mathbb{R}^n)$ ce qui entraîne que les fonctions polynomiales sont de classe $C^1(\mathbb{R}^n)$.

Comment montre-t-on qu'une fonction est de classe C^1 ?

Ceci est un exercice classique. L'idée est la même que pour la continuité de fonctions, partout où il n'y a pas de problème de définition, on dit que notre fonction est C^1 comme somme/produit/quotient de fonctions C^1. En revanche, là où il y a un problème de définition, il faut faire un travail particulier. Reprenons l'exemple de

$$f : (x, y) \mapsto \begin{cases} \dfrac{xy^2}{x^2 + y^2} & \text{si} \quad (x, y) \neq (0, 0) \\ 0 & \text{si} \quad (x, y) = (0, 0) \end{cases}$$

et voyons si elle est de classe $C^1(\mathbb{R}^2)$.

Il s'agit d'une expression possédant une incertitude en $(0, 0)$. Résoudre cet exercice se décline en les étapes suivantes.

1. Expliquer pourquoi la fonction est de classe $C^1(\mathbb{R}^2 - \{(0, 0)\})$ et calculer les dérivées partielles sur $\mathbb{R}^2 - \{(0, 0)\}$.
2. Démontrer ou infirmer l'existence des dérivées partielles en $(0, 0)$.
3. Démontrer ou infirmer la continuité des dérivées partielles en $(0, 0)$.

Notez que les deux dernières étapes suivent exactement la définition des fonctions de classe C^1. Prenons les étapes précédentes une par une :

1. La fonction est le quotient de deux fonctions de classe $C^1(\mathbb{R}^2 - \{(0, 0)\})$ dont le dénominateur ne s'annule pas. Donc elle est de classe $C^1(\mathbb{R}^2 - \{(0, 0)\})$. En tous ces points, on peut donc calculer les dérivées partielles :

$$\forall (x, y) \in \mathbb{R}^2 - \{(0, 0)\}, \quad \partial_x f(x, y) = \frac{y^2(y^2 - x^2)}{(x^2 + y^2)^2}, \quad \partial_y f(x, y) = \frac{2x^3 y}{(x^2 + y^2)^2}.$$

Notez d'ailleurs qu'on voit qu'en $(0,0)$, ces expressions de dérivées partielles ne sont pas valables. Pour montrer leur existence en $(0,0)$, on n'a pas d'autre choix que de revenir au taux d'accroissement. Notez qu'on pourrait également pour calculer les dérivées partielles en (x,y) non nul utiliser le taux d'accroissement. Cependant, c'est fastidieux et il est plus aisé d'utiliser les dérivées usuelles comme on le fait pour les fonctions d'une variable réelle.

2. Pour les dérivées partielles en $(0,0)$ revenons donc aux taux d'accroissements. On a

$$\forall h \in \mathbb{R}, \quad \frac{f(h,0)-f(0,0)}{h} = 0$$

qui tend vers 0. Donc $\partial_x f(0,0) = 0$. De même

$$\forall h \in \mathbb{R}, \quad \frac{f(0,h)-f(0,0)}{h} = 0$$

qui tend vers 0. Donc $\partial_y f(0,0) = 0$.

3. Ces dérivées partielles sont-elles continues ? **Cette question est exactement celle qu'on s'est posée dans le paragraphe continuité sauf que cette fois-ci on l'applique aux dérivées partielles.** Prenons la dérivée partielle en x qui est donnée par

$$\partial_x f : (x,y) \mapsto \begin{cases} \frac{y^2(y^2-x^2)}{(x^2+y^2)^2} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0) \end{cases}$$

Le fait d'avoir du degré 4 au numérateur et au dénominateur nous fait sentir que la fonction risque de ne pas être continue.

$$\forall y \neq 0, \quad \partial_x f(0,y) = \frac{y^2(y^2-0^2)}{(0^2+y^2)^2} = 1$$

donc la limite dans cette direction quand y tend vers 0 est 1. On en déduit que la dérivée partielle en x n'est pas continue en $(0,0)$. Donc f n'est pas de classe $C^1(\mathbb{R}^2)$.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- En quoi la définition de classe C^1 pour les fonctions à plusieurs variables est-elle analogue à celle pour les fonctions d'une variable réelle ?
- Quelle est la méthodologie pour montrer qu'une fonction est de classe C^1 ?
- Qcm associé.

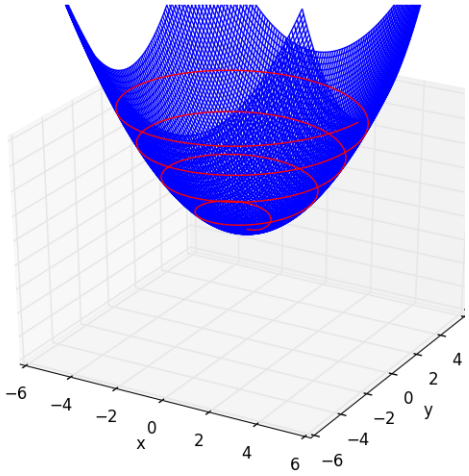
4.2.5 Règle de la chaîne

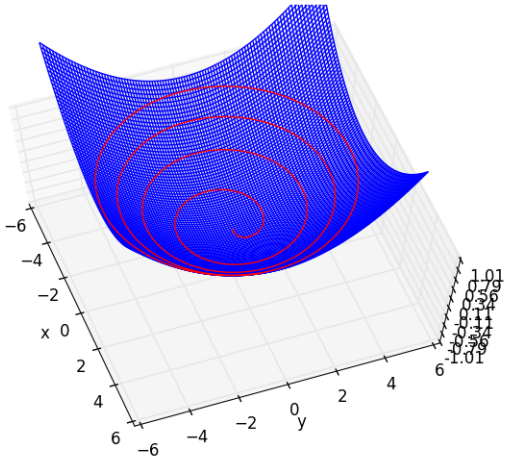
Un problème géométrique :

Voici en bleu la courbe, de $f : (x,y) \mapsto x^2+y^2$. En rouge, nous traçons la spirale $\mathcal{S}$ en rouge constituée des points donnés par les coordonnées suivantes

$$(x(t), y(t), z(t)) = (t\cos(t), t\sin(t), f(x(t), y(t))), \quad t \in \mathbb{R}_+.$$

Déterminer le vecteur tangent à $\mathcal{S}$ à l'instant $t_0 \in \mathbb{R}_+$.





Avant toute chose, pourquoi ces coordonnées donnent bien une spirale ?

Si $x(t) = t\cos(t)$ et $y(t) = t\sin(t)$ alors $||(x(t), y(t))||_2 = \sqrt{x^2(t) + y^2(t)} = t$. Ainsi la distance de $(x(t), y(t))$ à l'origine vaut t, ce qui signifie qu'elle croit quand t croit. Par ailleurs, la courbe tourne autour de l'origine à cause des termes en cos et sin. On peut le comprendre en se rappelant que le paramétrage $(\cos(t), \sin(t))$ correspond au cercle trigonométrique et donc à une rotation d'angle t autour de l'origine.

Comment déterminer le vecteur tangent en t_0 ?

Comme la tangente pour une fonction d'une variable est représentée par la dérivée, le vecteur tangent à la courbe en t_0 est donné par $(x(t), y(t), z(t)) = (t\cos(t), t\sin(t), f(x(t), y(t))), \quad t \in \mathbb{R}_+$ est donnée par le vecteur des dérivées : $(x'(t_0), y'(t_0), z'(t_0))$.

Comme $x(t) = t\cos(t)$ alors $x'(t_0) = \cos(t_0) - t_0\sin(t_0)$. Comme $y(t) = t\sin(t)$ alors $y'(t_0) = \sin(t_0) + t_0\cos(t_0)$. Enfin $z(t) = (t\cos(t))^2 + (t\sin(t))^2 = t^2$ alors $z'(t_0) = 2t_0$.

Et si on ne connaît pas l'expression de f ?

Il s'agit alors de dériver $f(x(t), y(t))$ sans connaître f. Vous savez dériver $f(x(t))$ comme fonction composée mais avec deux variables, c'est quelque chose de nouveau. Alors essayons de deviner la formule en reprenant l'exemple précédent. Supposons qu'on ne sache pas que $\cos(t)^2 + \sin(t)^2 = 1$ et dérivons à nouveau $z(t) = (t\cos(t))^2 + (t\sin(t))^2 = x(t)^2 + y(t)^2$. On a

$$z'(t) = 2x'(t)x(t) + 2y'(t)y(t) = x'(t)(2x(t)) + y'(t)(2y(t)).$$

Observons méticuleusement cette expression : le terme $2x(t)$ n'est autre que la dérivée partielle en x de f, de même $2y(t)$ est la dérivée partielle en y. Ainsi

$$z'(t) = x'(t)\partial_x f(x(t), y(t)) + y'(t)\partial_y f(x(t), y(t)).$$

Cette formule pour deux variables se généralise à n variables : il s'agit de la règle de la chaîne.

Proposition 12: Règle de la chaîne

Soient $f \in C^1(A, \mathbb{R})$ où A est un sous-ensemble de $\mathbb{R}^n$, $\forall i, x_i \in C^1(I, \mathbb{R})$ où I est un intervalle de $\mathbb{R}$

telle que $\forall t \in I, (x_1(t), \dots, x_n(t)) \in A$. Alors $g : t \in I \mapsto f(x_1(t), \dots, x_n(t)) \in C^1(I, \mathbb{R})$ et

$$\forall t \in I, g'(t) = \sum_i \partial_{x_i} f(x_1(t), \dots, x_n(t)) x_i'(t).$$

Exemple 20:

Calculons la dérivée par rapport à x de $g : x \mapsto f(e^x, x^3, x^2)$ pour $f : \mathbb{R}^3 \to \mathbb{R}$ de classe $C^1(\mathbb{R}^3)$?

La fonction g est dérivable sur $\mathbb{R}$ comme composée de $f : \mathbb{R}^3 \to \mathbb{R}$ de classe $C^1(\mathbb{R}^3)$ et de $x \in \mathbb{R} \mapsto (e^x, x^3, x^2) \in \mathbb{R}^3$ de classe $C^1(\mathbb{R})$ et

$$\forall x \in \mathbb{R}, \quad g'(x) = e^x \partial_x f(e^x, x^3, x^2) + 3x^2 \partial_y f(e^x, x^3, x^2) + 2x \partial_z f(e^x, x^3, x^2).$$

Voici une autre application : la résolution d'équations aux dérivées partielles(EDP) par changement de variable. De rares EDP sont explicitement résolvables. Considérons par exemple

$$y\partial_x f(x, y) - x\partial_y f(x, y) = 0.$$

Effectuons le changement de variable polaire :

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases}$$

et considérons la nouvelle fonction

$$\begin{array}{rccc} g & : & \mathbb{R}_+ \times [0, 2\pi[& \to & \mathbb{R}^2 \\ & & (r, \theta) & \mapsto & f(r\cos(\theta), r\sin(\theta)) \end{array}$$

Alors g est de classe $C^1(\mathbb{R}_+ \times [0, 2\pi[)$ comme composée de deux fonctions de classe C^1 et d'après la règle de la chaîne

$$\partial_\theta g(r, \theta) = \partial_\theta(f(r\cos(\theta), r\sin(\theta))) = -r\sin(\theta)\partial_x f(x, y) + r\cos(\theta)\partial_y f(x, y)$$

Donc

$$\partial_\theta g(r, \theta) = -y\partial_x f(x, y) + x\partial_y f(x, y) = 0.$$

Donc $\forall r, \theta, \quad \partial_\theta g(r, \theta) = 0$, et donc $g(r, \theta) = \psi(r)$ où ψ est une fonction C^1 d'une variable réelle. Finalement

$$\forall (x, y) \in \mathbb{R}^2, \quad f(x, y) = g(r, \theta) = \psi(r) = \psi(\sqrt{x^2 + y^2}).$$

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Refaire le calcul de l'exemple 18.
- A quel concept que vous connaissiez, la règle de la chaîne appliquée à une fonction d'une seule variable réelle correspond-elle ?
- Qcm associé.

4.2.6 ⋆ Retour à la membrane 1d ⋆

Nous avons vu dans le chapitre ev euclidien une stratégie pour résoudre le problème de la membrane 1D. Cette résolution passait par la résolution d'un système linéaire de type $Au = b$ grâce à la notion de distance. Nous proposons ici une autre manière de résoudre ce système à l'aide d'extremas (voir section suivante). Considérons l'application suivante pour le produit scalaire canonique de $\mathbb{R}^2$:

$$\begin{array}{rccc} J & : & \mathbb{R}^2 & \to & \mathbb{R} \\ & & u = (x, y) & \mapsto & \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle. \end{array}$$

On peut explicitement calculer

$$J(u) = \frac{1}{2}(a_{11}x^2 + a_{12}xy + a_{21}xy + a_{22}y^2) - b_1x - b_2y = \frac{1}{2}(a_{11}x^2 + 2a_{12}xy + a_{22}y^2) - b_1x - b_2y,$$

la dernière égalité étant due au fait que A est une matrice symétrique.

Dérivons J par rapport à x et y : $\partial_x J(x,y) = a_{11}x + a_{12}y - b_1$ et $\partial_y J(x,y) = a_{12}x + a_{22}y - b_2 = a_{21}x + a_{22}y - b_2$.

Ainsi

$$Au = b \iff \begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases} \iff \begin{cases} \partial_x J(x,y) = 0 \\ \partial_y J(x,y) = 0 \end{cases}$$

Trouver une solution au système linéaire revient à annuler les dérivées partielles de J. Nous allons voir que cela a à voir avec la recherche des extremums de la fonction J.

4.3 Extremas d'une fonction 1

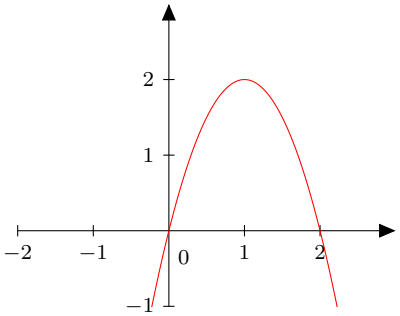
L'objectif de cette partie est d'introduire les résultats permettant de résoudre le problème de minimisation associé à la membrane élastique. Nous introduisons ici les notions d'extremas (maximum/minimum d'une fonction) et donnons la première clé pour les déterminer. La seconde clé sera obtenue ultérieurement car elle nécessite la notion d'endomorphisme symétrique.

Pour construire la première clé, revenons encore une fois aux fonctions de la variable réelle. Simplifier un problème mathématique est toujours formateur pour sa compréhension.

4.3.1 Fonction de variable réelle

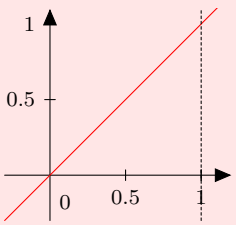
Qu'est-ce qu'un maximum/minimum ?

Etant donnée une fonction, on a naturellement envie de dire que le maximum est la plus grande valeur de cette fonction : sur le dessin il s'agit de 2. Le minimum serait alors la plus petite (ici -1). On appellerait alors extremum un minimum ou un maximum : extremum comme valeur extrême d'une fonction.



Remarque 10:

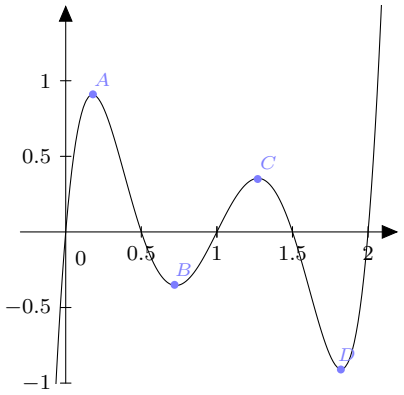
Il faut s'ôter l'idée que toute fonction admet un maximum. Par exemple la fonction $x \mapsto x$ n'admet pas de maximum sur $]0,1[$. La raison est que 1 n'est pas atteint par cette fonction sur $]0,1[$: 1 est ce qu'on appelle la borne supérieure de la fonction. En revanche cette même fonction sur $[0,1]$ a pour maximum 1 !



Il existe un résultat plus général assurant qu'une fonction continue sur un **segment** admet un maximum et un minimum. Notre dernier exemple entre d'ailleurs dans ce cas de figure : $[0,1]$ est un segment au contraire de $]0,1[$.

Qu'est-ce qu'un maximum/minimum local ?

Regardons le dessin : les points A et C sont des maximums locaux tandis que les points B et D sont des minimums locaux. On va dire que f admet un maximum local en un point a si on peut construire un intervalle autour de a tel que $f(a)$ soit le maximum sur cet intervalle. Certains de ces intervalles sont représentés sur le dessin. Cette phrase de français se traduit mathématiquement par la définition suivante :



Définition 16:

Soient I un intervalle de $\mathbb{R}$, $f : I \to \mathbb{R}$, $a \in \mathbb{R}$ on dit que f admet :

- un minimum local en a si et seulement si

$$\exists \delta > 0, \forall x \in I, |x - a| \leq \delta \Longrightarrow f(x) \geqslant f(a).$$

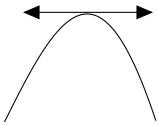
- un maximum local en a si et seulement si

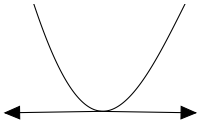
$$\exists \delta > 0, \forall x \in I, |x - a| \leq \delta \Longrightarrow f(x) \leqslant f(a).$$

- un extremum local en a si et seulement si f admet un maximum ou un minimum local en a.

Comment déterminer ces extremas locaux ?

Nous aimerions une technique pratique pour déterminer les extremas locaux d'une fonction. Pour intuiter le résultat, regardons la forme des extremas locaux. Ils sont de deux sortes :





Dans les deux cas, ils sont le lieu d'un changement de variation : la fonction passe de croissante à décroissante ou l'inverse. Un tel changement se matérialise par l'annulation de la dérivée en l'extremum local (tangente horizontale).

Proposition 13:

Soient I un intervalle ouvert de $\mathbb{R}$, $a \in I$, f de classe $C^1(I)$, si f a un extremum local en a alors $f'(a) = 0$. Un point où la dérivée de f est nulle est appelé un **point critique** de f.

Preuve :

Supposons que f ait un maximum local en a. Alors dans un voisinage de a, $f(x)$ est inférieur à $f(a)$:

$$\exists \delta > 0, \forall x \in I, |x-a| \leq \delta \Longrightarrow f(x) \leqslant f(a).$$

Rappelons que $f'(a) = \lim\limits_{x \to a} \dfrac{f(x)-f(a)}{x-a}$. Or pour x dans $]a-\delta, a]$, $x-a \leq 0$ et $f(x)-f(a) \leq 0$ donc $\dfrac{f(x)-f(a)}{x-a} \geqslant 0$. En faisant tendre x vers a. On a alors que $f'(a) \geqslant 0$.

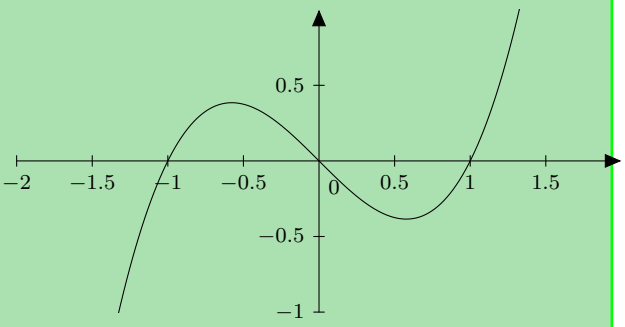
A l'inverse, pour x dans $[a, a+\delta[$, $x-a \geq 0$ et $f(x)-f(a) \leq 0$ donc $\dfrac{f(x)-f(a)}{x-a} \leqslant 0$. En faisant tendre x vers a. On a alors que $f'(a) \leqslant 0$.

Finalement $f'(a) = 0$.

La démo pour un minimum local est symétrique : entraînez-vous à la faire ?

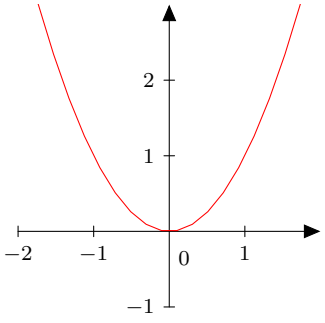
Exemple 21:

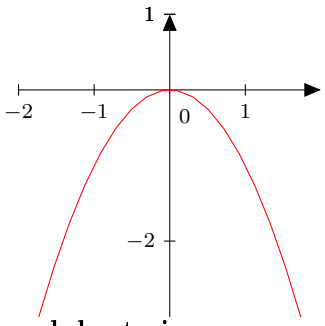
Imaginons qu'on souhaite déterminer les extremas locaux de $f : x \mapsto x^3 - x$. On cherche alors les points critiques. On a $f'(x) = 3x^2 - 1$ donc $x = \pm\dfrac{1}{\sqrt{3}}$. Donc s'il existe des extremas locaux, ils seront situés en $x = \pm\dfrac{1}{\sqrt{3}}$. En revanche, la proposition ne nous garantit pas que ce sont des extremas locaux. En effet elle dit : Si on est un extremum local alors notre dérivée est nulle.

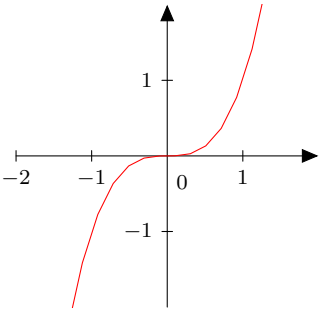


La réciproque est-elle vraie ?

Ceci serait très pratique : il suffirait de calculer une dérivée, trouver ses points d'annulation et on aurait fini. Seulement il existe pour une fonction de la variable réelle trois manières d'avoir une tangente horizontale en un point : être un minimum local ($f(x) = x^2$ en $x = 0$), être un maximum local ($f(x) = -x^2$ en $x = 0$) et être un point d'inflexion ($f(x) = x^3$ en $x = 0$).







Alors comment déterminer dans lequel des trois cas on se trouve ?

Nous souhaitons déterminer les extremas locaux d'une fonction. Le caractère local fait qu'un DL (qui est local) est une source d'information intéressante pour conclure. Imaginons que la fonction soit deux fois dérivable : alors son DL à l'ordre 2 en a existe et s'écrit

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2} + o((x-a)^2).$$

Mais si a est un point critique alors $f'(a) = 0$ et donc

$$f(x) - f(a) = \frac{f''(a)(x-a)^2}{2} + o((x-a)^2).$$

Ainsi, quand x est proche de a, $f(x) - f(a)$ est du signe de $\frac{f''(a)(x-a)^2}{2}$ (le reste étant négligeable au voisinage de a) et donc du signe de $f''(a)$.

Ainsi si $f''(a) > 0$, $f(x) \geq f(a)$ donc $f(a)$ est un minimum local. Si $f''(a) < 0$, $f(x) \leq f(a)$ donc $f(a)$ est un maximum local.

Et si $f''(a) = 0$? Il faut pour aller plus loin dans le DL et étudier le terme d'ordre 3. Quand x est proche de a, $f(x) - f(a)$ est alors du signe de $\frac{f'''(a)(x-a)^3}{6}$: si $f'''(a) \neq 0$, cela signifie que $f(x) - f(a)$ change de signe suivant que $x > a$ ou $x < a$. On a alors un point d'inflexion (exemple x^3). S'il est nul il faut aller à l'ordre 4.

Exemple 22:

Achevons de déterminer les extremas locaux de $f : x \mapsto x^3 - x$. Les points critiques sont $x = \pm\frac{1}{\sqrt{3}}$. On a $f''(x) = 6x$. Donc $f''(\frac{1}{\sqrt{3}}) > 0$. Donc $f(\frac{1}{\sqrt{3}})$ est un minimum local. De même $f(-\frac{1}{\sqrt{3}})$ est un maximum local (ce qu'on aurait pu deviner car f est impaire).

Est-ce que le maximum global d'une fonction est forcément un point critique ?

Reprenons l'exemple de $x \mapsto x$ sur $[0, 1]$, le maximum est 1 mais la tangente n'est pas horizontale. Donc la réponse est non. En revanche si le maximum n'est pas situé au bord du domaine alors c'est un maximum local (on l'a vu avant) et donc c'est un point critique.

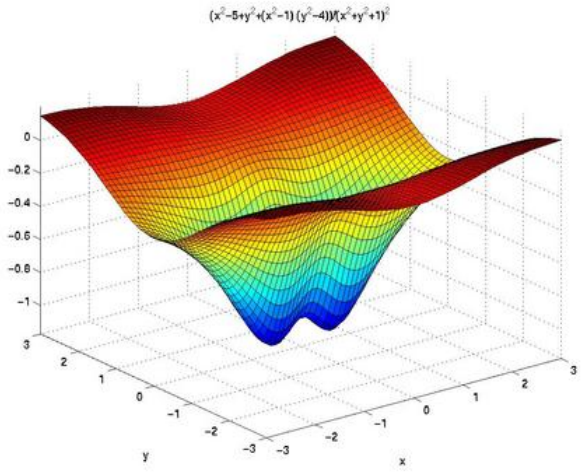
Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Pouvez-vous expliquer le lien entre la définition "en français" de minimum local et celle avec les quantificateurs ?
- Expliquer avec un dessin pourquoi un point où il y a un extremum local est nécessairement un point où la dérivée est nulle. Pourquoi la réciproque est-elle fausse ?
- Dans la démonstration mathématique de ce résultat, où est utilisée l'hypothèse "f est un maximum local" ?
- Pour quelle raison, la dérivée seconde est-elle précieuse pour savoir si un point critique est un extremum local ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Qcm associé.

4.3.2 Fonctions à plusieurs variables

L'objectif de la section est de déterminer des conditions pour déterminer les extremas locaux d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$. Il s'agit de généraliser la théorie déjà connue pour les fonctions de la variable réelle.

Le dessin ci-contre illustre bien la notion d'extremum local pour une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$: f on voit deux minimas locaux (qui sont en réalité globaux). Pour définir un minimum local, l'idée est de dire que sur un voisinage du point où le minimum est atteint, toutes les valeurs de f sont supérieurs. Attention il ne s'agit pas de dire que c'est un minimum pour tout point, la notion de voisinage est cruciale. Par voisinage, nous attendons pour les points qui sont à une distance suffisamment proche du point en question. Pour formaliser cette notion de distance, c'est la norme qui entre en jeu de la même manière que la valeur absolue était utilisée dans $\mathbb{R}$:



Définition 17:

Soient A un sous-ensemble de $\mathbb{R}^n$ et $||\cdot||$ une norme quelconque sur $\mathbb{R}^n$, $f : A \to \mathbb{R}$, $a \in \mathbb{R}^n$ on dit que f admet :

- un minimum local en a si et seulement si

$$\exists \delta > 0, \forall u \in A, ||u - a|| \leq \delta \Longrightarrow f(u) \geqslant f(a).$$

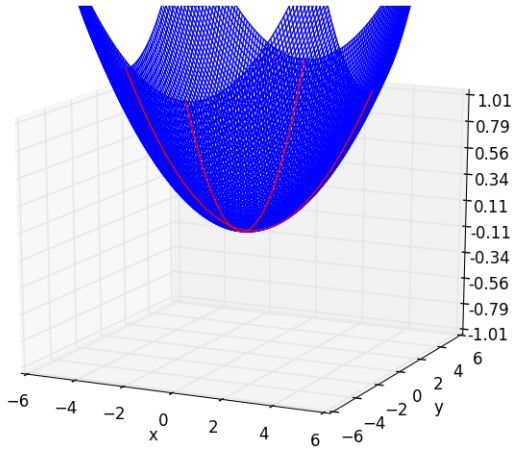
- un maximum local en a si et seulement si

$$\exists \delta > 0, \forall u \in A, ||u - a|| \leq \delta \Longrightarrow f(u) \leqslant f(a).$$

- un extremum local en a si et seulement si f admet un maximum ou un minimum local en a.

Comment déterminer en pratique les extremas locaux ?

Dans $\mathbb{R}$, on pouvait localiser les seuls endroits possibles en regardant les points d'annulation de la dérivée : les lieux où la tangente est horizontale. Considérons la fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ ci-contre, et $a = (0, 0)$ son extremum local. Fixons $y = 0$, et traçons $f_1 : x \mapsto f(x, 0)$ en rouge, de même fixons $x = 0$ et traçons $f_2 : y \mapsto f(0, y)$ en rouge. On voit que ces deux fonctions ont leur minimum local respectif en $x = 0$ et $y = 0$. Ces deux fonctions sont des fonctions de la variable réelle donc on peut dire utiliser le critère d'extremum pour ce type de fonctions. D'un côté $f_1'(0) = 0$. Or $f_1'(0) = \partial_x f(0, 0)$. Donc $\partial_x f(0, 0) = 0$. Un raisonnement analogue avec f_2 mène à $\partial_y f(0, 0) = 0$. Ainsi on a démontré que si f admet un extremum local en a, alors ses dérivées partielles s'y annulent. Ce résultat est vrai en toute généralité pour une fonction définie sur $\mathbb{R}^n$.



Théorème 4: Condition nécessaire

Soient A un sous-ensemble ouvert de $\mathbb{R}^n$, $a = (a_1, \ldots, a_n) \in A$, f de classe $C^1(A)$, si f a un extremum local en a alors

$$\forall i \in \{1, \ldots, n\}, \quad \partial_{x_i} f(a) = 0.$$

Un point où les dérivées partielles de f s'annulent est appelé un **point critique** de f.

Exemple 23: Déterminer les points critiques d'une fonction.

Soit $f : (x, y, z) \in \mathbb{R}^3 \mapsto x^2 + y^2 - z^2 - y$. Quels sont ses points critiques ?

La fonction est de classe $C^1(\mathbb{R}^3)$ car il s'agit d'une fonction polynomiale. Calculons ses dérivées partielles : $\partial_x f(x, y, z) = 2x$, $\partial_y f(x, y, z) = 2y - 1$ et $\partial_z f(x, y, z) = 2z$. Donc les points critiques sont les solutions du système linéaire :

$$\left\{\begin{array}{l}\partial_x f(x,y,z)=0\\ \partial_y f(x,y,z)=0\\ \partial_z f(x,y,z)=0\end{array}\right. \iff \left\{\begin{array}{rl}2x&=0\\ 2y-1&=0\\ 2z&=0\end{array}\right. \iff \left\{\begin{array}{l}x=0\\ y=\dfrac{1}{2}\\ z=0\end{array}\right.$$

La fonction a pour unique point critique $(0, 1/2, 0)$.

Remarque 11:

L'information donnée par ce théorème est très importante : si un point est un extremum local alors c'est un point où toutes les dérivées partielles sont nulles (appelé **point critique**). Ainsi quand on cherchera un extremum local, on commencera par déterminer les points critiques.

En revanche, comme pour les fonctions de $\mathbb{R}$ dans $\mathbb{R}$, ce n'est qu'une condition nécessaire : la réciproque est fausse.

Comment déterminer si un point critique est bien un extremum local ?

Inspirons-nous de ce qui a été fait pour f de $\mathbb{R}$ dans $\mathbb{R}$: regardons le développement limité à l'ordre 1 d'une fonction f de $\mathbb{R}^n$ dans $\mathbb{R}$ au voisinage de a point critique de f. Seulement, on n'a jamais déterminé le DL d'une fonction à plusieurs variables. La section suivante introduit cela.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- En quoi la définition et le théorème de cette partie sont-ils analogues à ceux dans le cas d'une fonction d'une variable réelle ?
- Pour une fonction d'une variable réelle, avoir en extremum local en a c'est avoir une tangente horizontale en a. Que signifie géométriquement d'avoir en extremum local en a pour une fonction à plusieurs variables ?
- Qcm associé.

4.3.3 DL et matrice jacobienne

Le DL de f vu précédemment

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)(x - a)^2}{2} + o((x - a)^2),$$

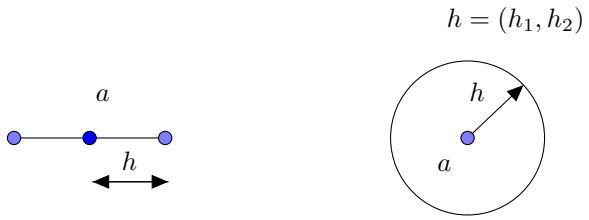
peut se récrire en posant $h = x - a$:

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

Ainsi quand x tend vers a, le paramètre h tend vers 0. Nous adopterons ce point de vue dans la suite.

Il faut alors se demander si on peut écrire des DL analogues pour des fonctions à plusieurs variables. Commençons sur un exemple par se placer sur $\mathbb{R}^2$: cette fois le paramètre h n'est plus un réel mais un

vecteur de taille 2. Sur le dessin à gauche, vous avez un dessin d'un voisinage de a dans $\mathbb{R}$ et sur celui de droite dans $\mathbb{R}^2$: dans $\mathbb{R}^2$, le vecteur h peut prendre une infinité de directions.



Afin de deviner la forme du DL, commençons par un exemple concret : soit $f : (x, y) \in \mathbb{R}^2 \mapsto x^2 - y^2 \in \mathbb{R}$. On se place au voisinage de $a = (a_1, a_2)$. Calculons $f(a + h)$.

$$f(a+h) = (a_1 + h_1)^2 - (a_2 + h_2)^2 = a_1^2 - a_2^2 + 2a_1h_1 - 2a_2h_2 + h_1^2 - h_2^2.$$

Donc $f(a + h) = f(a) + 2a_1h_1 - 2a_2h_2 + o(||h||_2)$.

Pourquoi $h_1^2 - h_2^2 = o(||h||_2)$?

Il s'agit de vérifier que $\epsilon(h) = \dfrac{h_1^2 - h_2^2}{||h||_2} = \dfrac{h_1^2 - h_2^2}{\sqrt{h_1^2 + h_2^2}}$ est de limite nulle quand h tend vers $(0, 0)$.

Or

$$0 \leq |\epsilon(h)| = \left| \frac{h_1^2 - h_2^2}{\sqrt{h_1^2 + h_2^2}} \right| \leq \frac{|h_1^2| + |h_2^2|}{\sqrt{h_1^2 + h_2^2}},$$

par inégalité triangulaire. On en déduit que

$$0 \leq |\epsilon(h)| \leq \frac{2||h||_\infty^2}{||h||_2} \leq \frac{2||h||_2^2}{||h||_2} \leq 2||h||_2$$

la dernière inégalité venant du fait que $||h||_\infty \leq ||h||_2$ Donc par encadrement, $\epsilon(h)$ tend bien vers 0 quand h tend vers $(0, 0)$.

Pourquoi est-ce utile ?

L'expression $f(a + h) = f(a) + 2a_1h_1 - 2a_2h_2 + o(||h||_2)$ resssemble furieusement à un DL à l'ordre 1. On a le terme d'ordre 0 donné par $f(a)$, on a le reste d'ordre 1. Ne reste plus qu'à comprendre à quoi correspond $2a_1h_1 - 2a_2h_2$.

Pour les fonctions d'une variable réelle le terme d'ordre 1 fait intervenir la dérivée appliquée en a. Ici la notion de dérivée n'existe pas mais il serait cohérent que les dérivées partielles interviennent. Calculons-les donc.

On a $\partial_x f(a_1, a_2) = 2a_1$ et $\partial_y f(a_1, a_2) = -2a_2$. Ainsi,

$$f(a + h) = f(a) + \partial_x f(a)h_1 + \partial_y f(a)h_2 + o(||h||_2).$$

Ainsi l'analogue de $hf'(a)$ dans le DL de fonction à une variable est $\partial_x f(a)h_1 + \partial_y f(a)h_2$. De ce DL pour deux variables, on peut intuiter celui pour une fonction à nombre quelconque de variables :

Théorème 5: DL et différentielle

Soit A un sous-ensemble de $\mathbb{R}^n$. Soit $f \in C^1(A)$, alors $\forall a \in A, \forall h$ tq $a+h \in A$,

$$f(a+h) = f(a) + \sum_{k=1}^{n} \partial_{x_k} f(a) h_k + o(||h||_2).$$

L'application $df(a) : h \in \mathbb{R}^n \mapsto \sum_{k=1}^{n} \partial_{x_k} f(a) h_k \in \mathbb{R}$ est une application linéaire appelée différentielle de f en a.

Remarque 12:

Pour les fonctions de $\mathbb{R}$ dans $\mathbb{R}$, la différentielle est l'application $df(a) : h \in \mathbb{R} \mapsto f'(a)h \in \mathbb{R}$.

Regardons de plus près l'application différentielle $df(a)$. Si on définit les vecteurs colonne et ligne suivant

$$\begin{pmatrix} \partial_{x_1} f(a) & \dots & \partial_{x_n} f(a) \end{pmatrix}, \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix},$$

on peut noter que

$$df(a)(h) = \sum_{k=1}^{n} \partial_{x_k} f(a) h_k = \begin{pmatrix} \partial_{x_1} f(a) & \dots & \partial_{x_n} f(a) \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$$

Définition 18: Jacobienne

Soient A un sous-ensemble de $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $C^1(A)$, $a \in A$, on appelle matrice jacobienne de f en a et on note $J_f(a)$ la matrice donnée par

$$\begin{pmatrix} \partial_{x_1} f(a) & \dots & \partial_{x_n} f(a) \end{pmatrix}$$

Quelle est la matrice jacobienne pour :

- $f : x \in \mathbb{R} \mapsto x^3$ en 1 ? C'est une matrice à 1 ligne 1 colonne qui contient $f'(1) = 3$
- $f : (x, y) \in \mathbb{R}^2 \mapsto x^2 - y^2$ en $(1, 2)$? C'est la matrice ligne composée des dérivées partielles appliquées en $(1, 2)$. Cette matrice est $J_f(1, 2) = \begin{pmatrix} 2 & -4 \end{pmatrix}$.

Voici une forme plus synthétique du DL d'ordre 1.

$$f(a+h) = f(a) + J_f(a)h + o(||h||_2).$$

Il est important de noter la cohérence des objets : $f(a)$ est un réel donc $J_f(a)h$ doit en être un aussi. C'est le cas car $J_f(a)$ est un vecteur ligne et h est un vecteur colonne de même taille.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quel est l'analogue pour les fonctions à plusieurs variables de $f'(a)$?
- Qcm associé.

4.3.4 Retour aux extremas

Exploitons le DL à des fins de détermination d'extremas locaux. Pour h un vecteur tel que $a+h$ soit dans le domaine de définition de f.

$$f(a+h) = f(a) + \partial_x f(a) h_1 + \partial_y f(a) h_2 + o(||h||_2)$$

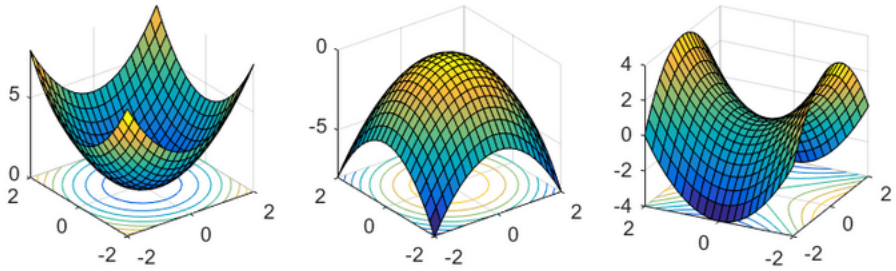
S'agissant d'un point critique, les dérivées partielles sont nulles si bien que

$$f(a+h) = f(a) + o(||h||_2)$$

Finalement

$$f(a+h) - f(a) = o(||h||_2)$$

Si pour tout h, $f(a+h) \geqslant f(a)$ (resp. $f(a+h) \leqslant f(a)$), alors $f(a)$ est un minimum local (resp. un maximum local). Si pour un certain h_1, on a $f(a+h_1) \geqslant f(a)$ et pour un certain h_2, on a $f(a+h_2) \leqslant f(a)$ alors f n'admet pas d'extremum local en a. Ces trois cas de figure sont représentés par les dessins ci-dessous.



Ainsi comme pour les fonctions d'une variable réelle, l'information cruciale du DL est le signe du reste $o(||h||_2)$ suivant le choix de h. Cependant $o(||h||_2)$ peut être de tout signe. Il est donc indispensable de pousser le DL à l'ordre suivant (l'ordre 2), ce qui justifie l'introduction de la notion de dérivée partielle seconde.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Pourquoi étudier le DL en a permet de savoir s'il y a un extremum local en a ?
- Qcm associé.

4.4 Classe C^2

4.4.1 Définition

Comment définit-on la dérivée seconde d'une fonction de $\mathbb{R}$ dans $\mathbb{R}$?

C'est simple : si la dérivée est elle même dérivable, la dérivée seconde f'' est définir comme la dérivée de la dérivée : $f'' = (f')'$.

Comment définir les dérivées partielles du second ordre ?

Le principe est identique. On sait définir les dérivées partielles d'ordre 1 à l'aide du taux d'accroissement. Les dérivées partielles sont elles aussi des fonctions de plusieurs variables donc on peut regarder si leur taux d'accroissement a une limite. Ainsi, une dérivée partielle d'ordre 2 va être défini comme "une dérivée partielle d'ordre 1 d'une dérivée partielle d'ordre 1".

Définition 19:

Soient A un sous-ensemble de $\mathbb{R}^n$, $f : A \to \mathbb{R}$ admettant des dérivées partielles sur A. On appelle dérivée partielle seconde **si elles existent** :

$$\forall (i,j) \in \{1,\dots,n\}^2, \quad \partial^2_{x_j x_i} f := \partial_{x_j}(\partial_{x_i} f).$$

Forts de cette définition, nous pouvons définir ce qu'est une fonction de classe C^2 : pour C^1, il fallait que les dérivées partielles d'ordre 1 existent et soient continues, pour C^2, il faut que les dérivées partielles d'ordre 1 et 2 existent et que les dérivées partielles d'ordre 2 soient continues.

Définition 20:

Soit A un sous-ensemble de $\mathbb{R}^2$, on dit que $f : A \to \mathbb{R}$ est de classe $C^2(A)$ si et seulement si
— $\forall i \in \{1, \dots, n\}, \partial_{x_i} f$ existent en tout point de A.
— $\forall (i,j) \in \{1, \dots, n\}^2, \partial^2_{x_j x_i} f$ existent en tout point de A et sont continues sur A.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Qcm associé.

4.4.2 DL au second ordre

Rappelons notre objectif : nous souhaitons obtenir un critère déterminant si un point est un extremum local. Pour cela, nous avons vu qu'il était indispensable d'aller à l'ordre 2 au moins dans le DL autour de ce point. Cette section donne la forme générale du DL d'une fonction de classe C^2 au second ordre.

Quelle est l'expression pour une fonction de $\mathbb{R}$ dans $\mathbb{R}$?

Voici l'expression du DL à l'ordre 2 autour de a (h petit) :

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

Comment généraliser cela à $\mathbb{R}^2$?

Nous avons déjà vu que le h réel devient un vecteur $h = (h_1, \dots, h_n)$, que le terme $f'(a)h$ devient $\partial_x f(a)h_1 + \partial_y f(a)h_2$. Le terme $f''(a)h^2$ va quant à lui devenir $\partial^2_{xx} f(a)h_1^2 + \partial^2_{xy} f(a)h_1h_2 + \partial^2_{yx} f(a)h_1h_2 + \partial^2_{yy} f(a)h_2^2$.

Souvenez-vous le terme d'ordre 1 pouvait être écrit sous forme matricielle comme le produit de la jacobienne $J_f(a)$ par h. Pour le terme d'ordre 2, vous pouvez vérifier que :

$$\begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} \partial^2_{xx} f(a) & \partial^2_{xy} f(a) \\ \partial^2_{yx} f(a) & \partial^2_{yy} f(a) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \partial^2_{xx} f(a)h_1^2 + \partial^2_{xy} f(a)h_1h_2 + \partial^2_{yx} f(a)h_1h_2 + \partial^2_{yy} f(a)h_2^2.$$

Ainsi le DL à l'ordre 2 s'écrit :

$$f(a+h) = f(a) + J_f(a) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} h_1 & h_2 \end{pmatrix} H_f(a) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + o(||h||_2^2),$$

où $H_f(a)$ est la matrice hessienne donnée par

$$H_f(a) = \begin{pmatrix} \partial^2_{xx} f(a) & \partial^2_{xy} f(a) \\ \partial^2_{yx} f(a) & \partial^2_{yy} f(a) \end{pmatrix}$$

Comment généraliser cela à $\mathbb{R}^n$?

Nous avons déjà vu que le h réel devient un vecteur $h = (h_1, \dots, h_n)$, que le terme $f'(a)h$ devient $\sum_{k=1}^{n} \partial_{x_k} f(a)h_k$. Le terme $f''(a)h^2$ va quant à lui devenir $\sum_{k=1}^{n} \sum_{l=1}^{n} \partial^2_{x_k x_l} f(a)h_k h_l$. Ces termes s'écrivent également de manière matricielle et sont regroupés dans le théorème admis suivant.

Théorème 6: Taylor Young

Soient A un sous-ensemble de $\mathbb{R}^n$, $f \in C^2(A), a = (a_1, \dots, a_n) \in A$ alors pour tout $h = (h_1, \dots, h_n)$

tel que $a+h \in A$

$$f(a+h) = f(a) + J_f(a) \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} + \frac{1}{2} \begin{pmatrix} h_1 & \dots & h_n \end{pmatrix} H_f(a) \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} + o(||h||_2^2),$$

où $H_f(a)$ est la matrice hessienne donnée par

$$H_f(a) = [\partial^2_{x_i x_j} f(a)]_{1 \leq i,j \leq n}.$$

Réfléchissons à l'objectif dans le cas de $\mathbb{R}^2$.

Pour un point critique a,

$$f(a+h) - f(a) = \frac{1}{2} \begin{pmatrix} h_1 & h_2 \end{pmatrix} H_f(a) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + o(||h||_2^2)$$

Le fait que a soit ou non un extremum local dépend donc du signe du terme d'ordre 2 :

$$\begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} \partial^2_{xx} f(a) & \partial^2_{xy} f(a) \\ \partial^2_{yx} f(a) & \partial^2_{yy} f(a) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \partial^2_{xx} f(a) h_1^2 + \partial^2_{xy} f(a) h_1 h_2 + \partial^2_{yx} f(a) h_1 h_2 + \partial^2_{yy} f(a) h_2^2.$$

Pas si évident en apparence ! Imaginons dans un monde idéal que nous arrivions à annuler les deux termes extradiagonaux de la matrice, on aurait alors :

$$\begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} \partial^2_{xx} f(a) & 0 \\ 0 & \partial^2_{yy} f(a) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \partial^2_{xx} f(a) h_1^2 + \partial^2_{yy} f(a) h_2^2.$$

Alors si $\partial^2_{xx} f(a)$ et $\partial^2_{yy} f(a)$ étaient du même signe, on aurait alors le signe du terme d'ordre 2 et donc de $f(a+h) - f(a)$.

Quel genre d'opération permet de mettre des 0 hors de la diagonale ?

La diagonalisation ! Ainsi si on arrive à diagonaliser la matrice hessienne, on pourra en tirer de précieuses informations sur les extremas locaux. Intéressons-nous donc à cette matrice. Elle comporte hors de sa diagonale $\partial^2_{xy} f(a)$, la dérivée en y puis en x et $\partial^2_{yx} f(a)$ la dérivée en y puis en x. Une question vient alors : ces termes sont-ils égaux. Il semblerait naturel que dérivée dans un ordre ou dans l'autre ne change rien. Par exemple pour $f(x,y) = x^2 y$, $\partial^2_{xy} f(a) = 2a_1 = \partial^2_{yx} f(a)$.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Quel est l'analogue pour les fonctions à plusieurs variables de $f''(a)$?
- En quoi le fait que la matrice hessienne soit diagonale nous aiderait pour savoir si un point critique est un extremum local ?
- Qcm associé.

4.4.3 Théorème de Schwarz

La matrice hessienne est donc symétrique. Remarquez que si la matrice était diagonale, le pb des extremas s'en trouverait simplifié. Comment s'y ramener : en diagonalisant. On va voir que les matrices sym ont de très bonnes prop de diago.

Est-ce la même chose de dériver d'abord par rapport à x puis par rapport à y, ou de dériver d'abord par rapport à y puis par rapport à x ? Voici un théorème admis qui répond partiellement :

Théorème 7: de Schwarz (admis)

Soient A un sous-ensemble de $\mathbb{R}^2$, $f \in C^2(A)$, alors $\forall a \in A, \partial^2_{xy} f(a) = \partial^2_{yx} f(a)$.

Et si on enlève l'hypothèse C^2 ?

On peut avoir des dérivées partielles secondes qui existent mais qui ne sont pas continues auquel cas la fonction n'est pas C^2.

▸ **Exercice 10**. Soit

$$f : (x, y) \mapsto \begin{cases} \dfrac{xy(x^2 - y^2)}{x^2 + y^2} & \text{si} \quad (x, y) \neq (0, 0) \\ 0 & \text{si} \quad (x, y) = (0, 0) \end{cases}$$

Démontrer que $\partial^2_{xy} f(0, 0) \neq \partial^2_{yx} f(0, 0)$.

Quelles conséquences a le théorème de Schwarz ?

Il nous donne une information très importante sur la matrice hessienne. Sous réserve que la fonction soit de classe C^2, la matrice hessienne est **symétrique**. Rappelons que nous souhaitons diagonaliser la matrice hessienne. Nous allons voir que la symétrie va s'avérer être un atout majeur dans cet objectif. Ceci motive l'introduction du chapitre suivant sur les endomorphismes symétriques.

5 Endomorphismes symétriques

5.1 Définitions

Nous souhaitons avoir une méthodologie pour déterminer les extremums d'une fonction afin de résoudre le problème de la membrane. Nous avons vu que la matrice hessienne joue un rôle important dans la recherche d'extremum et que la diagonaliser nous aiderait considérablement. Par ailleurs, nous avons constaté que cette matrice est symétrique. L'objectif du chapitre endomorphismes symétriques est d'expliquer pourquoi et comment on peut diagonaliser une matrice symétrique réelle.

Pourquoi le titre "endomorphismes symétriques" alors qu'on parle de matrice ?

Vous avez vu en L2 et L3 qu'à toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ est associée une application linéaire naturelle donnée par

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^n & \to & \mathbb{R}^n \\ & & u & \mapsto & Au \end{array}.$$

La matrice A est alors la matrice de φ_A dans la base canonique. Par exemple, l'application linéaire

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & & \begin{pmatrix} x \\ y \end{pmatrix} & \mapsto & \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x-y \\ x+2y \end{pmatrix} \end{array}$$

a pour matrice en base canonique $\begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}$. Cette application linéaire est appelée un **endomorphisme**, "endo" signifie qu'elle envoie un ev sur lui-même et "morphisme" signifie "application linéaire". La correspondance entre application linéaire et matrice est telle qu'on peut "identifier ces deux objets" (on peut construire un **isomorphisme** entre $\mathcal{M}_n(\mathbb{R})$ et $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$) : ainsi l'étude de l'endomorphisme donne des informations sur la matrice et réciproquement. Pour cette raison, dans l'optique d'étudier les propriétés de matrices symétriques, on va se pencher sur l'étude des endomorphismes qui leur sont associés.

Soit $\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ une matrice symétrique et soit l'endomorphisme canoniquement associé

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & & \begin{pmatrix} x \\ y \end{pmatrix} & \mapsto & \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x-y \\ -x+2y \end{pmatrix} \end{array}$$

Effectuons le produit scalaire de $\varphi(x,y) = (x-y, -x+2y)$ par un vecteur (x_2, y_2) quelconque : ce produit scalaire vaut alors

$$(x-y)x_2 + (-x+2y)y_2 = xx_2 - yx_2 - xy_2 + 2yy_2 = x(x_2 - y_2) + y(-x_2 + 2y_2).$$

Ainsi, ce calcul est "symétrique" : on peut échanger les rôles de (x,y) et (x_2,y_2) et ceci est possible grâce aux signes - (en rouge dans le calcul) qui sont là grâce à la symétrie de la matrice. Essayez de remplacer -1 par 2 dans le coefficient première ligne, deuxième colonne : vous verrez que le calcul n'est plus symétrique.

Ainsi l'endomorphisme φ vérifie :

$$\langle \varphi(x,y), (x_2,y_2) \rangle = \langle (x,y), \varphi(x_2,y_2) \rangle$$

C'est cette propriété qui caractérise ce qu'on appelle les endomorphismes symétriques :

Définition 21:

Soit E un espace euclidien, $f \in \mathcal{L}(E)$ l'ensemble des endomorphismes de E, alors f est dit symétrique si et seulement si

$$\forall (u,v) \in E^2, \quad \langle f(u), v \rangle = \langle u, f(v) \rangle.$$

Réfléchissons un peu sur la façon dont nous avons amené la définition : nous avons choisi un exemple de matrice symétrique et nous avons montré que l'endomorphisme canoniquement associé à cette matrice est un endomorphisme symétrique. Deux questions viennent alors à l'esprit :

• Est-ce que ce résultat est général ? Autrement dit, est-ce que si je prends une matrice symétrique quelconque, l'endomorphisme canonique

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^n & \to & \mathbb{R}^n \\ & & u & \mapsto & Au \end{array}$$

est symétrique ?

• La réciproque est-elle vraie ? Autrement dit, est-ce que si on prend un endomorphisme symétrique, la matrice qui lui est associé est symétrique ?

La seconde question est ambigue car à un endomorphisme on peut associer plusieurs matrices, chaque matrice étant associée à une base bien précise. Rappelons que pour une application linéaire f sur un ev E dont une base est $(e_1, \dots, e_n)$, la matrice de f dans cette base est fabriquée à l'aide des expressions des $f(e_i)$ dans la base $(e_1, \dots, e_n)$:

$$\begin{array}{c} \begin{array}{ccc} f(e_1) & \dots & f(e_n) \end{array} \\ \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pn} \end{pmatrix} \begin{array}{c} e_1 \\ \\ e_n \end{array} \end{array}$$

La question se reformule alors ainsi : est-ce que la matrice dans n'importe quelle base d'un endomorphisme symétrique est symétrique ? Pour le comprendre considérons un exemple dans $\mathbb{R}^2$:

$$\begin{array}{rccc} \varphi & : & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & & \begin{pmatrix} x \\ y \end{pmatrix} & \mapsto & \begin{pmatrix} x+y \\ x+y \end{pmatrix} \end{array}$$

Sa matrice dans la base canonique $((1,0),(0,1))$ est $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Elle est bien symétrique. Choisissons l'autre base $((1,1),(0,1))$: la matrice de φ dans cette base est $\begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix}$ et elle n'est pas symétrique !

Qu'est-ce qui différencie ces deux matrices ?

La base canonique est une base orthonormale au contraire de la seconde. Ceci semble jouer un rôle.

Et si on prend une base orthogonale mais non orthonormale ?

Essayons avec $((2,0),(0,1))$ qui est orthogonale mais non orthonormale : la matrice dans cette base est alors $\begin{pmatrix} 1 & 1/2 \\ 2 & 1 \end{pmatrix}$ et n'est pas symétrique : il semble donc que la **bon** soit le bon choix de base. C'est ce qu'atteste la proposition suivante.

Proposition 14:

Soit $\mathcal{B}$ une bon de E, $f \in \mathcal{L}(E)$, $F = Mat_{\mathcal{B}}(f)$, alors f est symétrique $\iff {}^tF = F$.
Autrement dit, un endomorphisme est symétrique si et seulement si sa matrice **dans une bon** de E est symétrique réelle.

Preuve :

Le principe de cette preuve consiste à écrire l'égalité $\langle f(u), v\rangle = \langle u, f(v)\rangle$ de manière matricielle.

Soit une bon $\mathcal{B}$ de E, on note U (resp. V) le vecteur colonne correspondant au vecteur u (resp. v) exprimé dans $\mathcal{B}$. On note $F = Mat_{\mathcal{B}}(f)$. On a alors $f(u) = FU$ et $f(v) = FV$.

Par ailleurs, d'après l'exercice 4 4 p 19, le produit scalaire de deux vecteurs, s'ils sont exprimés dans une bon, est donné par $\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i = {}^tUV$.

Ainsi si on combine les deux questions précédentes, l'égalité $\langle f(u), v\rangle = \langle u, f(v)\rangle$ devient ${}^t(FU)V = {}^tU(FV)$. Ainsi on obtient ${}^tU{}^tFV = {}^tUFV$. Comme ceci est vrai pour tout U, V puisque l'endomorphisme est symétrique, on a ${}^tF = F$. Donc la matrice F est symétrique.

Remarquez qu'on peut très bien partir de ${}^tF = F$ et refaire tout le raisonnement à l'envers. Ceci explique l'équivalence dans la proposition.

Revenons à notre objectif de départ : diagonaliser les matrices symétriques réelles. Pour cela, intéressons-nous aux endomorphismes symétriques naturellement liés à ces matrices. Nous allons voir que ces endomorphismes ont une structure matricielle très particulière. En particulier, leurs éléments propres ont des propriétés remarquables qui les rend diagonalisables dans $\mathbb{R}$. Le paragraphe suivant égrène ces propriétés remarquables, celui d'après les combine pour former le théorème fondamental de ce chapitre : le théorème spectral de diagonalisation des endomorphismes symétriques.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- En quoi la preuve de la proposition ne marche plus si l'hypothèse "bon" est supprimée ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Qcm associé.

5.2 Eléments propres d'un endomorphisme symétrique

Pour diagonaliser une matrice, il est nécessaire de s'intéresser à ses éléments propres (valeurs et vecteurs propres). L'objectif de cette partie est de comprendre ce que le caractère "symétrique" apporte en plus par rapport à une endomorphisme quelconque.

Donnons-nous deux matrices de taille 2 :

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Le polynôme caractéristique de la première est $X^2 - 1 = (X-1)(X+1)$, celui de la seconde est $X^2 + 1 = (X+i)(X-i)$. Ainsi la première a pour valeurs propres -1 et 1, la seconde i et $-i$. Ainsi, lorsque vous calculez le spectre d'une matrice réelle, rien ne garantit que ses valeurs propres soient réelles : certaines peuvent être réelles mais vous pouvez aussi avoir des paires complexes conjuguées. Pour un endomorphisme symétrique, ce dernier cas de figure est exclu :

Proposition 15: Spectre réel

Soit f un endomorphisme symétrique de E alors toutes les valeurs propres de f sont réelles.

Preuve :

Idée de la preuve : Démontrer que toute valeur propre λ vérifie $\lambda = \overline{\lambda}$.

On se donne une bon de E et on écrit la représentation matricielle de f dans cette bon : $F = Mat_{\mathcal{B}}(f)$.

On sait que F est une matrice symétrique réelle d'après la proposition de la section 5.1. Donnons-nous un vecteur propre u (associé à une valeur propre λ) s'écrivant matriciellement U dans cette bon. Remarquez tout d'abord que λ et U sont a priori complexes.

On a $FU = \lambda U$ par définition des vecteurs propres. On en déduit que

$$ {}^t\overline{U}FU = \lambda {}^t\overline{U}U = \lambda \sum_{i=1}^{n} \overline{u_i} u_i = \lambda \sum_{i=1}^{n} |u_i|^2. $$

Comme F est symétrique alors ${}^tF = F$, on en déduit que ${}^tUF\overline{U} = {}^tU{}^tF\overline{U} = {}^t(FU)\overline{U}$. Donc

$$ {}^tUF\overline{U} = {}^t(\lambda U)\overline{U} = \lambda {}^tU\overline{U} = \lambda \sum_{i=1}^{n} |u_i|^2. $$

Comme $\overline{{}^tUF\overline{U}} = {}^t\overline{U}\overline{F}U = {}^t\overline{U}FU$, on en déduit que

$$ \overline{\lambda \sum_{i=1}^{n} |u_i|^2} = \lambda \sum_{i=1}^{n} |u_i|^2. $$

Comme $\sum_{i=1}^{n} |u_i|^2 \in \mathbb{R}$ alors

$$ \overline{\lambda} \sum_{i=1}^{n} |u_i|^2 = \lambda \sum_{i=1}^{n} |u_i|^2. $$

Par ailleurs, comme U est un vecteur propre, il est non nul donc $\sum_{i=1}^{n} |u_i|^2 \neq 0$. On en déduit que $\lambda = \overline{\lambda}$ c'est-à-dire que $\lambda \in \mathbb{R}$.

Que vérifient les espaces propres d'un endomorphisme quelconque ?

Donnons-nous un endomorphisme f et deux valeurs propres λ et μ. Supposons qu'un vecteur u soit dans l'intersection des deux espaces propres associés à λ et μ alors par définition : $f(u) = \lambda u$ et $f(u) = \mu u$. Donc $(\lambda - \mu)u = 0_E$. Comme $\lambda \neq \mu$ alors u est le vecteur nul. Donc l'intersection de deux espaces propres différents est réduite à l'élément neutre ! Autrement dit les espaces propres sont en somme directe.

Que vérifient en plus les espaces propres d'un endomorphisme symétrique ?

Prenons les trois matrices symétriques

$$ A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} -2 & -2 & 1 \\ -2 & 1 & -2 \\ 1 & -2 & -2 \end{pmatrix}. $$

A a pour valeur propre 1 et 2 et une base de vecteurs propres est $((1,0),(0,1))$ (la matrice est déjà diagonale). B a pour valeurs propres 0 et 2. Si on soustrait les deux colonnes, on obtient la colonne nulle donc $(1,-1)$ est vecteur propre associé à la valeur propre 0. Par ailleurs

$$ B - 2I_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, $$

si on somme les deux colonnes, on obtient la colonne nulle donc $(1,1)$ est vecteur propre associé à la valeur propre 2. Une base de vecteurs propres de B est donc $((1,-1),(1,1))$.

Les trois valeurs propres de C sont $-3, -3, 3$.

$$C-3I_3=\begin{pmatrix}-5&-2&1\\-2&-2&-2\\1&-2&-5\end{pmatrix}$$

On voit que si on ajoute la première colonne à la troisième et qu'on lui enlève deux fois la seconde, on obtient le vecteur nul : on en déduit que $e_1+e_3-2e_2=(1,-2,1)\in\ker(C-3I_3)$. Donc $(1,-2,1)$ est vecteur propre de C pour la valeur propre 3.

$$C+3I_3=\begin{pmatrix}1&-2&1\\-2&4&-2\\1&-2&1\end{pmatrix}$$

On peut trouver une base de vecteurs propres : la première colonne moins la troisième donne le vecteur nul, deux fois la première plus la seconde vaut le vecteur nul. Donc $e_1-e_3=(1,0,-1)$ et $2e_1+e_2=(2,1,0)$ sont deux vecteurs propres de C pour la valeur propre -3.

Observons ces trois bases : à chaque fois, deux vecteurs propres associés à deux valeurs propres différentes sont orthogonaux. Pour C, $(1,-2,1)$ est bien orthogonal à $(1,0,-1)$ et $(2,1,0)$. Est-ce un hasard ? La réponse est non comme le démontre le raisonnement suivant :

Preuve :

de la proposition qui suit : Prenons un élément u de E_λ l'espace propre associé à λ, alors $f(u)=\lambda u$. Prenons un élément de E_μ l'espace propre associé à μ, alors $f(v)=\mu v$. Exploitons le fait que f soit symétrique, c'est-à-dire que

$$\langle f(u),v\rangle=\langle u,f(v)\rangle$$

D'un côté, on a $\langle f(u),v\rangle=\langle\lambda u,v\rangle=\lambda\langle u,v\rangle$. Par ailleurs, $\langle u,f(v)\rangle=\langle u,\mu v\rangle=\mu\langle u,v\rangle$. Ainsi $\mu\langle u,v\rangle=\lambda\langle u,v\rangle$, donc pour deux valeurs propres différentes λ et μ, on a $\langle u,v\rangle=0$. Donc deux vecteurs propres d'espaces propres différents sont nécessairement orthogonaux !

Proposition 16: Orthogonalité des espaces propres

Soit f un endomorphisme symétrique de E, λ,μ deux valeurs propres différentes de f, alors les espaces propres $E_\lambda=\ker(f-\lambda id_E)$ et $E_\mu=\ker(f-\mu id_E)$ sont orthogonaux. Autrement dit,

$$\forall u\in E_\lambda,\forall v\in E_\mu,\quad \langle u,v\rangle=0.$$

Remarque 13: très pratique

• Lorsque vous calculerez les vecteurs propres d'un endomorphismes symétrique, deux vecteurs propres associés à deux valeurs propres différentes devront nécessairement être orthogonaux. Dans le cas contraire, vous vous êtes trompés !

• Attention, au sein d'un même espace propre, deux vecteurs propres pris au hasard n'ont en revanche aucune raison d'être orthogonaux.

Passons maintenant à une troisième propriété importante des endomorphismes symétriques : on la décline en "Pour aller plus loin". J'encourage vivement à aller voir les derniers "Pour aller plus loin" du chapitre : ils permettent de comprendre en quoi les trois propositions de cette partie s'articulent pour démontrer le théorème phare du chapitre, le théorème spectral.

▸ **Pour aller plus loin**.

Proposition 17: stabilité de l'orthogonal

Soit f un endomorphisme symétrique de E, λ une valeur propre de F alors $E_\lambda^\perp$ est stable par f.

Preuve :

La stabilité d'un sev F par un endomorphisme f de E signifie que $f(F) \subset F$. La preuve consiste donc ici à démontrer que $f(E_\lambda^\perp) \subset E_\lambda^\perp$. Soit donc $v \in f(E_\lambda^\perp)$: le but est de prouver qu'il appartient à $E_\lambda^\perp$.

Par définition $\exists u \in E_\lambda^\perp, v = f(u)$.
Soit $w \in E_\lambda$, si on montre que $\langle v, w\rangle = 0$ on a gagné car on aura alors prouvé que v est bien dans $E_\lambda^\perp$. Or $\langle v, w\rangle = \langle f(u), w\rangle = \langle u, f(w)\rangle$ la dernière égalité étant due au fait que f est symétrique. Comme $w \in E_\lambda$, on a alors que $\langle v, w\rangle = \langle u, \lambda w\rangle = \lambda\langle u, w\rangle = 0$, ce qui achève la preuve.

Remarque 14:

Ce résultat est vrai pour un sev stable F quelconque (il n'est pas nécessaire que F soit un espace propre). Si F est stable par u symétrique, son orthogonal l'est aussi. Faites la démo !

Théorème 8: Théorème spectral

Soit E un espace euclidien de dimension n.

1. Soit f un endomorphisme symétrique, alors f admet une base orthonormale de vecteurs propres.

2. Soit $A \in S_n(\mathbb{R})$, alors A est diagonalisable **en bon** dans $\mathbb{R}$. Autrement dit

$$\exists P \in O_n(\mathbb{R}), \qquad P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \lambda_n \end{pmatrix}$$

tel que $\forall i \in \{1, \dots, n\}, \lambda_i \in \mathbb{R}$. L'ev $O_n(\mathbb{R})$ est l'ensemble des matrices dites orthogonales c'est-à-dire vérifiant $P^{-1} = {}^tP$. Une matrice orthogonale a ses colonnes qui forment une bon de $\mathbb{R}^n$.

▸ **Pour aller plus loin**.

Preuve :

La démonstration s'effectue par récurrence sur la dimension de l'espace E.

• **Initialisation :** Si E de dimension 1, alors le théorème est vrai puisque la matrice est déjà diagonale.

Hérédité : Supposons maintenant que le théorème soit vrai pour tout endomorphisme symétrique sur un espace de dimension n et prouvons-le pour un endomorphisme symétrique f sur un espace E de dimension $n+1$. On sait (admis) que tout endomorphisme réel admet au moins une valeur propre complexe. Notons $\lambda \in \mathbb{C}$ cette valeur propre. On peut supposer que c'est une valeur propre simple, la démonstration suivante restant valable pour une valeur propre de multiplicité supérieure à 2.

Comme f est symétrique, la première proposition implique que $\lambda \in \mathbb{R}$.

Par ailleurs, on sait aussi d'après le cours d'espaces euclidiens que $E = E_\lambda \oplus E_\lambda^\perp$. On sait également que l'image d'un vecteur de $E_\lambda^\perp$ par f est dans $E_\lambda^\perp$ grâce à la proposition sur la stabilité de l'orthogonal. Ainsi la restriction de l'endomorphisme f : $f_{E_\lambda^\perp} : u \in E_\lambda^\perp \mapsto f(u) \in E_\lambda^\perp$ est un endomorphisme de $E_\lambda^\perp$.

Or la dimension de $E_\lambda^\perp$ vaut $n+1-1=n$ puisque E_λ est de dimension 1. On peut donc appliquer l'hypothèse de récurrence à $f_{E_\lambda^\perp}$ qui est un endomorphisme symétrique sur un ev de dimension n. Donc d'après l'hypothèse de récurrence on peut fabriquer une bon $\mathcal{B}^\perp$ de vecteurs propres de $f_{E_\lambda^\perp}$. Prenons alors un vecteur unitaire e_1 de E_λ et posons $\mathcal{B}$ la famille constituée de e_1 et des vecteurs de $\mathcal{B}^\perp$. Il s'agit d'une base de E car elle contient $n+1$ vecteurs dans un ev de dimension $n+1$ et car elle est orthonormale. C'est donc bien une bon et par construction c'est une bon de vecteurs propres de f.

La preuve précédente démontre le point 1) du théorème spectral : en quoi cela démontre-t-il le point 2) ?

Vous savez qu'à toute application linéaire on peut associer une matrice dans une base donnée (voir partie 5.1). Si cette base qu'on note $(e_1, \dots, e_n)$ est une base de vecteurs propres alors il existe des réels $\lambda_1, \dots, \lambda_n$ tels que $\forall i \in \{1, \dots, n\}, f(e_i) = \lambda_i e_i$ (définition de vecteur propre). La matrice dans cette base est alors bien diagonale :

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \lambda_n \end{pmatrix}$$

Maintenant, pour passer de la base canonique à la bon de vecteurs propres, il faut faire un changement de base donné par une certaine matrice P inversible vérifiant l'identité $P^{-1}AP = D$. Le fait que cette matrice vérifié $P^{-1} = {}^tP$ est dû au fait que la base soit orthonormale : si vous souhaitez comprendre pourquoi il faut étudier un chapitre niveau 2ème année de prépa appelé "endomorphismes orthogonaux" ou "isométries".

Exemple 24:

Diagonaliser en bon la matrice $S = \begin{pmatrix} -2 & -2 & 1 \\ -2 & 1 & -2 \\ 1 & -2 & -2 \end{pmatrix}$.

• **Première étape :** La matrice est symétrique en réelle donc elle est diagonalisable en bon.

• **Deuxième étape : Déterminer le spectre.** Fait précédemment dans le cours.

• **Troisième étape : déterminer les vecteurs propres de S.** Fait précédemment dans le cours.

• **Quatrième étape : déterminer la bon de vecteurs propres.**
On admet pour cette question que E_3 a pour base $(1,-2,1)$ et que E_{-3} a pour base $((1,0,-1),(2,1,0))$. On rappelle qu'il suffit d'en fabriquer une à l'intérieur de chaque espace propre, les espaces propres étant orthogonaux deux à deux !
Pour E_3 qui est de dimension 1, il suffit de normaliser le vecteur : $\epsilon_1 = \frac{1}{\sqrt{6}}(1,-2,1)$.

Pour E_{-3}, qui est de dimension 2, il faut orthonormaliser la base : Notons $u_2 = (1,0,-1), u_3 = (2,-1,0)$.
On normalise le premier vecteur : $\epsilon_2 = \frac{1}{\sqrt{2}}(1,0,-1)$.
On crée un vecteur orthogonal : $f_3 = u_3 - \langle u_3, \epsilon_2\rangle \epsilon_2 = (1,1,1)$.
On normalise le second vecteur : $\epsilon_3 = \frac{1}{\sqrt{3}}(1,1,1)$.

Donc $(\epsilon_1, \epsilon_2, \epsilon_3)$ est une bon de vecteurs propres.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- En quoi la preuve de l'orthogonalité des espaces propres ne marche plus si l'hypothèse "f symétrique" est supprimée ?
- En quoi la preuve du caractère réel des valeurs propres ne marche plus si l'hypothèse "f symétrique" est supprimée ?
- Pour toutes les preuves de ce paragraphe dire où intervient chacune des hypothèses s'il y en a et être capable d'expliquer l'origine de chaque étape/égalité.
- Qcm associé.
- **Pour aller plus loin :** Faire un plan de la démonstration du théorème spectral et expliquer comment s'articulent les différentes propositions de ce paragraphe dans la preuve du théorème spectral.

6 Extremas d'une fonction

6.1 Sur un exemple

L'objectif de cette section est de deviner sur un exemple, un critère permettant de déterminer si un point critique est un extremum local ou non. A terme, vous utiliserez en général ce critère pour déterminer si un point est un extremum local ou non.

Considérons la fonction $\begin{array}{rccc} f & : & \mathbb{R}^2 & \to & \mathbb{R} \\ & & (x,y) & \mapsto & ax^2+2bxy+cy^2+1 \end{array}$ où a,b,c sont trois réels fixés.

Quels sont les points critiques de f ?

La fonction est polynomiale donc de classe $C^1(\mathbb{R}^2)$. En supposant que $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ est inversible, on a que

$$\left\{ \begin{array}{l} \partial_x f(x,y)=0 \\ \partial_y f(x,y)=0 \end{array} \right. \iff \left\{ \begin{array}{l} 2ax+2by=0 \\ 2bx+2cy=0 \end{array} \right. \iff \left\{ \begin{array}{l} x=0 \\ y=0 \end{array} \right.$$

L'unique point critique est alors $(0,0)$?

Pour quels a, b, c ce point critique est-il un extremum ?

Pour cela, il est nécessaire d'étudier le signe de DL d'ordre 2 au voisinage de 0. Remarquons que f étant polynomiale d'ordre 2, elle est son propre DL au voisinage de 0 (sans reste). En effet,

$$f(x,y) = 1 + ax^2 + 2bxy + cy^2 = f(0,0) + \frac{1}{2}(\partial^2_{xx} f(0,0)x^2 + 2\partial^2_{xy} f(0,0)xy + \partial^2_{yy} f(0,0)y^2)$$

Le terme $f(0,0)$ est celui d'ordre 0, les autres forment le terme d'ordre 2. Le terme d'ordre 1 est nul puisque $(0,0)$ est un point critique de f. Ainsi la matrice hessienne n'est autre que $H_f(0,0) = \begin{pmatrix} 2a & 2b \\ 2b & 2c \end{pmatrix}$ de sorte que

$$f(x,y) - f(0,0) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Rappelons que notre objectif est l'étude du signe de $f(x,y) - f(0,0)$ afin de savoir si f admet un extremum local en $(0,0)$. Pour cela, nous allons diagonaliser la matrice hessienne dont le polynôme caractéristique est $P = \lambda^2 - \mathrm{tr}(H_f(0,0))\lambda + \det(H_f(0,0)) = \lambda^2 - 2(a+c)\lambda + 4(ac - b^2)$. Nous savons qu'elle est diagonalisable en bon d'après le théorème spectral puisqu'elle est symétrique. Nous allons nous intéresser maintenant à deux cas particuliers pour a, b, c.

Cas 1 : $a = 2, b = 1, c = 2$

Le polynôme $P = \lambda^2 - 8\lambda + 12 = (\lambda - 2)(\lambda - 6)$ donc les valeurs propres sont 1 et 3. Autrement dit, il existe une matrice P inversible telle que

$$P^{-1} H_f(0,0) P = {}^t P H_f(0,0) P = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$$

On en déduit alors que pour tout (x,y) dans $\mathbb{R}^2$

$$f(x,y) - f(0,0) = \frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} H_f(0,0) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} P \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} {}^t P \begin{pmatrix} x \\ y \end{pmatrix} = {}^t \begin{pmatrix} x \\ y \end{pmatrix} P \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} {}^t P \begin{pmatrix} x \\ y \end{pmatrix}.$$

Ainsi

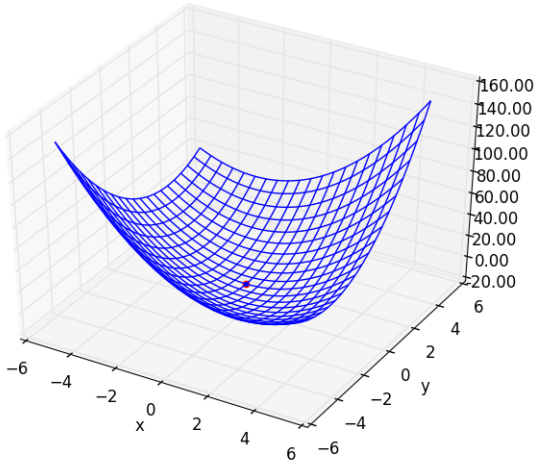
$$f(x,y) - f(0,0) = {}^t({}^t P \begin{pmatrix} x \\ y \end{pmatrix}) \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} {}^t P \begin{pmatrix} x \\ y \end{pmatrix}. = {}^t \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \begin{pmatrix} X \\ Y \end{pmatrix} = {}^t P \begin{pmatrix} x \\ y \end{pmatrix}$$

Développez le produit matriciel, vous verrez alors que

$$f(x,y) - f(0,0) = X^2 + 3Y^2 \geqslant 0$$

On alors montré que f admet un minimum global en $(0,0)$ (car le DL est exact sur $\mathbb{R}^2$). Plus généralement si les deux valeurs propres λ_1, λ_2 sont strictement positives, on a un minimum global car alors

$$f(x,y) - f(0,0) = \lambda_1 X^2 + \lambda_2 Y^2 \geqslant 0$$

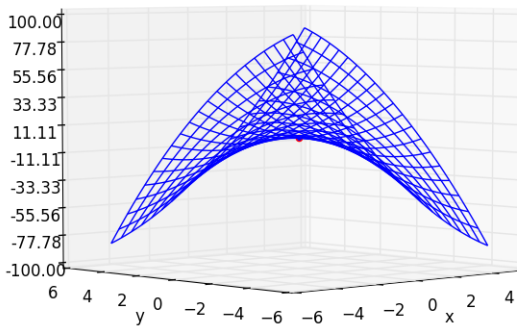


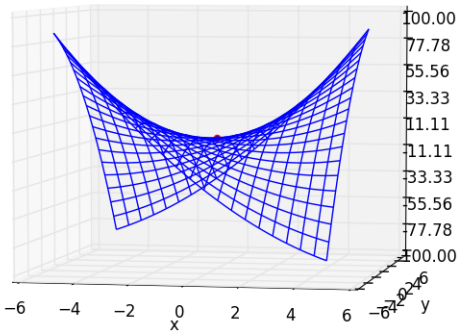
A l'inverse si les deux valeurs propres sont strictement négatives alors

$$f(x,y) - f(0,0) = \lambda_1 X^2 + \lambda_2 Y^2 \leqslant 0$$

Cas 2 : $a = 1, c = -1, b = \sqrt{3}$

Le polynôme $P = \lambda^2 - 16 = (\lambda - 4)(\lambda + 4)$ donc les valeurs propres sont -4 et 4. Donc le même raisonnement que ci-dessus mène à $f(x,y) - f(0,0) = 2X^2 - 2Y^2$. Ainsi pour $Y = 0$, $f(x,y) - f(0,0) = 2X^2 \geqslant 0$ et pour $X = 0$, $f(x,y) - f(0,0) = -2Y^2 \leqslant 0$. On a donc deux comportements différents suivants deux directions donc f n'admet pas d'extremum local en $(0,0)$.





Faisons un bilan : Le fait qu'il y ait un extremum ou non semble dépendre du signe des valeurs propres de la matrice hessienne. Lorsque les deux valeurs propres sont strictement positives il y a un minimum local, lorsqu'elles sont strictement négatives il y a un maximum local, lorsqu'elles sont non nulles mais de signe opposé il n'y a pas d'extremum. Ceci nous mène à une condition suffisante que nous exposons dans la section suivante.

6.2 Condition suffisante d'extremum

D'après la section précédente, il apparaît que la recherche d'extremum dépend du signe des valeurs propres de la matrice hessienne. On s'attend donc à devoir déterminer ces valeurs propres en pratique. Cependant

pour les fonctions de $\mathbb{R}^2$ dans $\mathbb{R}$, on peut ruser en souvenant de deux choses :

- le déterminant d'une matrice est égal au produit de ses valeurs propres.
- la trace d'une matrice est égal à la somme de ses valeurs propres.

Ainsi si les deux valeurs propres sont strictement positives (cas du minimum), le déterminant et la trace le sont. Si les deux valeurs propres sont strictement négatives, le déterminant est strictement positif et la trace est strictement négative (cas du maximum). Si les deux valeurs propres sont de signe opposé (pas d'extremum), le déterminant est strictement négatif. Ceci rend plus naturel la proposition suivante :

Proposition 18: Condition suffisante d'extremum

Soient A un sous-ensemble ouvert de $\mathbb{R}^2$, $f \in C^2(A)$, notons $H_f(a)$ sa matrice hessienne en a un point critique de A (dérivées partielles nulles).

- Si $\det(H_f(a)) > 0$ alors
  - si $\mathrm{tr}(H_f(a)) > 0$, f admet un minimum local en a.
  - si $\mathrm{tr}(H_f(a)) < 0$, f admet un maximum local en a.
- Si $\det(H_f(a)) < 0$ alors f n'a pas d'extremum local en a.
- Si $\det(H_f(a)) = 0$ alors on ne peut rien dire.

Exemple 25: Méthode pratique de recherche d'extremas

Considérons la fonction $f : (x, y) \in \mathbb{R}^2 \mapsto x^2 + y^2$.

Etape 0 : Donner la régularité de f

f est de classe $C^2(\mathbb{R}^2)$ car c'est une fonction polynomiale.

Etape 1 : Déterminer les points critiques

$$\begin{cases} \partial_x f(x,y)=0 \\ \partial_y f(x,y)=0 \end{cases} \iff \begin{cases} 2x=0 \\ 2y=0 \end{cases} \iff \begin{cases} x=0 \\ y=0 \end{cases}$$

Donc $(0, 0)$ est le seul point critique.

Etape 2 : Déterminer si les points critiques sont des extremas

Pour cela on applique le théorème suivant : on détermine la matrice hessienne. On a $\partial^2_{xx} f(x, y) = \partial^2_{yy} f(x, y) = 2$ et $\partial^2_{yx} f(x, y) = \partial^2_{xy} f(x, y) = 0$, la matrice hessienne en $(0, 0)$ est donc

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Le déterminant vaut 4 et $\mathrm{tr}(H_f(0)) = 4 > 0$ donc f admet un minimum local en $(0, 0)$.

Remarque 15:

Dans le cas où une des valeurs propres est nulle (où le déterminant est nul) on ne peut pas conclure. Il faut raisonner au cas par cas.

Questions :

- Quels sont les points essentiels de ce paragraphe ?
- Qcm associé.