Mathématiques pour l'informatique 1

(Notes provisoires)

Émilie Charlier

20 septembre 2021

Informations générales

— **Code cours :** MATH2019
— **Titulaire :** Émilie Charlier.
— **Assistants :** Célia Cisternino et Christophe Dozot.
— **Contact :** Campus du Sart-Tilman - zone Polytech 1
Institut de Mathématique B37, bureau 1/28
echarlier@uliege.be
ccisternino@uliege.be
c.dozot@uliege.be
— **Page web :** http://www.discmath.ulg.ac.be/charlier
— **Horaire du cours :** Q1, mardi après-midi.
— **Locaux :** Voir sur Celcat.
— **But du cours :** Il s'agit de donner au futur informaticien une palette d'outils mathématiques utiles à sa discipline. Tout en assurant la rigueur propre aux mathématiques de chacun des sujets présentés, nous souhaitons maintenir une interaction constante entre problèmes concrets et leur formalisation.
— **Contenu du cours :** Nous couvrirons les thèmes suivants.
- Logique propositionnelle et techniques de démonstrations associées.
- Démonstration par récurrence.
- Théorie naïve des ensembles et ensembles dénombrables.
- Arithmétique : algorithme d'Euclide et PGCD, arithmétique modulaire, numérations en bases entières, code de Gray.
- Calcul matriciel et systèmes linéaires.
— **Support :** Un syllabus est disponible sur la page web. Une bonne prise de note est indispensable.
— **Évaluation :** Une évaluation continue via le logiciel WIMS, une interrogation à la moitié du quadrimestre, un examen écrit en janvier portant sur la théorie et les exercices. La répartition est la suivante :

WIMS	10%
Interrogation de mi-quadri	15%
Examen de janvier	85%

— **Remédiation :** Deux séances de remédiation seront organisées les 29 novembre et 6 décembre. En cas d'échec à l'interrogation de mi-quadrimestre, ces séances sont obligatoires.
— **Modulation :** 6 crédits.

Chapitre 1

Mathématiques discrètes

1.1 Logique propositionnelle

Le programmeur, quel que soit le langage de programmation qu'il utilise, fait constamment usage de boucles "tant que" et d'instructions conditionnelles. Dans ces deux cas de figure (et bien d'autres encore), il est indispensable de pouvoir tester la véracité de conditions, parfois imbriquées les unes dans les autres, rendant compliquée la tâche d'expliquer (à son patron ou à son client, par exemple) dans quel état se trouve le système après exécution d'une partie quelconque de son programme. Il est donc primordial de pouvoir passer de façon systématique d'un état à un autre, et de pouvoir expliquer sa démarche. Dans cette optique, nous démarrerons ce cours de "Mathématiques pour l'informatique 1" par une introduction à la logique et à différentes techniques de preuve.

Propositions logiques et tables de vérité

Appelons *assertion* ou *proposition* toute phrase d'un langage donné dont on peut déterminer sans ambiguïté sa vérité ou sa fausseté. Par exemple, les phrases en langue française "je suis allée au cinéma hier soir" ou "il a neigé ce matin" peuvent être considérées comme des propositions. En logique mathématique et en informatique, on souhaite formaliser ce qu'est une proposition.

Définition 1.1. En logique propositionnelle classique (LPC), une *proposition*[1] est formée à partir de propositions atomiques en nombre fini arbitraire, appelées *variables propositionnelles* (ou simplement *variables*), en respectant des règles précises appelées des *opérations logiques* ou encore *connecteurs logiques*. Ces opérations logiques sont

- la négation ("non"), notée $\neg$;
- la conjonction ("et") , notée $\wedge$;
- la disjonction ("ou"), notée $\vee$;
- l'implication ("implique"), notée $\Rightarrow$;
- la bi-implication ("si et seulement si"), notée $\Leftrightarrow$.

Ainsi, une proposition sera toujours formée d'un nombre n quelconque, mais fini, de variables propositionnelles, qu'on pourra noter par exemple $x_1, x_2, x_3, \ldots, x_n$. Les propositions seront généralement notées par des lettres grecques $\alpha, \beta, \gamma, \ldots, \varphi, \psi, \theta, \ldots$

Puisqu'une proposition est par principe soit vraie, soit fausse, on lui attribuera une *valeur de vérité* "vrai" ou "faux", que l'on notera en général "1" ou "0" respectivement.

La construction récursive des propositions induit une notion de vérité qui dépend des valeurs de vérité prises par les variables propositionnelles. À chaque opération logique est

1. On parle parfois de *formule de la logique propositionnelle.*

associée une *table de vérité* répertoriant tous les cas possibles des valeurs de vérité de la nouvelle proposition ainsi créée en fonction des valeurs de vérités prises par les sous-propositions qui la composent.

Table de vérité de la négation

φ	$\neg\varphi$
0	1
1	0

Table de vérité de la conjonction

φ	ψ	$\varphi \wedge \psi$
0	0	0
0	1	0
1	0	0
1	1	1

Remarquons que la conjonction est commutative, c'est-à-dire que les tables de vérité de $\varphi \wedge \psi$ et de $\psi \wedge \varphi$ sont les mêmes.

Table de vérité de la disjonction

Remarquons que le "ou" de la langue française est ambigu en général. Il désigne selon les cas le "ou" dit *exclusif*, c'est-à-dire, lorsqu'on suppose que les deux parties de la phrase composant le "ou" ne sont pas simultanément vraies, ou bien le "ou" dit *non exclusif*, lorsqu'on autorise les deux parties à être vraies en même temps. Si l'on veut être précis, on peut dire "soit. . . , soit. . . , mais pas les deux" pour le "ou exclusif" et "soit. . . , soit. . . , voire les deux en même temps" pour le "ou" non exclusif. En logique propositionnelle, la **disjonction** désigne par définition le **"ou" non exclusif**, et il n'y a donc jamais d'ambiguïté.

φ	ψ	$\varphi \vee \psi$
0	0	0
0	1	1
1	0	1
1	1	1

Remarquons que, tout comme la conjonction, la disjonction est commutative : les tables de vérité de $\varphi \vee \psi$ et de $\psi \vee \varphi$ coïncident.

Table de vérité de l'implication

φ	ψ	$\varphi \Rightarrow \psi$
0	0	1
0	1	1
1	0	0
1	1	1

Remarquons que l'implication n'est pas commutative : les tables de vérité de $\varphi \Rightarrow \psi$ et de $\psi \Rightarrow \varphi$ ne sont pas les mêmes.

Table de vérité de la bi-implication

φ	ψ	$\varphi \Leftrightarrow \psi$
0	0	1
0	1	0
1	0	0
1	1	1

Condition nécessaire et suffisante

En mathématiques, on parle très souvent de "condition nécessaire et suffisante". Vous avez certainement tous déjà vu ou utilisé la formulation "si et seulement si". Ces mots ont un sens précis et méritent une définition formelle.

Définition 1.2. Supposons que φ et ψ soient des propositions. On lira l'implication $\varphi \Rightarrow \psi$ en disant "φ implique ψ" ou encore "si φ, alors ψ". De plus, lorsque cette implication $\varphi \Rightarrow \psi$ est vraie, on dira que φ est une *condition suffisante* pour ψ. Inversement, on dira que ψ est une *condition nécessaire* pour φ.

Supposons que "$\varphi \Leftrightarrow \psi$", ou "$\varphi$ si et seulement si ψ", soit l'énoncé d'un théorème duquel on souhaite donner une démonstration. Cette démonstration doit toujours contenir deux parties indépendantes, dont l'ordre n'a pas d'importance.

La première est la démonstration de $\varphi \Rightarrow \psi$. On qualifie cette partie de démonstration de la *condition nécessaire* : comprenez, du fait que ψ est une condition nécessaire pour φ. C'est à cette première partie que fait référence le "seulement si" de la formulation "φ si et seulement si ψ". On pourrait également dire : "pour que φ soit vrai, il est nécessaire que ψ soit vrai".

La deuxième partie de la démonstration, qui doit être obtenue indépendamment de la première, est celle de $\varphi \Leftarrow \psi$, ou encore $\psi \Rightarrow \varphi$. Cette deuxième partie est la *condition suffisante*, c'est-à-dire la démonstration du fait que ψ est une condition suffisante pour φ. C'est à cette deuxième partie que fait référence le "si" de la formulation "φ si et seulement si ψ". On pourrait également dire : "si ψ est vrai, alors φ est vrai aussi" ou encore, "il suffit que ψ soit vrai pour que φ soit vrai aussi".

Ainsi, on aura démontré que ψ est une condition nécessaire et suffisante pour φ. Remarquons qu'en dépit de la symétrie de la notation $\varphi \Leftrightarrow \psi$, on attribue bien les noms de condition nécessaire et de condition suffisante par rapport à la proposition de gauche (ici, de φ).

Parenthèses

Grâce à ces règles de formation de nouvelles propositions logiques à partir d'anciennes, on peut construire des propositions logiques de plus en plus compliquées. Il conviendra donc d'utiliser des parenthèses pour indiquer l'ordre dans lequel s'effectuent ces opérations logiques.

Remarquons par exemple que les tables de vérité de $\varphi \wedge (\psi \vee \theta)$ et de $(\varphi \wedge \psi) \vee \theta$ sont

différentes :

φ	ψ	θ	$\psi \vee \theta$	$\varphi \wedge (\psi \vee \theta)$	$\varphi \wedge \psi$	$(\varphi \wedge \psi) \vee \theta$
0	0	0	0	0	0	0
0	0	1	1	0	0	1
0	1	0	1	0	0	0
0	1	1	1	0	0	1
1	0	0	0	0	0	0
1	0	1	1	1	0	1
1	1	0	1	1	1	1
1	1	1	1	1	1	1

Équivalence logique

Définition 1.3. Deux propositions φ et ψ sont dites *logiquement équivalentes* lorsqu'elles ont les mêmes tables de vérité, ce que nous noterons $\varphi \equiv \psi$.

Remarquez la symétrie de cette définition. On note donc indifféremment $\varphi \equiv \psi$ ou $\psi \equiv \varphi$ pour dire que φ et ψ sont logiquement équivalentes.
Donnons tout de suite une liste de familles de propositions équivalentes. Pour vérifier que c'est bien le cas, il faudra donc, et ce pour chaque famille de propositions donnée, vérifier qu'elles ont effectivement bien les mêmes tables de vérités.

Proposition 1.4. *Soient φ, ψ, θ des propositions. Alors on a les équivalences logiques suivantes.*

1. $\varphi \equiv \neg(\neg\varphi)$
2. $\varphi \wedge \psi \equiv \psi \wedge \varphi$
3. $\varphi \vee \psi \equiv \psi \vee \varphi$
4. $\neg(\varphi \wedge \psi) \equiv \neg\varphi \vee \neg\psi$
5. $\neg(\varphi \vee \psi) \equiv \neg\varphi \wedge \neg\psi$
6. $\varphi \wedge (\psi \wedge \theta) \equiv (\varphi \wedge \psi) \wedge \theta$
7. $\varphi \vee (\psi \vee \theta) \equiv (\varphi \vee \psi) \vee \theta$
8. $\varphi \wedge (\psi \vee \theta) \equiv (\varphi \wedge \psi) \vee (\varphi \wedge \theta)$
9. $\varphi \vee (\psi \wedge \theta) \equiv (\varphi \vee \psi) \wedge (\varphi \vee \theta)$
10. $\varphi \Leftrightarrow \psi \equiv (\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$
11. $\varphi \Rightarrow \psi \equiv \neg\varphi \vee \psi$
12. $\neg(\varphi \Rightarrow \psi) \equiv \varphi \wedge \neg\psi$
13. $\varphi \Rightarrow \psi \equiv \neg\psi \Rightarrow \neg\varphi$
14. $\varphi \equiv \neg\varphi \Rightarrow (\psi \wedge \neg\psi)$
15. $\varphi \Rightarrow (\psi \vee \theta) \equiv (\varphi \wedge \neg\psi) \Rightarrow \theta$
16. $(\varphi \vee \psi) \Rightarrow \theta \equiv (\varphi \Rightarrow \theta) \wedge (\psi \Rightarrow \theta)$.

Remarquons que ces équivalences logiques montrent que tous les connecteurs logiques que nous avons vus s'obtiennent à partir de $\neg$ et de $\vee$.

Proposition 1.5. *Les connecteurs $\wedge$, $\Rightarrow$ et $\Leftrightarrow$ s'obtiennent à partir de $\neg$ et de $\vee$.*

Preuve. En utilisant les équivalences logiques 1 et 4, nous avons successivement $\varphi \wedge \psi \equiv \neg(\neg(\varphi \wedge \psi)) \equiv \neg(\neg\varphi \vee \neg\psi)$. Le reste est donné par les équivalences logiques 10 et 11. □

Exercice 1.6. Bien que complète, la preuve de la proposition 1.5 ne fournit pas explicitement de proposition équivalente à $\varphi \Leftrightarrow \psi$ utilisant uniquement les connecteurs $\neg$ et de $\vee$. Pouvez-vous en donner une ?

Tautologie et contradiction

Définition 1.7. Une *tautologie* est une proposition toujours vraie, quelles que soient les valeurs de vérité attribuées aux variables propositionnelles qui la composent. Autrement dit, la dernière colonne de la table de vérité d'une tautologie ne contient que la valeur 1.

Au vu de la proposition précédente, les propositions $\varphi \Leftrightarrow \neg(\neg\varphi)$ et $\big(\varphi \wedge (\psi \vee \theta)\big) \Leftrightarrow \big((\varphi \wedge \psi) \vee (\varphi \wedge \theta)\big)$ sont des tautologies. Il en est de même pour les propositions ainsi construites à partir de la proposition 1.4. Voici une liste de propositions, dont on vérifiera qu'il s'agit de tautologies en construisant leurs tables de vérité.

Proposition 1.8. *Soient φ, ψ, θ des propositions. Alors les propositions suivantes sont des tautologies.*

1. $\big((\varphi \Rightarrow \psi) \wedge \varphi\big) \Rightarrow \psi$
2. $\big((\varphi \vee \psi) \wedge \neg\varphi\big) \Rightarrow \psi$
3. $(\varphi \wedge \psi) \Rightarrow \varphi$
4. $\varphi \vee \neg\varphi$
5. $\neg(\varphi \wedge \neg\varphi)$
6. $\neg\varphi \Rightarrow (\varphi \Rightarrow \psi)$
7. $\psi \Rightarrow (\varphi \Rightarrow \psi)$
8. $(\varphi \wedge \neg\varphi) \Rightarrow \psi$
9. $\big(\varphi \Rightarrow (\psi \Rightarrow \theta)\big) \Rightarrow \big((\varphi \Rightarrow \psi) \Rightarrow (\varphi \Rightarrow \theta)\big)$

Définition 1.9. La négation d'une tautologie est appelée une *contradiction* ou une *absurdité*. Il s'agit donc d'une proposition toujours fausse, quelles que soient les valeurs de vérité attribuées aux variables propositionnelles qui la composent. Autrement dit, la dernière colonne de la table de vérité d'une contradiction ne contient que la valeur 0.

Problème SAT

Définition 1.10. Une proposition est dite *satisfaisable* si la dernière colonne de la table de vérité contient au moins un 1. Autrement dit, il doit exister une distribution des valeurs de vérité des variables la composant qui la rende vraie.

Définition 1.11. Une *clause* est une disjonction de variables propositionnelles ou de négation de variables propositionnelles.

Par exemple, $x \vee y \vee \neg z$ est une clause.

Proposition 1.12. *Toute proposition est logiquement équivalente à une conjonction de clauses.*

Preuve. Soit φ une proposition dont les variables propositionnelles sont $x_1, \ldots, x_n$. La table de vérité de $\neg\varphi$ est constituée de 2^n lignes, chaque ligne correspondant à une distribution des valeurs de vérité des n variables propositionnelles $x_1, \ldots, x_n$. À la ligne i $(1 \leq i \leq 2^n)$, on fait correspondre la conjonction $f_i(x_1) \wedge \cdots \wedge f_i(x_n)$, où pour chaque j $(1 \leq j \leq n)$, $f_i(x_j) = x_j$ si la valeur de vérité de x_j à la ligne i est 1 et $f_i(x_j) = \neg x_j$ sinon. La proposition $\neg\varphi$ est équivalente à la disjonction des conjonctions de $x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n$ ainsi construites et provenant des lignes ayant un 1 dans la dernière colonne. On obtient une conjonction de clauses logiquement équivalentes à φ en niant la disjonction ainsi obtenue et en utilisant les équivalences logiques 4 et 5 de la propostion 1.4 (étendues à un nombre quelconque de termes). □

Définition 1.13.
- Une *forme normale conjonctive* d'une proposition φ est une conjonction de clauses logiquement équivalente à φ.
- Une *forme normale disjonctive* d'une proposition φ est une disjonction de conjonctions de variables propositionnelles ou de négation de variables propositionnelles logiquement équivalente à φ.

Prenons comme exemple la proposition

$$\varphi \equiv \neg(w \Rightarrow (x \Rightarrow (y \wedge z))) \Leftrightarrow (\neg y \wedge (x \vee (y \Leftrightarrow (w \vee \neg z)))).$$

Nous commençons par construire la table de vérité de $\neg\varphi$.

w	x	y	z	$\neg(w \Rightarrow (x \Rightarrow (y \wedge z)))$	$\neg y \wedge (x \vee (y \Leftrightarrow (w \vee \neg z)))$	φ	$\neg\varphi$
0	0	0	0	0	0	1	0
0	0	0	1	0	1	0	1
0	0	1	0	0	0	1	0
0	0	1	1	0	0	1	0
0	1	0	0	0	1	0	1
0	1	0	1	0	1	0	1
0	1	1	0	0	0	1	0
0	1	1	1	0	0	1	0
1	0	0	0	0	0	1	0
1	0	0	1	0	0	1	0
1	0	1	0	0	0	1	0
1	0	1	1	0	0	1	0
1	1	0	0	1	1	1	0
1	1	0	1	1	1	1	0
1	1	1	0	1	0	0	1
1	1	1	1	0	0	1	0

On obtient une forme normale disjonctive de $\neg\varphi$:

$$\neg\varphi \equiv (\neg w \wedge \neg x \wedge \neg y \wedge z) \vee (\neg w \wedge x \wedge \neg y \wedge \neg z) \vee (\neg w \wedge x \wedge \neg y \wedge z) \vee (w \wedge x \wedge y \wedge \neg z).$$

En prenant la négation de cette forme normale disjonctive, on obtient une forme normale conjonctive de φ :

$$\varphi \equiv (w \vee x \vee y \vee \neg z) \wedge (w \vee \neg x \vee y \vee z) \wedge (w \vee \neg x \vee y \vee \neg z) \wedge (\neg w \vee \neg x \vee \neg y \vee z). \quad (1.1)$$

Le problème SAT est un *problème de décision*. Sans entrer dans les détails de la théorie de la calculabilité, objet d'un autre cours de votre cursus "Introduction to the theory of computation", un problème (bien posé) pour lequel la réponse est "oui" ou "non" est décidable si, étant donné n'importe quelle instance (parmi une infinité d'instances) de ce problème, il existe un algorithme qui permet de déterminer si cette instance est "positive" ou "négative", c'est-à-dire est une solution du problème ou non. Le problème SAT est un problème fondamental de l'informatique théorique et à la base de la théorie de la complexité. Il se formule comme suit.

Définition 1.14 (Problème SAT)**.** Étant donné une proposition sous forme normale conjonctive, déterminer si elle est ou si elle n'est pas satisfaisable.

Remarquez que vous disposez déjà d'un algorithme pour répondre à cette question, et ce, quelle que soit l'instance du problème SAT que l'on vous donne, c'est-à-dire, quelle que

soit la proposition (mise sous forme normale conjonctive ou pas d'ailleurs) [2]. Il s'agit donc d'un problème *décidable*. L'intérêt de SAT réside dans le fait qu'il s'agit d'un problème *difficile*, au sens où il n'existe pas d'algorithme rapide pour le résoudre [3]. Par exemple, que pensez-vous du temps mis par votre algorithme en fonction du nombre de variables propositionnelles de la formule de départ (même lorsque la proposition vous est donnée sous forme normale conjonctive) ?

Table de Karnaugh

Nous avons vu dans la proposition 1.12 que toute proposition était équivalente à une conjonction de clauses et que pour obtenir cette forme normale, il suffisait de se baser sur la table de vérité. Ce procédé est intéressant car tout à fait général. Cependant, il a un désavantage de taille, et c'est le cas de le dire, puisque la forme normale conjonctive est généralement bien plus longue que la proposition de départ (pensez aux 2^n de la table de vérité !). Il est donc important de réfléchir à des procédés qui tentent de fournir des propositions équivalentes les plus compactes possibles. L'un d'entre eux est donné par les tables de Karnaugh.

Définition 1.15. Une table de Karnaugh d'une proposition φ à $2n$ (resp. $2n+1$) variables est une table à deux dimensions construites comme suit. On sépare les variables propositionnelles en deux parties de taille n (resp. n et $n+1$). Les différentes distributions des valeurs de vérité de chaque groupe, sous forme de n-uplets (resp. n-uplets et $(n+1)$-uplets), sont listés verticalement et horizontalement avec la contrainte que deux entrées consécutives ne diffèrent qu'en une composante.

La table contient donc n^2 (resp. $n(n+1)$) cases, dans chacune desquelles on inscrit la valeur de vérité attribuées de φ correspondant aux valeurs de vérité de ses variables propositionnelles données par la ligne et la colonne sur lesquelles la case se trouve. Le but est de faire apparaître visuellement des simplifications possibles. Voici un exemple d'une table de Karnaugh de la formule

$$\varphi \equiv \neg(w \Rightarrow (x \Rightarrow (y \wedge z))) \Leftrightarrow (\neg y \wedge (x \vee (y \Leftrightarrow (w \vee \neg z))),$$

dont la table de vérité a déjà été donnée plus haut. Nous choisirons de séparer les variables propositionnelles w, x, y, z en les deux sous-groupes (w, x) (verticalement) et (y, z) (horizontalement). Remarquez qu'on aurait pu faire d'autres choix [4].

w x \ y z	0 0	0 1	1 1	1 0
0 0	1	0	1	1
0 1	0	0	1	1
1 1	1	1	1	0
1 0	1	1	1	1

Deux méthodes de simplification peuvent être utilisées selon qu'on essaie de former une disjonction ou une conjonction.

Recherche d'une forme normale disjonctive de longueur minimale

On souhaite obtenir une forme normale disjonctive logiquement équivalente qui soit la plus courte possible. Dans une table de Karnaugh, la dernière colonne (resp. ligne) sera

2. Lequel ?
3. La définition rigoureuse d'un problème difficile vous sera donnée dans votre cours "Introduction to the theory of computation". Vous y verrez que SAT est *NP-complet*.
4. Combien de tels choix sont-ils possibles ?

considérée comme étant adjacente à gauche (en dessous) de la première colonne (resp. ligne). On regroupe les valeurs de φ égales à 1 en rectangles contenant un nombre de cases égal à une puissance de 2 (c'est-à-dire $1, 2, 4, 8, 16, 32, \ldots$). Un *rectangle* de taille $\ell \times c$ est obtenu en sélectionnant ℓ lignes et c colonnes adjacentes [5]. On essaie trouver le plus petit nombre possible de rectangles les plus grands possibles, et ce, en n'oubliant aucun 1. Rien n'empêche qu'un même 1 fasse partie de plusieurs rectangles à la fois, mais aucun 0 ne doit appartenir à un rectangle.

Chaque rectangle sera décrit par une conjonction de variables propositionnelles et de négations de variables propositionnelles. Ceci est dû du fait qu'on ait imposé dans la définition d'une table de Karnaugh que les n-uplets de lignes (ou colonnes) adjacentes ne diffèrent qu'en une seule composante. La disjonction de ces conjonctions donnera une proposition équivalente à la proposition φ de départ.

Avec la procédure suivante, on assure que la forme normale disjonctive obtenue soit de *longueur minimale*. On commence par tester tous les rectangles de taille maximale (où la taille est donnée par le nombre de cases [6]), puis on diminue les tailles considérées jusqu'à finalement considérer les carrés de taille 1.

Remarquons qu'en général, plusieurs choix de rectangles respectant ces contraintes sont possibles. On peut donc obtenir autant de formes normales disjonctives que de choix de rectangles sont possibles et, évidemment, toutes les formes normales disjonctives ainsi obtenues sont logiquement équivalentes.

Dans notre exemple, nous testerons d'abord les rectangles de taille 16, puis $8, 4, 2$, et enfin les rectangles de taille 1, c'est à dire les cases seules. Aucun rectangle de taille 16 ou 8 n'est possible. On peut alors vérifier que les quatre rectangles de taille 4 suivants respectent les contraintes décrites ci-dessus.

— L'intersection des première et dernière lignes et des première et dernière colonnes :

w x \ y z	0 0	0 1	1 1	1 0
0 0	**1**	0	1	**1**
0 1	0	0	1	1
1 1	1	1	1	0
1 0	**1**	1	1	**1**

Ce rectangle est décrit par la conjonction $\neg x \wedge \neg z$.

— L'intersection des première et deuxième lignes et des troisième et quatrième colonnes :

w x \ y z	0 0	0 1	1 1	1 0
0 0	1	0	**1**	**1**
0 1	0	0	**1**	**1**
1 1	1	1	1	0
1 0	1	1	1	1

Ce rectangle est décrit par la conjonction $\neg w \wedge y$.

— L'intersection des troisième et quatrième lignes et des première et deuxième co-

5. Avec nos règles d'adjacence particulières, un rectangle peut être coupé en deux, puisqu'il peut contenir par exemple des cases des deux premières colonnes ainsi que des cases de la dernière colonne. En fait, on voit une table de Karnaugh comme un *tore*, c'est-à-dire que la colonne la plus à droite serait suivie de la colonne la plus à gauche, et de même la dernière ligne serait suivie de la première.

6. Un rectangle de taille 16 peut être composé de ℓ lignes et c colonnes pour les couples de valeurs de (ℓ, c) suivants : $(16, 1), (8, 2)$ et $(4, 4)$.

lonnes :

w x \ y z	0 0	0 1	1 1	1 0
0 0	1	0	1	1
0 1	0	0	1	1
1 1	**1**	**1**	1	0
1 0	**1**	**1**	1	1

Ce rectangle est décrit par la conjonction $w \wedge \neg y$.

— La troisième colonne :

w x \ y z	0 0	0 1	1 1	1 0
0 0	1	0	**1**	1
0 1	0	0	**1**	1
1 1	1	1	**1**	0
1 0	1	1	**1**	1

Ce rectangle est décrit par la conjonction $y \wedge z$.

Ainsi, nous obtenons la forme normale disjonctive équivalente à φ de longueur 8

$$(\neg x \wedge \neg z) \vee (\neg w \wedge y) \vee (w \wedge \neg y) \vee (y \wedge z).$$

À titre de comparaison, la forme normale disjonctive obtenue directement à partir de la table de vérité est de longueur 48. En effet, la table de vérité de φ (voir page 7) possède 12 fois la valeur 1, et chacun de ces 1 est décrit par une conjonction de longueur 4 (puisqu'il y a 4 variables propositionnelles).

Recherche d'une forme normale conjonctive de longueur minimale

Soit une proposition φ. En appliquant la méthode précédente à partir d'une table de Karnaugh de $\neg\varphi$, on obtient une forme normale disjonctive de $\neg\varphi$. Ensuite, en prenant sa négation et en utilisant les équivalences logiques 4 et 5 de la proposition 1.4, on obtient une forme normale conjonctive de φ. On peut montrer que cette procédure donne encore une longueur minimale.

Illustrons cette deuxième méthode sur notre exemple. En échangeant les 0 et les 1 dans la table de Karnaugh de φ, on obtient une table de Karnaugh correspondant à $\neg\varphi$:

w x \ y z	0 0	0 1	1 1	1 0
0 0	0	1	0	0
0 1	1	1	0	0
1 1	0	0	0	1
1 0	0	0	0	0

Aucun rectangle de 1 de taille $16, 8$ ou 4 n'est possible. On voit facilement que les deux rectangles de taille 2 avec le rectangle de taille 1 suivants respectent les contraintes exposées précédemment.

— L'intersection de la deuxième ligne et des première et deuxième colonnes :

w x \ y z	0 0	0 1	1 1	1 0
0 0	0	1	0	0
0 1	**1**	**1**	0	0
1 1	0	0	0	1
1 0	0	0	0	0

Ce rectangle est décrit par la conjonction $\neg w \wedge x \wedge \neg y$.

— L'intersection des première et deuxième lignes et de la deuxième colonne :

w x \ y z	0 0	0 1	1 1	1 0
0 0	0	**1**	0	0
0 1	1	**1**	0	0
1 1	0	0	0	1
1 0	0	0	0	0

Ce rectangle est décrit par la conjonction $\neg w \wedge \neg y \wedge z$.

— L'intersection de la quatrième ligne et de la troisième colonne :

w x \ y z	0 0	0 1	1 1	1 0
0 0	0	1	0	0
0 1	1	1	0	0
1 1	0	0	0	**1**
1 0	0	0	0	0

Ce rectangle est décrit par la conjonction $w \wedge x \wedge y \wedge \neg z$.

Ainsi, nous obtenons la forme normale disjonctive équivalente à $\neg\varphi$ de longueur 10

$$(\neg w \wedge x \wedge \neg y) \vee (\neg w \wedge \neg y \wedge z) \vee (w \wedge x \wedge y \wedge \neg z)$$

En passant à la négation de cette conjonction, on obtient la forme normale conjonctive

$$(w \vee \neg x \vee y) \wedge (w \vee y \vee \neg z) \wedge (\neg w \vee \neg x \vee \neg y \vee z)$$

équivalente à φ et plus compacte que celle (1.1) de longueur 16 obtenue précédemment.

En programmation, l'utilisation des tables de Karnaugh permet de réduire les conditions testées lors des instructions conditionnelles. De cette démarche résultent de nouvelles conditions qui peuvent être parfois non intuitives au premier abord, mais qui réduisent, souvent considérablement, la taille du code ainsi que son temps d'exécution en diminuant le nombre d'évaluations nécessaires.

1.2 Quelques techniques de démonstration

Équivalence logique et substitution

Il est important de noter que lorsque deux propositions sont logiquement équivalentes, on peut substituer l'une par l'autre dans n'importe quelle proposition sans que cela ne change sa table de vérité. Autrement dit, en substituant dans une proposition une "sous-proposition" par une autre "sous-proposition" logiquement équivalente, on obtient une nouvelle proposition logiquement équivalente. C'est le concept de substitution, qui est utilisé constamment dans les preuves mathématiques.

Démonstration directe d'une implication

Lorsque qu'un théorème s'énonce par "si A, alors B", la méthode la plus employée pour le démontrer est de supposer A vrai et, avec cette hypothèse, d'ensuite démontrer que B doit être vrai aussi. Ceci est justifié par la table de vérité associée au connecteur logique $\Rightarrow$. Quand on fait cela, on démontre bien une *implication*, c'est-à-dire, comme déjà vu précédemment, que A est une *condition suffisante* pour B. Il est important de bien comprendre que montrer une implication ne donne aucune information sur la vérité de A et B indépendamment.

Contraposition

Certaines équivalences logiques correspondent à des techniques de démonstration bien connues, qui méritent d'être mises en évidence. C'est le cas de $\varphi \Rightarrow \psi \equiv \neg\psi \Rightarrow \neg\varphi$.

Définition 1.16. La proposition $\neg\psi \Rightarrow \neg\varphi$ est appelée la *contraposition* (on dit aussi la *contraposée*) de $\varphi \Rightarrow \psi$.

Nous allons illustrer le principe de démonstration par contraposition avec le résultat suivant.

Proposition 1.17. *Si n est un nombre entier tel que n^2 est pair, alors n est pair.*

Preuve. Soit n un nombre entier quelconque. Nous devons montrer que si n^2 est pair, alors n est pair. Nous allons montrer la contraposée [7], c'est-à-dire que si n est impair, alors n^2 est impair. Supposons donc que n est impair. Dans ce cas, nous savons qu'il existe un entier k tel $n = 2k + 1$. On calcule $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$, prouvant ainsi que n^2 est impair. □

Ainsi, pour démontrer l'implication $A \Rightarrow B$, on montre que B est une *condition nécessaire* pour A. Autrement dit, si B est faux, alors A ne peut être vrai.

Remarque 1.18. La *réciproque* d'une implication $\varphi \Rightarrow \psi$ est l'implication $\psi \Rightarrow \varphi$. Il s'agit d'une notion différente de celle de contraposée. En effet, une implication et sa contraposée sont toujours vraies simultanément, alors que les valeurs de vérité d'une implication et de sa réciproque sont indépendantes.

Démonstration par l'absurde

L'équivalence logique $\varphi \equiv \neg\varphi \Rightarrow (\psi \wedge \neg\psi)$ donne lieu à la technique de *démonstration par l'absurde*. En effet, la proposition $(\psi \wedge \neg\psi)$ est une contradiction (pour rappel, cela signifie que sa table de vérité ne contient que des 0). Dès lors, l'implication $\neg\varphi \Rightarrow (\psi \wedge \neg\psi)$ ne peut être vraie que si $\neg\varphi$ est faux, c'est-à-dire, que si φ est vrai.

En mathématique, le raisonnement par l'absurde est très souvent utilisé pour montrer qu'un ensemble est vide. On suppose qu'il ne l'est pas, et on essaie de parvenir à une absurdité. Souvent le raisonnement par l'absurde a comme désavantage de ne pas être constructif. Par exemple, on peut montrer par l'absurde qu'une équation possède une solution (car si elle n'en avait pas, on obtiendrait une contradiction) sans pour autant être capable de fournir explicitement une solution de cette équation ! Voici un exemple d'une telle démonstration.

Proposition 1.19. *Il existe une infinité de nombres premiers.*

Pour rappel, un nombre naturel est premier s'il possède exactement deux diviseurs naturels. Les premiers nombres premiers sont $2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots$.

Preuve. Procédons par l'absurde et supposons qu'il n'y en ait qu'un nombre fini n. Notons-les $p_1, \ldots, p_n$. On considère maintenant le nombre $p_1 p_2 \cdots p_n + 1$. Ce nouveau nombre est un nombre naturel différent de 1 et qui n'est divisible par aucun des nombres $p_1, \ldots, p_n$. Il s'agit donc d'un nombre premier [8]. Mais il est aussi strictement plus grand que tous les nombres premiers supposés $p_1, \ldots, p_n$, ce qui impossible. □

7. Pourquoi est-ce judicieux de passer à la contraposée dans cet exemple ?
8. Pourquoi ?

Démonstration d'une alternative

Supposons que l'on veuille démontrer une *alternative*, c'est-à-dire une implication de la forme $\varphi \Rightarrow (\psi \vee \theta)$. Dans ce cas, l'équivalence logique $\varphi \Rightarrow (\psi \vee \theta) \equiv (\varphi \wedge \neg\psi) \Rightarrow \theta$ donne lieu une technique de démonstration courante et bien commode. Ainsi, pour démontrer que "si j'aime le chocolat, alors je prends comme dessert un moelleux au chocolat ou une crêpe au chocolat", on pourra montrer que "si j'aime le chocolat et que je n'ai pas choisi comme dessert le moelleux au chocolat, alors mon dessert est une crêpe au chocolat". Ceci nous permet de faire une hypothèse supplémentaire (ici, que je n'ai pas choisi le moelleux au chocolat), généralement fort utile pour notre démonstration. Remarquez que, de façon équivalente [9], on peut également montrer que "si j'aime le chocolat et que je n'ai pas choisi comme dessert une crêpe au chocolat, alors mon dessert est un moelleux au chocolat".

Voici un exemple d'une telle démonstration.

Proposition 1.20. *Si un entier n est divisible par 3, alors n est divisible par 6 ou $n-3$ est divisible par 6.*

Preuve. Supposons que n est un entier divisible par 3 mais pas par 6. Nous devons montrer que $n-3$ est divisible par 6. On sait que $n = 3k$, pour un entier k. Puisque n n'est pas divisible par 6, k n'est pas divisible par 2. Donc $k = 2\ell + 1$, pour un entier ℓ. On obtient que $n - 3 = 3k - 3 = 3(k-1) = 3 \cdot 2\ell = 6\ell$ et donc que $n-3$ est divisible par 6. □

Disjonction des cas

Lorsque l'on doit démontrer une implication de la forme $(\varphi \vee \psi) \Rightarrow \theta$, on procède généralement par *disjonction des cas*. Cette technique est traduite par l'équivalence logique $(\varphi \vee \psi) \Rightarrow \theta \equiv (\varphi \Rightarrow \theta) \wedge (\psi \Rightarrow \theta)$. Par exemple, si je dois prouver que "si cette fleur est jaune ou rouge, alors elle est à moi", je montrerai que "si cette fleur est jaune, alors elle est à moi" et aussi que "si cette fleur est rouge, alors elle est à moi".

Voici un exemple d'une démonstration par disjonction des cas. Remarquons que le fait d'être divisible par 6 n'implique pas d'être divisible par 10 et réciproquement, le fait d'être divisible par 10 n'implique pas d'être divisible par 6. On doit donc bien considérer ces deux cas de figure indépendamment l'un de l'autre.

Proposition 1.21. *Si un entier n est divisible par 6 ou par 10, alors n est pair.*

Preuve. Supposons d'abord que n est un entier divisible par 6. Alors $n = 6k$, pour un entier k. On obtient que $n = 2 \cdot 3k$, donc que n est un nombre pair.

Supposons à présent que n est un entier divisible par 10. Alors $n = 10\ell$, pour un entier ℓ. On obtient que $n = 2 \cdot 5\ell$, donc que n est un nombre pair. □

1.3 Ensembles et relations

S'il existe un concept primordial en mathématiques, c'est bien celui d'ensemble. Pourtant, la définition d'un ensemble est difficile à formuler simplement, et bien des problèmes peuvent résulter d'une approche trop naïve de ce qu'on appelle la *théorie des ensembles*. Je vous conseille par exemple de vous renseigner un peu sur le paradoxe de Russel. Heureusement, il ne faut pas trop s'en inquiéter a priori, puisque ces notions naïves seront suffisantes pour la plupart des cours de votre cursus en sciences informatiques, et en particulier, pour

9. Puisque $\psi \vee \theta \equiv \theta \vee \psi$.

les cours de "Mathématiques pour l'informatique 1" et "Mathématiques pour l'informatique 2". Nous nous contenterons donc de considérer qu'un ensemble est bien défini dans les conditions suivantes.

Définition 1.22. Un ensemble est bien défini s'il est donné par une collection d'*éléments* qui satisfont une propriété *caractéristique* explicite, c'est-à-dire commune à tous les éléments de l'ensemble et à eux seuls. Autrement dit, un ensemble est bien défini lorsqu'on peut explicitement donner une propriété telle qu'un élément appartient à l'ensemble que nous voulons décrire *si et seulement si* il satisfait cette propriété. Si une telle propriété est notée P, on notera l'ensemble correspondant par $\{x\colon P(x)\}$, c'est à dire l'ensemble des x vérifiant la propriété P.

Pour être un peu plus précis, nous supposerons toujours que les éléments constituant nos ensembles font partie d'un *référentiel* (qui peut être, par exemple, les étudiants inscrits à ce cours, les nombres entiers, les nombres réels, les villes de Belgique, ...) et que la propriété *sélectionne* certains éléments de ce référentiel (par exemple, les étudiants inscrits au cours qui mesurent moins d'1m70, les nombres pairs, les nombres réels qui sont irrationnels, les villes de Flandre, ...). S'il n'y a pas d'ambiguïté sur le référentiel, on gardera la notation implicite $\{x\colon P(x)\}$. Si par contre, on souhaite distinguer deux référentiels, par exemple, ceux des entiers $\mathbb{Z}$ et des réels $\mathbb{R}$, on écrira $\{x \in \mathbb{Z}\colon x \leq 0\}$ et $\{x \in \mathbb{R}\colon x \leq 0\}$. Bien sûr, cette notation n'a de sens que lorsque la propriété (ici $x \leq 0$) est définie dans notre référentiel. Par exemple, la notation $\{x \in \mathbb{C}\colon x \leq 0\}$ n'a pas de sens ![10]

Nous introduisons maintenant quelques notations importantes, qui ont un sens précis. Ces signes sont peu nombreux et, tout comme les connecteurs logiques de la section précédente, doivent être utilisés à bon escient en toutes circonstances.

Définition 1.23. Soient A et B deux ensembles.

- Si x est un élément (de l'univers, le plus grand référentiel possible, et supposé connu de tous...), on écrit $x \in A$ pour signifier que x est un élément de A. On note aussi $x \notin A$ pour signifier que x n'est pas un élément de A.
- On note l'*inclusion* de A dans B par $A \subseteq B$. Ceci signifie que tout élément de A est aussi un élément de B. On dit que A est un *sous-ensemble* ou une *partie* de B.
- L'*égalité* de deux ensembles est bien définie. On écrit $A = B$ lorsque $A \subseteq B$ et $B \subseteq A$. Autrement dit, A et B sont égaux lorsque tout élément de A est aussi dans B et, inversement, tout élément de B appartient également à A.
- L'*union* de A et de B, notée $A \cup B$, est l'ensemble qui contient à la fois les éléments de A et de B. On a donc $A \cup B = \{x\colon x \in A \text{ ou } x \in B\}$.
- L'*intersection* de A et de B, notée $A \cap B$, est l'ensemble qui contient les éléments qui appartiennent à la fois à A et à B. On a donc $A \cap B = \{x\colon x \in A \text{ et } x \in B\}$.
- La *différence* de deux ensembles [11] A et B (on dit aussi "A moins B"), notée $A \backslash B$, est l'ensemble qui contient les éléments de A n'appartenant pas à B. On a donc $A \backslash B = \{x\colon x \in A \text{ et } x \notin B\}$.
- L'*ensemble des parties* [12] d'un ensemble A, noté $\mathcal{P}(A)$, est [13] l'ensemble de toutes les parties de A : $\mathcal{P}(A) = \{B\colon B \subseteq A\}$.
- On distingue un ensemble particulier, appelé l'*ensemble vide* et noté $\emptyset$. Il s'agit de l'ensemble qui ne contient pas d'élément.

Mentionnons ici quelques propriétés correspondant à des tautologies vues précédemment (excepté la dernière). En particulier, ces propriétés peuvent être démontrées en utilisant des tables de vérité.

10. Pourquoi ?
11. Attention que cette notion n'est pas symétrique.
12. *The power-set* en anglais.
13. sans suprise

Proposition 1.24. *Soit X un ensemble et soient $A, B, C \in \mathcal{P}(X)$. Alors*

1. $A \cap B = B \cap A$
2. $A \cup B = B \cup A$
3. $X \setminus (A \cap B) = (X \setminus A) \cup (X \setminus B)$
4. $X \setminus (A \cup B) = (X \setminus A) \cap (X \setminus B)$
5. $A \cap (B \cap C) = (A \cap B) \cap C$
6. $A \cup (B \cup C) = (A \cup B) \cup C$
7. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
8. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
9. $A \setminus B = A \cap (X \setminus B)$
10. *Si $A \subseteq B$, alors $\mathcal{P}(A) \subseteq \mathcal{P}(B)$.*

Preuve. Démontrons 8 de deux façons différentes.

La première consiste à montrer une égalité d'ensembles par *double inclusion*. Prenons tout d'abord un élément de l'ensemble de gauche : soit $x \in A \cup (B \cap C)$. Nous devons montrer que x appartient à l'ensemble de droite $(A \cup B) \cap (A \cup C)$. Nous savons que $x \in A$ ou $x \in B \cap C$. Nous allons procéder par disjonction des cas (voir la section précédente). Si $x \in A$, alors comme $A \subseteq A \cup B$, nous obtenons que $x \in A \cup B$. De la même façon en remplaçant B par C, nous obtenons que $x \in A \cup C$. Donc $x \in (A \cup B) \cap (A \cup C)$. Si maintenant $x \in B \cap C$, alors $x \in B$ et $x \in C$. Comme $B \subseteq A \cup B$ et $C \subseteq A \cup C$, nous obtenons que $x \in A \cup B$ et $x \in A \cup C$, et donc que $x \in (A \cup B) \cap (A \cup C)$. À ce stade, nous avons démontré l'inclusion $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

À présent, prenons un élément de l'ensemble de droite : soit $x \in (A \cup B) \cap (A \cup C)$ [14]. Remarquez que montrer $x \in A \cup (B \cap C)$ revient à montrer une alternative (voir la section précédente) : $x \in A$ ou $x \in B \cap C$. Supposons donc que $x \notin A$ et montrons qu'avec cette hypothèse supplémentaire, on doit avoir $x \in B \cap C$. Comme $x \in A \cup B$ et que $(A \cup B) \setminus A \subseteq B$, on obtient que $x \in B$. De la même façon en remplaçant B par C, on obtient que $x \in C$. On a donc bien obtenu que $x \in B \cap C$. Nous avons donc également démontré l'inclusion $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$. Ces deux inclusions simultanées nous donnent l'égalité désirée.

Voici une deuxième démonstration plus directe, qui utilise la logique propositionnelle. Étant donné un élément quelconque x, on considère trois variables propositionnelles a, b, c dont les valeurs de vérité représentent les différentes situations possibles : a (resp. b et c) prend la valeur de vérité 1 si $x \in A$ (resp. $x \in B$ et $x \in C$) et la valeur de vérité 0 si $x \notin A$ (resp. $x \notin B$ et $x \notin C$). Par définition de l'égalité de deux ensembles, il suffit [15] de vérifier que les tables de vérité correspondant aux deux propositions $x \in A \cup (B \cap C)$ et $x \in (A \cup B) \cap (A \cup C)$ sont identiques. Vu nos conventions, celle-ci sont représentées par $a \vee (b \wedge c)$ et $(a \vee b) \wedge (a \vee c)$ et nous savons déjà que ces deux propositions sont logiquement équivalentes. □

Diagrammes de Venn

Les *diagrammes de Venn* [16] sont utilisés pour représenter des ensembles et leurs relations. Vous avez probablement déjà tous utilisé cette technique, sans nécessairement lui avoir donné de nom. Un grand rectangle (ou à défaut, la feuille ou le tableau) représente votre

14. Il n'y a pas de problème à lui donner le même nom que dans la première partie dela preuve, ces parties devant être indépendantes. Ceci dit, rien ne vous empêche de l'appeler autrement non plus.
15. Détaillez, au besoin.
16. Ou plus communément, les patates.

référentiel. Chaque ensemble est représenté à l'intérieur du rectangle par une courbe fermée [17]. On indique un élément de l'ensemble par un point à l'intérieur de la zone délimitée par la courbe fermée représentant cet ensemble. Un diagramme de Venn représentant n ensembles doit toujours contenir 2^n zones. Ceci se traduit en logique propositionnelle par les 2^n lignes de la table de vérité d'une proposition constituée de n variables propositionnelles. En effet, chaque zone d'un diagramme de Venn de n ensembles doit correspond à une exactement une distribution des valeurs de vérité des n variables propositionnelles $x \in A_1, \ldots, x \in A_n$. En général, ceci est très difficile à réaliser pour de nombreux ensembles et en pratique, on n'utilise les diagrammes de Venn que pour représenter les relations de un, deux ou trois ensembles, parfois quatre [18].

Essayez par exemple de représenter la zone $A \cup (B \cap C)$. Votre diagramme doit vous permettre de visualiser l'égalité $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. Il est également formateur de représenter toutes les relations de la proposition 1.24 à l'aide de diagrammes de Venn.

Produits cartésiens, relations et fonctions

Nous concluons cette courte section sur les ensembles par trois définitions importantes. La première est celle du produit cartésien d'ensembles. Par exemple, la notation $\mathbb{R}^2$ désigne le produit cartésien de deux ensembles, qui sont dans ce cas deux copies du même ensemble, celui des nombres réels $\mathbb{R}$. Cette notation n'est pas toujours bien assimilée et pourtant, elle est omniprésente en mathématiques.

Définition 1.25. Le *produit cartésien* de deux ensembles A et B, noté $A \times B$, est l'ensemble des couples d'éléments dont la première composante appartient à A et la seconde composante appartient à B :

$$A \times B = \{(x, y) \colon x \in A \text{ et } y \in B\}.$$

Lorsque $A = B$, on écrit $A \times A = A^2$.

De façon plus générale, si $A_1, \ldots, A_n$ sont n ensembles, on définit leur produit cartésien $A_1 \times \cdots \times A_n$ comme l'ensemble des n-uples dont la i-ième composante appartient à A_i pour chaque indice $i \in \{1, \ldots, n\}$:

$$A_1 \times \cdots \times A_n = \{(x_1, \ldots, x_n) \colon x_1 \in A_1, \ldots x_n \in A_n\}.$$

Lorsque les n ensembles $A_1, \ldots, A_n$ sont les mêmes, on leur donne un nom commun, par exemple A, et on écrit $A_1 \times \cdots \times A_n = A^n$.

La deuxième est la définition d'une relation.

Définition 1.26. Soient A et B deux ensembles. Une *relation* R de A dans B est simplement une partie de $A \times B$. Autrement dit, une relation est un ensemble de couples d'éléments de A et de B. Lorsque $(x, y) \in R$, on dit que x est en relation avec y. On note le plus souvent $(x, y) \in R$ par xRy.

Il est important de remarquer qu'on peut avoir $(x, y) \in R$ sans pour autant avoir $(y, x) \in R$. Autrement dit, une relation R n'est pas forcément symétrique.

Le choix de la notation xRy peut paraître étonnant au premier abord. Mais pensez à la relation d'ordre $<$ sur les nombres réels par exemple. On peut en effet voir l'ordre $<$ comme une relation (au sens de notre définition) de $\mathbb{R}$ dans $\mathbb{R}$. Cette relation est exactement donnée

17. Une patate, donc.

18. C'est un excellent exercice que de réaliser un diagramme de Venn pour quatre ensembles. Celui-ci devra donc contenir exactement seize zones différentes.

par l'ensemble $\{(x, y) \in \mathbb{R}^2 : x < y\}$. Dans ce cas, on écrira bien plutôt $2 < 3$ que $(2, 3) \in <$. Remarquons que la relation $<$ est un exemple de relation non symétrique : en effet nous avons $2 < 3$ mais pas $3 < 2$.
Voici un autre exemple de relation, cette fois dans $\mathbb{Z}$. La relation "être multiple de" se traduit par la partie $\{(x, y) \in \mathbb{Z}^2 :$ il existe $m \in \mathbb{Z}$ tel que $y = mx\}$ de $\mathbb{Z}^2$. Cet ensemble contient le couple $(2, -6)$ mais pas le couple $(3, 10)$. Cette relation est-elle symétrique ? Qu'en est-il de la relation $\{(x, y) \in (\mathbb{Z}_0)^2 :$ il existe $m \in \mathbb{Q}$ tel que $y = mx\}$ (où la notation $\mathbb{Z}_0$ désigne l'ensemble $\mathbb{Z} \setminus \{0\}$).
Enfin, nous définissons les fonctions de A dans B comme des relations particulières.

Définition 1.27. Soient A et B des ensembles. Une *fonction* (ou *application*) f de A dans B est une relation de A dans B telle que pour tout $x \in A$, il existe un unique $y \in B$ tel que $(x, y) \in f$. Dans ce cas, on note $y = f(x)$, et la relation f est vue comme une "loi de transformation". On note également $f : A \to B$ pour signifier que f est une fonction. On dit également que A est le *domaine* de f et que B est le *but* (ou le *co-domaine*) de f. Enfin, on note également $f : A \to B$, $x \mapsto f(x)$ pour définir f.

Remarquons qu'avec cette définition, une fonction vient toujours avec son domaine et son but. On considérera donc les fonctions $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ et $g : [-1, 1] \to \mathbb{R}$, $x \mapsto x^2$ comme différentes. Dans ce cas, on note $g = f_{|[-1,1]}$ pour indiquer que g est la restriction de f au domaine $[-1, 1]$.

1.4 Suites et sommes

Définition 1.28. Une *suite de points d'un ensemble* A est une fonction $x : \mathbb{N} \to A$. Lorsqu'on manipule des suites, on utilise souvent la notation x_i plutôt que $x(i)$ pour désigner l'image du naturel i, appelée aussi *terme indicé par* i de la suite x. On trouve aussi souvent les notations $(x_i)_{i\in\mathbb{N}}$ ou $(x_i)_{i\geq 0}$ pour désigner une suite $x : \mathbb{N} \to A$.

Une suite de réels est alors simplement une fonction $x : \mathbb{N} \to \mathbb{R}$ (l'ensemble A choisi est l'ensemble $\mathbb{R}$).

Exemple 1.29. Voici quelques exemples de suites de nombres.

1. $(i)_{i\in\mathbb{N}} = (0, 1, 2, 3, 4 \ldots)$
2. $(i^2)_{i\in\mathbb{N}} = (0, 1, 4, 9, 16, \ldots)$
3. $(\frac{1}{i+1})_{i\in\mathbb{N}} = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$

Dans le troisième exemple, il est parfois plus commode de désigner cette suite par $(\frac{1}{i})_{i\geq 1}$. Cet abus de notation sera justifié lorsque nous aborderons la notion d'ensembles dénombrables.

Remarquons que les suites ne sont pas des ensembles, mais des ensembles ordonnés. Les termes d'une suite sont énumérés dans l'ordre, alors que ce n'est pas le cas pour les éléments d'un ensemble.

Définition 1.30. Soit une suite $(x_i)_{i\in\mathbb{N}}$ et soient $j, k \in \mathbb{N}$. Pour écrire la somme des termes consécutifs d'une suite en commençant par le terme indicé par j jusqu'au terme indicé par k, on peut écrire

$$x_j + x_{j+1} + \cdots + x_k,$$

en convenant que cette somme vaut 0 dans le cas où $j > k$. Remarquons que l'usage des points de suspension est inévitable dans cette écriture puisqu'on ne sait pas exactement combien de termes cette somme contient.

On introduit la notation suivante, plus compacte, pour désigner la même somme :

$$\sum_{i=j}^{k} x_i.$$

Le symbole $\sum$ est appelé le *signe sommatoire* et i est appelé l'*indice sommatoire*.

Exercice 1.31 (Distributivité)**.** Soient des suites réelles $(x_i)_{i\in\mathbb{N}}$ et $(y_i)_{i\in\mathbb{N}}$, soient $j, k \in \mathbb{N}$ et soient $a, b \in \mathbb{R}$. On a

$$\sum_{i=j}^{k}(ax_i + by_i) = a\sum_{i=j}^{k} x_i + b\sum_{i=j}^{k} y_i.$$

1.5 Quantificateurs

Le langage de la logique propositionnelle ne contient pas les symboles $\forall$ and $\exists$. Ces symboles, appelés *quantificateurs*, sont les ingrédients de base de la *logique du premier ordre*.

Définition 1.32. Le symbole $\forall$ se lit "pour tout" et est appelé le *quantificateur universel*. Le symbole $\exists$ se lit "il existe" et est appelé le *quantificateur existentiel*.

Commençons par considérer quelques exemples d'énoncés mathématiques contenant des quantificateurs. Il s'agit surtout de bien comprendre comment se comporte la négation de tels énoncés.
Par exemple, la phrase "toutes les personnes dans la salle ont moins de 30 ans" est une assertion. La négation de cette phrase est "il existe une personne dans la salle qui a au moins 30 ans". Formellement, on écrira

$$\neg\big((\forall x)P(x)\big) \equiv (\exists x)\neg P(x).$$

Considérons maintenant la phrase "il y a des enfants qui ne sont pas sages"[19]. Nier cette phrase revient à dire "tous les enfants sont sages". Formellement, on écrira

$$\neg\big((\exists x)P(x)\big) \equiv (\forall x)\neg P(x).$$

Dans ces écritures, x n'est pas une variable propositionnelle mais on suppose que x est une variable représentant un élément du domaine D de la structure dans laquelle est définie le prédicat P, et on met ce fait en évidence en écrivant $P(x)$.
La logique du premier ordre est aussi appelée *calcul des prédicats*.

Définition 1.33. Un *prédicat* sur un domaine D est une partie de D^p pour un certain naturel p. Lorsqu'on veut être précis, on parle de prédicat d'*arité* p. Si P est un prédicat d'arité p sur D et si $(x_1, \ldots, x_p) \in D^p$, on dit que $(x_1, \ldots, x_p)$ *vérifie* P lorsque $(x_1, \ldots, x_p) \in P$, ce qu'on l'on écrit également $P(x_1, \ldots, x_p)$.

On peut également voir un prédicat comme une fonction de D^p à valeurs dans $\{0, 1\}$, ou encore $\{\text{vrai}, \text{faux}\}$. On a $P(x_1, \ldots, x_p) = 1$ si et seulement si $(x_1, \ldots, x_p) \in P$. On écrit alors simplement $P(x_1, \ldots, x_p)$ pour signifier que $P(x_1, \ldots, x_p) = 1$.
Le terme "premier ordre" vient du fait qu'on a le droit de quantifier uniquement sur les variables du domaine D et non sur les prédicats eux-mêmes.

19. Si tant est que le fait d'être sage soit bien défini...

Dans nos deux exemples, le domaine était précisé dans le premier cas (l'ensemble des personnes de la salle) et non précisé dans le deuxième (à défaut de précision, on considérera les enfants du monde entier). En mathématique, on écrira par exemple $\forall x \in \mathbb{R}, (|x| \leq 1) \wedge (x \geq \sqrt{|x|})$ [20] lorsque l'on veut spécifier que l'on considère uniquement des x réels.

Définition 1.34. Lorsqu'une variable d'une formule n'est pas quantifiée (ou *liée* par un quantificateur), on dit que c'est une variable *libre*. Une formule qui ne contient aucune variable libre est appelée une *formule close*. Les variables quantifiées sont dites *liées* ou *muettes*.

On peut modifier les variables liées (partout dans une même formule) sans en changer le sens. Ainsi, lorsque P est un prédicat qui contient une seule variable libre x, les formules $\forall x P(x)$ et $\forall y P(y)$ sont équivalentes. Remarquez que, dans ce cas, les formules $\forall x P(x)$ et $\forall y P(y)$ sont closes.

Contre-exemple

L'équivalence logique $\neg\big((\forall x)P(x)\big) \equiv (\exists x)\neg P(x)$ conduit à la technique de démonstration par le contre-exemple. Plus précisément, lorsque l'on souhaite prouver qu'une formule close du type $\forall x P(x)$ est fausse, il suffit d'exhiber un contre-exemple, c'est-à-dire un x tel que $\neg P(x)$ est vrai. Ainsi, pour démontrer que l'affirmation que "tout réel x vérifie l'inégalité $x^2 + 3x + 1 \geq 0$" est fausse, il suffit de trouver un réel x tel que cette inégalité n'est pas vérifiée. Par exemple, le réel -1 est tel que $(-1)^2 + 3(-1) + 1 = 1 - 3 + 1 = -1 < 0$ et est donc un contre-exemple de l'affirmation $\forall x \in \mathbb{R}, x^2 + 3x + 1 \geq 0$.

Voici un autre exemple, rencontré au cours "Mathématique". On dit qu'une fonction $f\colon \mathbb{R} \to \mathbb{R}$ est *paire* lorsque pour tout $x \in \mathbb{R}$, on a $f(-x) = f(x)$. C'est par exemple le cas de la fonction $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$. Pour montrer que la fonction $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$ n'est pas paire, on montrera que la formule "pour tout $x \in \mathbb{R}$, on a $f(-x) = f(x)$" est fausse pour cette fonction f particulière, c'est-à-dire que l'affirmation "pour tout $x \in \mathbb{R}$, on a $(-x)^3 = x^3$" est fausse. Il suffit donc de donner un contre-exemple, prouvant ainsi qu'il existe $x \in \mathbb{R}$ tel que $(-x)^3 \neq x^3$. Par exemple, le réel -2 est tel que $(-2)^3 = -8 \neq 2^3 = 8$. Dans ce cas précis, on notera que n'importe quel autre réel négatif aurait également été un contre-exemple.

Ordre des quantificateurs

Changer l'ordre des quantificateurs modifie en général le sens d'une formule. Lorsqu'on écrit $\forall x\, \exists y\, P(x, y)$, ce qui se lit "pour tout x, il existe y tel que $P(x, y)$", le choix de y dépend de celui de x. Inversement, lorsqu'on écrit $\exists y\, \forall x\, P(x, y)$, ce qui se lit "il existe y tel que pour tout x, la propriété $P(x, y)$ a lieu, le choix de y ne dépend pas de celui de x. La propriété est *plus forte* dans le second cas : on a toujours $\exists y\, \forall x\, P(x, y) \implies \forall x\, \exists y\, P(x, y)$.

Voici un exemple parlant. Chaque personne est né d'une mère. On peut donc dire que *pour tout être humain, il existe une femme qui est sa mère*. Inverser l'ordre des quantificateurs revient à dire qu'*il existe une femme telle que pour tout être humain, cette femme est sa mère*. Autrement dit, en bon français, *il existe une femme qui est la mère de tous les êtres humains*. On comprend bien que ces deux phrases n'ont pas le même sens !

Néanmoins, il est utile de remarquer que deux quantificateurs universels successifs commutent toujours entre eux. Il en est de même pour deux quantificateurs existentiels successifs. Ainsi, on a $\forall x\, \forall y\, P(x, y) \equiv \forall y\, \forall x\, P(x, y)$ et $\exists x\, \exists y\, P(x, y) \equiv \exists y\, \exists x\, P(x, y)$.

20. Ceci n'est pas un théorème ! Pourquoi ?

1.6 La démonstration par récurrence

L'ensemble des naturels est noté $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$. Nous utiliserons également la notation $\mathbb{N}_0$ pour désigner l'ensemble des entiers positifs : $\mathbb{N}_0 = \{1, 2, 3, \ldots\}$. Dans les sections précédentes, nous avons détaillé différentes techniques de démonstration basées sur des tautologies. Ici, nous rappelons le *principe d'induction*, ou de *démonstration par récurrence*.

Théorème 1.35 (Principe de récurrence)**.** *Soient P un prédicat défini sur $\mathbb{N}$ et $m \in \mathbb{N}$ tels que les deux conditions suivantes soient vérifiées.*

1. *L'entier m vérifie P.*
2. *Pour tout entier $n \geq m$, si n vérifie P, alors $n+1$ vérifie P.*

Alors tous les entiers $n \geq m$ vérifient P.

Dans ce théorème, la condition 1 est appelée le *cas de base* ou l'*initialisation* et la condition 2 est appelée l'*induction* ou le *pas de récurrence*.
En guise d'illustration, voici un exemple de propriété se démontrant par récurrence. C'est aussi l'occasion de se familiariser avec l'utilisation du signe somme Σ.

Proposition 1.36. *Pour tout $n \in \mathbb{N}_0$, on a*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \tag{1.2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}. \tag{1.3}$$

Preuve. Nous procédons par récurrence sur n. Démontrons les cas de base. Ici, il s'agit de $n = 1$. Dans ce cas, on a d'une part $\sum_{i=1}^{1} i = 1$ et $\sum_{i=1}^{1} i^2 = 1$ et d'autre part $\frac{1(1+1)}{2} = 1$ et $\frac{1(1+1)(2\cdot 1+1)}{6} = 1$. Le cas de base est donc vérifié.
Nous montrons à présent l'induction elle-même. Soit $n \geq 1$ quelconque, mais fixé. Supposons que les égalités (1.2) et (1.3) soient vérifiées. C'est ce que l'on appelle l'*hypothèse de récurrence*. Nous devons montrer que $\sum_{i=1}^{n+1} i = \frac{(n+1)(n+2)}{2}$ et $\sum_{i=1}^{n+1} i^2 = \frac{(n+1)(n+2)(2(n+1)+1)}{6}$.
Allons-y. Pour la première égalité, nous avons successivement

$$\begin{aligned}
\sum_{i=1}^{n+1} i &= \sum_{i=0}^{n} i + (n+1) \\
&= \frac{n(n+1)}{2} + (n+1) \\
&= \frac{n(n+1) + 2(n+1)}{2} \\
&= \frac{(n+1)(n+2)}{2},
\end{aligned}$$

où l'on a utilisé l'hypothèse de récurrence (1.2) à la deuxième ligne. Pour la deuxième égalité, en développant le membre de gauche, nous obtenons d'une part

$$\begin{aligned}
\sum_{i=1}^{n+1} i^2 &= \sum_{i=0}^{n} i^2 + (n+1)^2 \\
&= \frac{n(n+1)(2n+1)}{6} + (n+1)^2 \\
&= \frac{n(n+1)(2n+1) + 6(n+1)^2}{6}
\end{aligned}$$

$$= \frac{(n+1)\big(n(2n+1)+6(n+1)\big)}{6}$$
$$= \frac{(n+1)(2n^2+7n+6)}{6},$$

où l'on a utilisé l'hypothèse de récurrence (1.3) à la deuxième ligne. D'autre part, en développant le membre de droite, nous avons

$$\frac{(n+1)(n+2)(2(n+1)+1)}{6} = \frac{(n+1)(n+2)(2n+3)}{6} = \frac{(n+1)(2n^2+7n+6)}{6}.$$

□

Remarque 1.37. Les nombres $\frac{n(n+1)}{2}$ sont appelés les nombres triangulaires. Nous allons comprendre pourquoi avec le théorème 1.50.

Vous utilisez ce principe d'induction dans votre cours "Mathématique" pour démontrer le binôme de Newton, la formule de Leibniz et la formule intégrale de Taylor [21].

On parle aussi parfois de *récurrence forte*. Il s'agit en fait d'un cas particulier du principe de récurrence classique où l'hypothèse utilisée pour démontrer le pas de récurrence est plus forte que dans le cas d'une récurrence classique. Pour démontrer que $n+1$ vérifie P, on suppose ici que i vérifie P pour tout $i \leq n$, et non pas seulement que n vérifie P.

Théorème 1.38 (Principe de récurrence forte)**.** *Soient P un prédicat défini sur $\mathbb{N}$ et $m \in \mathbb{N}$ tels que les deux conditions suivantes soient vérifiées.*

1. *L'entier m vérifie P.*
2. *Pour tout entier $n > m$, si tout entier $i \in \{m, m+1 \ldots, n-1\}$ vérifie P, alors n vérifie P.*

Alors tous les entiers $n \geq m$ vérifient P.

Preuve. Soit Q le prédicat défini sur $\mathbb{N}$ par [22]

$$n \text{ vérifie } Q \iff (n \geq m \text{ et tout } i \in \{m, m+1, \ldots, n\} \text{ vérifie } P).$$

Nous allons appliquer le principe de récurrence classique sur le prédicat Q, c'est-à-dire le théorème 1.35, pour montrer que tous les entiers $n \geq m$ vérifient Q. Tout d'abord, puisque m vérifie P par hypothèse, on a aussi que m vérifie Q (par définition de Q). Soit maintenant un entier $n \geq m$ qui vérifie Q. Alors par définition de Q, tout $i \in \{m, m+1, \ldots, n\}$ vérifie P. Par hypothèse sur P, on obtient que $n+1$ vérifie P. Ainsi, nous avons bien que $n+1 \geq m$ et que tout $i \in \{m, m+1, \ldots, n+1\}$ vérifie P, c'est-à-dire que $n+1$ vérifie Q. En utilisant le théorème 1.35, nous obtenons que tout $n \geq m$ vérifie Q. En particulier, tout $n \geq m$ vérifie P également. □

Le théorème fondamental de l'arithmétique est un exemple d'application du principe de récurrence forte.

Pour le démontrer, nous admettons momentanément le lemme d'Euclide, dont la démonstration sera donnée dans la section 1.8 (voir corollaire 1.68).

Lemme 1.39 (Lemme d'Euclide)**.** *Soient a et b des nombres entiers. Si un nombre premier p divise ab, alors p divise a ou b.*

21. Voir les notes du cours "Mathématique" de F. Bastin, année académique 2017-2018.
22. En notation ensembliste, on a $Q = \{n \geq m \colon \forall i \in \{m, m+1, \ldots, n\},\ i \in P\}$.

Théorème 1.40 (Théorème fondamental de l'arithmétique)**.** *Tout nombre entier $n \geq 2$ admet une unique décomposition (à l'ordre des facteurs près) de la forme*

$$n = p_1^{i_1} p_2^{i_1} \cdots p_k^{i_k} \tag{1.4}$$

où $k, i_1, i_2, \ldots, i_k \in \mathbb{N}_0$ et $p_1, p_2, \ldots, p_k$ sont des nombres premiers deux à deux distincts.

Preuve. Nous montrons tout d'abord l'existence de la décomposition, c'est-à-dire que tout nombre entier $n \geq 2$ admet une telle décomposition. Procédons par récurrence (forte) sur n. Pour $n = 2$, il suffit de prendre $k = 1$, $p_1 = 2$ et $i_1 = 1$. Supposons à présent que $n > 2$ et que tout nombre entier m compris entre 2 et $n - 1$ admette une telle décomposition. Soit p le plus petit diviseur de n compris entre 2 et n. Par hypothèse de récurrence appliquée à $\frac{n}{p}$, on obtient que $\frac{n}{p}$ admet une décomposition de la forme (1.4). Pour conclure, il suffit donc de montrer que p est un nombre premier. Ceci est clair car s'il existait un diviseur d de p compris entre 2 et $p - 1$, cet entier d diviserait aussi n, contredisant le fait que p soit le plus petit tel diviseur de n.

Nous montrons maintenant l'unicité de la décomposition. Procédons par récurrence (forte) sur n. Pour $n = 2$, le choix $k = 1$, $p_1 = 2$ et $i_1 = 1$ est le seul possible. Supposons à présent que $n > 2$ et que tout nombre entier m compris entre 2 et $n - 1$ admette au plus une telle décomposition. Supposons que n admette deux décomposition de la forme (1.4), c'est-à-dire que

$$n = p_1^{i_1} p_2^{i_1} \cdots p_k^{i_k} = q_1^{j_1} q_2^{j_1} \cdots q_\ell^{j_\ell}$$

où $k, \ell \in \mathbb{N}_0$, $p_1, \ldots, p_k, q_1, \ldots, q_\ell$ sont des nombres premiers et $i_1, \ldots, i_k, j_1, \ldots, j_\ell \in \mathbb{N}_0$. Alors p_1 divise le produit $q_1^{j_1} q_2^{j_1} \cdots q_\ell^{j_\ell}$. Comme p_1 est premier, par le lemme d'Euclide, il doit exister un indice s tel que $p_1 = q_s$. D'où

$$\frac{n}{p_1} = p_1^{i_1 - 1} p_2^{i_1} \cdots p_k^{i_k} = q_1^{j_1} q_2^{j_1} \cdots q_s^{j_s - 1} \cdots q_\ell^{j_\ell}.$$

Par hypothèse de récurrence appliquée à $\frac{n}{p_1}$, ces deux décompositions sont identiques et donc les deux décompositions de n aussi. □

La décomposition donnée par le théorème fondamental de l'arithmétique est appelée la *décomposition en facteurs premiers.*

1.7 Ensembles dénombrables

En informatique et en mathématiques discrètes, on travaille le plus souvent avec des ensembles dénombrables (voire finis), avec pour paradigme l'ensemble des naturels $\mathbb{N}$.

Injection, surjection, bijection, fonction réciproque

Avant de donner la définition d'ensemble dénombrable, nous rappelons celles d'injection, de surjection et de bijection.

Définition 1.41. Soit f une fonction de A dans B.

1. On dit que la fonction f est *injective* lorsque pour tous $a, a' \in A$, on a $a \neq a' \Rightarrow f(a) \neq f(a')$. Dans ce cas, on dit que f est une *injection* de A dans B.
2. On dit que la fonction f est *surjective* lorsque pour tous $b \in B$, il existe $a \in A$ tel que $b = f(a)$ [23]. Dans ce cas, on dit que f est une *surjection* de A dans B.

23. Ou de façon équivalente tel que $(a, b) \in f$. Voir la définition que nous avons donnée d'une fonction page 17.

3. On dit que la fonction f est *bijective* lorsqu'elle est à la fois injective et surjective. Dans ce cas, on dit que f est une *bijection* de A dans B.

Lorsque $f\colon A \to B$ est une bijection, alors f admet une fonction réciproque, qui est elle-même une bijection.

Définition 1.42. Soit $f\colon A \to B$ une bijection. Alors la *fonction réciproque* (ou *fonction inverse*) de f est la fonction $f^{-1}\colon B \to A$ qui à $b \in B$ associe l'unique $a \in A$ tel que $f(a) = b$. Autrement dit, $f^{-1}(b) = a$ si et seulement si $b = f(a)$.

Proposition 1.43. *Si f est une bijection de A dans B, alors f^{-1} est une bijection de B dans A.*

Preuve. Ce découle du fait que, pour tout $a \in A$ et tout $b \in B$, on ait $f^{-1}(f(a)) = a$ et $f(f^{-1}(b)) = b$. □

Ceci justifie qu'on puisse parler d'*ensembles en bijection* puisque si f est une bijection de A dans B, alors f^{-1} est une bijection de B dans A.
Rappelons la définition d'une fonction composée.

Définition 1.44. La fonction composée de $f\colon A \to B$ et $g\colon B \to C$ est la fonction $g \circ f\colon A \to C$, $a \mapsto g(f(a))$.

Le résultat suivant est souvent très utile.

Proposition 1.45. *Soient des fonctions $f\colon A \to B$ et $g\colon B \to C$.*

1. *Si f et g sont des injections, alors $g \circ f\colon A \to C$ est une injection.*
2. *Si f et g sont des surjections, alors $g \circ f\colon A \to C$ est une surjection.*
3. *Si f et g sont des bijections, alors $g \circ f\colon A \to C$ est une bijection.*

Preuve. 1. Supposons que f et g sont des injections. Soient $a, a' \in A$ tels que $a \neq a'$. Alors $f(a) \neq f(a')$ puisque f est une injection. Comme g est aussi une injection, on obtient bien que $g \circ f(a) = g(f(a)) \neq g(f(a')) = g \circ f(a')$. Ainsi, $g \circ f$ est bien une injection.
2. Supposons que f et g sont des surjections. Soit $c \in C$. Puisque g est une surjection, il existe $b \in B$ tel que $g(b) = c$. Ensuite, puisque f est une surjection également, il existe $a \in A$ tel que $f(a) = b$. On obtient que $g \circ f(a) = g(f(a)) = g(b) = c$. Ceci montre que $g \circ f$ est une surjection.
3. Supposons que f et g sont des bijections. La fonction $g \circ f$ est une injection par le point 1 et une surjection par le point 2, donc une bijection. □

Ensembles dénombrables : définition et exemples

Définition 1.46. Un ensemble A est *dénombrable* s'il existe une injection de A dans $\mathbb{N}$.

En particulier, $\mathbb{N}$ est dénombrable[24]. Montrons que l'ensemble $2\,\mathbb{N} = \{2n\colon n \in \mathbb{N}\}$ des naturels pairs et l'ensemble $2\,\mathbb{N}+1 = \{2n+1\colon n \in \mathbb{N}\}$ des naturels impairs sont tous deux en bijection avec $\mathbb{N}$.

Proposition 1.47. *Les ensembles $2\,\mathbb{N}$ et $2\,\mathbb{N}+1$ sont dénombrables.*

Preuve. Il suffit de vérifier que les fonctions $f\colon 2\,\mathbb{N} \to \mathbb{N}$, $n \mapsto \frac{n}{2}$ et $g\colon 2\,\mathbb{N}+1 \to \mathbb{N}$, $n \mapsto \frac{n-1}{2}$ sont injectives[25]. □

24. Pourquoi ?
25. Il s'agit en fait de bijections.

La proposition précédente est en fait un cas particulier du résultat suivant.

Proposition 1.48. *Toute partie d'un ensemble dénombrable est dénombrable.*

Preuve. Soit A un ensemble dénombrable et soit $B \subseteq A$. Par définition, il existe une injection $f\colon A \to \mathbb{N}$. La fonction $i\colon B \to A$, $x \mapsto x$ étant injective, on obtient par la proposition 1.45 que la fonction $f \circ i\colon B \to \mathbb{N}$ est une injection de B dans $\mathbb{N}$, ce qui suffit. □

Proposition 1.49. *L'ensemble $\mathbb{Z}$ est dénombrable.*

Preuve. Il suffit de vérifier que la fonction

$$f\colon \mathbb{Z} \to \mathbb{N},\ z \mapsto \begin{cases} 2z-1 & \text{si } z > 0 \\ -2z & \text{si } z \leq 0 \end{cases}$$

est une injection [26]. □

Théorème 1.50. *L'ensemble $\mathbb{N} \times \mathbb{N}$ est dénombrable. Plus précisément, la fonction*

$$\pi\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N},\ (m,n) \mapsto \frac{(m+n)(m+n+1)}{2} + n$$

est une bijection.

Preuve. Montrons que π est une injection. Soient (m,n) et (m',n') deux éléments distincts de $\mathbb{N} \times \mathbb{N}$. Nous devons prouver que $\pi(m,n) \neq \pi(m',n')$.
Soit $T\colon \mathbb{N} \to \mathbb{N},\ k \mapsto \sum_{i=1}^{k} i$. Pour tout $k \in \mathbb{N}$, on a $T(k+1) = T(k) + k + 1$. La fonction T est donc strictement croissante, c'est-à-dire que pour tous $k, \ell \in \mathbb{N}$ tels que $k < \ell$, on a $T(k) < T(\ell)$.
Avec la définition de π et la proposition 1.36, on vérifie facilement que $T(m+n) \leq \pi(m,n) < T(m+n+1)$ et $T(m'+n') \leq \pi(m',n') < T(m'+n'+1)$.
Si $m+n < m'+n'$, alors $m+n+1 \leq m'+n'$ et en utilisant la croissance de T, on obtient $\pi(m,n) < T(m+n+1) \leq T(m'+n') \leq \pi(m',n')$, prouvant que $\pi(m,n) < \pi(m',n')$. Si $m+n > m'+n'$, on obtient de la même manière que $\pi(m,n) > \pi(m',n')$. Considérons maintenant le cas où $m+n = m'+n'$. Dans ce cas, comme on a supposé $(m,n) \neq (m',n')$, on sait que $m \neq m'$ *et* $n \neq n'$ [27]. Ainsi, $\pi(m,n) = T(m+n) + n \neq T(m+n) + n' = T(m'+n') + n' = \pi(m',n')$. Dans tous les cas, on a bien $\pi(m,n) \neq \pi(m',n')$.
Montrons à présent que π est une surjection. Soit $a \in \mathbb{N}$. Nous devons montrer qu'il existe $(m,n) \in \mathbb{N} \times \mathbb{N}$ tel que $\pi(m,n) = a$. Soit k l'unique entier tel que $T(k) \leq a < T(k+1)$. Alors $(m,n) = (k-a+T(k), a-T(k))$ convient. En effet, puisque $a - T(k) < T(k+1) - T(k) = k+1$, on obtient que $m = k - a + T(k) \geq 0$ et

$$\pi(m,n) = T(m+n) + n = T(k) + a - T(k) = a.$$

□

Une autre façon d'obtenir que $\mathbb{N} \times \mathbb{N}$ est dénombrable est donnée par le résultat suivant.

Théorème 1.51. *La fonction*

$$\rho\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N},\ (m,n) \mapsto 2^m(2n+1) - 1$$

est une bijection.

26. Il s'agit en fait d'une bijection.

27. Attention : $(m,n) \neq (m',n')$ est équivalent à $m \neq m'$ *ou* $n \neq n'$. Mais ici, on a l'information supplémentaire que $m+n = m'+n'$.

Preuve. Montrons que ρ est une injection. Soient (m,n) et (m',n') deux éléments de $\mathbb{N}\times\mathbb{N}$ tels que $\rho(m,n)\neq\rho(m',n')$. On a donc $2^m(2n+1)-1=2^{m'}(2n'+1)-1$, et donc $2^m(2n+1)=2^{m'}(2n'+1)$. Si $m<m'$, alors $2n+1=2^{m'-m}(2n'+1)$, ce qui est impossible puisque le membre de gauche est un nombre impair et celui de droite un nombre pair. De manière symétrique, on ne peut pas avoir $m>m'$ non plus. D'où $m=m'$ et $2n+1=2n'+1$. Ainsi $n=n'$, et on a bien $(m,n)=(m',n')$.
Montrons que ρ est une surjection. Soit $a\in\mathbb{N}$. Soit $k\in\mathbb{N}$ tel que 2^k soit la plus grande puissance de 2 qui divise $a+1$. Alors $\frac{a+1}{2^k}$ est impair. Il existe donc $\ell\in\mathbb{N}$ tel que $\frac{a+1}{2^k}=2\ell+1$. On a donc $\rho(k,\ell)=2^k(2\ell+1)-1=2^k\left(\frac{a+1}{2^k}\right)-1=a+1-1=a$, ce qui suffit. □

Proposition 1.52. *Si A et B sont dénombrables, alors $A\times B$ est dénombrable.*

Preuve. Supposons que A et B soient dénombrables. Par hypothèse, il existe des injections $f\colon A\to\mathbb{N}$ et $g\colon B\to\mathbb{N}$.
Montrons que la fonction $h\colon A\times B\to\mathbb{N}\times\mathbb{N}$, $(a,b)\mapsto(f(a),g(b))$ est injective. Soient (a,b) et (a',b') des éléments de $A\times B$ tels que $h(a,b)=h(a',b')$. Alors, par définition de h, on a $f(a)=f(a')$ et $g(b)=g(b')$. Comme f et g sont des injections, on obtient que $a=a'$ et $b=b'$, et donc $(a,b)=(a',b')$, montrant que la fonction h est bien injective.
Au vu de la proposition 1.45 et du théorème 1.50, la fonction composée $\pi\circ h\colon A\times B\to\mathbb{N}$ est injective. Ainsi, $A\times B$ est dénombrable. □

Théorème 1.53. *L'ensemble $\mathbb{Q}$ est dénombrable.*

Preuve. Il suffit de montrer qu'il existe une injection de $\mathbb{Q}$ dans $\mathbb{Z}\times\mathbb{N}_0$ (justifiez pourquoi c'est suffisant). Ceci découle du fait que tout rationnel r non nul s'écrit de façon unique sous la forme $r=\frac{p}{q}$ où $(p,q)\in\mathbb{Z}\times\mathbb{N}_0$ et $\operatorname{pgcd}(p,q)=1$. □

Théorème 1.54. *L'ensemble*

$$\{0,1\}^{\mathbb{N}}=\left\{(a_n)_{n\in\mathbb{N}}\colon \forall n\in\mathbb{N},\, a_n\in\{0,1\}\right\}$$

des suites infinies de 0 et de 1 est non dénombrable.

Preuve. Nous allons utiliser le célèbre argument de la diagonale, aussi dit *de Cantor*. Nous procédons par l'absurde et nous supposons que l'ensemble $\{0,1\}^{\mathbb{N}}$ soit dénombrable. Il existe donc une injection $f\colon\{0,1\}^{\mathbb{N}}\to\mathbb{N}$. Considérons maintenant la suite particulière $(c_k)_{k\in\mathbb{N}}$ de $\{0,1\}^{\mathbb{N}}$ définie par $c_k=1-a_k$ s'il existe[28] une suite $(a_n)_{n\in\mathbb{N}}$ telle que $f((a_n)_{n\in\mathbb{N}})=k$ et $c_k=0$ sinon. Notons à présent $j=f((c_k)_{k\in\mathbb{N}})$. Mais alors on obtient que $c_j=1-c_j$, une absurdité. □

Mentionnons enfin les deux résultats suivants (sans preuve).

Fait 1.55. *L'ensemble $\mathbb{R}$ est non dénombrable.*

Fait 1.56. *L'ensemble $\mathcal{P}(\mathbb{N})$ des parties de $\mathbb{N}$ est non dénombrable.*

Suites sur des ensembles dénombrables et sommes à multi-indices

On peut étendre la définition de suite à n'importe quel ensemble dénombrable. En particulier une suite $x\colon\mathbb{N}\times\mathbb{N}\to A$ est une suite indicée par des couples de naturels plutôt que par des naturels. On note $x_{i,j}$ l'image du couple (i,j), appelée le terme de la suite x indicé

28. Notez bien que si une telle suite existe, alors elle est unique puisque f est une injection.

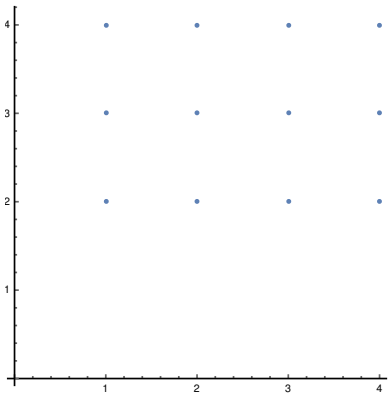

FIGURE 1.1 – Représentation de l'ensemble $\{1,2,3,4\} \times \{2,3,4\}$

par (i,j). On trouve aussi souvent des sommes de termes de telles suites. Par exemple, on a

$$\begin{aligned}\sum_{i=1}^{4}\sum_{j=2}^{4} x_{i,j} &= \sum_{i=1}^{4}(x_{i,2}+x_{i,3}+x_{i,4}) \\ &= (x_{1,2}+x_{1,3}+x_{1,4})+(x_{2,2}+x_{2,3}+x_{2,4}) \\ &\quad +(x_{3,2}+x_{3,3}+x_{3,4})+(x_{4,2}+x_{4,3}+x_{4,4}).\end{aligned}$$

On somme donc sur tous les couples d'indices (i,j) appartenant à l'ensemble $\{1,2,3,4\} \times \{2,3,4\}$. Cet ensemble est représenté dans la figure 1.1. Dans le calcul précédent, on a d'abord développé la somme sur l'indice j, et ensuite sur l'indice i. Remarquons qu'on aurait pu faire le calcul dans l'autre sens :

$$\begin{aligned}\sum_{i=1}^{4}\sum_{j=2}^{4} x_{i,j} &= \sum_{j=2}^{4} x_{1,j} + \sum_{j=2}^{4} x_{2,j} + \sum_{j=2}^{4} x_{3,j} + \sum_{j=2}^{4} x_{4,j} \\ &= (x_{1,2}+x_{1,3}+x_{1,4})+(x_{2,2}+x_{2,3}+x_{2,4}) \\ &\quad +(x_{3,2}+x_{3,3}+x_{3,4})+(x_{4,2}+x_{4,3}+x_{4,4}).\end{aligned}$$

Remarquons aussi que puisque la somme est commutative, on a aussi

$$\begin{aligned}\sum_{j=2}^{4}\sum_{i=1}^{4} x_{i,j} &= \sum_{j=2}^{4}(x_{1,j}+x_{2,j}+x_{3,j}+x_{4,j}) \\ &= (x_{1,2}+x_{2,2}+x_{3,2}+x_{4,2})+(x_{1,3}+x_{2,3}+x_{3,3}+x_{4,3}) \\ &\quad +(x_{1,4}+x_{2,4}+x_{3,4}+x_{4,4}) \\ &= \sum_{i=1}^{4}\sum_{j=2}^{4} x_{i,j}.\end{aligned}$$

Il se peut aussi que la deuxième somme dépende de la première :

$$\begin{aligned}\sum_{i=2}^{4}\sum_{j=2}^{i} x_{i,j} &= \sum_{j=2}^{2} x_{2,j} + \sum_{j=2}^{3} x_{3,j} + \sum_{j=2}^{4} x_{4,j} \\ &= x_{2,2} + (x_{3,2}+x_{3,3}) + (x_{4,2}+x_{4,3}+x_{4,4}).\end{aligned}$$

Dans ce cas, si l'on veut permuter les sommes, il faut faire attention ! En effet, il faut sommer sur tous les couples d'indices appartenant à l'ensemble

$$\{(i,j) \in \mathbb{N}\times\mathbb{N} \colon 2 \le i \le 4 \text{ et } 2 \le j \le i\} = \{(2,2),(3,2),(3,3),(4,2),(4,3),(4,4)\}$$

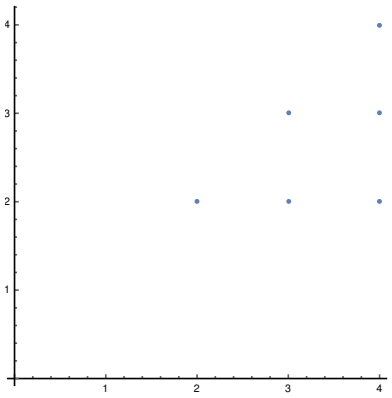

FIGURE 1.2 – Représentation de l'ensemble $\{(i,j) \in \mathbb{N} \times \mathbb{N} \colon 2 \leq i \leq 4 \text{ et } 2 \leq j \leq i\}$

et sur aucun autre couple. Cet ensemble est représenté dans la figure 1.2. On a l'égalité d'ensembles

$$\{(i,j) \in \mathbb{N} \times \mathbb{N} \colon 2 \leq i \leq 4 \text{ et } 2 \leq j \leq i\} = \{(i,j) \in \mathbb{N} \times \mathbb{N} \colon j \leq i \leq 4 \text{ et } 2 \leq j \leq 4\}.$$

Dans la première description, on énumère les éléments par colonnes :
- colonne $1 = \{(2,2)\}$: pour $i = 2$, on fait varier j entre 2 et 2
- colonne $2 = \{(3,2),(3,3)\}$: pour $i = 3$, on fait varier j entre 2 et 3
- colonne $3 = \{(4,2),(4,3),(4,4)\}$: pour $i = 4$, on fait varier j entre 2 et 4.

Dans la deuxième description, on énumère les éléments par lignes :
- ligne $1 = \{(2,2),(3,2),(4,2)\}$: pour $j = 2$, on fait varier i entre 2 et 4
- ligne $2 = \{(3,3),(4,3)\}$: pour $j = 3$, on fait varier i entre 3 et 4
- ligne $3 = \{(4,4)\}$: pour $j = 4$, on fait varier i entre 4 et 4.

On a donc l'égalité de somme suivante :

$$\sum_{i=2}^{4} \sum_{j=2}^{i} x_{i,j} = \sum_{j=2}^{4} \sum_{i=j}^{4} x_{i,j}.$$

Ces remarques de permutations de sommes sont vraies en général.

Proposition 1.57 (Permutation du signe somme lorsque la deuxième somme de dépend pas de la première)**.** *Soit une suite réelle* $(x_{i,j})_{i,j\in\mathbb{N}}$ *et soient* $I_1, I_2, J_1, J_2 \in \mathbb{N}$*. On a*

$$\sum_{i=I_1}^{I_2} \sum_{j=J_1}^{J_2} x_{i,j} = \sum_{j=J_1}^{J_2} \sum_{i=I_1}^{I_2} x_{i,j}.$$

Proposition 1.58 (Permutation du signe somme lorsque la deuxième somme dépend de la première)**.** *Soit une suite réelle* $(x_{i,j})_{i,j\in\mathbb{N}}$ *et soient* $I_1, I_2, J \in \mathbb{N}$*. On a*

$$\sum_{i=I_1}^{I_2} \sum_{j=J}^{i} x_{i,j} = \sum_{j=J}^{I_2} \sum_{i=\max\{I_1,j\}}^{I_2} x_{i,j}$$

et

$$\sum_{i=I_1}^{I_2} \sum_{j=i}^{J} x_{i,j} = \sum_{j=I_1}^{J} \sum_{i=I_1}^{\min\{I_2,j\}} x_{i,j}.$$

1.8 Division euclidienne et PGCD

Théorème 1.59 (Division euclidienne dans $\mathbb{Z}$)**.** *Soient $n \in \mathbb{Z}$ et $d \in \mathbb{Z}_0$. Alors n se décompose de façon unique sous la forme*

$$n = qd + r, \text{ avec } q \in \mathbb{Z} \text{ et } r \in \{0, \ldots, |d| - 1\}. \tag{1.5}$$

Preuve. Montrons tout d'abord l'existence d'une telle décomposition. La suite $(k|d|)_{k\in\mathbb{Z}}$ est strictement croissante puisque $(k+1)|d| - k|d| = |d| > 0$ pour tout $k \in \mathbb{Z}$. Il existe donc $k \in \mathbb{Z}$ tel que $k|d| \leq n < (k+1)|d|$. Posons $r = n - k|d|$. De plus, posons $q = k$ si $d > 0$ et $q = -k$ si $d < 0$. Alors $n = qd + r$, $q \in \mathbb{Z}$ et $r \in \{0, \ldots, |d| - 1\}$.

Montrons maintenant l'unicité de la décomposition. Si $n = qd + r = q'd + r'$, avec $q, q' \in \mathbb{Z}$ et $r, r' \in \{0, \ldots, |d| - 1\}$, alors on a $(q - q')d = r' - r$. Dans ce cas, on a $0 \leq |q - q'||d| = |(q - q')d| = |r' - r| < |d|$ et donc $0 \leq |q - q'| < 1$. Puisque $|q - q'|$ est un entier, cela entraîne que $|q - q'| = 0$. Ceci démontre que $q = q'$, et par conséquent, que $r = r'$. □

Les notations des théorèmes précédents n'ont pas été choisies au hasard puisque d est appelé le *diviseur*, q le *quotient* et r le *reste* de la division euclidienne de n par d. Nous adoptons les notations classiques suivantes pour le quotient et le reste d'une division euclidienne.

Définition 1.60. Soient $a \in \mathbb{Z}$ et $b \in \mathbb{Z}_0$. On note respectivement $\mathrm{DIV}(a, b)$ et $\mathrm{MOD}(a, b)$ le quotient et le reste de la division euclidienne de a par b.

La division euclidienne porte son nom en raison de l'algorithme d'Euclide [29] qui permet de calculer le PGCD (plus grand commun diviseur) de deux naturels [30]. En effet, cet algorithme calcule le PGCD en réalisant des divisions euclidiennes successives, jusqu'à arriver à une condition d'arrêt. Nous allons présenter et démontrer cet algorithme essentiel.

Algorithm 1 Algorithme d'Euclide

Require: $a, b \in \mathbb{N}_0$
Ensure: $\mathrm{pgcd}(a, b)$
  $r \leftarrow \max(a, b)$, $s \leftarrow \min(a, b)$
  while $s > 0$ **do**
    $(r, s) \leftarrow (s, \mathrm{MOD}(r, s))$
  end while
  return r

Remarquons que le calcul du PGCD effectué par l'algorithme d'Euclide n'utilise pas la décomposition des nombres en leurs facteurs premiers.

Exemple 1.61. Calculons le PGCD de 1078 et de 322 à l'aide de l'algorithme d'Euclide. On calcule successivement les divisions euclidiennes suivantes :

$$\begin{aligned} 1078 &= 3 \cdot 322 + 112 \\ 322 &= 2 \cdot 112 + 98 \\ 112 &= 1 \cdot 98 + 14 \\ 98 &= 7 \cdot 14 + 0. \end{aligned}$$

La sortie de l'algorithme est le dernier reste non nul de cette suite de divisions euclidiennes, soit 14. Sans connaître l'algorithme d'Euclide, nous aurions calculé la décomposition de 1078 et de 322 en facteurs premiers : $1078 = 2 \cdot 7^2 \cdot 11$ et $322 = 2 \cdot 7 \cdot 23$. Le PGCD de 1078 et de 322 est le produit de leurs facteurs premiers communs (répétés autant de fois qu'ils apparaissent dans les deux décompositions à la fois), soit $2 \cdot 7 = 14$.

29. Datant d'environ 300 avant J.C. Pas besoin d'un ordinateur pour concevoir un algorithme !
30. Le PGCD de deux entiers relatifs a et b est simplement défini comme le PGCD de $|a|$ et $|b|$.

Théorème 1.62. *L'algorithme d'Euclide est correct et se termine toujours.*

Preuve. L'algorithme d'Euclide se termine toujours. En effet, la variable s contient toujours un nombre naturel et à chaque étape de la boucle, la valeur de s décroît strictement puisque $\text{MOD}(r, s) < s$.
Détaillons les divisions euclidiennes successives de l'algorithme d'Euclide (en supposant que $a \geq b$) :

$$\begin{aligned} a &= q_1 \cdot b + r_1 \\ b &= q_2 \cdot r_1 + r_2 \\ r_1 &= q_3 \cdot r_2 + r_3 \\ r_2 &= q_4 \cdot r_3 + r_4 \\ &\vdots \\ r_{j-2} &= q_j \cdot r_{j-1} + r_j \\ r_{j-1} &= q_{j+1} \cdot r_j + 0 \end{aligned}$$

où les q_i et r_i sont les quotient et reste de la division euclidienne de l'étape i (où $1 \leq i \leq j+1$), avec $r_1, \ldots, r_j$ non nuls[31]. On pose $r_{-1} = a$ et $r_0 = b$. La sortie de l'algorithme est le dernier reste non nul des divisions euclidiennes successives, soit r_j.
Pour montrer que l'algorithme d'Euclide est correct, nous devons démontrer que $r_j = \text{pgcd}(a, b)$. Pour cela, on doit montrer deux choses. Premièrement, que r_j est un diviseur de a et de b. Deuxièmement, que r_j est plus grand que tous les autres diviseurs de a et de b.
Le fait que r_j divise a et b s'obtient de proche en proche, en remontant les égalités. En effet, la dernière égalité montre que r_j divise r_{j-1}. Ensuite, de l'avant-dernière égalité, on obtient que r_j divise r_{j-2}, puis $r_{j-3}, \ldots, r_1, r_0 = b$ et enfin $r_{-1} = a$.
Supposons à présent que d soit un diviseur commun de a et de b. De la première égalité, on obtient que d divise $r_1 = a - q_1 \cdot b$. Ensuite, de la deuxième égalité, on obtient que d divise $r_2 = b - q_2 \cdot r_1$. En continuant de proche en proche vers le bas jusque l'avant-dernière égalité, on obtient que d divise r_j. On a donc bien $r_j \geq d$. □

Théorème 1.63 (Bachet-Bézout)**.** *Pour tous $a, b \in \mathbb{N}_0$, il existe $m, n \in \mathbb{Z}$ tels que*

$$ma + nb = \text{pgcd}(a, b).$$

Preuve. Avec les notations de la preuve du théorème 1.62, il suffit d'observer que l'on peut exprimer chaque r_i sous la forme $r_i = m_i a + n_i b$ où $m_i, n_i \in \mathbb{Z}$. En effet, puisque le PGCD de a et b est donné par r_j, les entiers $m = m_j$ et $n = n_j$ conviendront pour la thèse. Formellement, on montre ceci par récurrence sur $i \geq -1$. Premièrement on a $r_{-1} = a = 1 \cdot a + 0 \cdot b$ et $r_0 = b = 0 \cdot a + 1 \cdot b$. Supposons maintenant que i soit tel que $1 \leq i \leq j$, que $r_{i-2} = m_{i-2} a + n_{i-2} b$ et que $r_{i-1} = m_{i-1} a + n_{i-1} b$, avec $m_{i-2}, n_{i-2}, m_{i-1}, n_{i-1} \in \mathbb{Z}$. Alors on calcule

$$\begin{aligned} r_i &= r_{i-2} - q_i r_{i-1} \\ &= m_{i-2} a + n_{i-2} b - q_i (m_{i-1} a + n_{i-1} b) \\ &= (m_{i-2} - q_i m_{i-1}) a + (n_{i-2} - q_i n_{i-1}) b. \end{aligned}$$

Ainsi les entiers $m_i = m_{i-2} - q_i m_{i-1}$ et $n_i = n_{i-2} - q_i n_{i-1}$ sont tels que $r_i = m_i a + n_i b$. □

31. Remarquez que le j n'est pas connu à l'avance, mais son existence est garantie par la décroissance des restes. C'est le principe d'une boucle "tant que". Il faut toujours garantir que la condition de boucle sera nécessairement violée après un certain nombre (non connu a priori) d'itérations.

Il faut cependant remarquer que les coefficients m et n du théorème 1.63 ne sont pas uniques puisque l'égalité $ma + nb = \text{pgcd}(a, b)$ implique que pour tout $k \in \mathbb{Z}$, on a aussi $(m + kb)a + (n - ka)b = \text{pgcd}(a, b)$.

Exemple 1.64. Continuons l'exemple 1.61 pour obtenir des entiers m et n tels $m \cdot 1078 + n \cdot 322 = 14$. On calcule successivement :

$$\begin{aligned} 14 &= 112 - 98 \\ &= 112 - (322 - 2 \cdot 112) \\ &= 3 \cdot 112 - 322 \\ &= 3 \cdot (1078 - 3 \cdot 322) - 322 \\ &= 3 \cdot 1078 - 10 \cdot 322. \end{aligned}$$

On appelle "algorithme d'Euclide étendu" l'algorithme décrit dans la preuve du théorème 1.63 qui permet d'obtenir le PGCD de deux naturels a et b non nuls ainsi que des entiers m et n tels que $ma + nb = \text{pgcd}(a, b)$. Les coefficients obtenus par cet algorithme sont appelés les *coefficients de Bézout*. Ainsi, dans notre exemple, on a obtenu que les coefficients de Bézout de 1078 et 322 sont $m = 3$ et $n = -10$.

Exercice 1.65. Modifier l'algorithme 1 pour obtenir l'algorithme d'Euclide étendu. La sortie attendue est le triplet de nombres $(\text{pgcd}(a, b), m, n)$.

Rappelons que deux naturels sont premiers entre eux lorsque leur PGCD vaut 1. Le théorème suivant est très utile en général (dans notre cas, nous l'utiliserons en arithmétique modulaire pour la recherche d'inverses, mais il a également de nombreuses autres applications), et surtout très joli.

Théorème 1.66 (Bézout)**.** *Deux naturels non nuls a et b sont premiers entre eux si et seulement s'il existe des entiers (relatifs) m et n tels que $ma + nb = 1$.*

Preuve. La condition est nécessaire par le théorème 1.63. Montrons que la condition est suffisante. Soient $a, b \in \mathbb{N}_0$ et supposons qu'il existe $m, n \in \mathbb{Z}$ tels que $ma + nb = 1$. Nous devons montrer que $\text{pgcd}(a, b) = 1$. Comme $\text{pgcd}(a, b)$ divise à la fois a et b, on obtient de cette égalité que $\text{pgcd}(a, b)$ divise 1, ce qui implique que $\text{pgcd}(a, b) = 1$. □

Théorème 1.67 (Lemme de Gauss)**.** *Si a, b, c sont des entiers tels que c divise ab et $\text{pgcd}(a, c) = 1$, alors c divise b.*

Preuve. Soient $a, b, c \in \mathbb{Z}$ tels que c divise ab et $\text{pgcd}(a, c) = 1$. D'une part, il existe $q \in \mathbb{Z}$ tel que $ab = qc$ et d'autre part, par le théorème de Bézout, il existe $m, n \in \mathbb{Z}$ tels que $ma + nc = 1$. On obtient que

$$b = (ma + nc)b = mab + ncb = mqc + ncb = (mq + nb)c,$$

ce qui montre que c divise b. □

Nous obtenons comme corollaire le lemme d'Euclide, déjà énoncé avant la démonstration du théorème fondamental de l'arithmétique à la section 1.6.

Corollaire 1.68 (Lemme d'Euclide)**.** *Soient a et b des entiers. Si un nombre premier p divise ab, alors p divise a ou b.*

Preuve. Il s'agit de la démonstration d'une alternative. Supposons que p est un nombre premier divisant ab et ne divisant pas a. Alors $\text{pgcd}(a, p) = 1$ et par le lemme de Gauss, on obtient que p divise b. □

1.9 Numération en bases entières

Enfants, nous avons tous appris à compter sur nos doigts. À l'école, nous avons appris à représenter les nombres en base 10. C'est cette représentation qui est utilisée dans la vie de tous les jours. Nous écrivons les nombres de cette manière depuis toujours, si bien que nous *pensons* les nombres de cette manière. Pourtant, les nombres existent indépendamment de la façon dont on les représente ![32] Dans l'histoire des mathématiques, il a existé de nombreuses manières de représenter les nombres (pensez par exemple aux chiffres romains, ou à notre manière de lire l'heure). Dans les ordinateurs, c'est le binaire qui est utilisé [33]. Représenter les nombres en binaire, c'est en fait décomposer les nombres dans la base 2 (au lieu de la base 10 comme nous en avons l'habitude). En fait, on peut définir des systèmes de numération en n'importe quelle bases entières $b \geq 2$ [34]. C'est l'objet du résultat suivant. La démonstration est constructive et utilise à nouveau une répétition de divisions euclidiennes. L'algorithme suivant permet d'obtenir une suite de nombres $(c_\ell, c_{\ell-1}, \ldots, c_0)$ telle que $n = \sum_{i=0}^{\ell} c_i\, b^i$. Nous montrerons ensuite que, en ajoutant certaines contraintes sur les coefficients c_i, cette décomposition est unique. Ceci nous permettra de définir les représentations des entiers en base b.

Algorithm 2 Algorithme glouton de décomposition en base b

Require: $n, b \in \mathbb{N}_0$ avec $b \geq 2$
Ensure: $(c_\ell, \ldots, c_0)$ tels que $n = \sum_{i=0}^{\ell} c_i\, b^i$, $c_\ell \neq 0$ et pour tout $i \in \{0, \ldots, \ell\}$, $c_i \in \{0, \ldots, b-1\}$
  $i \leftarrow \lfloor \log_b(n) \rfloor$, $r \leftarrow n$, liste_chiffres $\leftarrow ()$
  while $i \geq 0$ **do**
    $c \leftarrow \mathrm{DIV}(r, b^i)$
    $r \leftarrow \mathrm{MOD}(r, b^i)$
    liste_chiffres $\leftarrow$ (liste_chiffres, c)
    $i \leftarrow i - 1$
  end while
  return liste_chiffres

Exemple 1.69. Illustrons l'action de l'algorithme glouton sur l'entrée $n = 177$ en base $b = 2$ et en base 3.

Considérons tout d'abord le cas de la base $b = 2$. Puisque $2^7 = 128 \leq 177 < 2^8 = 256$, on a $\lfloor \log_2(177) \rfloor = 7$. On a donc successivement

$$\begin{aligned} 177 &= 1 \cdot 128 + 49 \\ 49 &= 0 \cdot 64 + 49 \\ 49 &= 1 \cdot 32 + 17 \\ 17 &= 1 \cdot 16 + 1 \\ 1 &= 0 \cdot 8 + 1 \\ 1 &= 0 \cdot 4 + 1 \\ 1 &= 0 \cdot 2 + 1 \\ 1 &= 1 \cdot 1 + 0. \end{aligned}$$

La liste des chiffres produite par l'algorithme est $(1, 0, 1, 1, 0, 0, 0, 1)$. On a donc $\ell = 7$, $c_\ell = 1 \neq 0$, tous les chiffres sont bien compris entre 0 et $1 = b - 1$ et

$$177 = 1 \cdot 2^7 + 0 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0.$$

32. À méditer...
33. Il y a 10 sortes de personnes : celles qui connaissent le binaire et celles qui ne le connaissent pas.
34. Qu'en est-il de la numération unaire, c'est-à-dire de la base 1 ?

Considérons maintenant le cas de la base $b = 3$. Puisque $3^4 = 81 \leq 177 < 3^5 = 243$, on a $\lfloor \log_3(177) \rfloor = 4$. On a donc successivement

$$\begin{aligned} 177 &= 2 \cdot 81 + 15 \\ 15 &= 0 \cdot 27 + 15 \\ 15 &= 1 \cdot 9 + 6 \\ 6 &= 2 \cdot 3 + 0 \\ 0 &= 0 \cdot 1 + 0 \end{aligned}$$

La liste des chiffres produite par l'algorithme est $(2, 0, 1, 2, 0)$. On a donc $\ell = 4$, $c_\ell = 2 \neq 0$, tous les chiffres sont bien compris entre 0 et $2 = b - 1$ et

$$177 = 2 \cdot 3^4 + 0 \cdot 3^3 + 1 \cdot 3^2 + 2 \cdot 3^1 + 0 \cdot 3^0.$$

Théorème 1.70. *Soit b un entier plus grand ou égal à 2. L'algorithme glouton de décomposition en base b est correct et se termine toujours.*

Preuve. L'algorithme se termine toujours. En effet, la variable i est initialisée à $\lfloor \log_b(n) \rfloor$ et diminue d'une unité à chaque étape de la boucle. La condition $i \geq 0$ sera donc violée après exactement $\lfloor \log_b(n) \rfloor + 1$ étapes.
Détaillons les divisions euclidiennes successives de l'algorithme glouton. Soit $\ell = \lfloor \log_b(n) \rfloor$ et $r_\ell = n$. On a successivement

$$\begin{aligned} r_\ell &= c_\ell \cdot b^\ell + r_{\ell-1} \\ r_{\ell-1} &= c_{\ell-1} \cdot b^{\ell-1} + r_{\ell-2} \\ r_{\ell-2} &= c_{\ell-2} \cdot b^{\ell-2} + r_{\ell-3} \\ &\vdots \\ r_1 &= c_1 \cdot b^1 + r_0 \\ r_0 &= c_0 \cdot b^0 + 0 \end{aligned}$$

où les c_i et r_{i-1} sont les quotient et reste de la division euclidienne de l'étape $\ell - i + 1$ (où $0 \leq i \leq \ell$). La sortie de l'algorithme est la suite des chiffres $(c_\ell, c_{\ell-1}, \ldots, c_1, c_0)$.
Pour montrer que l'algorithme est correct, nous devons démontrer que $n = \sum_{i=0}^{\ell} c_i \, b^i$, $c_\ell \neq 0$ et $c_i \in \{0, \ldots, b-1\}$ pour tout $i \in \{0, \ldots, \ell\}$. Puisque $\ell = \lfloor \log_b(n) \rfloor$, nous avons $b^\ell \leq n < b^{\ell+1}$ et $1 \leq c_\ell = \mathrm{DIV}(n, b^\ell) < b$. De plus, pour tout $i \in \{0, \ldots, \ell - 1\}$, nous avons $r_i = \mathrm{MOD}(r_{i+1}, b^{i+1}) < b^{i+1}$, et donc $c_i = \mathrm{DIV}(r_i, b^i) < b$. Ainsi tous les coefficients c_i sont compris entre 0 et $b - 1$ et, de plus, $c_\ell \neq 0$. Enfin, en remontant toutes les égalités jusqu'à la première, nous obtenons

$$\begin{aligned} r_0 &= c_0 \\ r_1 &= c_1 b + r_0 \;=\; c_1 b + c_0 \\ r_2 &= c_2 b^2 + r_1 \;=\; c_2 b^2 + c_1 b + c_0 \\ &\vdots \\ r_\ell &= c_\ell b^\ell + r_{\ell-1} \;=\; c_\ell b^\ell + \sum_{i=0}^{\ell-1} c_i b^i \;=\; \sum_{i=0}^{\ell} c_i b^i. \end{aligned}$$

□

Définition 1.71. Les *chiffres* de la base b sont $0, \ldots, b-1$. La *représentation de n en base b* est la suite de chiffres $(c_\ell, c_{\ell-1}, \ldots, c_1, c_0)$ produits par l'algorithme glouton. Quand il n'y a pas d'ambiguïté, on écrit simplement $c_\ell \cdots c_0$.

$b=10$	$b=2$	$b=3$	$b=10$	$b=2$	$b=3$
1	1	1	11	1011	102
2	10	2	12	1100	110
3	11	10	13	1101	111
4	100	11	14	1110	112
5	101	12	15	1111	120
6	110	20	16	10000	121
7	111	21	17	10001	122
8	1000	22	18	10010	200
9	1001	100	19	10011	201
10	1010	101	20	10100	202

TABLE 1.1 – Représentations en bases 2, 3 et 10

En base 10, on retrouve bien nos 10 chiffres de toujours : $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. En base 2, seuls deux chiffres sont autorisés : 0 et 1.

Exemple 1.72. La figure 1.1 reprend les premières représentations des naturels non nuls en bases 2 ,3 et 10. Les chiffres apparaissant dans ces représentations correspondent bien aux chiffres produits par l'algorithme glouton.

Proposition 1.73. *Soit $n, b \in \mathbb{N}_0$ avec $b \geq 2$. Le nombre de chiffres de la représentation de n en base b est égal à $\lfloor \log_b(n) \rfloor + 1$.*

Preuve. Ceci découle de l'algorithme glouton. □

Si l'on impose que les coefficients c_i soient compris entre 0 et $b-1$ et que de plus le coefficient c_ℓ de b^ℓ soit différent de 0, alors la décomposition en base entière fournie par l'algorithme glouton est en fait la seule possible. C'est l'objet du théorème suivant.

Théorème 1.74 (Décomposition en base entière)**.** *Soit b une base entière, c'est-à-dire un naturel plus grand ou égal à 2. Alors tout nombre $n \in \mathbb{N}_0$ se décompose de façon unique sous la forme*

$$n = \sum_{i=0}^{\ell} c_i \, b^i, \quad \textit{avec pour tout } i \in \{0, \ldots, \ell\},\ c_i \in \{0, \ldots, b-1\} \textit{ et } c_\ell \neq 0.$$

Preuve. L'existence de la décomposition est donnée par l'algorithme glouton. Montrons à présent l'unicité de la décomposition. Nous allons montrer que les coefficients c_i sont univoquement déterminés par n et b par récurrence sur n. Si $n < b$, puisque $b \geq 2$, la seule décomposition possible est $n = n \cdot b^0$, de sorte que $\ell = 0$ et $c_0 = n$. Supposons maintenant que $n \geq b$ et que l'unicité est vérifiée pour tout entier m tel que $1 \leq m < n$. On a $n = (\sum_{i=1}^{\ell} c_i \, b^{i-1}) \cdot b + c_0$. Comme $c_0 \in \{0, \ldots, b-1\}$, nous obtenons que c_0 est le reste de la division euclidienne de n par b, et est donc univoquement déterminé par n et b. Le nombre $m = \sum_{i=1}^{\ell} c_i \, b^{i-1} = \sum_{i=0}^{\ell-1} c_{i+1} \, b^i$, quant à lui, est le quotient de la division de n par b, et est donc aussi univoquement déterminé par n et b. De plus, $b \geq 2$ implique que $m < n$, et donc par hypothèse de récurrence, on obtient que les coefficients $c_\ell, \ldots, c_1$ sont univoquement déterminés par m et b, et donc par n et b (puisque n et b déterminent m). □

Remarquez qu'on aurait pu également utiliser l'algorithme glouton pour démontrer l'unicité de la décomposition en base entière, en montrant successivement que les coefficients c_ℓ, puis $c_{\ell-1}, \cdots, c_0$ étaient univoquement déterminés par n et b. La démonstration n'aurait pas

0	0000	8	1100
1	0001	9	1101
2	0011	10	1111
3	0010	11	1110
4	0110	12	1010
5	0111	13	1011
6	0101	14	1001
7	0100	15	1000

TABLE 1.2 – Codes de Gray de longueur 4 des entiers de 0 à 15

été plus ou moins difficile. Celle que nous avons présentée ci-dessus a l'avantage de mettre en lumière un deuxième algorithme de décomposition.

Exercice 1.75. Adapter la preuve de l'unicité de la décomposition en base entière présentée ci-dessus pour obtenir un algorithme de décomposition qui produit les chiffres de la représentation de n en base b de droite à gauche, c'est-à-dire dans l'ordre $(c_0, c_1, \ldots, c_\ell)$.

Exercice 1.76. Montrer l'unicité de la décomposition en base entière en utilisant l'algorithme glouton.

1.10 Code de Gray

Le codage de Gray est une autre façon de représenter les naturels avec les chiffres 0 et 1. Le codage de Gray est construit de telle sorte que les représentations de deux entiers consécutifs diffèrent en exactement un chiffre. Nous avons en fait déjà utilisé le codage de Gray lorsque nous avons vu les tables de Karnaugh.

Il y a plusieurs façons d'obtenir les codes de Gray des naturels. Voici la première façon, qui constituera notre définition. Nous la donnons sous forme d'algorithme. On note $\bar{0} = 1$ et $\bar{1} = 0$.

Remarquons que pour les tables de Karnaugh, nous avions besoin qu'il n'y ait qu'un chiffre différent entre la dernière ligne de la table et la première puisque nous considérions que la première ligne "suivait" la dernière. Ceci est bien vérifié par le codage de Gray. C'est ce que nous dit notamment le résultat suivant, que nous énonçons sans preuve.

Fait 1.77. *Pour $n \in \mathbb{N}_0$ donné, l'algorithme 3 produit 2^n suites binaires de longueur n, chacune apparaissant une et une seule fois. De plus, entre deux suites consécutives w_i et w_{i+1} (avec $0 \leq i \leq 2^n - 2$), il y a exactement un élément qui est modifié. Enfin, on a toujours*

$$w_0 = (\underbrace{0, \ldots, 0}_{n}) \quad \text{et} \quad w_{2^n-1} = (1, \underbrace{0, \ldots, 0}_{n-1}).$$

En particulier, exactement un élément est modifié entre w_{2^n-1} et w_0.

Définition 1.78. Le *code de Gray* d'un naturel k est la k-ième suite binaire produite par l'algorithme 3 avec $\lfloor \log_2(k) \rfloor + 1$ en entrée.

Une deuxième méthode de calcul des codes de Gray justifie qu'on l'appelle aussi parfois le *code binaire réfléchi*. Pour obtenir les codes de Gray de longueur n des entiers de 0 à $2^n - 1$, on peut procéder comme suit. On commence par les codes de longueur 1 de 0 et de 1, qui sont 0 et 1 respectivement. Ensuite, pour chaque $i \in \{2, \ldots, n\}$, pour obtenir les codes de Gray de longueur i des entiers de 0 à $2^i - 1$ à partir de ceux de longueur $i - 1$ des entiers de 0 à $2^{i-1} - 1$, on symétrise (en miroir) la liste des codes de Gray de longueur

Algorithm 3 Code de Gray des naturels de 0 à $2^n - 1$

Require: $n \in \mathbb{N}_0$

Ensure: $(w_0, \ldots, w_{2^n-1})$, où chaque w_i est une liste de n chiffres binaires

```
code ← (0, ..., 0)   (n zeros)
liste_codes ← (code)
i ← 1
while i < 2^n do
    j ← n
    while j ≥ 1 do
        code_temp ← inverser le j-ième élément de code
        if code_temp n'apparaît pas dans liste_codes then
            code ← code_temp
            liste_codes ← (liste_codes, code)
            j ← 0
        else
            j ← j − 1
        end if
    end while
    i ← i + 1
end while
return liste_codes
```

$n-1$, créant une nouvelle liste de longueur doublée, ensuite on écrit 0 devant la première moitié des éléments de la nouvelle liste et on écrit 1 devant la deuxième moitié.

Exemple 1.79. La table 1.3 illustre la deuxième méthode d'obtention des codes de Gray avec $n = 3$.

0	0
1	1

0	00
1	01
2	11
3	10

0	000
1	001
2	011
3	010
4	110
5	111
6	101
7	100

TABLE 1.3 – Codes de Gray des entiers de 0 à 7 avec la méthode du miroir

Les méthodes de calculs des codes de Gray données jusqu'ici possèdent le désavantage qu'il faut connaître la liste complète des codes de Gray de longueurs inférieures pour générer un code de Gray d'une longueur n donnée. La troisième méthode que nous présentons permet de calculer le code de Gray d'un naturel directement à partir de sa représentation en base 2. Nous donnons une nouvelle fois cette méthode sans preuve [35]. La notation $\oplus$ désigne le

35. Bien que la preuve ne soit pas difficile, en considérant les codes de Gray donnés à partir de la méthode du miroir. Rédiger une preuve de ceci est un très bon exercice.

"ou exclusif", dont voici la table de vérité :

φ	ψ	$\varphi \oplus \psi$
0	0	0
0	1	1
1	0	1
1	1	0

Fait 1.80. *Soit $n \in \mathbb{N}_0$. Si $c_\ell \cdots c_0$ est la représentation en base 2 de n, alors le code de Gray de n est $(c_\ell \oplus 0, c_{\ell-1} \oplus c_\ell, \ldots, c_0 \oplus c_1)$.*

Au vu de cette proposition, on dit parfois que le code de Gray de n peut être obtenu en effectuant l'addition sans retenue de $c_\ell \cdots c_0$ et $c_\ell \cdots c_1$.

Exemple 1.81. La représentation de 6 en base 2 est 110 et nous avons vu que le code de Gray de 6 est 101. La représentation de 11 en base 2 est 1011 et nous avons vu que le code de Gray de 11 est 1110. La table 1.4 montre comment obtenir ces codes en utilisant la proposition 1.80.

1	1	0
0	1	1
1	0	1

1	0	1	1
0	1	0	1
1	1	1	0

TABLE 1.4 – Codes de Gray de 6 et de 11 calculés à partir de leurs représentations en base 2

1.11 Arithmétique modulaire

L'arithmétique modulaire, en plus d'être une jolie curiosité, est aussi un excellent moyen de repenser les mathématiques que nous avons apprises jusque là d'une façon différente. De plus, pour ne rien gâcher, elle possède de nombreuses applications très importantes, comme par exemple la cryptographie. Nous n'aurons malheureusement pas le temps d'aborder ces applications dans ce cours. Mais prenons du plaisir à jouer avec les modulos. Nous commençons tout de suite.

Définition 1.82. Pour tout naturel $m \geq 2$, nous notons $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$. Nous définissons deux opérations binaires sur cet ensemble, appelée *addition modulo m* et *multiplication modulo m*. On les note $+_m$ et $\cdot_m$ et on les définit comme suit :

$$+_m \colon \mathbb{Z}_m \times \mathbb{Z}_m,\ (i,j) \mapsto \mathrm{MOD}(i+j, m)$$

et

$$\cdot_m \colon \mathbb{Z}_m \times \mathbb{Z}_m,\ (i,j) \mapsto \mathrm{MOD}(i \cdot j, m).$$

Autrement dit, pour tous $i, j \in \mathbb{Z}_m$, on a $i +_m j = \mathrm{MOD}(i+j, m)$ et $i \cdot_m j = \mathrm{MOD}(ij, m)$.

Dans la suite de cette section, m désigne toujours un entier plus grand ou égal à 2. Les tables d'addition et de multiplication dans $\mathbb{Z}_5$ et $\mathbb{Z}_6$ sont données aux figures 1.5 et 1.6.

Faire des calculs dans $\mathbb{Z}_m$ peut rapidement s'avérer fastidieux si l'on ne remarque pas qu'il revient au même d'effectuer les calculs dans $\mathbb{Z}$ et de "réduire modulo m" à la fin (ou à n'importe quel moment qui nous arrange d'ailleurs). Ceci est l'objet du résultat pratique suivant.

$+_5$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\cdot_5$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

TABLE 1.5 – Addition et multiplication modulo 5

$+_6$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\cdot_6$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

TABLE 1.6 – Addition et multiplication modulo 6

Proposition 1.83. *Soient $x, y, k \in \mathbb{Z}$. Alors*

1. $\mathrm{MOD}(x, m) = \mathrm{MOD}(y, m) \iff x - y$ *est multiple de* m.
2. $\mathrm{MOD}(x + km, m) = \mathrm{MOD}(x, m)$.
3. $\mathrm{MOD}(x + y, m) = \mathrm{MOD}(\mathrm{MOD}(x, m) + y, m)$
4. $\mathrm{MOD}(x \cdot y, m) = \mathrm{MOD}(\mathrm{MOD}(x, m) \cdot y, m)$

Preuve. Supposons que $x = qm + r$ et $y = q'm + r'$, avec $q, q' \in \mathbb{Z}$ et $r, r' \in \mathbb{Z}_m$. On a donc $r = \mathrm{MOD}(x, m)$ et $r' = \mathrm{MOD}(y, m)$. On a $x - y = (q - q')m + r - r'$ et comme $|r - r'| < m$, on obtient bien que $x - y$ est multiple de m si et seulement si $r - r' = 0$. Le point 1 est démontré. Le point 2 découle directement du point 1.

Montrons les points 3 et 4. On a $x + y = qm + r + y$ et $xy = qmy + ry$. En utilisant le point 2, on obtient $\mathrm{MOD}(x + y, m) = \mathrm{MOD}(r + y, m)$ et $\mathrm{MOD}(xy, m) = \mathrm{MOD}(ry, m)$, comme souhaité. □

Ce résultat implique toute égalité dans $\mathbb{Z}$ est aussi vérifiée "modulo m". Par exemple, l'égalité $27 = 2 \cdot 8 + 11$ donne $1 = 0 \cdot_2 0 +_2 1$ dans $\mathbb{Z}_2$, $0 = 2 \cdot_3 2 +_3 2$ dans $\mathbb{Z}_3$, $3 = 2 \cdot_4 0 +_4 3$ dans $\mathbb{Z}_4$, $2 = 2 \cdot_5 3 +_5 1$ dans $\mathbb{Z}_5$, etc.

L'addition et la multiplication modulaires vérifient les propriétés habituelles de l'addition et de la multiplication, que nous énumérons ci-dessous. Nous n'en faisons pas ici la démonstration, bien qu'il s'agisse de simples vérifications [36]. On utilise la priorité des opérations habituelles (d'abord $\cdot_m$ puis $+_m$). Remarquons que seul le dernier point est propre à $\mathbb{Z}_m$.

Proposition 1.84. *Pour tout $i, j, k \in \mathbb{Z}_m$, nous avons*

1. $(i +_m j) +_m k = i +_m (j +_m k)$ — *associativité de* $+_m$
2. $(i \cdot_m j) \cdot_m k = i \cdot_m (j \cdot_m k)$ — *associativité de* $\cdot_m$
3. $i \cdot_m (j +_m k) = i \cdot_m j \ +_m \ i \cdot_m k$ — *distributivité de* $\cdot_m$ *sur* $+_m$
4. $i +_m j = j +_m i$ — *commutativité de* $+_m$
5. $i \cdot_m j = j \cdot_m i$ — *commutativité de* $\cdot_m$
6. $0 +_m i = i +_m 0 = i$ — 0 *est neutre pour* $+_m$

36. C'est d'ailleurs un bon exercice que de s'atteler à ces vérifications.

7. $1 \cdot_m i = i \cdot_m 1 = i$ — *1 est neutre pour* $\cdot_m$
8. $i +_m (m - i) = 0$ *si* $i \neq 0$ — *l'opposé de* $i \neq 0$ *est* $m - i$

Comme d'habitude, l'associativité permet de donner du sens aux écritures $i +_m j +_m k$ et $i \cdot_m j \cdot_m k$ puisque l'ordre dans lequel on effectue ces opérations n'a pas d'importance.
Il est utile de remarquer que l'addition par un élément de $\mathbb{Z}_m$ est injective.

Proposition 1.85. *Pour tout $i \in \mathbb{Z}_m$, la fonction $\mathbb{Z}_m \to \mathbb{Z}_m$, $j \mapsto i +_m j$ est injective.*

Preuve. Soient $j, j' \in \mathbb{Z}_m$ tels que $i +_m j = i +_m j'$. Alors $j = (m - i) +_m i +_m j = (m - i) +_m i +_m j' = j'$. □

Nous avons facilement identifié les opposés des éléments de $\mathbb{Z}_m$: l'opposé d'un élément i de $\mathbb{Z}_m$ est simplement $m - i$ si $i \neq 0$ et 0 sinon. Déterminer les inverses est par contre plus délicat.

Lemme 1.86. *Soient $i, j, j' \in \mathbb{Z}_m$. Alors $i \cdot_m j = i \cdot_m j' = 1 \implies j = j'$.*

Preuve. Supposons que $i \cdot_m j = i \cdot_m j' = 1$. En utilisant la proposition précédente, on en déduit que $j = j \cdot_m 1 = j \cdot_m (i \cdot_m j') = (j \cdot_m i) \cdot_m j' = (i \cdot_m j) \cdot_m j' = 1 \cdot_m j' = j'$. □

Le lemme précédent nous dit que si un élément possède un inverse, alors celui-ci est unique. Au vu de son importance, nous mettons en évidence la définition de la notion d'inverse.

Définition 1.87. Un élément i de $\mathbb{Z}_m$ est *inversible modulo m* s'il existe j dans $\mathbb{Z}_m$ tel que $i \cdot_m j = 1$. Au vu du lemme 1.86, il ne peut exister qu'un seul tel élément j et lorsqu'il existe, celui-ci est appelé l'*inverse de i modulo m*. On dit aussi que i est un *élément inversible de* $\mathbb{Z}_m$.

Théorème 1.88. *Un élément i de $\mathbb{Z}_m$ est inversible modulo m si et seulement si i et m sont premiers entre eux.*

Preuve. Montrons d'abord la condition nécessaire. Soit i un élément inversible de $\mathbb{Z}_m$. Alors il existe $j \in \mathbb{Z}_m$ tel que $i \cdot_m j = 1$, c'est-à-dire tel que $\mathrm{MOD}(ij, m) = 1$. Il existe donc $q \in \mathbb{N}$ tel que $ij = qm + 1$, et donc tel que $ij - qm = 1$. En appliquant le théorème de Bézout, on obtient que i et m sont premiers entre eux.
Montrons à présent que la condition est suffisante. Soit i un élément de $\mathbb{Z}_m$ premier avec m. Par le théorème de Bézout, il existe des entiers a et b tels que $ai + bm = 1$. En utilisant la proposition 1.83, on obtient que $i \cdot_m \mathrm{MOD}(a, m) = \mathrm{MOD}(ia, m) = \mathrm{MOD}(1 - bm, m) = 1$. L'élément $\mathrm{MOD}(a, m)$ est donc l'inverse de i dans $\mathbb{Z}_m$. □

Remarquons que la preuve du théorème 1.88 montre que la recherche d'un inverse modulaire peut de faire à l'aide de l'algorithme d'Euclide. Ceci sera très utile pour résoudre des équations dans $\mathbb{Z}_m$.
Attention : la multiplication par un élément quelconque de $\mathbb{Z}_m$ n'est pas toujours injective ! Par exemple, on a $2 \cdot_6 1 = 2 \cdot_6 4$ dans $\mathbb{Z}_6$. Néanmoins, il est vrai que la multiplication par un élément inversible de $\mathbb{Z}_m$ est injective.

Proposition 1.89. *Pour tout i inversible modulo m, la fonction $\mathbb{Z}_m \to \mathbb{Z}_m$, $j \mapsto i \cdot_m j$ est injective.*

Preuve. Soit k l'inverse de i dans $\mathbb{Z}_m$ et soient $j, j' \in \mathbb{Z}_m$ tels que $i \cdot_m j = i \cdot_m j'$. Alors $j = k \cdot_m i \cdot_m j = k \cdot_m i \cdot_m j' = j'$. □

Dans le cas de $\mathbb{Z}_6$, obtient que seuls 1 et 5 sont inversibles modulo 6. On vérifie en effet que dans la table de multiplication modulo 6, les lignes correspondants aux éléments 0, 2, 3 et 4 ne contiennent pas l'élément 1, mais que 1 apparaît bien dans les lignes correspondants à 1 et 5.

Dans le cas de $\mathbb{Z}_5$, on observe que toutes les lignes de la table de multiplication modulo 5, excepté celle de 0, contient 1. Ceci montre que tous les éléments non nuls de $\mathbb{Z}_5$ sont inversibles. En fait, ceci est un cas particulier du résultat plus général suivant puisque 5 est un nombre premier.

Théorème 1.90. *Tous les éléments non nuls de $\mathbb{Z}_m$ sont inversibles si et seulement si m est un nombre premier.*

Preuve. Ceci découle directement du théorème 1.88 et du fait que m est premier avec $1, 2, \ldots, m-1$ si et seulement si m est un nombre premier. □

Exercice 1.91. Pour résoudre les exercices, vous n'avez pas le droit d'utiliser de calculatrice ! Il faudra donc d'être efficace, et ne pas passer en revue toutes les valeurs possibles pour x. De plus, il est entendu que lorsqu'on demande de résoudre une équation du type $ax + b$ dans $\mathbb{Z}_m$, les opérations de multiplication et d'addition doivent être interprétées comme étant réellement $\cdot_m$ et $+_m$.

1. Résoudre l'équation $10x + 8 = 0$ dans $\mathbb{Z}_{21}$.
   Comme $\text{pgcd}(10, 21) = 1$, nous savons que 10 est inversible dans $\mathbb{Z}_{21}$. Puisque $21 - 2 \cdot 10 = 1$, on obtient que $\text{MOD}(-2, 21) = 19$ est l'inverse de 10 dans $\mathbb{Z}_{21}$. Ainsi, en supposant que $x \in \mathbb{Z}_{21}$, on a les équivalences suivantes :
   $$\begin{aligned} 10 \cdot_{21} x +_{21} 8 = 0 &\iff 10 \cdot_{21} x = 13 \\ &\iff x = 19 \cdot_{21} 13 \\ &\iff x = \text{MOD}(19 \cdot 13, 21) \\ &\iff x = \text{MOD}((-2) \cdot (-8), 21) \\ &\iff x = 16. \end{aligned}$$
   Remarquez que l'utilisation de la proposition 1.83 a grandement facilité le calcul de $19 \cdot_{21} 13$. L'équation $10x + 8 = 0$ a donc 16 comme unique solution dans $\mathbb{Z}_{21}$.
2. Résoudre l'équation $10x + 8 = 0$ dans $\mathbb{Z}_{12}$.
   Comme $\text{pgcd}(10, 12) = 2$, 10 n'est pas inversible dans $\mathbb{Z}_{12}$. En supposant que $x \in \mathbb{Z}_{12}$, on a les équivalences suivantes :
   $$\begin{aligned} 10 \cdot_{12} x +_{12} 8 = 0 &\iff 10 \cdot_{12} x = 4 \\ &\iff \text{MOD}(10x, 12) = 4 \\ &\iff \exists q \in \mathbb{Z},\ 10x = 12q + 4 \\ &\iff \exists q \in \mathbb{Z},\ 5x = 6q + 2 \\ &\iff \text{MOD}(5x, 6) = 2 \\ &\iff 5 \cdot_6 \text{MOD}(x, 6) = 2. \end{aligned}$$
   Comme $\text{pgcd}(5, 6) = 1$, nous savons que 5 est inversible dans $\mathbb{Z}_6$. L'inverse de 5 dans $\mathbb{Z}_6$ est 5 puisque $5 \cdot_6 5 = 1$. On obtient les équivalences suivantes :
   $$\begin{aligned} 5 \cdot_6 \text{MOD}(x, 6) = 2 &\iff \text{MOD}(x, 6) = 5 \cdot_6 2 \\ &\iff \text{MOD}(x, 6) = 4 \\ &\iff \exists q \in \mathbb{Z},\ x = 6q + 4 \\ &\iff x = 4 \ \text{ ou } \ x = 10. \end{aligned}$$
   Les solutions de l'équation $10x + 8 = 0$ dans $\mathbb{Z}_{12}$ sont donc exactement 4 et 10.

3. Résoudre l'équation $10x + 8 = 0$ dans $\mathbb{Z}_{15}$.
   Comme $\text{pgcd}(10, 15) = 5$, 10 n'est pas inversible dans $\mathbb{Z}_{15}$. En supposant que $x \in \mathbb{Z}_{15}$, on a les équivalences suivantes :

$$\begin{aligned} 10 \cdot_{15} x +_{15} 8 = 0 &\iff 10 \cdot_{15} x = 7 \\ &\iff \text{MOD}(10x, 15) = 7 \\ &\iff \exists q \in \mathbb{Z},\ 10x = 15q + 7 \end{aligned}$$

   Mais si on a $7 = 10x - 15q$ avec $x, q \in \mathbb{Z}$, alors 7 est nécessairement un multiple de $\text{pgcd}(10, 15) = 5$. Comme ce n'est pas le cas (7 n'est pas multiple de 5), on obtient donc qu'il ne peut exister d'entiers x et q tels que $10x = 15q + 7$, et par conséquent que l'équation $10x + 8 = 0$ n'a pas de solution dans $\mathbb{Z}_{15}$.

Chapitre 2

Calcul matriciel

Cette partie du cours est librement inspirée des note de cours de "Mathématique" de Mélanie Bertelson, de "Algèbre linéaire" de Georges Hansoul, de "Algèbre linéaire" de Pascal Laubin et de "Algèbre linéaire" de Michel Rigo.

2.1 Premières définitions et exemples

Exemple 2.1. Une matrice est une table rectangulaire de nombres complexes, qu'on place généralement entre de grandes parenthèses.

Plus généralement, on peut définir une matrice comme une table rectangulaire de nombres d'un ensemble de nombres différent de $\mathbb{C}$. Cet ensemble doit alors être spécifié. Cela peut par exemple être $\mathbb{R}, \mathbb{C}, \mathbb{Z}, \mathbb{Q}, \mathbb{N}$ ou même $\mathbb{Z}_m$. Dans un premier temps, nous travaillerons uniquement dans $\mathbb{C}$.

Exemple 2.2. Voici quelques matrices :

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 4 & 2 \\ 3 & 9 & 5 \end{pmatrix} \quad \begin{pmatrix} -1 \\ i \\ 3+i \\ i\pi \end{pmatrix} \quad \begin{pmatrix} \pi & e^2 & \sqrt{2} \\ 2 & 4+\pi & 2 \end{pmatrix} \tag{10}$$

Définition 2.3. On parle, de façon évidente, de *lignes, colonnes* et *éléments* d'une matrice. Une matrice est de *taille* $\ell \times c$ si ℓ est le nombre de ses lignes et c le nombre de ses colonnes. On note $\mathbb{C}^{\ell \times c}$ l'ensemble des matrices de taille $\ell \times c$. [1] Une matrice dans $\mathbb{C}^{1 \times c}$ est appelée une *matrice-ligne*, tandis qu'une matrice dans $\mathbb{C}^{\ell \times 1}$ est appelée une *matrice-colonne*. L'élément à l'intersection de la i-ième ligne et de la j-ième colonne d'une matrice A est noté A_{ij}. On écrit

$$A = (A_{ij})_{\substack{1 \leq i \leq \ell \\ 1 \leq j \leq c}} = \begin{pmatrix} A_{11} & \cdots & A_{1c} \\ \vdots & & \vdots \\ A_{\ell 1} & \cdots & A_{\ell c} \end{pmatrix}.$$

Exemple 2.4 (suite)**.** Commentons les exemples donnés précédemment. La première matrice a 9 éléments naturels, possède 3 lignes et 3 colonnes, et est donc de taille 3×3. La deuxième matrice possède 4 éléments complexes, possède 4 lignes et 1 colonne, et est donc de taille 4×1. Il s'agit d'une matrice-colonne. La troisième matrice possède 6 éléments réels et est de taille 2×3. Enfin, la quatrième matrice ne possède qu'un seul élément et est donc de taille 1×1.

1. De même, on note $\mathbb{R}^{\ell \times c}$ l'ensemble des matrices de taille $\ell \times c$ et ayant tous leurs éléments dans $\mathbb{R}$.

Définition 2.5. Une matrice est dite *carrée* si elle possède le même nombre de lignes et de colonnes. Si A est une matrice carrée de taille $m \times m$, ses éléments *diagonaux* sont $A_{11}, \ldots, A_{mm}$. Une matrice carrée est *diagonale* si tous ses éléments non diagonaux sont nuls. Elle est dite *triangulaire supérieure* si tous les éléments situés en-dessous de sa diagonale sont nuls, i.e. $A_{ij} = 0$ si $i > j$, et *triangulaire inférieure* si tous les éléments situés au-dessus de sa diagonale sont nuls, i.e. $A_{ij} = 0$ si $i < j$.

Exemple 2.6 (suite)**.** Parmi les quatre matrices données en exemple ci-dessus, seules la première et la dernière sont des matrices carrées. Voici trois nouveaux exemples. La première matrice est une matrice diagonale, la deuxième une matrice triangulaire supérieure et la troisième une matrice triangulaire inférieure.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 2 \\ 0 & 0 & 5 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 2 & 4 & 0 \\ 0 & 9 & 5 \end{pmatrix}.$$

Remarquons qu'avec ces définitions, deux matrices A et B de tailles différentes ne sont jamais égales, et que deux matrices A et B de même taille $\ell \times c$ sont égales si et seulement si $A_{ij} = B_{ij}$ pour tous $i \in \{1, \ldots, \ell\}$ et $j \in \{1, \ldots, c\}$.
Nous introduisons quelques notations importantes, faisant référence à des matrices particulières qu'on rencontre très souvent.

Définition 2.7.

— On note I_m la matrice carrée de taille $m \times m$ possédant des 1 partout sur sa diagonale et des 0 ailleurs :

$$I_m = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Ainsi $I_2 \neq I_3$! Lorsqu'il n'y a pas de confusion possible concernant la taille de la matrice, on s'autorise à noter simplement I. Cette matrice est appelée la *matrice identité de taille* m, ou simplement la *matrice identité* lorsque la taille est supposée connue.

— On note $0_{\ell \times c}$ la matrice de taille $\ell \times c$ ne possédant que des éléments nuls :

$$0_{\ell \times c} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

À nouveau, on s'autorise à noter simplement 0 lorsqu'il n'y a pas de confusion possible concernant la taille de la matrice. Cette matrice est appelée la *matrice nulle de taille* $\ell \times c$, ou simplement la *matrice nulle* lorsque la taille est supposée connue.

2.2 Opérations sur les matrices

Matrices associées : transposée, conjuguée et adjointe

Définition 2.8. La matrice *transposée* (ou simplement la *transposée*) d'une matrice A de taille $\ell \times c$ est la matrice de taille $c \times \ell$, notée $A^\intercal$ dont les lignes sont les colonnes de A.

Autrement dit, on a

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1c} \\ A_{21} & A_{22} & \cdots & A_{2c} \\ \vdots & \vdots & & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & A_{\ell c} \end{pmatrix}^{\intercal} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{\ell 1} \\ A_{12} & A_{22} & \cdots & A_{\ell 2} \\ \vdots & \vdots & & \vdots \\ A_{1c} & A_{2c} & \cdots & A_{\ell c} \end{pmatrix},$$

ou encore

$$A^{\intercal} = (A_{ji})_{\substack{1 \leq j \leq c \\ 1 \leq i \leq \ell}}.$$

Remarquez que les éléments diagonaux ne changent pas lors de la transposition d'une matrice carrée. Voici un exemple de taille 3×3.

Exemple 2.9. On a $\begin{pmatrix} 1 & 0 & 1 \\ 2 & 4 & 7 \\ 3 & 0 & 32 \end{pmatrix}^{\intercal} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 0 \\ 1 & 7 & 32 \end{pmatrix}$.

Définition 2.10. Une matrice carrée est dite *symétrique* lorsqu'elle est égale à sa transposée.

Exemple 2.11. La matrice $\begin{pmatrix} 3 & -1 & 8 \\ -1 & 32 & 7 \\ 8 & 7 & 5 \end{pmatrix}$ est symétrique.

Définition 2.12. La matrice *conjuguée* d'une matrice (complexe) A est la matrice de même taille, notée $\overline{A}$, obtenue en remplaçant chacun des éléments de A par leur conjugué : si A est une matrice de taille $\ell \times c$, alors

$$\overline{A} = \left(\overline{A_{ij}}\right)_{\substack{1 \leq i \leq \ell \\ 1 \leq j \leq c}}.$$

Exemple 2.13. On a $\overline{\begin{pmatrix} 1 & i \\ \pi+3i & 0 \\ 3-i & -9i \end{pmatrix}} = \begin{pmatrix} 1 & -i \\ \pi-3i & 0 \\ 3+i & 9i \end{pmatrix}$.

Remarquez que $A = \overline{A} \iff A$ a tous ses éléments dans $\mathbb{R}$.

Définition 2.14. La matrice *adjointe* d'une matrice (complexe) A est la matrice $\overline{A^{\intercal}}$. On la note A^*.

Remarquez qu'on a toujours $\overline{A^{\intercal}} = \overline{A}^{\intercal}$. Autrement dit, la transposée de la matrice conjuguée est égale à la matrice conjuguée de la transposée.

Somme et produits

Définition 2.15. Si A et B sont deux matrices de même taille $\ell \times c$, alors on définit leur *somme* $A + B$ comme étant la matrice de taille $\ell \times c$ obtenue en additionnant les éléments se correspondant :

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1c} \\ A_{21} & A_{22} & \cdots & A_{2c} \\ \vdots & \vdots & & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & A_{\ell c} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1c} \\ B_{21} & B_{22} & \cdots & B_{2c} \\ \vdots & \vdots & & \vdots \\ B_{\ell 1} & B_{\ell 2} & \cdots & B_{\ell c} \end{pmatrix} = \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} & \cdots & A_{1c}+B_{1c} \\ A_{21}+B_{21} & A_{22}+B_{22} & \cdots & A_{2c}+B_{2c} \\ \vdots & \vdots & & \vdots \\ A_{\ell 1}+B_{m1} & A_{\ell 2}+B_{\ell 2} & \cdots & A_{\ell c}+B_{\ell c} \end{pmatrix}.$$

Proposition 2.16. *Si A, B, C sont des matrices de même taille, alors*

1. $(A + B) + C = A + (B + C)$ — *associativité de la somme*
2. $A + B = B + A$ — *commutativité de la somme*
3. $A + 0 = 0 + A = A$ — *0 est neutre pour la somme*

Autrement dit, la somme de matrices est associative et commutative. De plus la matrice nulle est neutre pour la somme.

Preuve. Ce découle du fait que la somme s'effectue "composante à composante" et que nous travaillons dans $\mathbb{C}$, où les mêmes propriétés de la somme sont vérifiées. □

Nous allons définir deux produits différents. Il est important de ne pas les confondre ! Le premier est le produit d'une matrice par un nombre.

Définition 2.17. Soit A une matrice de taille $\ell \times c$ et λ un nombre complexe. Le *produit de A par λ* est la matrice $\ell \times c$ obtenue en multipliant chaque élément de A par λ :

$$\lambda A = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \cdots & \lambda A_{1n} \\ \lambda A_{21} & \lambda A_{22} & \cdots & \lambda A_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda A_{\ell 1} & \lambda A_{\ell 2} & \cdots & \lambda A_{mn} \end{pmatrix}$$

Proposition 2.18. *Si $\lambda \in \mathbb{C}$ et $A, B \in \mathbb{C}^{\ell \times c}$, alors*

1. $\lambda(A+B) = \lambda A + \lambda B$ — *distributivité du produit par un nombre sur +*
2. $A + (-1)\cdot A = (-1)\cdot A + A = 0$ — *l'opposé de A est $(-1)\cdot A$*

Preuve. Ceci découle du fait que le produit d'une matrice par un nombre et la somme de deux matrices s'effectuent "composante à composante" et que nous travaillons dans $\mathbb{C}$, où les mêmes propriétés de la somme et du produit sont vérifiées. □

Comme d'habitude, on écrit l'opposé de A par $-A$. Ceci donne sens aux écritures $(-1)\cdot A = -A$. De plus, on écrit $-\lambda A$ pour désigner la matrice $-(\lambda A) = (-\lambda)A$. Enfin, on écrit également $A - B$ au lieu de $A + (-B)$. L'associativité de la somme, quant à elle, permet de donner du sens à l'écriture $A + B + C$.

Définition 2.19. Une *combinaison linéaire* de matrices $A_1, \ldots, A_k$ de même taille est une expression de la forme

$$\sum_{i=1}^{k} \lambda_i A_i$$

où $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$.

Voici la définition importante du produit de deux matrices. C'est lui que l'on appelle le *produit matriciel*.

Définition 2.20. On peut former le produit $A \cdot B$ de deux matrices A et B dans le cas où le nombre de colonnes de A est égal au nombre de lignes de B. Autrement dit, on peut former le produit d'une matrice A de taille $\ell \times c$ avec une matrice B de taille $\ell' \times c'$ lorsque $c = \ell'$. On obtient alors une matrice de taille $\ell \times c'$ notée $A \cdot B$, ou simplement AB, en procédant comme suit : pour chaque $i \in \{1, \ldots, \ell\}$ et chaque $j \in \{1, \ldots, c'\}$, l'élément $(AB)_{ij}$ est donné par

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

où $n = c = \ell'$.

Remarquons que l'élément $(AB)_{ij}$ du produit matriciel AB est en fait le résultat du produit scalaire [2]

$$(A_{i1}, \ldots, A_{in}) \bullet (B_{1j}, \ldots, B_{nj})$$

2. La notion de produit scalaire de deux vecteurs est vue au cours "Mathématique" de F. Bastin. Pour rappel, le *produit scalaire* de deux vecteurs $(a_1, \ldots, a_n)$ et $(b_1, \ldots, b_n)$ de $\mathbb{R}^n$ est noté $(a_1, \ldots, a_n) \bullet (b_1, \ldots, b_n)$ et est égal au nombre $a_1 b_1 + a_2 b_2 + \cdots a_n b_n$.

du vecteur formé des éléments de la i-ième ligne de A et du vecteur formé des éléments de la j-ième colonne de B.
Nous avons donc

$$A \cdot B = \begin{pmatrix} \sum_{k=1}^{n} A_{1k}B_{k1} & \sum_{k=1}^{n} A_{1k}B_{k2} & \cdots & \sum_{k=1}^{n} A_{1k}B_{kc'} \\ \sum_{k=1}^{n} A_{2k}B_{k1} & \sum_{k=1}^{n} A_{2k}B_{k2} & \cdots & \sum_{k=1}^{n} A_{2k}B_{kc'} \\ \vdots & & & \vdots \\ \sum_{k=1}^{n} A_{\ell k}B_{k1} & \sum_{k=1}^{n} A_{\ell k}B_{k2} & \cdots & \sum_{k=1}^{n} A_{\ell k}B_{kc'} \end{pmatrix}.$$

Exemple 2.21. On a $\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix}\begin{pmatrix} B_{11} \\ B_{21} \\ B_{31} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11}+A_{12}B_{21}+A_{13}B_{31} \\ A_{21}B_{11}+A_{22}B_{21}+A_{23}B_{31} \end{pmatrix}$.

Exemple 2.22. Soient $A = \begin{pmatrix} 2 & 1 & 3 \end{pmatrix}$ et $B = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$. Alors

$$AB = \begin{pmatrix} 5 \end{pmatrix} \quad \text{et} \quad BA = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 0 & 0 \\ 2 & 1 & 3 \end{pmatrix}.$$

Attardons-nous à présent sur une observation importante : **le produit matriciel n'est pas commutatif**. Autrement dit, on a $AB \neq BA$ en général ! Premièrement, ce n'est pas parce qu'on peut former le produit AB qu'on peut aussi former le produit BA. Par exemple, si A est de taille 2×3 et B est de taille 3×7, alors le produit AB a du sens alors que le produit BA n'en a pas. Deuxièmement, même dans le cas où l'on peut former les deux produits AB et BA, l'égalité $AB = BA$ peut ne pas être vérifiée. Dans l'exemple précédent, on peut former les produits AB et BA, mais AB est une matrice de taille 1×1 tandis que BA est une matrice de taille 3×3. Et troisièmement, même si les deux produits AB et BA ont du sens et sont des matrices de même taille (ceci se produit uniquement lorsque A et B sont des matrices carrées), on n'a pas nécessairement $AB = BA$ non plus. En effet, on a par exemple $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ et $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$.
Après cette remarque négative, étudions à présent les propriétés du produit matriciel.

Proposition 2.23. *Soient λ, μ des nombres complexes et A, B, C des matrices. Si les opérations suivantes ont du sens (c'est-à-dire si les matrices ont des tailles compatibles pour pouvoir effectuer les produits matriciels), alors*

1. $\lambda(\mu A) = (\lambda\mu)A$ — *associativité des produits*
   $(\lambda A)B = A(\lambda B) = \lambda(AB)$
2. $(AB)C = A(BC)$ — *associativité du produit matriciel*
3. $A(B + C) = AB + AC$ — *distributivité du produit matriciel sur la somme*
   $(A + B)C = AC + BC$
4. $(AB)^\intercal = B^\intercal A^\intercal$ — *transposée d'un produit matriciel*
5. $AI = IA = A$ — *I est neutre pour le produit matriciel de matrices carrées*
6. $A0 = 0$ *et* $0A = 0$ — *0 est absorbant pour le produit matriciel (quelconque)*

Preuve. Démontrons l'associativité du produit matriciel. Soient $A \in \mathbb{C}^{\ell \times c}, B \in \mathbb{C}^{c \times c'}, C \in \mathbb{C}^{c' \times c''}$ et soient $i \in \{1, \ldots, \ell\}$ et $j \in \{1, \ldots, c''\}$. D'une part, on a

$$\big((AB)C\big)_{ij} = \sum_{k=1}^{c'} (AB)_{ik}C_{kj}$$

$$= \sum_{k=1}^{c'} \Big(\sum_{n=1}^{c} A_{in} B_{nk} \Big) C_{kj}$$
$$= \sum_{k=1}^{c'} \sum_{n=1}^{c} A_{in} B_{nk} C_{kj}.$$

D'autre part,

$$\big(A(BC)\big)_{ij} = \sum_{n=1}^{c} A_{in} (BC)_{nj}$$
$$= \sum_{n=1}^{c} A_{in} \Big(\sum_{k=1}^{c'} B_{nk} C_{kj} \Big)$$
$$= \sum_{n=1}^{c} \sum_{k=1}^{c'} A_{in} B_{nk} C_{kj}.$$

On a donc bien $\big((AB)C\big)_{ij} = \big(A(BC)\big)_{ij}$ comme souhaité.

Montrons également que si $A \in \mathbb{C}^{\ell \times c}$, alors $AI_c = A$. Par définition de la matrice identité, on a $(I_c)_{kj} = 1$ si $k = j$ et $(I_c)_{kj} = 0$ si $k \neq j$. Par conséquent, pour tous $i \in \{1, \ldots, \ell\}$ et $j \in \{1, \ldots, c\}$, nous avons

$$(AI_c)_{ij} = \sum_{k=1}^{c} A_{ik} (I_c)_{kj} = A_{ij}.$$

□

Cette proposition permet de donner du sens aux écritures $\lambda\mu A$, λAB et ABC.

Voici maintenant un exemple concret d'application du calcul matriciel. Une matrice comme celle de cet exemple s'appelle une *matrice technologique*.

Exemple 2.24. Une entreprise produit quatre sortes d'articles (output) A_1, A_2, A_3 et A_4, ce pour quoi elle utilise trois sortes d'input : matières premières, énergie et main d'œuvre. On suppose que la quantité d'articles produits est proportionnelle à la quantité d'input fourni. On peut représenter cette situation au moyen de la table suivante :

	A_1	A_2	A_3	A_4
matières premières	1	6	1	4
énergie	0	1	2	2
main-d'œuvre	1	1	1	2

où le nombre situé à l'intersection de la i-ème ligne et la j-ème colonne représente la quantité de l'input correspondant à la ligne i qui est nécessaire pour produire une unité de l'article A_j. Considérons à présent D la matrice à 3 lignes et 4 colonnes obtenues à partir de ces données :

$$D = \begin{pmatrix} 1 & 6 & 1 & 4 \\ 0 & 1 & 2 & 2 \\ 1 & 1 & 1 & 2 \end{pmatrix}.$$

Si Q est une matrice-colonne à 4 éléments donnant la quantité commandée de chacun des 4 articles et si P est une matrice-ligne à 3 éléments donnant le coût de chaque unité d'input, alors le prix total de la commande en question est le produit PDQ :

$$\begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} 1 & 6 & 1 & 4 \\ 0 & 1 & 2 & 2 \\ 1 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix}$$

$$= \begin{pmatrix} p_1 + p_3 & 6p_1 + p_2 + p_3 & p1 + 2p2 + p3 & 4p_1 + 2p_2 + 2p_3 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix}$$

$$= (p_1 + p_3)q_1 + (6p_1 + p_2 + p_3)q_2 + (p1 + 2p2 + p3)q_3 + (4p_1 + 2p_2 + 2p_3)q_4.$$

Inverse d'une matrice

Lemme 2.25. *Si A, B, C sont des matrices carrées de même taille telles que $BA = AC = I$, alors $B = C$.*

Preuve. Supposons que $BA = AC = I$. Alors $B = BI = BAC = IC = C$. □

Cette propriété s'énonce également comme suit : si A admet un inverse *à gauche* B et un inverse *à droite* C, alors nécessairement $B = C$.

Définition 2.26. Une matrice carrée A est dite *inversible* s'il existe une matrice carrée B de même taille que A telle que $AB = BA = I$. Au vu du lemme 2.25, il ne peut exister qu'une seule telle matrice B et celle-ci est appelée la matrice inverse de A. Lorsqu'elle existe, la matrice inverse de A est notée A^{-1}.

Comparez la définition de l'inverse d'une matrice avec la définition de l'inverse modulaire vu précédemment.

Exemple 2.27. On a

$$\begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix}^{-1} = \frac{1}{5} \begin{pmatrix} -1 & 2 \\ -3 & 1 \end{pmatrix}$$

car

$$\begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix} \cdot \frac{1}{5} \begin{pmatrix} -1 & 2 \\ -3 & 1 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} -1 & 2 \\ -3 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

2.3 Déterminant d'une matrice carrée

Nous allons voir que l'inverse d'une matrice n'existe pas toujours. Plus précisément, nous allons déterminer dans quelles conditions une matrice admet un inverse. Pour cela, nous avons besoin de définir une notion importante : le déterminant. Dans cette section, nous fixons une matrice carrée A de taille $m \times m$.

Définition 2.28. Le *déterminant* de A est un nombre, noté $\det(A)$ ou $|A|$, que l'on définit récursivement comme suit :

1. Si $m = 1$, alors $\det(A) = A_{11}$.
2. Si $m > 1$, alors
$$\det(A) = \sum_{k=1}^{m} A_{1k}\, \mathcal{C}_{1k},$$
où $\mathcal{C}_{1k} := (-1)^{1+k} \det(M_{1k})$, avec M_{1k} la matrice $(m-1) \times (m-1)$ obtenue en supprimant la première ligne et la k-ième colonne de A.

On appelle ce procédé le *calcul du déterminant par développement suivant la première ligne.*

Exemple 2.29. On a

$$\begin{vmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{vmatrix} = 4 \cdot (-1)^{1+1} \cdot \begin{vmatrix} 8 & 9 \\ 11 & 12 \end{vmatrix} + 5 \cdot (-1)^{1+2} \cdot \begin{vmatrix} 7 & 9 \\ 10 & 12 \end{vmatrix} + 6 \cdot (-1)^{1+3} \cdot \begin{vmatrix} 7 & 8 \\ 10 & 11 \end{vmatrix}.$$

Définition 2.30. Le *mineur* de l'élément A_{ij} est le déterminant de la matrice $M_{ij}(A)$ de taille $(m-1) \times (m-1)$ obtenue en supprimant de A la i-ième ligne et la j-ième colonne. Le nombre $\mathcal{C}_{ij}(A) := (-1)^{i+j} \det(M_{ij}(A))$ est appelé le *cofacteur* de l'élément A_{ij}. Si la matrice A est clairement identifiée, on s'autorise à écrire simplement M_{ij} et $\mathcal{C}_{ij}$.

En réalité, la définition du déterminant est indépendante de la ligne choisie. Nous admettons l'important résultat suivant.

Fait 2.31 (Première loi des mineurs pour les lignes)**.** *Pour chaque $i \in \{1, \ldots, m\}$, on a*

$$\sum_{k=1}^{m} A_{ik}\, \mathcal{C}_{ik} = \det(A).$$

On appelle ce procédé le *calcul du déterminant par développement suivant la i-ème ligne.*

Et si on se trompait de ligne pour calculer les cofacteurs ? Et bien, cela change tout ! Dans ce cas, on obtient toujours la même réponse : 0. En effet, on a également l'important résultat suivant, que nous démontrons pas non plus.

Fait 2.32 (Deuxième loi des mineurs pour les lignes)**.** *Soient $i, i' \in \{1, \ldots, m\}$ tels que $i \neq i'$. Alors*

$$\sum_{k=1}^{m} A_{ik}\, \mathcal{C}_{i'k} = 0.$$

On peut également calculer le déterminant en calculant les mineurs relatifs aux éléments d'une colonne. À nouveau, nous admettons ce résultat.

Fait 2.33 (Première loi des mineurs pour les colonnes)**.** *Pour chaque $j \in \{1, \ldots, m\}$, on a*

$$\sum_{k=1}^{m} A_{kj}\, \mathcal{C}_{kj} = \det(A).$$

On appelle ce procédé le *calcul du déterminant par développement suivant la j-ème colonne.*

On a également le résultat analogue de la proposition 2.32 dans le cas des colonnes [3].

Fait 2.34 (Deuxième loi des mineurs pour les colonnes)**.** *Soient $j, j' \in \{1, \ldots, m\}$ tels que $j \neq j'$. Alors*

$$\sum_{k=1}^{m} A_{kj}\, \mathcal{C}_{kj'} = 0.$$

Grâce aux lois des mineurs, nous obtenons les propriétés suivantes du déterminant.

Proposition 2.35. *Si A possède deux lignes ou deux colonnes identiques, alors* $\det(A) = 0$.

Preuve. Nous faisons la démonstration dans le cas des lignes, celle sur les colonnes se faisant de façon similaire. Supposons que les lignes i et i' de A soient identiques, avec $i \neq i'$. Cela signifie que pour tout $k \in \{1, \ldots, m\}$, on a $A_{ik} = A_{i'k}$. Au vu des lois des mineurs sur les lignes, on a

$$\det(A) = \sum_{k=1}^{m} A_{ik}\, \mathcal{C}_{ik} = \sum_{k=1}^{m} A_{i'k}\, \mathcal{C}_{ik} = 0.$$

□

3. On peut aussi se tromper de colonne !

Proposition 2.36. *Pour toute matrice carrée A, on a* $\det(A) = \det(A^\intercal)$.

Preuve. On procède par récurrence sur la taille des matrices. Le résultat est évident pour les matrices de taille 1×1. Supposons à présent que $n > 1$ et que le résultat soit vrai pour toutes les matrices de taille $m \times m$ avec $m < n$. Soit A une matrice de taille $n \times n$. Remarquez que si M_{ij} désigne la matrice obtenue en supprimant la i-ième ligne et la j-ième colonne de A, alors $(M_{ij})^\intercal$ désigne la matrice obtenue en supprimant la j-ième ligne et la i-ième colonne de $A^\intercal$. Autrement dit, on a $(M_{ij}(A))^\intercal = M_{ji}(A^\intercal)$. Ainsi, par hypothèse de récurrence (utilisée à la deuxième égalité), pour tous $i, j \in \{1, \dots, n\}$, on a

$$\begin{aligned}\mathcal{C}_{ij}(A) &= (-1)^{i+j} \det(M_{ij}(A)) \\ &= (-1)^{i+j} \det((M_{ij}(A))^\intercal) \\ &= (-1)^{i+j} \det(M_{ji}(A^\intercal)) \\ &= \mathcal{C}_{ji}(A^\intercal).\end{aligned}$$

On obtient

$$\det(A^\intercal) \;=\; \sum_{k=1}^{n} (A^\intercal)_{1k}\, \mathcal{C}_{1k}(A^\intercal) \;=\; \sum_{k=1} A_{k1}\, \mathcal{C}_{k1}(A) \;=\; \det(A)$$

comme annoncé. □

La proposition précédente a pour conséquence que toute propriété du déterminant sur les lignes d'une matrice est également valable pour les colonnes, et vice-versa.

La conséquence suivante des lois des mineurs exprime la multilinéarité du déterminant sur les lignes et les colonnes.

Proposition 2.37.

1. *Soient $i \in \{1, \dots, m\}$, $L_1, \dots, L_{i-1}, L_{i+1}, \dots, L_m, M, N, \in \mathbb{C}^{1\times m}$ et $a, b \in \mathbb{C}$. Alors*

$$\det \begin{pmatrix} L_1 \\ \vdots \\ L_{i-1} \\ aM + bN \\ L_{i+1} \\ \vdots \\ L_m \end{pmatrix} = a \cdot \det \begin{pmatrix} L_1 \\ \vdots \\ L_{i-1} \\ M \\ L_{i+1} \\ \vdots \\ L_m \end{pmatrix} + b \cdot \det \begin{pmatrix} L_1 \\ \vdots \\ L_{i-1} \\ N \\ L_{i+1} \\ \vdots \\ L_m \end{pmatrix}.$$

2. *Soient $j \in \{1, \dots, m\}$, $C_1, \dots, C_{j-1}, C_{j+1}, \dots, C_m, D, E \in \mathbb{C}^{m\times 1}$ et $a, b \in \mathbb{C}$. Alors*

$$\begin{aligned}&\det\left(C_1 \cdots C_{j-1}\ aD + bE\ C_{j+1} \cdots C_m\right) \\ &= a \cdot \det\left(C_1 \cdots C_{j-1}\ D\ C_{j+1} \cdots C_m\right) + b \cdot \det\left(C_1 \cdots C_{j-1}\ E\ C_{j+1} \cdots C_m\right).\end{aligned}$$

Preuve. Il suffit de calculer le déterminant en développant sur la i-ième ligne ou la j-ième colonne. □

Exemple 2.38. On a toujours

$$\begin{vmatrix} a & b+c & d \\ e & f+g & h \\ i & j+k & \ell \end{vmatrix} = \begin{vmatrix} a & b & d \\ e & f & h \\ i & j & \ell \end{vmatrix} + \begin{vmatrix} a & c & d \\ e & g & h \\ i & k & \ell \end{vmatrix}.$$

Corollaire 2.39. *Soient $A \in \mathbb{C}^{m\times m}$ et $\lambda \in \mathbb{C}$. Alors*

$$\det(\lambda A) = \lambda^m \det(A).$$

Preuve. C'est une conséquence directe de la multilinéarité du déterminant. □

Exemple 2.40. On a toujours

$$\begin{vmatrix} 3a & 3b & 3c \\ 3d & 3e & 3f \\ 3g & 3h & 3i \end{vmatrix} = 27 \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}.$$

Proposition 2.41. *Le déterminant d'une matrice carrée vaut 0 si une ligne est une combinaison linéaire des autres lignes ou si une colonne est une combinaison linéaire des autres colonnes.*

Preuve. Par la proposition 2.35, nous savons que le déterminant d'une matrice qui possède deux lignes ou deux colonnes identiques vaut 0. Le résultat découle alors de la multilinéarité du déterminant sur les lignes et les colonnes. □

Exemple 2.42. Au vu de la proposition précédente, on a toujours

$$\begin{vmatrix} a & b & c \\ 3a+2d & 3b+2e & 3c+2f \\ d & e & f \end{vmatrix} = 0.$$

Le calcul suivant illustre la preuve de cette proposition :

$$\begin{vmatrix} a & b & c \\ 3a+2d & 3b+2e & 3c+2f \\ d & e & f \end{vmatrix} = 3 \underbrace{\begin{vmatrix} a & b & c \\ a & b & c \\ d & e & f \end{vmatrix}}_{=0} + 2 \underbrace{\begin{vmatrix} a & b & c \\ d & e & f \\ d & e & f \end{vmatrix}}_{=0} = 0.$$

Corollaire 2.43. *Le déterminant d'une matrice carrée est inchangé si l'on ajoute à une ligne une combinaison linéaire des <u>autres</u> lignes. De même, le déterminant d'une matrice carrée est inchangé si l'on ajoute à une colonne une combinaison linéaire des <u>autres</u> colonnes.*

Preuve. Ceci découle des propositions 2.37 et 2.41. □

En fait, la réciproque de la proposition 2.41 est vraie également et nous pouvons donc formuler le critère suivant (que nous admettons pour moitié, donc).

Fait 2.44. *Soit $A \in \mathbb{C}^{m\times m}$. Les assertions suivantes sont équivalentes.*

1. *Le déterminant de A vaut 0.*
2. *Une ligne de A est combinaison linéaire des autres.*
3. *Une colonne de A est combinaison linéaire des autres.*

Voici quelques remarques concernant le calcul du déterminant.

1. La matrice obtenue en supprimant de A la $i$ème ligne et la j-ème colonne est

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1j-1} & A_{1j+1} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2j-1} & A_{2j+1} & \cdots & A_{2m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ A_{i-11} & A_{i-12} & \cdots & A_{i-1j-1} & A_{i-1j+1} & \cdots & A_{i-1,m} \\ A_{i+11} & A_{i+12} & \cdots & A_{i+1j-1} & A_{i+1j+1} & \cdots & A_{i+1,m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{m,j-1} & A_{m,j+1} & \cdots & A_{mm} \end{pmatrix}$$

2. L'expression $(-1)^{i+j}$ vaut $+1$ ou -1 selon que la puissance $i+j$ est paire ou impaire. En particulier quand on passe du coefficient A_{ij} au coefficient suivant $A_{i,j+1}$, le signe change. Cela donne une répartition des coefficients $(-1)^{i+j}$ dans une matrice en damier.
$$\begin{pmatrix} +1 & -1 & +1 & \cdots \\ -1 & +1 & -1 & \cdots \\ +1 & -1 & +1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$
3. On a intérêt à développer le déterminant selon, si possible, une ligne ou une colonne avec un grand nombre de 0 ou dont les coefficients sont relativement simples.
4. On a intérêt à d'abord regarder si on peut se servir des propositions 2.37 et 2.41 et du corollaire 2.43. Lorsque c'est le cas, les calculs du déterminant sont grandement facilités.

Étudions à présent quelques cas particuliers où le calcul du déterminant est facilité.

1. Lorsque A est une matrice carrée de taille 2×2, on a $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ et
$$\det(A) = (-1)^{1+1} A_{11} A_{22} + (-1)^{1+2} A_{12} A_{21} = A_{11} A_{22} - A_{12} A_{21}.$$
On parle de produit "en croix".
2. Lorsque A est une matrice carrée de taille 3×3, on a $A = \left(\begin{smallmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{smallmatrix}\right)$. En développant suivant la première ligne, on trouve
$$\begin{aligned} \det(A) &= (-1)^{1+1} A_{11} \left|\begin{smallmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{smallmatrix}\right| + (-1)^{1+2} A_{12} \left|\begin{smallmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{smallmatrix}\right| + (-1)^{1+3} A_{13} \left|\begin{smallmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{smallmatrix}\right| \\ &= A_{11}(A_{22}A_{33} - A_{23}A_{32}) - A_{12}(A_{21}A_{33} - A_{23}A_{31}) + A_{13}(A_{21}A_{32} - A_{22}A_{31}) \\ &= A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{13}A_{21}A_{32} - A_{11}A_{23}A_{32} - A_{12}A_{21}A_{33} \\ &\quad - A_{13}A_{22}A_{31}. \end{aligned}$$
On parle de la *règle de Sarrus*. Attention que celle-ci n'est pas valide en dimension supérieure !
3. Lorsque A est une matrice triangulaire (inférieure ou supérieure), son déterminant est égal au produit des coefficients situés sur la diagonale principale :
$$\det \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix} = A_{11} \cdot A_{22} \cdots A_{mm}.$$
et
$$\det \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ 0 & A_{22} & & A_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & A_{mm} \end{pmatrix} = A_{11} \cdot A_{22} \cdots A_{mm}.$$
En particulier, on a $\det(I_m) = 1$.

Exemple 2.45. Considérons la matrice
$$A = \begin{pmatrix} 1 & 0 & 0 & 4 \\ 0 & 5 & 6 & 0 \\ 0 & 0 & 3 & 0 \\ 2 & 0 & 0 & 7 \end{pmatrix}.$$

Comme la troisième ligne ne comprend qu'un seul coefficient non nul, il est judicieux de choisir de développer le déterminant selon celle-ci (on aurait pu tout aussi bien choisir la deuxième colonne pour la même raison). On obtient

$$\det(A) = (-1)^{3+3} \cdot 3 \cdot \begin{vmatrix} 1 & 0 & 4 \\ 0 & 5 & 0 \\ 2 & 0 & 7 \end{vmatrix}.$$

On développe à présent suivant la deuxième ligne (ou colonne) :

$$\det(A) = 3 \cdot 5 \cdot (-1)^{2+2} \cdot \begin{vmatrix} 1 & 4 \\ 2 & 7 \end{vmatrix} = 15(1 \cdot 7 - 4 \cdot 2) = 15 \cdot (-1) = -15.$$

Pour terminer cette section sur le déterminant, nous admettons l'important résultat suivant :

Fait 2.46 (Déterminant d'un produit de matrices carrées)**.** *Soient A, B des matrices carrées de même taille. Alors* $\det(AB) = \det(A)\det(B)$.

On énonce généralement le résultat précédent comme ceci : le déterminant d'un produit de matrices carrées est égal au produit des déterminants de ces matrices.

2.4 Lien entre déterminant et inverse

Comme annoncé, les notions d'inverse et de déterminant sont liées. En effet, le théorème suivant donne une caractérisation des matrices inversibles en fonction de leurs déterminants, ainsi qu'une formule de calcul de l'inverse.

Théorème 2.47. *Soit A une matrice carrée de taille $m \times m$. Alors A est inversible si et seulement si* $\det(A) \neq 0$*, auquel cas son inverse est donné par*

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \cdots & \mathcal{C}_{1m} \\ \mathcal{C}_{21} & \mathcal{C}_{22} & \cdots & \mathcal{C}_{2m} \\ \vdots & & & \vdots \\ \mathcal{C}_{m1} & \mathcal{C}_{m2} & \cdots & \mathcal{C}_{mm} \end{pmatrix}^{\mathsf{T}}.$$

Preuve. Supposons tout d'abord que A est inversible. Il existe donc une matrice B de taille $m \times m$ telle que $AB = BA = I_m$. En utilisant le fait 2.46, on a $\det(A)\det(B) = \det(AB) = \det(I_m) = 1$. On obtient ainsi que $\det(A) \neq 0$.

Montrons à présent la réciproque ainsi que la formule annoncée de l'inverse. Nous supposons donc que $\det(A) \neq 0$. Alors par les faits 2.31 et 2.32, la matrice A^{-1} de l'énoncé est telle que $AA^{-1} = I_m$. En effet, pour tous $i, j \in \{1, \ldots, m\}$, on a

$$\begin{aligned} (AA^{-1})_{ij} &= \sum_{k=1}^{m} A_{ik}(A^{-1})_{kj} \\ &= \frac{1}{\det(A)} \sum_{k=1}^{m} A_{ik}\mathcal{C}_{jk}. \end{aligned}$$

Ainsi, par le fait 2.31 (première loi des mineurs pour les lignes), on obtient $(AA^{-1})_{ij} = 1$ lorsque $i = j$ et par le fait 2.32 (deuxième loi des mineurs pour les lignes) on obtient $(AA^{-1})_{ij} = 0$ lorsque $i \neq j$. De la même façon, en utilisant les faits 2.33 et 2.34 (lois des mineurs pour les colonnes), on obtient que $A^{-1}A = I_m$. Ceci montre bien que A est inversible et que A^{-1} est son inverse. □

Exemple 2.48. Considérons la matrice

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & 4 \\ 0 & 5 & -1 \end{pmatrix}.$$

Alors on a

$$\det(A) = 1 \cdot (1 \cdot (-1) - 4 \cdot 5) + 2 \cdot (3 \cdot 5) = 9.$$

Donc A est inversible et son inverse est donné par

$$A^{-1} = \frac{1}{9} \begin{pmatrix} -21 & 3 & 15 \\ 10 & -1 & -5 \\ -2 & 2 & 1 \end{pmatrix}^{\intercal} = \begin{pmatrix} \frac{-21}{9} & \frac{10}{9} & \frac{-2}{9} \\ \frac{3}{9} & \frac{-1}{9} & \frac{2}{9} \\ \frac{15}{9} & \frac{-5}{9} & \frac{1}{9} \end{pmatrix}.$$

On vérifie bien que $AA^{-1} = I_3$ (alors $A^{-1}A$ est automatiquement vrai aussi).

Puisque nous avons vu l'arithmétique modulaire dans ce cours, nous sommes à même de comprendre l'énoncé suivant. La démonstration de ce résultat est une adaptation directe de celle du théorème 2.47.

Théorème 2.49. *Soit $m \geq 2$ un entier. Alors une matrice carrée A de taille $n \times n$ dont les éléments sont dans $\mathbb{Z}_m$ est inversible si et seulement si* $\det(A)$ *est inversible dans $\mathbb{Z}_m$, auquel cas son inverse est donné par*

$$A^{-1} = (\det(A))^{-1} \begin{pmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \cdots & \mathcal{C}_{1n} \\ \mathcal{C}_{21} & \mathcal{C}_{22} & \cdots & \mathcal{C}_{2n} \\ \vdots & & & \vdots \\ \mathcal{C}_{n1} & \mathcal{C}_{n2} & \cdots & \mathcal{C}_{nn} \end{pmatrix}^{\intercal},$$

où $(\det(A))^{-1}$ *désigne l'inverse de* $\det(A)$ *dans $\mathbb{Z}_m$.*

Si m est un nombre premier, nous savons que tout élément non nul de $\mathbb{Z}_m$ est inversible dans $\mathbb{Z}_m$, et nous obtenons donc directement le corollaire suivant. Un tel résultat illustre bien le fait que les structures de $\mathbb{C}$ et de $\mathbb{Z}_m$ lorsque m est premier sont très semblables.

Corollaire 2.50. *Soit m un nombre premier. Alors une matrice carrée A dont les éléments sont dans $\mathbb{Z}_m$ est inversible si et seulement si* $\det(A) \neq 0$.

2.5 Systèmes d'équations linéaires

Commençons par un exemple introductif, qui généralise l'exemple 2.24.

Exemple 2.51. On peut décrire un système de production de la façon suivante. À partir de ℓ types d'entrées (inputs), on produit c types de sorties (outputs). On assigne des unités de mesure aux différents inputs et outputs (pensez aux matrices technologiques introduites dans l'exemple 2.24). Soit A_{ij} la quantité de l'input i nécessaire pour produire une unité du produit j. On suppose deux choses :

1. Pour produire x_j unités du produit j il faut $A_{ij}x_j$ unités de l'input i (hypothèse de rendement constant).
2. Pour produire $x_1, x_2, \ldots, x_c$ unités des différents outputs, il faut $A_{i1}x_1 + A_{i2}x_2 + \cdots + A_{ic}x_c$ unités de l'input i (hypothèse d'additivité).

La quantité $b_1, b_2, \ldots, b_\ell$ des différents inputs consommée pour fabriquer x_1, x_2, ..., x_c unités des différents outputs est donnée par les relations :

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1c}x_c &= b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2c}x_c &= b_2 \\ & \vdots \\ A_{\ell 1}x_1 + A_{\ell 2}x_2 + \cdots + A_{\ell c}x_c &= b_\ell \end{cases}$$

Ici les quantités supposées connues sont les nombres A_{ij} et b_i et les inconnues sont les quantités x_j que l'on peut produire à partir des inputs fournis.

Définition 2.52. L'ensemble des équations ci-dessus est appelé un *système de ℓ équations linéaires à c inconnues*. Les nombres b_i sont *les termes indépendants*. Si $b_i = 0$ pour tout i, le système est dit *homogène*. *Résoudre* le système signifie trouver l'ensemble de toutes les *solutions* du système, c'est-à-dire l'ensemble de tous les c-uplets $(x_1, x_2, \ldots, x_c) \in \mathbb{C}^c$ pour lesquels les égalités ci-dessus sont vérifiées. Le système est dit *compatible* s'il admet au moins une solution et *incompatible* (ou *impossible*) sinon. Deux systèmes linéaires sont *équivalents* s'ils possèdent le même ensemble de solutions. On écrit $S_1 \iff S_2$ pour signifier que les systèmes linéaires S_1 et S_2 sont équivalents.

Le système précédent peut s'écrire au moyen de matrices. On écrit les c inconnues sous forme de matrice-colonne $\begin{pmatrix} x_1 & x_2 & \cdots & x_c \end{pmatrix}^\intercal$ de taille $c \times 1$ et les ℓ termes indépendants sous forme de matrice-colonne $\begin{pmatrix} b_1 & b_2 & \cdots & b_\ell \end{pmatrix}^\intercal$ de taille $\ell \times 1$. On écrit les nombres A_{ij} qui relient input – output sous forme d'une matrice $\ell \times c$, appelée la *matrice des coefficients du système* :

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1c} \\ A_{21} & A_{22} & \cdots & A_{2c} \\ \vdots & \vdots & & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & A_{\ell c} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_\ell \end{pmatrix}.$$

Ainsi, le système peut s'écrire sous forme d'une équation matricielle :

$$Ax = b.$$

Une *solution* du système est alors une matrice-colonne $x \in \mathbb{C}^{c \times 1}$ telle que l'égalité $Ax = b$ est vérifiée.

Exemple 2.53. On voudrait écrire l'élément $(1, 2, 3)$ de $\mathbb{C}^3$ comme combinaison linéaire des éléments $\{(1, 0, -1), (0, 1, 1), (1, -1, 0)\}$, c'est-à-dire trouver des nombres complexes a, b, c tels que

$$a\,(1, 0, -1) + b\,(0, 1, 1) + c\,(1, -1, 0) = (1, 2, 3).$$

Cette égalité de vecteurs à trois composantes est équivalente au système de trois équations à trois inconnues suivant :

$$\begin{cases} a + c &= 1 \\ b - c &= 2 \\ -a + b &= 3 \end{cases}$$

qui, lui-même, s'écrit sous forme matricielle comme suit :

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Posons $A = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \end{smallmatrix}\right)$, $x = \left(\begin{smallmatrix} a \\ b \\ c \end{smallmatrix}\right)$ et $b = \left(\begin{smallmatrix} 1 \\ 2 \\ 3 \end{smallmatrix}\right)$. On doit donc résoudre l'équation matricielle $Ax = b$. Comme $\det(A) = 2 \neq 0$, la matrice A est inversible et en multipliant chaque membre de l'équation $Ax = b$ à gauche par A^{-1}, on obtient

$$A^{-1}Ax = A^{-1}b,$$

c'est-à-dire

$$x = A^{-1}b.$$

On a donc résolu le système et obtenu qu'il y a une unique solution $A^{-1}b$. Il ne reste plus qu'à la calculer. On calcule d'abord la matrice inverse de A :

$$A^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix}.$$

Ensuite, on obtient que l'unique solution est

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix}.$$

Définition 2.54. Un système d'équations linéaires est appelé un *système de Cramer* si le nombre d'équations est égal au nombre d'inconnues et si la matrice des coefficients est inversible.

Remarquez que la matrice des coefficients d'un système de Cramer est toujours carrée.

Théorème 2.55. *Un système de Cramer admet une solution unique.*

Preuve. En effet, puisque A est inversible, on a $Ax = b \iff x = A^{-1}b$. □

Corollaire 2.56. *Un système de Cramer homogène $Ax = 0$ admet pour unique solution la matrice-colonne* 0.

Exemple 2.57. Le système linéaire

$$\begin{cases} 2x + 6y &= 0 \\ x + 13y &= 0 \end{cases}$$

est de Cramer car $\left|\begin{smallmatrix} 2 & 6 \\ 1 & 13 \end{smallmatrix}\right| = 20 \neq 0$. Il n'admet donc que la solution nulle $(x, y) = (0, 0)$.

Pour résoudre un système de Cramer, on peut toujours procéder comme suit. On calcule A^{-1}, l'inverse de la matrice des coefficients A et on effectue le produit $A^{-1}b$ de A^{-1} avec la matrice-colonne des termes indépendants. Cette méthode de la matrice inverse est toutefois assez difficile en pratique car elle nécessite le calcul d'un grand nombre de déterminants. La méthode de Gauss décrite plus loin mène à des algorithmes de résolution des systèmes linéaires beaucoup plus efficaces.

Définition 2.58. Un système linéaire est *sous forme triangulaire* si sa matrice des coefficients est triangulaire, c'est-à-dire s'il est du type :

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1m}x_m &= b_1 \\ \quad\quad A_{22}x_2 + A_{23}x_3 + \cdots + A_{2m}x_m &= b_2 \\ \quad\quad\quad\quad A_{33}x_3 + \cdots + A_{3m}x_m &= b_3 \\ \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \vdots & \\ \quad\quad\quad\quad\quad\quad\quad\quad A_{mm}x_m &= b_m \end{cases}$$

Notons que comme le déterminant d'une matrice triangulaire est égal au produit des coefficients diagonaux, un système triangulaire est de Cramer si et seulement si les coefficients diagonaux sont tous non nuls. Un système de Cramer sous forme triangulaire est particulièrement facile à résoudre. En effet la dernière équation donne immédiatement x_m :

$$x_m = \frac{b_m}{A_{mm}}.$$

En remplaçant x_m par la valeur trouvée dans l'avant-dernière équation, on obtient x_{m-1} :

$$A_{m-1,m-1}x_{m-1} + A_{m-1,m}x_m = b_{m-1} \iff x_{m-1} = \frac{1}{A_{m-1,m-1}}(b_{m-1} - A_{m-1,m}\frac{b_m}{A_{mm}}).$$

En remontant de cette façon toutes les équations jusqu'à la première, on obtient successivement les valeurs de $x_m, x_{m-1}, \ldots, x_1$.

Exemple 2.59. Considérons le système

$$\begin{cases} -x_1 + x_2 - 3x_3 = 6 \\ 2x_2 + x_3 = 2 \\ -4x_3 = 0 \end{cases}$$

La dernière équation donne $x_3 = 0$, la deuxième donne $x_2 = 1$ et la première donne $x_1 = -5$. L'unique solution est donc $(x_1, x_2, x_3) = (-5, 1, 0)$.

2.6 Méthode de Gauss pour la résolution d'un système de Cramer

Les observations suivantes sont valables pour tous les systèmes linéaires :

1. On ne modifie pas les solutions d'un système en changeant l'ordre des lignes du système.
2. On ne modifie pas les solutions d'un système en multipliant une ligne du système (les deux membres de l'égalité) par un nombre complexe non nul.
3. On ne modifie pas les solutions d'un système en ajoutant à une ligne une combinaison linéaire des autres lignes. Autrement dit, pour tout $i \in \{1, \ldots, \ell\}$ et tout $\lambda \in \mathbb{C}$, on a l'équivalence

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1c}x_c = b_1 \\ \vdots \\ A_{i1}x_1 + A_{i2}x_2 + \cdots + A_{ic}x_c = b_i \\ \vdots \\ A_{\ell 1}x_1 + A_{\ell 2}x_2 + \cdots + A_{\ell c}x_n = b_\ell \end{cases}$$

$$\iff \begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1c}x_c = b_1 \\ \vdots \\ (A_{i1} + \lambda A_{11})x_1 + (A_{i2} + \lambda A_{12})x_2 + \cdots + (A_{in} + \lambda A_{1c})x_c = b_i + \lambda b_1 \\ \vdots \\ A_{\ell 1}x_1 + A_{\ell 2}x_2 + \cdots + A_{\ell c}x_c = b_\ell \end{cases}$$

La *méthode de Gauss*, également appelée *méthode du pivot*, est la procédure décrite ci-après. Étant donné un système de Cramer, on peut toujours se ramener à un système

triangulaire en procédant comme suit : on choisit une ligne dont le coefficient de x_1 est non nul et on la place en première place. Cela revient à supposer que $A_{11} \neq 0$.

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1m}x_m &= b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2m}x_m &= b_2 \\ & \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mm}x_m &= b_m \end{cases}$$

Ensuite on remplace la seconde ligne par elle-même plus un multiple de la première de manière à ce que le coefficient de x_1 dans la nouvelle seconde ligne, c'est-à-dire A'_{21}, soit nul : pour chaque $j \in \{1, \ldots, m\}$, on calcule

$$A'_{2j} = A_{2j} - \frac{A_{21}}{A_{11}} A_{1j} \quad \text{et} \quad b'_2 = b_2 - \frac{A_{21}}{A_{11}} b_1.$$

On procède ainsi jusqu'à avoir supprimé les coefficients de x_1 dans chaque ligne sauf la première. Ainsi, on obtient un système équivalent du type :

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1m}x_m &= b_1 \\ A'_{22}x_2 + \cdots + A'_{2m}x_m &= b'_2 \\ & \vdots \\ A'_{m2}x_2 + \cdots + A'_{mm}x_m &= b'_m \end{cases}$$

La matrice des coefficients a été modifiée, mais le déterminant n'a pas changé. On recommence de la même façon avec le système des $m-1$ équations à $m-1$ inconnues obtenu en ignorant la première ligne. On choisit donc une ligne pour laquelle le coefficient de x_2 est non nul (il y a en forcément une sinon le déterminant de la matrice de départ serait nul), et on la place en deuxième ligne. On procède ensuite comme précédemment pour supprimer les coefficients de x_2 de toutes les lignes suivantes. En continuant ce processus, on aboutit à un système de Cramer triangulaire équivalent au système de départ, pour lequel il est facile d'obtenir la solution.

Exemple 2.60. Considérons le système linéaire S suivant

$$\begin{cases} x + y - 2z = 1 \\ 2x + 3y + z = 4 \\ x + 4y + z = 1 \end{cases}$$

On obtient successivement

$$S \iff \begin{cases} x + y - 2z = 1 \\ y + 5z = 2 \\ 3y + 3z = 0 \end{cases} \iff \begin{cases} x + y - 2z = 1 \\ y + 5z = 2 \\ -12z = -6 \end{cases}$$

On en tire $z = \frac{1}{2}$, puis $y = 2 - 5z = -\frac{1}{2}$ et $x = 1 - y + 2z = \frac{5}{2}$. L'unique solution du système est donc $(x, y, z) = (\frac{5}{2}, -\frac{1}{2}, \frac{1}{2})$.

Remarquez qu'on n'a pas besoin de vérifier a priori qu'on est en présence d'un système de Cramer. En effet, la méthode de Gauss aboutira à un système triangulaire de Cramer si et seulement si le système S de départ est de Cramer (puisque les deux systèmes sont équivalents).

2.7 Rang et résolution d'un système quelconque

Nous commençons par définir la notion de rang d'une matrice, ainsi que celle de rang d'un système d'équations linéaires (quelconque).

Définition 2.61. Une *sous-matrice* S d'une matrice A est une matrice que l'on peut obtenir à partir de A en supprimant des lignes et des colonnes. Le *rang d'une matrice* A, noté $\mathrm{rg}(A)$, vaut m si la plus grande sous-matrice carrée de A ayant un déterminant non nul est de taille $m \times m$. Le *rang d'un système linéaire* $Ax = b$ est le rang de la matrice des coefficients A.

Proposition 2.62.

1. *Si A est une matrice carrée $m \times m$ de déterminant non nul, alors son rang vaut m.*
2. *Si une ligne d'une matrice A est une combinaison linéaire des autres lignes de A, alors la matrice obtenue en supprimant cette ligne a le même rang que A.*
3. *Si une colonne d'une matrice A est une combinaison linéaire des autres colonnes de A, alors la matrice obtenue en supprimant cette colonne a le même rang que A.*

Preuve. Montrons le premier point. Soit A est une matrice carrée $m \times m$ de déterminant non nul. Puisque A est de taille $m \times m$, on a forcément $\mathrm{rg}(A) \leq m$. Ensuite, puisque $\det(A) \neq 0$, on a également $\mathrm{rg}(A) \geq m$. En combinant ces deux inégalités, on obtient que $\mathrm{rg}(A) = m$.
Montrons à présent le deuxième point (le troisième s'obtient de façon similaire). Soit $A \in \mathbb{C}^{\ell \times c}$ de rang r. Notons $L_1, \ldots, L_\ell$ les lignes de A et supposons que L_i est une combinaison linéaire des autres lignes, avec $1 \leq i \leq \ell$. Il existe donc $\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_\ell \in \mathbb{C}$ tels que

$$L_i = \sum_{k \neq i} \lambda_k L_k. \tag{2.1}$$

On note A' la sous-matrice de A obtenue en enlevant la ligne L_i de A. On doit montrer que $\mathrm{rg}(A') = r$. Puisque A' est une sous-matrice de A, on a $\mathrm{rg}(A') \leq \mathrm{rg}(A) = r$. Soit S une sous-matrice de A de taille $r \times r$ de déterminant non nul. Si la ligne L_i ne traverse pas S, alors S est également une sous-matrice de A' et $\mathrm{rg}(A') \geq r$. Supposons à présent que L_i traverse S. En utilisant (2.1) et la multilinéarité du déterminant sur les lignes, on obtient que $\det(S) = \sum_{k \neq i} \lambda_k \det(S_k)$, où S_k est la matrice obtenue en remplaçant dans S les éléments correspondant à la ligne L_i par les éléments correspondant dans la ligne L_k. Puisque $\det(S) \neq 0$, il existe $k \neq i$ tel que $\det(S_k) \neq 0$. Par conséquent, la ligne L_k ne traverse pas la sous-matrice S de A, car sinon, on aurait deux lignes identiques dans S_k et son déterminant serait nul. Ceci implique que, à un échange de lignes près, S_k est une sous-matrice de A'. On a donc obtenu que, dans ce cas également, $\mathrm{rg}(A') \geq r$. En combinant les inégalités, on obtient que $\mathrm{rg}(A') = \mathrm{rg}(A)$, comme souhaité. □

Exemple 2.63. Si A est la matrice $\left(\begin{smallmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 2 & -1 \end{smallmatrix}\right)$, alors $\det(A) = 0$ mais le déterminant de la matrice $\left(\begin{smallmatrix} 4 & 6 \\ 2 & -1 \end{smallmatrix}\right)$ obtenue en supprimant la première ligne et la première colonne de A est non nul. Donc le rang de A vaut 2.

Nous allons à présent énoncer un critère qui permet de calculer le rang d'une matrice à moindre frais. On l'appelle la *règle des déterminants bordés*. Nous donnons préalablement une définition utile.

Définition 2.64. Soient S et S' deux sous-matrices d'une matrice A. On dit que S *borde* S' lorsque S' est une sous-matrice de S.

Remarquez que cette définition dépend bien de la place des éléments des sous-matrices S et S' parmi ceux de A. En effet, il se pourrait que deux choix différents de lignes et de colonnes de A produisent la même sous-matrice S'. Dans ce cas, une même sous-matrice S de A pourrait border S' ou ne pas border S' selon la place de S' dans A.

Exemple 2.65. Si A est la matrice $\left(\begin{smallmatrix} 3 & 4 & 0 \\ 2 & 2 & 0 \\ 1 & 1 & 0 \end{smallmatrix}\right)$, alors la sous-matrice $S' = \left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right)$ peut être obtenue de deux façons différentes : soit en supprimant la première ligne et les deuxième et troisième colonnes, soit en supprimant la première ligne et les première et troisième colonnes. Dans le premier cas, on dira que la sous-matrice $S = \left(\begin{smallmatrix} 3 & 0 \\ 2 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ de A borde S', et dans le deuxième cas, on dira que la même sous-matrice S ne borde pas S'.

Nous admettons le résultat pratique suivant.

Fait 2.66 (Règle des déterminants bordés)**.** *Une matrice A est de rang r si et seulement si les deux conditions suivantes sont satisfaites.*

1. *Il existe une sous-matrice carrée S de A de taille $r \times r$ ayant un déterminant non nul.*
2. *Toutes les sous-matrices de A de taille $(r+1) \times (r+1)$ qui bordent S ont un déterminant nul.*

Nous admettons également le fait suivant.

Fait 2.67. *Soit A une matrice de rang r et soit A' une sous-matrice de A de taille $r \times r$ de déterminant non nul. Toute ligne de A est combinaison linéaire des lignes de A qui traversent A'.*

Définition 2.68. Soient $A \in \mathbb{C}^{\ell \times c}$ et $b \in \mathbb{C}^{\ell \times 1}$. La *matrice augmentée* $A|b$ est la matrice

$$A|b = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1c} & b_1 \\ A_{21} & A_{22} & \cdots & A_{2c} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ A_{\ell 1} & A_{\ell 2} & \cdots & A_{\ell c} & b_\ell \end{pmatrix}.$$

obtenue en ajoutant la colonne b à la matrice A.

Nous donnons le résultat suivant sans démonstration. Celui-ci est non seulement utile d'un point de vue théorique, mais il peut également être utilisé pour décrire une méthode générale de résolution des systèmes linéaires. Celle-ci sera principalement basée sur la recherche de l'inverse d'une sous-matrice matrice $r \times r$ de la matrice A des coefficients, si r est le rang du système. Il est aussi utile pour déterminer rapidement dans quels cas des systèmes linéaires contenant des paramètres possèdent ou non des solutions (sans nécessairement devoir les donner). En effet, il se peut que pour certaines valeurs de ces paramètres, le système soit impossible, que pour d'autres, le système soit de Cramer, et que pour d'autres encore, le système possède une infinité de solutions.

Fait 2.69 (Critère de compatibilité d'un système linéaire)**.** *Les conditions suivantes sont équivalentes.*

1. *Le système linéaire $Ax = b$ est compatible.*
2. *Le rang de A est égal au rang de la matrice augmentée $A|b$.*

Une important conséquence du fait 2.44 et du critère de compatibilité est que si un système possède au moins une solution (ce que l'on peut tester à l'aide du critère de compatibilité), alors on peut, pour le résoudre, ne conserver que r équations bien choisies, où r est le rang du système. C'est ce qu'exprime le résultat suivant.

Proposition 2.70. *Un système linéaire compatible $Ax = b$ de rang r est équivalent à tout système $A'x = b'$ obtenu en sélectionnant r lignes de telle sorte que* $\mathrm{rg}(A') = r$.

Preuve. Soient $Ax = b$ et $A'x = b'$ deux systèmes comme dans l'énoncé. On suppose de plus que A est de taille $\ell \times c$. Nous devons montrer que ces deux systèmes sont équivalents. Toute solution de $Ax = b$ est aussi solution de $A'x = b'$ puisque retirer des équations revient à retirer des contraintes. Comme le système $Ax = b$ est compatible par hypothèse, on obtient que le système $A'x = b'$ l'est aussi. Puisque $\mathrm{rg}(A') = r$, le critère de compatibilité nous donne donc que $\mathrm{rg}(A'|b') = r$. Quitte à permuter les lignes de $A|b$, on peut supposer que $A'|b'$ est obtenu en sélectionnant les r premières lignes de $A|b$. Par le fait 2.67, les lignes de $A|b$ sont combinaisons linéaires des r premières lignes de $A'|b'$. Ainsi, pour tout $i \in \{1, \ldots, \ell\}$, il existe $\lambda_1^{(i)}, \ldots, \lambda_r^{(i)} \in \mathbb{C}$ tels que pour tout $j \in \{1, \ldots, c\}$, on a

$$A_{ij} = \sum_{k=1}^{r} \lambda_k^{(i)} A_{kj} \quad \text{et} \quad b_i = \sum_{k=1}^{r} \lambda_k^{(i)} b_k. \tag{2.2}$$

Montrons à présent que toute solution de $A'x = b'$ est aussi solution de $Ax = b$. Soit $x \in \mathbb{C}^{c\times 1}$ une solution de $A'x = b'$. Cela signifie que pour tout $i \in \{1, \ldots, r\}$, on a

$$\sum_{j=1}^{c} A_{ij} x_j = b_i. \tag{2.3}$$

Nous devons montrer que l'égalité (2.3) est vérifiée pour tout $i \in \{1, \ldots, \ell\}$. Soit donc $i \in \{1, \ldots, \ell\}$. En utilisant les égalités (2.2) et (2.3) (pour des indices convenables), on obtient successivement

$$\sum_{j=1}^{c} A_{ij} x_j = \sum_{j=1}^{c} \sum_{k=1}^{r} \lambda_k^{(i)} A_{kj} x_j = \sum_{k=1}^{r} \lambda_k^{(i)} \sum_{j=1}^{c} A_{kj} x_j = \sum_{k=1}^{r} \lambda_k^{(i)} b_k = b_i.$$

□

2.8 Méthode de Gauss pour la résolution d'un système quelconque

D'un point de vue algorithmique, la méthode de résolution fournie par le critère de compatibilité des systèmes linéaires (fait 2.69) n'est pas la plus efficace. En pratique, on lui préférera la méthode de Gauss que nous avions décrite pour des systèmes de Cramer, et qui, nous allons le voir, s'adapte pour des systèmes linéaires quelconques.

On considère un système linéaire sans aucune restriction, et on garde les mêmes notations que précédemment. On choisit une ligne dont le coefficient de x_1 est non nul que l'on place tout en haut et que l'on utilise pour supprimer les coefficients de x_1 de toutes les lignes suivantes. Dans le cas où tous les coefficients de x_1 sont nuls [4], on choisit une inconnue x_i dont le coefficient dans la première ligne est non nul et on la renumérote x_1 (et x_1 devient x_i). Ensuite on choisit une autre ligne dans laquelle le coefficient de x_2 est non nul, que l'on place en deuxième place et qui sert à supprimer les coefficients de x_2 des lignes suivantes. On continue tant que l'on peut trouver une inconnue dans les lignes suivantes dont le coefficient est non nul. On arrive finalement à un système dont la partie supérieure gauche est sous forme triangulaire et dont les coefficients des lignes suivantes sont tous nuls (sinon

4. Ceci n'arrive pas si le système est de Cramer.

on pourrait continuer).

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1r}x_r + A_{1r+1}x_{r+1} + ... + A_{1c}x_c &= b_1 \\ & \vdots \\ A_{rr}x_r + A_{rr+1}x_{r+1} + \cdots + A_{rc}x_c &= b_r \\ 0 &= b_{r+1} \\ & \vdots \\ 0 &= b_\ell \end{cases}$$

Ainsi, le rang du système est égal à r, c'est-à-dire au nombre d'équations qui ne sont pas de la forme $0 = b_i$ [5]. Si au moins un des nombres $b_{r+1}, ..., b_\ell$ est non nul, alors le système est impossible. Sinon, le système est équivalent au système obtenu en supprimant toutes les équations de la forme $0 = 0$ et en déplaçant les inconnues $x_{r+1}, ..., x_c$ dans les seconds membres des r premières équations :

$$\begin{cases} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1r}x_r &= b_1 - A_{1r+1}x_{r+1} - \cdots - A_{1c}x_c \\ & \vdots \\ A_{rr}x_r &= b_r - A_{rr+1}x_{r+1} - \cdots - A_{rc}x_c \end{cases}$$

Nous trouvons alors facilement les solutions $(x_1, \ldots, x_r)$ qui dépendent de $x_{r+1}, \ldots, x_c$. Remarquez qu'avec cette méthode, le rang du système se calcule a posteriori (pas besoin de le pré-calculer pour savoir si le système est impossible ou non avant de commencer).

Exemple 2.71. On a successivement

$$\begin{cases} 2x - 3y + z = 3 \\ x + y - 2z = 4 \\ 4x - y - 3z = 11 \\ x - 4y + 3z = -1 \end{cases} \iff \begin{cases} x + y - 2z = 4 \\ 2x - 3y + z = 3 \\ 4x - y - 3z = 11 \\ x - 4y + 3z = -1 \end{cases} \iff \begin{cases} x + y - 2z = 4 \\ -5y + 5z = -5 \\ -5y + 5z = -5 \\ -5y + 5z = -5 \end{cases}$$

$$\iff \begin{cases} x + y - 2z &= 4 \\ y - z &= 1 \\ 0 &= 0 \\ 0 &= 0 \end{cases} \iff \begin{cases} x + y = 4 + 2z \\ y = 1 + z \end{cases}$$

Le système est donc de rang 2. On résout par rapport à x et y avec z comme paramètre, ce qui donne

$$\begin{cases} x + (1 + z) &= 4 + 2z \\ y &= 1 + z \end{cases} \iff \begin{cases} x &= 3 + z \\ y &= 1 + z \end{cases}$$

Les solutions sont donc $(x, y, z) = (3 + \lambda, 1 + \lambda, \lambda)$, avec $\lambda \in \mathbb{C}$.

Exemple 2.72. On a

$$\begin{cases} x + 2y - z = 1 \\ 2x - y + z = 4 \\ -3x - y = 5 \end{cases} \iff \begin{cases} x + 2y - z = 1 \\ -5y + 3z = 2 \\ 5y - 3z = 8 \end{cases} \iff \begin{cases} x + 2y - z &= 1 \\ -5y + 3z &= 2 \\ 0 &= 10 \end{cases}$$

Le système est de rang 2 et impossible.

5. Pourquoi ?

2.9 Méthode de Gauss pour le calcul de l'inverse d'une matrice

Nous avons vu que la méthode de Gauss, en plus de résoudre des systèmes linéaires, permet de calculer le rang d'une matrice. On peut également utiliser la méthode de Gauss pour calculer l'inverse d'une matrice, lorsque celui-ci existe, évidemment. En fait, il s'agit d'un des algorithmes les plus efficaces pour le calcul de l'inverse (et certainement bien plus efficace que le calcul de tous les cofacteurs !).
Si $A \in \mathbb{C}^{m\times m}$ est une matrice inversible, alors son inverse est l'unique solution $X \in \mathbb{C}^{m\times m}$ de l'équation

$$AX = I_m. \tag{2.4}$$

Si l'on note $x_1, \dots, x_m$ les colonnes de X, alors l'équation matricielle (2.4) est équivalente aux m systèmes linéaires

$$Ax_j = e_j,$$

où $j \in \{1, \dots, m\}$ et $e_j \in \mathbb{C}^{m\times 1}$ est la matrice-colonne possédant un 1 sur la ligne j et des 0 sur les autres lignes. Comme ces m systèmes possèdent la même matrice des coefficients A, on peut tous les résoudre simultanément en utilisant la méthode de Gauss.

Exemple 2.73. Soit la matrice

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{pmatrix}.$$

On a

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 3 & 0 & 0 & 1 \end{array}\right) \iff \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -3 & 1 & -2 & 0 & 1 \end{array}\right)$$
$$\iff \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & -\frac{3}{2} & 1 \end{array}\right)$$

Il nous reste à présent à résoudre 3 systèmes linéaires triangulaires. Pour la première colonne de X, on obtient successivement $X_{31} = -\frac{1}{2}$, $X_{21} = \frac{-1}{-2} = \frac{1}{2}$ et enfin $X_{11} = 1 - 2X_{21} - X_{31} = \frac{1}{2}$. Pour la deuxième colonne de X, on obtient $X_{32} = -\frac{3}{2}$, $X_{22} = \frac{1}{-2} = -\frac{1}{2}$ et $X_{12} = 0 - 2X_{22} - X_{32} = \frac{5}{2}$. Enfin, pour la troisième colonne de X, on obtient $X_{33} = 1$, $X_{23} = \frac{0}{-2} = 0$ et $X_{13} = 0 - 2X_{23} - X_{33} = -1$. On a donc obtenu que

$$A^{-1} = \begin{pmatrix} \frac{1}{2} & \frac{5}{2} & -1 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & -\frac{3}{2} & 1 \end{pmatrix}.$$

Évidemment, dans le cas d'une matrice 3×3, la méthode des cofacteurs est compétitive par rapport à la méthode de Gauss, mais plus la taille de la matrice augmente, plus le gain de temps apporté par la méthode de Gauss est important.

Tout comme pour le calcul du rang, la méthode de Gauss permet de voir a posteriori si une matrice est ou non inversible. Il ne faut donc pas savoir avant de commencer si elle est inversible ou non (ce qui serait dommage puisque cela imposerait un calcul de déterminant inutilement lourd).

Table des matières