

Notes de cours de DEA
2000 - 2001

Algorithmes stochastiques pour l'optimisation

Olivier François
LMC - IMAG, BP 53, 38041 Grenoble cedex 9
Olivier.Francois@imag.fr

October 30, 2003

Contents

1 Algorithmes de recherche aléatoire	5
1.1 Introduction	5
1.2 Exemples d'application	6
1.2.1 Voyageur de commerce	6
1.2.2 Sac-à-dos	6
1.2.3 Modèle de Ising et images	7
1.3 Recherche sur un graphe	8
1.3.1 Graphes	8
1.3.2 Recherche aléatoire	10
1.4 Dynamique de Metropolis	12
1.4.1 Gradient	12
1.4.2 Algorithme de Metropolis	13
1.5 Algorithmes de population	14
1.5.1 description	14
1.5.2 Algorithme de Mutation+sélection	15
1.5.3 Algorithme de Mutation/sélection	16
2 Généralités sur les chaînes de Markov	19
2.1 Définitions	19
2.2 Convergence	20
2.3 Temps d'atteinte	24
2.3.1 Matrice fondamentale	24
2.3.2 Temps d'atteinte et spectre : quelques remarques.	25
2.3.3 Discussion	27
3 Chaînes de Markov réversibles	29
3.1 Justification : un point de vue stochastique sur l'optimisation	29
3.2 Chaînes réversibles	30
3.2.1 Définitions	30
3.2.2 Structure Hilbertienne. Représentation spectrale	31
3.2.3 Convergence	32
3.2.4 Valeurs propres et formes de Dirichlet	34
3.3 Inégalités géométriques	35
3.3.1 Définitions	35

3.3.2	Bornes de Cheeger	37
3.3.3	Bornes de Poincaré	40
3.4	Application à la dynamique de Metropolis	42
3.4.1	Choix des chemins	42
3.4.2	Bornes sur λ_2	43
3.4.3	Borne sur ρ	45
3.5	Un problème	46
4	Chaînes de Markov à transitions rares	51
4.1	Définition	51
4.2	Lemmes sur les chaînes de Markov	51
4.2.1	Sous-graphes	51
4.2.2	Mesure invariante	52
4.2.3	Temps d'atteinte	52
4.3	Application aux chaînes à transitions rares	53
4.3.1	Mesure invariante	53
4.3.2	Temps d'atteinte	54
4.4	Propriétés des chaînes à transitions rares	56
4.4.1	Altitudes	56
4.4.2	Convergence	56
5	L'algorithme de Mutation/Sélection	59
5.1	Préambule	59
5.2	Introduction	59
5.3	L'algorithme d'Exploration/Sélection	61
5.4	Résultats principaux	63
5.4.1	Définitions	63
5.4.2	Énoncés	64
5.5	Notations et rappels	65
5.6	Le rôle des populations uniformes	66
5.7	Démonstrations	67
5.7.1	Concentration	67
5.7.2	Temps d'atteinte	68
5.7.3	Cycles	70
5.7.4	Schémas optimaux	70
5.7.5	Exposant de convergence	72
5.8	Implantation et exemples	73
5.8.1	Implantation	73
5.8.2	Comparaisons	74

Chapter 1

Algorithmes de recherche aléatoire

1.1 Introduction

Soit E (on notera parfois X) un ensemble de cardinal fini, muni d'une structure de graphe (non orienté) (E, G) et f une fonction positive définie sur E , appelée *fonction objectif*

$$f : E \rightarrow \mathbb{R}_+.$$

Le problème que l'on cherche à résoudre est de trouver un minimum absolu de f , c'est à dire $i_* \in E$ tel que

$$f(i_*) = \min_{i \in E} f(i).$$

Nous décrivons, dans ce chapitre, un certain nombre d'algorithmes fondés sur le principe de recherche aléatoire. L'objectif de ce cours est d'étudier ces algorithmes. Les *chaînes de Markov* seront l'outil central de cette étude (bien que ce chapitre ne les mentionne que de manière allusive). L'ouvrage de Kemeny et Snell est une bonne référence pour les chaînes définies sur un ensemble fini, celui de Ross permet de se familiariser rapidement avec les notions associées sans trop de détails techniques.

L'accent sera mis sur l'étude de l'algorithme de Metropolis, ou *recuit simulé*. L'article de Bertsimas et Tsitsiklis (1989) fournit une présentation claire de cet algorithme. Azencott (1988) propose une présentation proche de l'esprit de ce cours. On pourra aussi consulter l'ouvrage de Aarst et Korst (1989) pour les applications. Nous étudierons aussi quelques algorithmes de population, où plusieurs chercheurs interagissent pour trouver la solution optimale. Ces algorithmes sont aussi connus sous le nom d'*algorithmes évolutionnaires* et génèrent actuellement une énorme littérature. Fogel (1994) et Bäck et Schwefel (1993, 1997) proposent une présentation synthétique de ces algorithmes.

Le cours est structuré de la manière suivante. Le chapitre 1 introduit aux principes de recherche aléatoire, et décrit informellement les principaux algorithmes en insistant sur les résultats de convergence les plus marquants. Le chapitre 2 est constitué de rappels concernant les chaînes de Markov sur un espace fini, et résume le prérequis des trois chapitres suivants. On cherchera à décrire la convergence des algorithmes vers une solution optimale du problème de minimisation. Pour cela, nous adoptons deux points de vue différents : analytique d'une part et probabiliste d'autre part. Dans

le chapitre 3, la convergence est étudiée à travers les valeurs propres de l'opérateur associé à la chaîne de Markov. Cette étude est développée dans un cadre hilbertien, et met en évidence l'importance de grandeurs géométriques calculées à partir du graphe de transition de la chaîne. La présentation s'inspire essentiellement des articles de Diaconis et Stroock (1991), et Ingrassia (1993,1994). Le chapitre peut être approfondi par la lecture du cours de Saloff-Coste (1997). Le chapitre 4 développe le point de vue probabiliste, en étudiant le temps d'atteinte d'une solution optimale. Il se fonde sur le formalisme des grandes déviations (transitions rares). Il s'inspire de Freidlin-Wentzell (1984, chap. 6) et des notes de cours de Catoni (1997) dont la lecture complète ce chapitre de manière intéressante. Le chapitre 5 est original, consacré à l'étude d'un algorithme de population particulier, mais nécessite un investissement bibliographique supplémentaire.

1.2 Exemples d'application

Les quelques problèmes décrits ci-dessous ont souvent servi de motivation pour les algorithmes stochastiques.

1.2.1 Voyageur de commerce

Un voyageur de commerce visite n villes pendant sa tournée (et doit éviter de passer deux fois dans la même ville). Un coût est associé à chaque trajet, et le coût total de la tournée est la somme de chacun des coûts associés aux trajets. Le problème est de déterminer une tournée de coût minimal (figure 1.1). Dans ce cas, l'ensemble E est constitué de toutes les tournées possibles. Si on numérote les villes de 1 à n , une tournée s'identifie à une permutation cyclique des n villes. Le cardinal de E est donc égal à $(n - 1)!$. Ce nombre croît exponentiellement vite avec n (cf. la formule de Stirling). Une recherche exhaustive est donc proscrite assez rapidement.

1.2.2 Sac-à-dos

Considérons n objets de poids respectifs w_1, \dots, w_n , pour $w_i \geq 0$, et de valeurs respectives c_1, \dots, c_n . Les objets doivent être rangés (tous ne pourront peut-être pas l'être) dans un sac-à-dos de capacité limitée au poids total K . Le problème est de maximiser la somme des valeurs des objets rangés dans le sac sous la contrainte que le poids reste inférieur à K . Dans cet exemple, E est l'ensemble des configurations possibles de n objets. Une telle configuration s'identifie à un élément de $\{0, 1\}^n$ en posant

$$e = (e_1, \dots, e_n) \in \{0, 1\}^n$$

où

$$e_i = \begin{cases} 1 & \text{si } i \text{ est rangé dans le sac-à-dos,} \\ 0 & \text{sinon.} \end{cases}$$

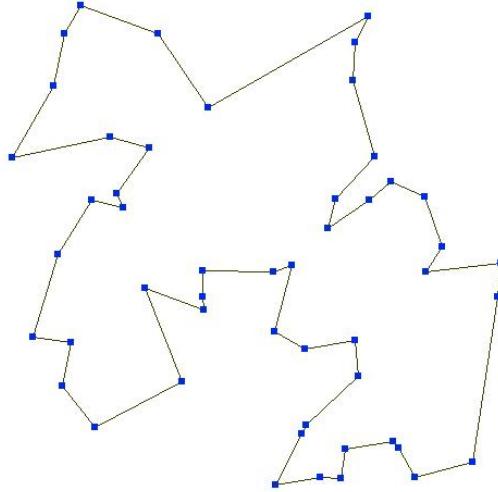


Figure 1.1: Tour optimal pour une instance du problème du voyageur de commerce comportant 50 villes. La solution provient d'une méthode de population.

Par conséquent, E possède 2^n éléments. Une recherche énumérative est aussi proscrite dans ce cas.

Exercice 1.2.1 Définir les fonctions “poids total” et “coût total” sur E et donner une forme analytique au problème de maximisation en question.

1.2.3 Modèle de Ising et images

L'exemple ci dessous est important. Il a ouvert la voie aux méthodes de Monte-Carlo en imagerie (figure 1.2). Il est dû à Geman et Geman (1985), (voir aussi Frigessi et al. (1993, 1997) pour des développements plus récents).

Une image binaire peut être représentée comme un élément de

$$E = \{-1, +1\}^S$$

où S est l'ensemble des pixels de l'image. Dans ce codage, la valeur -1 représente la couleur noire, tandis que $+1$ représente le blanc. On suppose qu'une image X est une variable aléatoire distribuée selon une certaine loi de probabilité :

$$\forall x \in E, \quad P(X = x) = \pi(x) = e^{U(x)}/Z$$

où $U(x)$ décrit les caractéristiques “géométriques” de l'image et Z est une constante de normalisation (modèle de Ising). En réalité, on observe une version bruitée de l'image,

que l'on note Y . Admettons que l'on modélise le bruit de la manière suivante. Pour tout $i \in S$, Y_i est égal à $\pm x_i$ et

$$\mathrm{P}(Y_i = -x_i \mid X_i = x_i) = \frac{e^{-\sigma}}{e^{-\sigma} + e^{+\sigma}}.$$

Cela signifie que l'on sait estimer la probabilité pour que la valeur d'un site soit modifiée par le bruit ($Y_i = \pm 1$). Débruyer revient à déterminer la réalisation x_* de X qui maximise la probabilité conditionnelle

$$L(x \mid y) = \mathrm{P}(X = x \mid Y = y).$$

D'après la formule de Bayes, nous avons

$$\mathrm{P}(Y = y)L(x \mid y) = \mathrm{P}(Y = y \mid X = x)\mathrm{P}(X = x)$$

soit, en utilisant l'indépendance (conditionnelle) du bruit,

$$\mathrm{P}(Y = y)L(x \mid y) = \prod_{i \in S} \frac{e^{\sigma x_i y_i}}{e^{-\sigma} + e^{+\sigma}} \frac{e^{U(x)}}{Z}$$

En passant au logarithme, cela revient à chercher l'image x_* qui maximise, pour y fixé,

$$U_y(x) = U(x) + \sum_{i \in S} \sigma x_i y_i.$$

En pratique, le *modèle d'Ising* est fréquemment choisi

$$U(x) = - \sum_{i \sim j} \beta x_i x_j + h \sum_i x_i$$

où $i \sim j$ signifie que i et j sont des pixels “voisins” dans l'image en un sens à préciser (plusieurs définitions sont possibles). Les paramètres β et h sont des paramètres fixés *a priori*.

1.3 Recherche sur un graphe

1.3.1 Graphes

Dans tout ce cours, l'ensemble de recherche E sera muni d'une structure de graphe (E, G) non orienté. L'ensemble E est l'ensemble des sommets du graphes, G est l'ensemble des arêtes

$$G = \{i \rightarrow j, \{i, j\} \subset E\}$$

Cette structure définit la manière de parcourir E lors de la recherche des minima absolus de f . Ce graphe est prédéfini et la manière de le construire dépend fortement de la connaissance que l'on a du problème de minimisation.

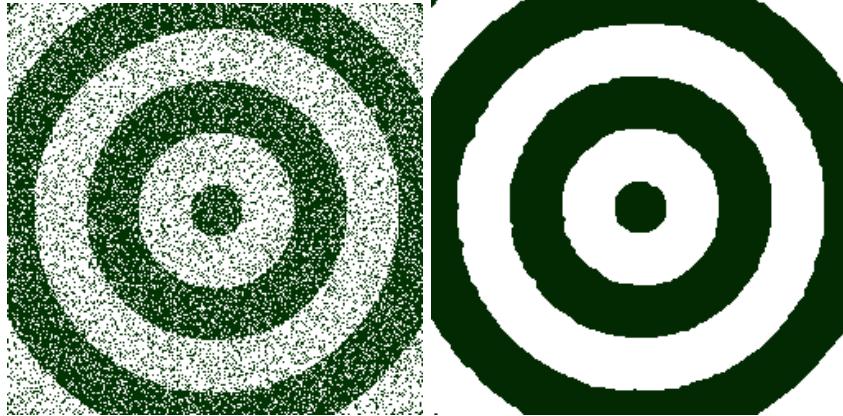


Figure 1.2: Une image binaire est débruitée (à droite) en utilisant un algorithme d'optimisation fondé sur le recuit simulé.

On dit que deux sommets sont *voisins* s'ils sont reliés par une arête. On note aussi $j \in N(i)$ pour $i \rightarrow j$. Le degré d'un sommet i , noté $\deg(i)$ est le nombre de voisins de i . Le degré du graphe est

$$\deg(G) = \max_{i \in E} \deg(i) .$$

Un *chemin* entre i et j est une suite de sommets $(i_k)_{k=0,\dots,r}$ reliés par des arêtes. On note

$$\gamma_{ij} : i_0 = i \rightarrow i_1 \rightarrow \dots \rightarrow i_r = j .$$

La longueur du chemin γ_{ij} est le nombre d'arêtes empruntées par le chemin :

$$|\gamma_{ij}| = r .$$

Exemple 1.3.1 $E = \{0, 1\}^n$, $n \geq 1$. Définissons une structure de graphe sur E pour résoudre le problème du sac-à-dos.

Solution. On peut définir une arête $e \rightarrow e'$ en choisissant i dans $\{1, \dots, n\}$ et e' tel que

$$e'_j = \begin{cases} e_j & \text{si } j \neq i \\ 1 - e_j & \text{sinon.} \end{cases}$$

Dans ce cas, $\deg(e) = \deg(G) = n$. ■

Exercice 1.3.1 Définir, comme dans l'exercice précédent, des relations de voisinage entre permutations cycliques de n éléments.

1.3.2 Recherche aléatoire

Notre premier algorithme (RA) est défini (informellement) de la manière suivante

1. Choisir un sommet initial i dans E .
2. Répéter
 - Choisir un voisin j dans $N(i)$ au hasard.
 - Si $f(j) \leq f(i)$, alors conserver j en mémoire.
 - $i := j$

Jusqu'à Arrêt.

3. Retourner la valeur en mémoire.

Il s'agit simplement de mémoriser la meilleure solution atteinte lors d'un parcours aléatoire du graphe (E, G) . Le critère d'arrêt est à définir. Disons de suite que cette méthode s'avère très peu efficace en général.

Exemple 1.3.2 Marche aléatoire sur le graphe complet. Supposons que le minimum absolu de f est unique, et notons ce point i_* . On suppose que

$$\forall i \in E, \quad \deg(i) = \#E.$$

Alors le temps d'atteinte de i_* , partant de $i \neq i_*$ est en moyenne égal au cardinal de E .

Démonstration. Le temps d'atteinte de i_* est défini par la variable aléatoire

$$T_{i_*} = \inf\{t \geq 1 ; X_t = i_*\}$$

où X_t est le sommet visité au temps t . Soit $i \neq i_*$. Le temps moyen d'atteinte partant de i est

$$m_{ii_*} = E[T_{i_*} \mid X_0 = i] = m_{i_*}.$$

L'indépendance en i vient du fait que le graphe est complet. En conditionnant, on obtient

$$E[T_{i_*} \mid X_0 = i] = E[E[T_{i_*} \mid X_1] \mid X_0 = i].$$

Mais

$$E[T_{i_*} \mid X_1 = j] = \begin{cases} 1 & \text{si } j = i_* \\ 1 + m_{i_*} & \text{sinon.} \end{cases}$$

De plus, nous avons

$$P(X_{t+1} = j \mid X_t = i) = p(i, j) = \frac{1}{\deg(i)} = \frac{1}{\#E}.$$

D'où

$$m_{i_*} = p(i, i_*) + \sum_{j \neq i_*} (1 + m_{i_*})p(i, j) = 1 + (1 - p(i, i_*))m_{i_*}.$$

Ceci entraîne que $m_{i_*} = \#E$. ■

Dans cet exemple, la méthode de recherche aléatoire n'a pas d'intérêt pratique. Notons toutefois le résultat suivant qui assure la convergence de la méthode en général.

Proposition 1.3.1 *On suppose que le graphe (E, G) est connexe. Soit (X_t) la suite de sommets définis par l'algorithme RA. Soit*

$$T_{i_*} = \inf\{t \geq 1 ; X_t = i_*\}$$

le temps d'atteinte d'un minimum absolu i_* . Alors,

$$P(T_{i_*} = \infty) = 0.$$

(RA atteint i_* presque sûrement.)

Démonstration. Comme dans l'exemple précédent, notons

$$p(i, j) = P(X_{t+1} = j | X_t = i)$$

les probabilités de transition associées à RA et $P = (p(i, j))$. Nous avons

$$p(i, j) = \begin{cases} 1/\deg(i) & \text{si } j \in N(i) \\ 0 & \text{sinon.} \end{cases}$$

Par connexité, il existe $\delta > 0$ et $D > 0$ tels que

$$\forall i \neq i_* \quad p_{ii_*}^D \geq \delta.$$

En fait, D n'est autre que la longueur du plus court chemin entre i_* et le sommet qui lui est le plus éloigné

$$D = \max_{i \neq i_*} \min_{\gamma_{ii_*}} |\gamma_{ii_*}|.$$

Effectuons la division euclidienne de l'entier t (t grand) par D

$$t = qD + r \quad r < D.$$

Alors, en supposant $X_0 = i$, nous avons

$$\begin{aligned} P(T_{i_*} > t) &= P(X_1 \neq i_* ; \dots ; X_t \neq i_*) \\ &\leq P(X_D \neq i_* ; X_{2D} \neq i_* ; \dots ; X_{qD} \neq i_*) \\ &= \sum_{j_1 \neq i_*} p_{ij_1}^D \sum_{j_2 \neq i_*} p_{j_1 j_2}^D \dots \sum_{j_q \neq i_*} p_{j_{q-1} j_q}^D. \end{aligned}$$

Par une récurrence évidente

$$\mathrm{P}(T_{i_*} > t) \leq (1 - \delta)^q \rightarrow 0 \quad \text{lorsque } t \rightarrow \infty.$$

Les événements $(T_{i_*} > t)$ sont décroissants et

$$\cap_t (T_{i_*} > t) = (T_{i_*} = \infty).$$

Un résultat classique de théorie de la mesure assure que

$$\mathrm{P}(T_{i_*} = \infty) = \lim_{t \rightarrow \infty} \mathrm{P}(T_{i_*} > t) = 0.$$

■

Commentaires. Cette proposition montre qu'un algorithme de recherche aléatoire "atteint" la solution minimale au prix d'hypothèses très faibles sur le graphe d'exploration (irréductibilité de la chaîne de Markov associée). Elle restera vraie pour les autres algorithmes définis dans ce chapitre. Bien entendu, elle ne fournit que peu de renseignements sur la vitesse de convergence (elle est géométrique) et sur une éventuelle condition d'arrêt de l'algorithme.

1.4 Dynamique de Metropolis

1.4.1 Gradient

Définition 1.4.1 Soit f une fonction définie sur E . On dit que i est un minimum local de f dans le graphe (E, G) si

$$\forall j \in N(i), \quad f(i) \leq f(j)$$

Définition 1.4.2 Soit f une fonction définie sur E . On appelle gradient de l'arête $i \rightarrow j$, la grandeur

$$\Delta f(i, j) = f(j) - f(i)$$

Il paraît naturel d'essayer d'améliorer l'algorithme RA en procédant à une "descente" dans le graphe (E, G) selon l'algorithme suivant.

1. Choisir un sommet initial i dans E .
2. Répéter
 - Choisir un voisin j dans $N(i)$ au hasard.
 - Calculer $\Delta f(i, j)$
 - Si $\Delta \leq 0$, $i := j$.

Jusqu'à Arrêt.

3. Retourner i .

Cet algorithme est fréquemment utilisé. Notons que le calcul de $\Delta f(i, j)$ est une étape cruciale qui détermine souvent le choix du graphe (E, G) . En effet, on aura souvent intérêt à calculer directement le gradient plutôt que de calculer $f(i)$ puis $f(j)$ et faire la différence. Cela repose sur un choix de graphe pour lequel l'expression analytique de la différence peut-être calculée explicitement.

Exercice 1.4.1 *Donner l'expression de Δf dans les problèmes de sac-à-dos et de reconstruction d'image pour le graphe décrit dans le premier paragraphe, exemple 1.3.1. Que peut-on faire pour le voyageur de commerce ?*

Le défaut majeur de cet algorithme est de ne pas résoudre le problème de minimisation initial. En effet, il converge en général vers un minimum local de f qui n'est pas (forcément) un minimum global.

1.4.2 Algorithme de Metropolis

Dans le but de pouvoir s'échapper des minima locaux, on s'autorise maintenant des transitions à l'encontre du gradient. Ces transitions seront rares pour ne pas trop perturber la descente. Cela conduit à la *dynamique de Metropolis* ou *recuit simulé*. On suppose que le graphe est régulier

$$\forall i \in E, \quad \deg(i) = \deg(G).$$

L'algorithme se décrit de la manière suivante.

1. Choisir un sommet initial i dans E .

2. Répéter

- Choisir un voisin j dans $N(i)$ au hasard.
- Calculer $\Delta f(i, j)$
- Si $\Delta \leq 0$, $i := j$.
- Si $\Delta > 0$, $i := j$ avec la probabilité $\epsilon(i, j)$.

Jusqu'à Arrêt.

3. Retourner i .

La probabilité $\epsilon(i, j)$ sera choisie petite. Naturellement, les transitions de gradient positif seront d'autant plus pénalisées que l'“amplitude” du gradient sera grande. Cela conduit à prendre

$$\epsilon(i, j) = p^{\Delta f(i, j)}$$

avec $0 < p < 1$, ou

$$\epsilon(i, j) = e^{-\frac{1}{T} \Delta f(i, j)}$$

avec $T > 0$. Les deux formulations deviennent identiques lorsque l'on pose

$$p = e^{-1/T}.$$

Une grande partie des outils présentés dans ce cours aura pour finalité l'étude du comportement asymptotique de la chaîne de Markov associée à cet algorithme, avec pour objectif de déterminer des conditions d'arrêt pour ce dernier (au moins d'un point de vue théorique). Nous mettrons en évidence l'importance d'une grandeur géométrique (définie à partir des valeurs de f sur le graphe) appelée hauteur critique, notée h_* , introduite par Hajek (1988). Cette constante se comprend intuitivement comme la hauteur qu'il faut franchir pour sortir du plus profond des minima locaux (non globaux) de f . Le résultat le plus important montrera que le temps de “convergence” vers i_* est de l'ordre de $\exp(h_*/T)$ lorsque T est petit.

1.5 Algorithmes de population

1.5.1 description

La dynamique de Metropolis se fonde sur la trajectoire d'une seule solution dans le graphe (E, G) . Les *algorithmes de population* proposent une alternative intéressante à la dynamique de Metropolis, car elle reposent sur l'évolution simultanée de plusieurs solutions en interaction. Il s'agit de méthodes d'ordre zéro : elles ne supposent aucune régularité de la fonction à minimiser. Ces stratégies, dites *évolutionnaires* se fondent sur les principes d'exploration aléatoire et de sélection des meilleures solutions. Pour définir l'évolution du vecteur de solutions, on lui applique de manière successive les opérateurs de *mutation* et de *sélection*. Parmi ces stratégies, les plus populaires sont les algorithmes génétiques (Holland, 1975) et les stratégies d'évolution (Rechenberg, 1973). Dans ces deux méthodes, les opérateurs de mutation et de sélection sont appliqués tour à tour au vecteur de solutions. Considérons l'ensemble produit

$$X = E^n$$

et un vecteur de solutions de dimension n (la taille de la population)

$$x = (x_1, \dots, x_n) \in X.$$

Les algorithmes de population reposent sur le principe algorithmique suivant.

1. Choisir un vecteur initial x dans X .
2. Répéter
 - a) Évaluer chaque solution : calculer $f(x_i)$, pour $i = 1, \dots, n$.
 - b) Proposer un changement pour chacune des solutions.
 - c) Utiliser un critère d'acceptation pour valider chacun des changements.

Jusqu'à Arrêt.

3. Retourner la meilleure solution rencontrée.

Remarquons que l'algorithme de Metropolis est construit sur ce schéma. Il en va de même lorsque l'on fait évoluer indépendamment n solutions selon leur propre dynamique de Metropolis. Dans ce cas, l'étape a) est inutile. L'étape b) consiste à choisir des voisins indépendamment, et c) consiste à appliquer le critère de Metropolis n fois.

Il va aussi sans dire que l'on peut encore se contenter de n'évaluer que des gradients. Lorsque le graphe est choisi de manière adéquate, le coût de calcul devient moins élevé. Dans une telle situation, a) est inutile pourvu que l'on évalue le vecteur à l'initialisation. L'étape b) revient en général à choisir des voisins dans le graphe et la différence Δf est calculée dans la phase d'acceptation/rejet.

1.5.2 Algorithme de Mutation+sélection

Nous présentons dans ce paragraphe un algorithme génétique, dont la dynamique a été étudiée en détails par Cerf (1994, 1996a, 1996b). La boucle d'un algorithme génétique fait intervenir deux opérateurs. L'opérateur d'exploration est tout simplement fondé sur l'exploration indépendante du graphe (E, G) par les n solutions. Cette étape est souvent appelée *mutation*. Le critère d'acceptation est appelé *sélection*. Il fait appel au rééchantillonnage de la population obtenue après mutation, selon une loi qui favorise les bonnes solutions. Dans ce mode de sélection, un "mutant" peut être accepté plusieurs fois. Pour ne pas trop perturber le processus de sélection, les mutations seront rares, dans le même esprit que la dynamique de Metropolis. L'algorithme s'écrit donc de la manière suivante

1. Choisir un vecteur initial x dans X .

2. Répéter

- **Mutation.** Pour $i = 1, \dots, n$, choisir x'_i voisin de x_i au hasard (choix indépendants) avec la probabilité

$$p_{\text{mut}} = e^{-\alpha/T}, \quad \alpha > 0, \quad T > 0.$$

Si x'_i n'est pas choisi, alors $x'_i := x_i$.

- **Sélection.** Simuler un échantillon (y_1, \dots, y_n) , avec $y_i \in \{x'_1, \dots, x'_n\}$ selon la loi

$$P(Y = x'_i) \propto \exp(-f(x'_i)/T).$$

- Remplacer: $x_i := y_i$, pour $i = 1, \dots, n$.

Jusqu'à Arrêt.

3. Retourner la meilleure solution rencontrée.

L'étape de sélection s'appelle *roulette de Boltzmann*. Lorsque T est petit, ce mode de sélection concentre les solutions autour de la meilleure solution courante. En effet, supposons f injective et notons

$$x'_* = \operatorname{argmin}\{f(x'_i) ; i = 1, \dots, n\}.$$

Alors, nous avons

$$P(Y = x'_i) = \frac{\exp((f(x'_*) - f(x'_i))/T)}{\sum_j \exp((f(x'_*) - f(x'_j))/T)}.$$

Ceci montre que

$$P(Y = x'_i) \rightarrow \delta_{x'_i x'_*}$$

lorsque T tend vers zéro ($\delta_{x'_i x'_*}$ désigne le symbole de Kronecker). Pour assurer la convergence, nous supposerons donc que T est petit. L'opérateur de sélection reproduit alors les meilleures solutions et l'opérateur de mutation introduit de la diversité en injectant de nouvelles solutions dans la compétition.

Les principaux résultats concernant cette dynamique s'énoncent informellement de la manière suivante. Soit X_t le vecteur de solutions obtenu à la génération t .

- Il existe une taille de population critique $n_* \geq 2$, telle que, pour $n > n_*$, on a

$$\lim_{T \rightarrow 0} \lim_{t \rightarrow \infty} P(i_* \in X_t) = 1.$$

- Il existe une constante H_* , analogue à la hauteur critique de Hajek, telle que le temps de convergence vers i_* est de l'ordre de $\exp(H_*/T)$ lorsque T est petit.

1.5.3 Algorithme de Mutation/sélection

Nous étudierons une variante de l'algorithme de Mutation+Sélection, dont la formulation peut sembler moins naturelle mais dont l'étude est plus aisée. Notons que cette variante est compétitive du point de vue calculatoire. Il s'agit de l'algorithme de Mutation/Sélection. Il est décrit par la boucle suivante.

1. Choisir un vecteur initial x dans X .

2. Répéter

- Déterminer $x_* = \operatorname{argmin}\{f(x_i) ; i = 1, \dots, n\}$
- **Mutation/Sélection.** Pour $i = 1, \dots, n$, choisir x'_i voisin de x_i au hasard (choix indépendants) avec la probabilité

$$p_{mut} = e^{-1/T}, \quad T > 0$$

et faire $x_i := x'_i$. Avec la probabilité $1 - p_{mut}$, faire $x_i := x_*$.

Jusqu'à Arrêt.

3. Retourner la meilleure solution rencontrée.

Cet algorithme repose sur le même principe que le précédent. Une sélection stricte maintient la présence de la meilleure solution, tandis que les mutants apportent la nouveauté. La probabilité de mutation évanescante garantit une certaine descente vers le minimum global.

La différence avec l'algorithme précédent est que mutation et sélection sont couplées en un seul opérateur. Ceci permet de décrire les transitions entre populations plus facilement. Par conséquent, les résultats obtenus seront plus précis. Comme pour les algorithmes génétiques, il est crucial de comprendre les relations liant la convergence de la chaîne de Markov associée à l'algorithme et la géométrie du problème de minimisation (Cerf, 1996). Pour cela, deux indices sont définis. Le premier est un indice de dimension

$$n_* = \max_{i \neq i^*} d(i, i^*) \quad (1.1)$$

où d est la distance sur le graphe. Cet indice correspond à la distance maximale entre i_* et un sommet arbitraire de E . Une grande valeur de n_* est intuitivement défavorable à la convergence de l'algorithme. Ceci peut être compensé par une taille de population élevée. Le second indice est un indice de convergence. A nouveau, il quantifie la difficulté du problème de minimisation en relation avec sa géométrie. Formellement, cet indice est décrit de la manière suivante

$$d_* = \max_{i \neq i_*} \min_{j: f(j) < f(i)} d(i, j) . \quad (1.2)$$

En fait, cet indice mesure le nombre minimal d'arêtes nécessaires pour définir un chemin "sortant" de n'importe quel minimum local (non global) de f . Nous démontrerons les résultats suivants.

- Soit $n > n_*$. Alors, nous avons

$$\lim_{T \rightarrow 0} \lim_{t \rightarrow \infty} P(i_* \in X_t) = 1 .$$

- Le temps de convergence vers i_* est de l'ordre de $\exp(d_*/T)$ lorsque T est petit.

Chapter 2

Généralités sur les chaînes de Markov

2.1 Définitions

Définition 2.1.1 Un processus aléatoire $\{X_t ; t \in \mathbb{N}\}$ à valeurs dans E est une chaîne de Markov homogène notée (X_t) si, pour tout $t \geq 0$,

$$\forall i_0, i_1, \dots, i, j \in E, \quad P(X_t = j \mid X_{t-1} = i, X_{t-2} = i_{t-2}, \dots, X_0 = i_0) = p(i, j).$$

Pour une chaîne de Markov, la loi conditionnelle de la variable X_t sachant le passé du processus au temps t ne dépend du passé qu'à travers la dernière observation du processus. Cette propriété est souvent appelée *propriété de Markov*. La valeur $p(i, j)$ représente la probabilité que le processus effectue une transition vers l'état j lorsqu'il se trouve dans l'état i . Puisque ces probabilités de transition sont des nombres positifs et que le processus doit obligatoirement faire une transition vers un état dans E , on a

$$\forall i, j \in E, \quad p(i, j) \geq 0 \quad \text{et} \quad \sum_{j \in E} p(i, j) = 1.$$

On note P la matrice des probabilités $p(i, j)$ de transition de l'état i vers l'état j . La matrice P s'appelle la *matrice de transition* de la chaîne. Nous pouvons décrire la loi de l'état de la chaîne en un nombre arbitraire de pas en utilisant les probabilités de transition.

Définition 2.1.2 On appelle matrice de transition en s étapes d'une chaîne de Markov homogène (X_t) , la matrice $P^{(s)}$ dont les coefficients sont égaux à

$$p^{(s)}(i, j) = P(X_{s+t} = j \mid X_t = i) = P(X_s = j \mid X_0 = i).$$

Il vient immédiatement de la définition précédente

$$P^{(1)} = P \quad ; \quad P^{(0)} = Id.$$

et

$$p^{(s)}(i, j) \geq 0 \quad ; \quad \sum_{j \in E} p^{(s)}(i, j) = 1, \quad j \in E.$$

Théorème 2.1.1 *La matrice de transitions en s étapes de la chaîne (X_t) est la puissance s -ième de P .*

Si l'on désire connaître la loi de la variable X_t , pour un $t \in \mathbb{N}$, il est nécessaire de préciser la loi de la variable X_0 . Cette variable est appelée condition initiale de la chaîne. On note

$$\forall k \in E, \quad P(X_0 = k) = \pi^0(k) \geq 0 ; \quad \sum_{k \in E} \pi^0(k) = 1.$$

La loi de X_t s'obtient en conditionnant à la variable initiale X_0

$$\begin{aligned} \forall i \in E, \quad P(X_t = i) &= \sum_{k \in E} P(X_t = i \mid X_0 = k) P(X_0 = k) \\ &= \sum_{k \in E} \pi(k) p^{(t)}(k, i) \end{aligned}$$

Pour tout $t \in \mathbb{N}$, on désigne par π^t la loi de la variable X_t

$$\forall i \in E, \quad P(X_t = i) = \pi^t(i).$$

En notant (vecteur ligne pour les lois de probabilité)

$$\pi^0 = (\pi^0(1), \dots, \pi^0(|E|))$$

et

$$\pi^t = (\pi^t(1), \dots, \pi^t(|E|))$$

nous obtenons, en notation matricielle, la loi de X_t

$$\pi^t = \pi^0 P^t.$$

2.2 Convergence

À la suite de l'étude du comportement de la chaîne en t transitions, nous désirons étudier le comportement asymptotique de celle-ci lorsque $t \rightarrow \infty$. Soit (X_t) une chaîne de Markov homogène de matrice de transition P sur un espace d'état E . L'équation (2.1) établit que la loi π^t s'obtient par récurrence à partir de la loi $\pi^0 = \pi$ de la variable initiale X_0

$$\pi^t = \pi^{t-1} P ; \quad \pi^0 = \pi.$$

Définition 2.2.1 *On appelle loi de probabilité invariante de la chaîne (X_t) toute loi π solution de l'équation*

$$\pi = \pi P. \tag{2.1}$$

Commentaires. Une telle loi existe toujours. Une matrice de transition admet la valeur propre 1 (le vecteur dont toutes les composantes sont égales à 1 est associé à cette valeur propre). Le vecteur associé à une loi de probabilité invariante π est vecteur propre à gauche associé à la valeur propre 1. On peut s'assurer qu'un tel vecteur propre existe bien. En effet l'équation (2.1) se formule de manière équivalente en disant que ${}^T\pi$ est vecteur propre de TP associé à la valeur propre 1. Soit ν un vecteur propre associé à TP . Montrons que $\mu = |\nu|$ est encore vecteur propre (la valeur absolue est appliquée composante par composante)

$$\mu P(j) = \sum_i |\nu(i)| p(i, j) \geq \left| \sum_i \nu(i) p(i, j) \right| = \mu(j)$$

Or, nous avons

$$\sum_j \sum_i |\nu(i)| p(i, j) = \sum_i |\nu(i)| .$$

Par conséquent, $\mu P = \mu$. ■

Exemple 2.2.1 Considérons la chaîne de matrice de transition

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix}$$

avec $\alpha, \beta \in (0, 1)$, et examinons la convergence de cette chaîne.

Solution. Une probabilité invariante $\pi = (\pi(0), \pi(1))$ est, pour cet exemple, solution du système

$$(\pi(0), \pi(1)) = (\pi(0), \pi(1)) \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix}$$

soit

$$\begin{aligned} \pi(0) &= \alpha\pi(0) + \beta\pi(1) \\ \pi(1) &= (1-\alpha)\pi(0) + (1-\beta)\pi(1) \\ \pi(0) + \pi(1) &= 1 . \end{aligned}$$

Ceci donne

$$\pi(0) = \frac{\beta}{1+\beta-\alpha} ; \quad \pi(1) = \frac{1-\alpha}{1+\beta-\alpha} .$$

Le système admet une unique solution qui soit une loi de probabilité. Diagonalisons la matrice P . Les valeurs propres sont 1 et $\alpha - \beta$. Remarquons que cette dernière valeur est en module strictement inférieure à un. Nous avons

$$\pi^t = \pi^0 P^t = \pi^0 V \begin{pmatrix} 1 & 0 \\ 0 & (\alpha - \beta)^t \end{pmatrix} V^{-1} \rightarrow \mu$$

où V est une matrice de vecteurs propres. La loi π^t définie à l'aide de la récurrence

$$\pi^t = \pi^{t-1} P$$

admet donc $\mu = \pi$ pour limite, quelle que soit la loi initiale π^0 , car la loi de probabilité π est le point fixe associé à la récurrence précédente. D'un point de vue de calcul approché, nous pouvons considérer que la loi invariante π est une approximation de la loi π^t . Un problème important est de savoir si cela constitue une bonne approximation. La convergence se fait ici à une vitesse exponentielle, contrôlée par la grandeur $|\alpha - \beta|$.

Dans cet exemple élémentaire nous voyons le rôle des valeurs propres de P . Notons que ces valeurs propres sont toujours de module plus petit que 1. En effet, si φ est vecteur propre associé à λ , de norme ℓ_∞ égale à $\|\varphi\|_\infty$ nous avons

$$\|P\varphi\|_\infty = |\lambda| \|\varphi\|_\infty \leq \|\varphi\|_\infty$$

et donc

$$|\lambda| \leq 1.$$

Définition 2.2.2 Une chaîne de Markov de matrice de transition P est irréductible si

$$\forall i, j \in E, \exists r \text{ t.q. } p^{(r)}(i, j) > 0.$$

Elle est apériodique si

$$\forall i \in E, \{t \text{ t.q. } p^{(t)}(i, i) = 0\} \text{ est fini.}$$

Il est facile de visualiser ces propriétés sur les diagrammes de transition construits sur E de la manière suivante

$$\forall i, j \in E, i \rightarrow j \text{ ssi } p(i, j) > 0.$$

La chaîne est irréductible si elle peut effectivement atteindre tout état j à partir de n'importe quel état i en un nombre fini de pas. La chaîne est apériodique si elle ne boucle pas indéfiniment sur quelques états. Ceci se traduit aussi par le fait que le p.g.c.d. des longueurs des cycles fermés du graphe est égal à 1. Ces deux propriétés signifient pour la chaîne qu'un certain mélange est présent. Lorsque ces propriétés sont réunies, la chaîne est *ergodique*.

Avant d'énoncer le théorème de convergence des chaînes de Markov, rappelons le célèbre résultat de Perron-Frobenius (cf Cinlar, 198x).

Théorème 2.2.1 Soit A une matrice carrée de dimension n t.q. $a_{ij} > 0$ alors A admet une valeur propre réelle strictement positive α de multiplicité 1 et telle que

$$\alpha > |\lambda|$$

pour toute autre valeur propre λ de A . À la valeur propre α correspond un vecteur propre dont les coordonnées sont positives.

Théorème 2.2.2 Soit (X_t) une chaîne de Markov irréductible et apériodique.

1) Il existe une unique loi de probabilité π invariante. De plus, $\pi(i) > 0$ pour tout $i \in E$.

2) Pour tout $i, j \in E$,

$$\lim_{t \rightarrow \infty} p^{(t)}(i, j) = \pi(j).$$

3) Quelle que soit la loi de X_0 , (X_t) converge en loi vers π .

Démonstration. Nous donnons une preuve abrégée. Cinlar donne une démonstration très claire des théorèmes de Perron-Frobenius et de convergence auxquelles le lecteur intéressé pourra se reporter. Les hypothèses du théorème garantissent qu'il existe un rang t_0 tel que

$$\forall t \geq t_0, \quad p^t(i, j) > 0 \quad \forall i, j.$$

Pour simplifier sans perte de généralité, nous supposerons que $t_0 = 1$. D'après le théorème de Perron-Frobenius, la valeur propre $\alpha = 1$ est de multiplicité 1. Ainsi, P s'écrit

$$P = V \begin{pmatrix} 1 & 0 \\ 0 & d_{11} & * \\ 0 & * & d_{|E|-1, |E|-1} \end{pmatrix} V^{-1}$$

où V désigne une matrice de vecteurs propres de P . Les valeurs propres λ_i de la matrice

$$D = \begin{pmatrix} d_{11} & * \\ & d_{|E|-1, |E|-1} \end{pmatrix}$$

ont un module strictement inférieur à 1. La norme matricielle de D est majorée de la manière suivante

$$\|D^t\| \leq c (\max |\lambda_i|)^t$$

où c est une constante positive. Nous avons

$$\lim_{t \rightarrow \infty} P^t = \lim_{t \rightarrow \infty} V \begin{pmatrix} 1 & 0 \\ 0 & D^t \end{pmatrix} V^{-1}$$

et le théorème découle du fait que l'on peut choisir la première colonne de V égale au vecteur dont toutes les coordonnées sont égales à 1. ■

Nous avons établi que

$$\lim_{t \rightarrow \infty} P(X_{t+s} = j \mid X_s = i) = \pi(j).$$

Cette limite est donc indépendante de i . Ceci nous amène à l'approximation suivante lorsque t est grand

$$P(X_{t+s} = j \mid X_s = i) \approx P(X_{t+s} = j).$$

En d'autres termes, nous pouvons considérer que deux états de la chaîne séparés par une longue histoire sont indépendants. A propos de la vitesse de convergence, nous avons vu, lors de la démonstration précédente que

$$|\Pr(X_t = j \mid X_0 = i) - \pi(j)| \leq c\rho^t$$

où c est une constante positive et

$$\rho = \max\{|\lambda_i| ; \lambda_i \text{ valeur propre de } P \neq 1\}.$$

La limite est atteinte à une vitesse exponentielle et cette vitesse de convergence se caractérise par le coefficient ρ .

2.3 Temps d'atteinte

Dans le premier chapitre, nous avons abordé (sommairement) le calcul du temps moyen d'atteinte d'un sommet de E par une marche aléatoire. Dans cette section, nous précisons la manière de calculer les temps d'atteinte. Puisque l'on s'intéresse à la première visite en un sommet donné, nous supposerons sans perte de généralité que la chaîne est absorbante en ce sommet.

2.3.1 Matrice fondamentale

Définition 2.3.1 *Un point $i_* \in E$ est absorbant si $p(i_*, i_*) = 1$.*

On suppose dans cette section que la chaîne de matrice de transition P admet une unique loi invariante et que i_* est un point absorbant. L'unique loi invariante π charge donc i_*

$$\pi = \delta_{i_*}.$$

Soit $j \neq i_*$. On considère la variable aléatoire $N(j)$ égale au nombre de visites en j avant l'absorption et

$$\forall i \neq i_* , n_{ij} = E[N(j) \mid X_0 = i].$$

Finalement, on note $N = (n_{ij})$ la matrice carrée de dimension $n - 1$ dont les coefficients sont définis grâce à la relation précédente. Cette matrice s'appelle la *matrice fondamentale* de la chaîne.

Lemme 2.3.1 *Considérons la matrice carrée Q de dimension $n - 1$ correspondant à la restriction de P à $E \setminus i_* \times E \setminus i_*$. Alors, nous avons*

$$N = (I - Q)^{-1} = \sum_{k=0}^{\infty} Q^k$$

où I est la matrice identité de dimension $n - 1$.

Démonstration. Exercice ■

Considérons le temps d'atteinte de i_* , que nous notons T_{i_*} . Nous avons

$$T_{i_*} = \sum_{j \neq i_*} N(j)$$

et bien entendu

$$m_{ii_*} = E[T_{i_*} \mid X_0 = i] = \sum_{j \neq i_*} n_{ij} .$$

Exercice 2.3.1 Calculer les temps d'atteinte pour un exemple de matrice de transition de dimension 4 que l'on aura construit soi-même.

2.3.2 Temps d'atteinte et spectre : quelques remarques.

Dans cette section, nous considérons quelques exemples de chaînes absorbantes qui permettent de mettre en relation la vitesse de convergence lue sur le spectre de la matrice de transition (la seconde plus grande valeur propre en module) et les temps moyens d'atteinte. Les chaînes étudiées sont à transitions rares.

Exemple 2.3.1 *Transitions rares indépendantes.* Soit $\epsilon > 0$ et

$$P_\epsilon = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 - \lambda_2(\epsilon) & \lambda_2(\epsilon) & 0 & 0 & 0 \\ 1 - \lambda_3(\epsilon) & 0 & \lambda_3(\epsilon) & 0 & 0 \\ 1 - \lambda_{|E|}(\epsilon) & 0 & 0 & 0 & \lambda_{|E|}(\epsilon) \end{pmatrix}$$

où

$$\forall i = 2, \dots, |E|, \quad \lambda_i(\epsilon) = 1 - e^{-c_i/\epsilon}$$

et $0 < c_{|E|} < \dots < c_2$.

Dans cet exemple, on séjourne pendant une durée de loi géométrique dans l'état initial de la chaîne avant d'effectuer une transition vers le sommet absorbant. Considérons une loi de probabilité initiale π^0 qui ne charge pas $i_* = 1$. Nous avons

$$E[T_{i_*}] = \sum_i \pi^0(i) E[T_{i_*} \mid X_0 = i] .$$

Or, pour $i \neq i_*$,

$$E[T_{i_*} \mid X_0 = i] = \frac{1}{1 - \lambda_i(\epsilon)}$$

et

$$E[T_{i_*}] = \sum_{i>1} \pi^0(i) e^{c_i/\epsilon} \sim \pi^0(2) e^{c_2/\epsilon}$$

Lorsque ϵ est petit cette somme se comporte comme le terme dominant qui est $e^{c_2/\epsilon}$ c'est à dire l'inverse de la différence entre la plus grande valeur propre (égale à 1) et la seconde valeur propre. Ceci peut se traduire à l'aide d'un équivalent logarithmique

$$\lim_{\epsilon \rightarrow 0} \epsilon \log E[T_{i_*}] = - \lim_{\epsilon \rightarrow 0} \epsilon \log(1 - \lambda_2(\epsilon)) .$$

■

Lorsque i_* est absorbant, ce résultat peut être facilement généralisé (en supposant les valeurs propres de la même forme). Diagonalisons la matrice fondamentale N , introduite dans le lemme 2.3.1. Nous obtenons

$$N = VDV^{-1}$$

où D est une matrice diagonale de terme général $1/(1 - \lambda_i)$, λ_i valeur propre de P . Nous avons

$$E[T_{i_*}] = \sum_{j \neq i_*} n_{ij} = \sum_{j \neq i_*} VDV_{ij}^{-1} \sim_{\log} \frac{1}{1 - \lambda_2(\epsilon)} .$$

Nous allons maintenant montrer de manière informelle qu'une telle relation entre le *trou spectral* $(1 - \lambda_2)$ et l'inverse du temps d'atteinte de i_* peut encore se vérifier lorsque la chaîne n'est plus absorbante mais lorsqu'elle se concentre sur l'état i_* .

Exemple 2.3.2 Considérons une chaîne de Markov ergodique de matrice de transition P_ϵ telle que la loi invariante π_ϵ se concentre sur i_*

$$\pi_\epsilon(i_*) \rightarrow 1 \quad \epsilon \rightarrow 0 .$$

et dont les valeurs propres vérifient

$$\lambda_i \sim 1 - b_i e^{-c_i/\epsilon}, \quad b_i > 0$$

pour $0 < c_{|E|} < \dots < c_2$. Alors $E[T_{i_*}]$ se comporte comme l'inverse du trou spectral lorsque ϵ est petit.

Pour justifier ce résultat, nous avons besoin du principe de couplage suivant.

Théorème 2.3.1 Soit P une matrice de transition ergodique. Considérons une chaîne (X_t) de matrice P et de loi initiale μ et une chaîne (Y_t) de matrice P , stationnaire et indépendante de (X_t) . Soit

$$T = \min\{t \geq 0 ; X_t = Y_t\}$$

l'instant de couplage des deux chaînes. Alors, nous avons

$$\|P_\mu^t(.,.) - \pi(.,)\|_\infty \leq 2 P(T > t) ,$$

où π est l'unique loi invariante de P .

Démonstration. Nous avons

$$\begin{aligned}
 |\mathbb{P}(X_t = j) - \pi(j)| &= |\mathbb{P}(X_t = j) - \mathbb{P}(Y_t = j)| \\
 &= |\mathbb{P}(X_t = j, T \leq t) + \mathbb{P}(X_t = j, T > t) \\
 &\quad - \mathbb{P}(Y_t = j, T \leq t) - \mathbb{P}(Y_t = j, T > t)| \\
 &\leq \mathbb{P}(X_t = j, T > t) + \mathbb{P}(Y_t = j, T > t) \\
 &\leq 2\mathbb{P}(T > t).
 \end{aligned}$$

L'argument de couplage intervient dans la première inégalité. En effet, nous avons, d'après la propriété de Markov

$$\mathbb{P}(X_t = j, T \leq t) = \mathbb{P}(Y_t = j, T \leq t).$$

■

Maintenant, fixons ϵ assez petit et diagonalisons la matrice de transition. Par représentation spectrale, $|p^t(i, j) - \pi(j)|$ se comporte comme $\lambda_2^t(\epsilon)$. D'autre part, puisque la loi invariante se concentre sur i_* , l'instant de couplage avec une chaîne stationnaire converge vers T_{i_*} (heuristiquement). Si on somme des deux cotés de l'inégalité du Théorème 2.3.1. On obtient d'une part $1/(1 - \lambda_2(\epsilon))$ et d'autre part

$$\sum_t \mathbb{P}(T > t) = E[T] \approx E[T_{i_*}] .$$

Ceci montre que l'on peut aborder le problème de la convergence pour de telles chaînes par le spectre ou par des arguments probabilistes.

2.3.3 Discussion

La chaîne de Markov présentée dans l'exemple 2.3.2 constitue un modèle pour tous les algorithmes de minimisation que nous avons rencontrés au chapitre 1. Cet exemple démontre qu'il y a deux manière d'aborder le problème de convergence de ces algorithmes vers la solution optimale. La première solution est l'étude du spectre de la matrice de transition et en particulier de la seconde valeur propre (en module). Nous illustrerons ce point de vue avec la dynamique de Metropolis, qui possède le mérite de définir un opérateur symétrique dans un certain espace fonctionnel. Le second point de vue est plus probabiliste et consiste à étudier les temps d'atteinte. Nous développerons ce point de vue pour les chaînes à transitions rares en reprenant en particulier l'argument de couplage ci-dessus. Enfin le traitement des deux points de vue reposera sur des arguments combinatoires portant sur les trajectoires des chaînes considérées et nous noterons de nombreuses similitudes entre les aspects fonctionnels et trajectoriels.

Chapter 3

Chaînes de Markov réversibles

3.1 Justification : un point de vue stochastique sur l'optimisation

Les algorithmes stochastiques rencontrés dans le chapitre 1 (dynamique de Metropolis, algorithmes de population) reposent sur la simulation d'une chaîne de Markov ergodique. La résolution du problème initial de minimisation est donc remplacée par la génération d'une loi (invariante) qui est concentrée sur les bonnes solutions. Nous nous focalisons dans ce chapitre sur la dynamique de Metropolis et nous reportons l'étude des algorithmes de population à un chapitre ultérieur (chapitre 5).

Supposons que le minimum de la fonction objectif f est atteint en un point unique, que l'on note i_* . Il est facile de définir une loi de probabilité sur l'ensemble E qui attribue une grande probabilité à i_* . Considérons un paramètre $T > 0$ et la loi de Gibbs définie par

$$\forall i \in E, \quad \pi_T(i) = \exp(-f(i)/T)/Z_T$$

où Z_T est une constante de normalisation. Le paramètre T est appelé *température*. Nous obtenons le résultat suivant.

Proposition 3.1.1 *Soit X une variable aléatoire de loi π_T , $T > 0$. Alors X converge en loi vers la masse de Dirac en i_**

$$P(X = i_*) \rightarrow 1 \quad \text{lorsque } T \rightarrow 0.$$

Exercice 3.1.1 *Rédiger la démonstration.*

L'idée sous-jacente à la dynamique de Metropolis est de remplacer le problème déterministe de minimisation par le problème de simulation de la loi π_T . La dynamique de Metropolis peut être modélisée par une chaîne de Markov qui converge (en loi) vers π_T .

Proposition 3.1.2 Soit (E, G) un graphe régulier connexe et f une fonction positive définie sur E . Les probabilités de transition associées à l'algorithme de Metropolis (paragraphe 1.4, chapitre 1) sont données par la matrice $P_T = (p_T(i, j))$ où

$$\forall i \neq j \in E, \quad p_T(i, j) = \begin{cases} 1/\deg(i) & \text{si } j \in N(i) \text{ et } f(j) \leq f(i) \\ e^{-\Delta f(i,j)/T}/\deg(i) & \text{si } j \in N(i) \text{ et } f(j) > f(i) \\ 0 & \text{sinon.} \end{cases}$$

et $p_T(i, i) = 1 - \sum_{j \neq i} p_T(i, j)$. De plus, la chaîne de Markov associée à P_T converge en loi vers π_T .

Démonstration. La chaîne est irréductible, apériodique. De plus, nous avons

$$\forall i, j \in E, \quad \pi_T(i)p_T(i, j) = \pi_T(j)p_T(j, i).$$

Conclure. ■

Exercice 3.1.2 Modifier la dynamique de Metropolis pour l'adapter à un graphe irrégulier de sorte que la convergence vers la loi de Gibbs soit encore vérifiée.

Commentaires. Il n'est pas nécessaire (formellement) de générer une chaîne de Markov pour obtenir un tirage aléatoire selon la loi π_T , d'autant plus que cette méthode possède le désavantage de ne pas simuler exactement la loi souhaitée. Toutefois, il est intéressant de noter que l'on ne calcule pas la constante Z_T dans l'algorithme de Metropolis, alors qu'une méthode directe (comme le classique algorithme d'inversion) induit ce calcul, souvent infaisable en pratique.

Commentaires. Dire un mot des méthodes de simulation exacte, un peu hors sujet mais intéressant (Propp et Wilson, Fill), et des généralisations de l'échantillonneur de Gibbs (Swendsen et Wang).

3.2 Chaînes réversibles

On considère dans la suite de ce chapitre uniquement des chaînes ergodiques. Soit π une loi de probabilité définie sur l'ensemble E .

3.2.1 Définitions

Définition 3.2.1 Une chaîne de Markov de matrice de transition P est π -réversible si

$$\forall i, j \in E, \quad \pi(i)p(i, j) = \pi(j)p(j, i)$$

Bien entendu, une chaîne réversible et ergodique converge en loi vers π . La question traitée dans ce chapitre est essentiellement la vitesse de convergence et sa quantification à l'aide de constantes géométriques. La dynamique de Metropolis est réversible pour la mesure de Gibbs. Voici quelques autres exemples de chaîne π -réversible.

Exemple 3.2.1 Soit P définie par

$$\forall i \neq j \in E, \quad p(i, j) = \begin{cases} \frac{1}{\#E} & \text{si } \pi(j) \geq \pi(i) \\ \frac{1}{\#E} \frac{\pi(j)}{\pi(i)} & \text{sinon.} \end{cases}$$

et $p(i, i) = 1 - \sum_{j \neq i} p(i, j)$.

Exemple 3.2.2 Soit P définie par

$$\forall i \neq j \in E, \quad p(i, j) = \frac{1}{\#E} \frac{\pi(j)}{\pi(i) + \pi(j)}$$

et $p(i, i) = 1 - \sum_{j \neq i} p(i, j)$.

Exemple 3.2.3 Soit P définie par

$$\forall i, j \in E, \quad p(i, j) = \pi(j) .$$

Le paragraphe suivant nous permettra de construire des exemples à volonté, à l'aide du théorème de représentation spectrale.

3.2.2 Structure Hilbertienne. Représentation spectrale

Nous adoptons pour l'étude des chaînes π -réversible un point de vue fonctionnel qui facilite la présentation de certains résultats. Mais le matériel présenté dans les paragraphes qui suivent peut aussi être vu à partir de considérations matricielles élémentaires. L'ensemble des fonctions définies sur E est muni d'un produit scalaire noté

$$\langle f, g \rangle = \sum_{i \in E} \pi(i) f(i) g(i) .$$

En référence à ce produit scalaire, l'espace vectoriel sera noté $\ell_2(\pi)$. La réversibilité se traduit par une propriété fonctionnelle de P , vu comme un opérateur sur $\ell_2(\pi)$.

Proposition 3.2.1 P est π -réversible si et seulement si l'opérateur P est auto-adjoint dans $\ell_2(\pi)$

$$\forall f, g \in \ell_2(\pi), \quad \langle Pf, g \rangle = \langle f, Pg \rangle .$$

Exercice 3.2.1 Réédiger la démonstration.

L'intérêt de ce point de vue est l'application du théorème de Hilbert-Schmidt.

Proposition 3.2.2 *Il existe une base orthonormale de fonctions propres de P dans $\ell_2(\pi)$, telle que la première fonction de base soit identiquement égale à 1. De plus, les valeurs propres de P sont réelles et comprises entre -1 et $+1$.*

Exercice 3.2.2 *Retrouver en utilisant la réversibilité que $\lambda^2 \leq 1$ pour toute valeur propre λ de P .*

Désormais les valeurs propres de P seront ordonnées de façon décroissante

$$-1 \leq \lambda_{|E|} \leq \dots \leq \lambda_2 < 1.$$

Théorème 3.2.1 Représentation spectrale. *Soit $f_1 \equiv 1, f_2, \dots, f_{|E|}$ une base orthonormale de fonctions propres de P dans $\ell_2(\pi)$. Alors,*

$$\forall i, j \in E, t \geq 0, \quad p^t(i, j) = \pi(j)(1 + \sum_{\ell=2}^{|E|} \lambda_\ell^t f_\ell(i) f_\ell(j))$$

où les λ_ℓ sont les valeurs propres de P .

Démonstration. Soit $i, j \in E$ et $t = 1$. Posons $f = \mathbb{1}_i$ et $g = \mathbb{1}_j$. Notons que

$$\langle Pf, g \rangle = \sum_l \pi(l) Pf(l) g(l) = \pi(j) p(j, i).$$

D'autre part,

$$Pf = \langle Pf, 1 \rangle 1 + \sum_{\ell=2}^{|E|} \langle Pf, f_\ell \rangle f_\ell$$

Or

$$\langle Pf, f_\ell \rangle = \begin{cases} \pi(i) \lambda_\ell f_\ell(i) & \text{si } \ell \geq 2 \\ \pi(i) & \text{si } \ell = 1. \end{cases}$$

Ainsi

$$\langle Pf, g \rangle = \pi(i) \pi(j) (1 + \sum_{\ell=2}^{|E|} \lambda_\ell f_\ell(i) f_\ell(j))$$

et le résultat suit pour $t = 1$. La même démarche s'applique à t quelconque. ■

3.2.3 Convergence

Définition 3.2.2 *Soient μ et ν deux mesures de probabilité sur E . On appelle distance en variation totale, la distance*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{i \in E} |\mu(i) - \nu(i)|$$

Remarque. Nous avons aussi

$$\|\mu - \nu\|_{TV} = \max_{A \subset E} |\mu(A) - \nu(A)|.$$

Pour une chaîne réversible, nous pouvons énoncer un résultat plus précis de convergence vers l'équilibre que celui donné lors de la démonstration du théorème de convergence des chaînes de Markov ergodiques (chapitre 2).

Théorème 3.2.2 Soit (X_t) une chaîne de Markov de matrice de transition P π -réversible. Alors

$$\|P(X_t = . | X_0 = i) - \pi(\cdot)\|_{TV}^2 \leq \frac{1 - \pi(i)}{4\pi(i)} \rho(P)^{2t}$$

où $\rho(P) = \max\{|\lambda_{|E|}|, \lambda_2\}$ et

$$-1 \leq \lambda_{|E|} \leq \dots \leq \lambda_2 < 1$$

désignent les valeurs propres de E rangées dans l'ordre décroissant.

Démonstration. D'après l'inégalité de Cauchy-Schwarz

$$\frac{1}{4} \left(\sum_{j \in E} |p^t(i, j) - \pi(j)| \right)^2 \leq \frac{1}{4} \sum_{j \in E} \pi(j) \left(\frac{p^t(i, j)}{\pi(j)} - 1 \right)^2.$$

Considérons une base orthonormale de fonctions propres $f_1 = 1, f_2, \dots$ et appliquons le théorème de représentation spectrale

$$\frac{1}{4} \sum_{j \in E} \pi(j) \left(\frac{p^t(i, j)}{\pi(j)} - 1 \right)^2 = \frac{1}{4} \sum_{j \in E} \pi(j) \left(\sum_{\ell \geq 2} \lambda_\ell^t f_\ell(i) f_\ell(j) \right)^2.$$

Ainsi, après permutation des symboles de sommation et par orthogonalité

$$\|P(X_t = . | X_0 = i) - \pi(\cdot)\|_{TV}^2 \leq \frac{1}{4} \sum_{\ell \geq 2} \lambda_\ell^{2t} f_\ell^2(i).$$

La matrice carrée de dimension $|E|$, $O = (o_{ij})$, définie par

$$\forall i, j \in E, \quad o_{ij} = \sqrt{\pi(i)} f_j(i)$$

(en identifiant les éléments de E et leurs indices) est une matrice orthogonale de $\mathbb{R}^{|E|}$. Par conséquent, nous avons

$$\sum_{\ell \geq 2} \pi(i) f_\ell^2(i) = 1 - \pi(i) f_1^2(i) = 1 - \pi(i).$$

Nous pouvons alors conclure à l'aide d'une majoration évidente. ■

3.2.4 Valeurs propres et formes de Dirichlet

Le théorème de représentation spectrale permet de formuler les valeurs propres qui nous intéressent (λ_2 et $\lambda_{|E|}$) de manière variationnelle. Posons

$$E[f] = \sum_{i \in E} \pi(i)f(i) = \langle 1, f \rangle$$

et

$$\text{Var}(f) = E[(f - E[f])^2] = \langle f - E[f], f - E[f] \rangle.$$

Définition 3.2.3 On appelle forme de Dirichlet la forme quadratique définie sur $\ell_2(\pi)$ par

$$\forall f \in \ell_2(\pi), \quad \mathcal{E}(f, f) = \langle (I - P)f, f \rangle$$

où I désigne l'opérateur identité. On note aussi

$$\forall f \in \ell_2(\pi), \quad \mathcal{F}(f, f) = \langle (I + P)f, f \rangle.$$

Lemme 3.2.1 Nous avons les identités suivantes. Pour tout $f \in \ell_2(\pi)$

$$\text{Var}(f) = \frac{1}{2} \sum_{i,j \in E} \pi(i)\pi(j)(f(i) - f(j))^2$$

et

$$\begin{aligned} \mathcal{E}(f, f) &= \frac{1}{2} \sum_{i,j \in E} \pi(i)p(i, j)(f(i) - f(j))^2 \\ \mathcal{F}(f, f) &= \frac{1}{2} \sum_{i,j \in E} \pi(i)p(i, j)(f(i) + f(j))^2. \end{aligned}$$

Exercice 3.2.3 Rédiger la démonstration.

Théorème 3.2.3 Soit P une matrice de transition π -réversible. Alors

$$1 - \lambda_2 = \inf\{\mathcal{E}(f, f); E[f] = 0; \text{Var}(f) = 1\}$$

et

$$1 + \lambda_{|E|} = \inf\{\mathcal{F}(f, f); E[f^2] = 1\}.$$

Démonstration. Ces résultats découlent du théorème de Hilbert-Schmidt. Considérons une base orthonormale (f_ℓ) telle que $f_1 \equiv 1$. Soit $f \in \ell_2(\pi)$ telle que $E[f] = 0$ et $E[f^2] = 1$. Alors

$$f = \sum_{\ell \geq 2} \langle f, f_\ell \rangle f_\ell$$

Nous avons

$$\mathcal{E}(f, f) = \sum_{\ell \geq 2} (1 - \lambda_\ell) \langle f, f_\ell \rangle^2$$

et

$$Var(f) = \sum_{\ell \geq 2} \langle f, f_\ell \rangle^2 = 1.$$

La grandeur $\mathcal{E}(f, f)$ est minimale pour $f = f_2$ sous la contrainte $Var(f) = 1$ et $E[f] = 0$ et $\mathcal{E}(f_2, f_2) = 1 - \lambda_2$ (pour le vérifier on peut introduire des multiplicateurs de Lagrange). Etablir la seconde identité est un exercice. ■

3.3 Inégalités géométriques

3.3.1 Définitions

L'objectif de ce paragraphe est de donner des bornes calculables pour les valeurs de λ_2 et $\lambda_{|E|}$, afin d'estimer la vitesse de convergence d'une chaîne réversible vers l'équilibre. De telles bornes feront intervenir des constantes géométriques calculées à partir du graphe de transition de la chaîne (pondéré par les probabilités de transition). Les constantes géométriques mettront en évidence certaines propriétés de la chaîne qui ralentissent l'accès à l'équilibre : goulots d'étranglement, pièges dont il est difficile de sortir, grandes distances à parcourir pour aller d'un état vers un autre, etc. Nous décrirons deux types de constante : les constantes isopérimétriques d'une part, et les constantes de Poincaré, d'autre part. Les constantes isopérimétriques sont directement inspirées de la géométrie des surfaces riemanniennes. Les constantes de Poincaré reposent sur les trajectoires de la chaîne. De nombreuses références étudient ces grandeurs dans le cas de marches aléatoires sur les graphes (Diaconis et Stroock, 1991), (Saloff-Coste, 1998). Nous étudierons plus particulièrement l'application à la dynamique de Metropolis (Ingrassia, 1994). Notons

$$\forall i, j \in E, \quad a(i, j) = \pi(i)p(i, j).$$

Définition 3.3.1 Soit S un sous-ensemble propre de E (i.e., $S \neq E$ et $S \neq \emptyset$). On appelle constante isopérimétrique de S la grandeur

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} a(i, j)}{\pi(S)},$$

où $\pi(S) = \sum_{i \in S} \pi(i)$. La constante isopérimétrique symétrique de S est

$$\bar{\phi}(S) = \frac{\phi(S)}{1 - \pi(S)}.$$

On note encore

$$\phi = \min_{S; 0 < \pi(S) < 1/2} \phi(S),$$

et

$$\bar{\phi} = \min_{S; 0 < \pi(S) < 1/2} \bar{\phi}(S).$$

La constante $\phi(S)$ s'interprète comme la probabilité pour que la chaîne sorte de S en un pas sachant qu'elle démarre d'un sommet de S , sous hypothèse de stationnarité. La constante ϕ sera petite s'il existe un “petit ensemble” (pour la loi invariante) à partir duquel il sera difficile de sortir. Cette constante représente donc un indice de la présence de pièges pour la chaîne. La constante $\bar{\phi}$ s'interprète de la même manière, mais $\bar{\phi}(S)$ possède l'avantage de la symétrie (qui la rend plus facile à manipuler, car le minimum peut être pris sur tous les sous-ensembles propres de E). Une petite valeur de ϕ ou de $\bar{\phi}$ indiquera que la chaîne est lente dans son accès à l'équilibre. Réciproquement, des chaînes avec de grandes constantes isopérimétriques seront relativement rapides.

Définissons maintenant un second type de constante géométrique, à partir de considérations trajectorielles sur la chaîne en question. Pour cela, fixons, pour tout couple de sommets (i, j) un chemin γ_{ij} entre i et j dans le graphe de transition. Nous notons

$$\Gamma = \{\gamma_{ij} ; i, j \in E\}.$$

De la sorte, Γ contient exactement un chemin entre tout couple de sommets. Nous choisirons l'ensemble Γ de sorte que les chemins ne bouclent pas sur eux-même.

Définition 3.3.2 La A -longueur d'un chemin $\gamma \in \Gamma$ est

$$|\gamma|_A = \sum_{e \in \gamma} a(e)^{-1}$$

où $e : e_- \rightarrow e_+$ désigne une arête entre deux sommets consécutifs de γ et $a(e) = a(e_-, e_+)$.

Définition 3.3.3 On note

$$\kappa = \max_e \sum_{\gamma_{ij} \ni e} |\gamma_{ij}|_A \pi(i) \pi(j)$$

où le maximum est pris sur toutes les arêtes du graphe de transition, et la somme est prise sur tous les chemins dans Γ qui empruntent l'arête e .

Commentaires. La constante κ à tendance à croître lorsque le nombre de chemins qui empruntent une arête donnée augmente. Elle peut donc être interprétée comme une mesure des goulets d'étranglement de la chaîne. Notons que la longueur des chemins a aussi une influence sur la valeur de cette constante.

Nous décrivons dans le même état d'esprit une seconde constante. De manière analogue, fixons $\Sigma = \{\sigma_i ; i \in E\}$ un ensemble de cycles ayant un nombre impair d'arêtes et tel que Σ contient un unique cycle passant par chaque sommet.

Définition 3.3.4 *On note*

$$\iota = \max_e \sum_{\sigma_i \ni e} |\sigma_i|_A \pi(i)$$

3.3.2 Bornes de Cheeger

Dans ce paragraphe, nous établissons des bornes inférieures et supérieures pour la seconde valeur propre. Ces bornes impliquent les constantes isopérimétriques ϕ et $\bar{\phi}$.

Proposition 3.3.1 *Soit S un sous ensemble propre de E . Alors, nous avons*

$$\lambda_2 \geq 1 - \bar{\phi}(S).$$

Démonstration. Par la formule variationnelle, nous avons

$$1 - \lambda_2 = \inf \left\{ \frac{\mathcal{E}(f, f)}{Var(f)} ; E[f] = 0 \right\}.$$

Considérons la fonction indicatrice de S et

$$f = \mathbb{1}_S - \pi(S)$$

Nous avons

$$Var(f) = \frac{1}{2} \sum_{i,j} \pi(i)\pi(j)(\mathbb{1}_S(i) - \mathbb{1}_S(j))^2$$

et, par un calcul élémentaire,

$$Var(f) = \pi(S)\pi(\bar{S}).$$

De plus

$$\mathcal{E}(f, f) = \frac{1}{2} \sum_{i,j} \pi(i)p(i,j)(\mathbb{1}_S(i) - \mathbb{1}_S(j))^2 = \sum_{i \in S, j \notin S} \pi(i)p(i,j).$$

Le résultat suit de manière évidente. ■

Corollaire 3.3.1

$$\begin{aligned}\lambda_2 &\geq 1 - \bar{\phi} \\ \lambda_2 &\geq 1 - 2\phi\end{aligned}$$

Exercice 3.3.1 *Rédiger la démonstration.*

La borne supérieure est plus difficile à établir que la borne inférieure. Elle provient d'une analogie avec un résultat plus ancien et remarquable en géométrie concernant le spectre d'une variété riemannienne (Cheeger, 1973). Pour les chaînes de Markov réversibles, le théorème est dû à Lawler et Sokal (1988) et Sinclair et Jerrum (1989) qui l'ont publié indépendamment (dans des revues différentes).

Proposition 3.3.2

$$\lambda_2 \leq 1 - \phi^2/2$$

La démonstration de cette inégalité repose sur deux observations intéressantes. Soit $f \in \ell_2(\pi)$. Posons

$$f_+ = \max(f, 0)$$

et

$$S(f) = \{i \in E ; f(i) > 0\}.$$

Lemme 3.3.1 *Soit $f \in \ell_2(\pi)$, $\beta \geq 0$ et supposons que $S(f)$ est non vide. Alors, nous avons*

$$\beta \langle f_+, f_+ \rangle \geq \mathcal{E}(f_+, f_+)$$

dès que $(I - P)f \leq \beta f$ sur $S(f)$.

Démonstration.

$$\begin{aligned}\frac{1}{2} \sum_{i,j} a(i,j)(f_+(i) - f_+(j))^2 &\leq \frac{1}{2} \sum_{i,j} a(i,j)(f_+(i) - f_+(j))(f(i) - f(j)) \\ &= \langle (I - P)f, f_+ \rangle\end{aligned}$$

Or, nous avons

$$\langle (I - P)f, f_+ \rangle = \sum_{i \in S(f)} \pi(i) f_+(i) (I - P)f(i)$$

et par hypothèse

$$\sum_{i \in S(f)} \pi(i) f_+(i) (I - P)f(i) \leq \sum_{i \in S(f)} \pi(i) f_+(i) \beta f(i) = \beta \langle f_+, f_+ \rangle.$$

■

Lemme 3.3.2 Pour tout $f \in \ell_2(\pi)$ telle que $S(f) \neq \emptyset$

$$\mathcal{E}(f_+, f_+) \geq \frac{\phi(f)^2}{2} < f_+, f_+ >$$

où $\phi(f) = \inf\{\phi(S) ; \emptyset \neq S \subset S(f)\}$.

Démonstration. On peut supposer sans perte de généralité que $f \geq 0$. Alors d'après l'inégalité de Cauchy-Schwarz (appliquée à deux reprises)

$$\begin{aligned} \sum_{i,j} a(i,j) |f^2(i) - f^2(j)| &\leq \sqrt{2}\mathcal{E}(f, f)^{1/2} \left(\sum_{i,j} (f(i) + f(j))^2 a(i,j) \right)^{1/2} \\ &\leq 2\mathcal{E}(f, f)^{1/2} \left(\sum_{i,j} (f^2(i) + f^2(j)) a(i,j) \right)^{1/2} \\ &= 2^{3/2} \mathcal{E}(f, f)^{1/2} \sqrt{< f, f >} \end{aligned}$$

D'autre part, nous avons

$$\begin{aligned} \sum_{i,j} a(i,j) |f^2(i) - f^2(j)| &= 2 \sum_{f(j) > f(i)} a(i,j) (f^2(j) - f^2(i)) \\ &= 4 \sum_{f(j) > f(i)} a(i,j) \int_{f(i)}^{f(j)} t dt \\ &= 4 \int_0^\infty t \sum_{f(j) \geq t > f(i)} a(i,j) dt. \end{aligned}$$

Or, nous avons

$$\sum_{f(j) \geq t > f(i)} a(i,j) = \sum_{i \in S_t, j \notin S_t} a(i,j)$$

avec $S_t = \{i ; f(i) > t\} \subset S(f)$. Une minoration évidente conduit à

$$4 \int_0^\infty t \sum_{f(j) \geq t > f(i)} a(i,j) dt \geq \phi(f) \int_0^\infty t \pi(S_t) dt$$

et par un calcul analogue au précédent,

$$\phi(f) \int_0^\infty t \pi(S_t) dt = \frac{1}{2} < f, f > .$$

■

En combinant les deux lemmes, nous pouvons donner la démonstration.

Exercice 3.3.2 *Rédiger la démonstration. On choisit $\beta = 1 - \lambda_2$ et pour f une fonction propre associée à λ_2 . Puisque $E[f] = 0$ (f est orthogonale à 1), nous avons $S(f) \neq \emptyset$ et on peut éventuellement choisir $-f$ à la place de f de sorte à ce que*

$$\pi(S_f) < \frac{1}{2}.$$

De manière similaire, on peut établir le résultat suivant, en impliquant $\bar{\phi}$ au lieu de ϕ .

Théorème 3.3.1

$$\lambda_2 \leq 1 - \frac{\bar{\phi}^2}{8}.$$

3.3.3 Bornes de Poincaré

Dans ce paragraphe, nous donnons des inégalités sur λ_2 et $\lambda_{|E|}$ impliquant les constantes κ et ι .

Proposition 3.3.3 (*Inégalité de Poincaré.*) *La seconde valeur propre d'une chaîne de Markov ergodique π -réversible satisfait*

$$\lambda_2 \leq 1 - 1/\kappa.$$

Démonstration. Soit $f \in \ell_2(\pi)$ telle que $E[f] = 0$ et introduisons l'ensemble de chemins Γ . Nous avons

$$Var(f) = \frac{1}{2} \sum_{i,j} (f(i) - f(j))^2 \pi(i) \pi(j)$$

et donc

$$\begin{aligned} Var(f) &= \frac{1}{2} \sum_{i,j} \left(\sum_{i_k \in \gamma_{ij}} f(i_{k-1}) - f(i_k) \right)^2 \pi(i) \pi(j) \\ &= \frac{1}{2} \sum_{i,j} \left(\sum_{e \in \gamma_{ij}} \sqrt{\frac{a(e)}{a(e)}} \Delta f(e) \right)^2 \pi(i) \pi(j) \end{aligned}$$

où $e : e_- \rightarrow e_+$ désigne une arête empruntée par le chemin γ_{ij} et

$$\Delta f(e) = f(e_-) - f(e_+).$$

D'après, l'inégalité de Cauchy-Schwarz, nous avons alors

$$\begin{aligned}
 Var(f) &\leq \frac{1}{2} \sum_{i,j} \left(\sum_{e \in \gamma_{ij}} \frac{1}{a(e)} \right) \left(\sum_{e \in \gamma_{ij}} a(e) \Delta^2 f(e) \right) \pi(i)\pi(j) \\
 &= \frac{1}{2} \sum_{i,j} |\gamma_{ij}|_A \pi(i)\pi(j) \sum_{e \in \gamma_{ij}} a(e) \Delta^2 f(e) \\
 &\leq \kappa \frac{1}{2} \sum_e a(e) \Delta^2 f(e) \\
 &= \kappa \mathcal{E}(f, f)
 \end{aligned}$$

Le résultat découle de la formulation variationnelle de la seconde valeur propre. ■

Exercice 3.3.3 Montrer que

$$\lambda_2 \leq 1 - 1/K$$

avec $K = \max_e \frac{1}{a(e)} \sum_{\gamma_{ij} \ni e} |\gamma_{ij}| \pi(i)\pi(j)$.

Proposition 3.3.4

$$\lambda_{|E|} \geq -1 + 2/\iota$$

Démonstration. Soit $f \in \ell_2(\pi)$. L'idée consiste à écrire

$$f(i) = \frac{1}{2} \{ (f(i) + f(i_1)) - (f(i_1) + f(i_2)) + \dots + (f(i_{r-1}) + f(i)) \}$$

où

$$\sigma_i : i \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_{r-1} \rightarrow i$$

est un cycle de longueur impaire dans Σ . Nous avons alors

$$E[f^2] = \sum_i \frac{\pi(i)}{4} \left(\sum_{e \in \sigma_i} \sqrt{\frac{a(e)}{a(e)}} (-1)^{k(e)} (f(e_+) + f(e_-)) \right)^2.$$

Ainsi, par l'inégalité de Cauchy-Schwarz

$$\begin{aligned}
 E[f^2] &\leq \sum_i \frac{\pi(i)}{4} |\sigma_i|_A \sum_{e \in \sigma_i} (f(e_+) + f(e_-))^2 a(e) \\
 &\leq \frac{\iota}{4} \sum_e (f(e_+) + f(e_-))^2 a(e) \\
 &\leq \frac{\iota}{2} \mathcal{F}(f, f)
 \end{aligned}$$

On conclut grâce à l'expression variationnelle de la dernière valeur propre. ■

Pour comparer λ_2 et $\lambda_{|E|}$, on utilise parfois une inégalité plus grossière que la précédente, mais plus facile à manipuler.

Proposition 3.3.5

$$\lambda_{|E|} \geq -1 + 2 \min_i p(i, i) .$$

Exercice 3.3.4 *Rédiger la démonstration.*

En particulier, lorsque $p(i, i) > 1/2$ pour tout $i \in E$, alors

$$\rho(P) = \lambda_2 .$$

Exercice 3.3.5 *A l'aide de la représentation spectrale, construire des exemples de chaînes réversibles à trois ou quatre états dont on connaît le spectre. Calculer les différentes constantes géométriques. Quand les bornes sont-elles bonnes ? mauvaises ?*

3.4 Application à la dynamique de Metropolis

Nous appliquons les résultats démontrés dans les paragraphes précédents à l'étude de la convergence de la dynamique de Metropolis. Les taux de transition de cette chaîne de Markov ont été calculés lors de la proposition 3.1.2. Notons, dans cette section, que f ne désigne plus une fonction test quelconque mais la fonction objectif. Nous supposons que son minimum absolu i_* est unique et $f(i_*) = 0$. Le processus est réversible pour la loi de probabilité de Gibbs ($T > 0$)

$$\forall i \in E , \quad \pi_T(i) = \exp(-f(i)/T)/Z_T$$

et

$$\forall i \neq j \in E , \quad p_T(i, j) = \begin{cases} 1/\deg(G) & \text{si } j \in N(i) \text{ et } \pi(j) \geq \pi(i) \\ \frac{\pi(j)}{\pi(i)\deg(G)} & \text{si } j \in N(i) \text{ et } \pi(j) < \pi(i) \\ 0 & \text{sinon.} \end{cases}$$

et $p_T(i, i) = 1 - \sum_{j \neq i} p_T(i, j)$.

Par manque d'espace, nous ne démontrerons que les résultats concernant la seconde valeur propre. Le lecteur intéressé pourra se reporter à l'article de Ingrassia (1994) pour les démonstrations concernant la plus petite valeur propre.

3.4.1 Choix des chemins

Les estimations faites pour la dynamique de Metropolis reposent sur un choix particulier de chemins du graphe (E, G) . En premier lieu, définissons l'*altitude* d'un chemin dans (E, G) .

Définition 3.4.1 Soit $i, j \in E$. Soit $\gamma_{ij} = (i_k)_{k=0, \dots, r}$ un chemin entre les sommets i et j dans le graphe (E, G) . L'altitude de ce chemin est

$$\text{alt}(\gamma_{ij}) = \max_{k=0, \dots, r} f(i_k).$$

On pose

$$H(i, j) = \min_{\gamma_{ij}} \text{alt}(\gamma_{ij}).$$

La grandeur $H(i, j)$ est la plus petite altitude qu'il faut dépasser pour aller de i à j .

Définition 3.4.2 (Hauteur critique de Hajek) On pose

$$h_* = \max_{i,j} \{H(i, j) - f(i) - f(j)\} = \max_{i \neq i_*} \{H(i, i_*) - f(i)\}.$$

Exercice 3.4.1 Calculer h_* pour des exemples construits à partir d'ensembles E de petite taille, puis pour le modèle d'Ising.

Définissons maintenant l'ensemble de chemins Γ qui nous servira à étudier la dynamique de Metropolis.

Définition 3.4.3 Soit $i, j \in E$. Le chemin γ_{ij} est admissible si

$$\text{alt}(\gamma_{ij}) - f(i) - f(j) \leq h_*.$$

L'ensemble Γ est choisi de sorte que tous les chemins sont admissibles.

3.4.2 Bornes sur λ_2

Nous considérons les constantes suivantes associées à l'ensemble Γ

$$b_\Gamma = \max_e \#\{\gamma \in \Gamma ; e \in \gamma\}$$

et

$$\gamma_\Gamma = \text{longueur maximale d'un chemin de } \Gamma.$$

La constante b_Γ s'interprète comme une mesure des goulots d'étranglement et γ_Γ comme une mesure d'éloignement.

Théorème 3.4.1 (Poincaré) Pour la dynamique de Metropolis, nous avons

$$\lambda_2 \leq 1 - \frac{Z_T}{b_\Gamma \gamma_\Gamma \deg(G)} e^{-h_*/T}.$$

Démonstration. Il s'agit d'appliquer la borne de Poincaré. Pour une arête $e \in G$, nous avons

$$a(e) = \pi(e_-) p_T(e_-, e_+) = \begin{cases} \pi(e_-)/\deg(G) & \text{si } \pi(e_-) \leq \pi(e_+) \\ \pi(e_+)/\deg(G) & \text{si } \pi(e_-) > \pi(e_+) \end{cases}$$

Posons

$$\pi(e_*) = \min(\pi(e_-), \pi(e_+))$$

alors, pour tout $\gamma_{ij} \in \Gamma$, nous avons

$$|\gamma_{ij}|_A = \sum_{e \in \gamma_{ij}} \frac{\deg(G)}{\pi(e_*)}.$$

Par conséquent,

$$|\gamma_{ij}|_A \pi(i) \pi(j) = \frac{\deg(G)}{Z_T} \sum_{e \in \gamma_{ij}} e^{(f(e_*) - f(i) - f(j))/T}$$

où $f(e_*) = \max(f(e_-), f(e_+))$. Par définition de Γ , cela implique que

$$\kappa \leq \frac{b_\Gamma \gamma_\Gamma \deg(G)}{Z_T} e^{h_*/T}$$

■

Théorème 3.4.2 (*Cheeger*) Pour la dynamique de Metropolis, nous avons

$$\lambda_2 \leq 1 - \frac{Z_T^2}{8(b_\Gamma \gamma_\Gamma)^2} e^{-2h_*/T}.$$

Démonstration. On utilise le fait que $\lambda_2 \leq 1 - \phi^2/2$ et que

$$\phi \geq \frac{1}{2\eta}$$

où $\eta = \max_e \left\{ \frac{1}{a(e)} \sum_{\gamma_{ij} \ni e} \pi(i) \pi(j) \right\}$. Exercice. ■

Exercice 3.4.2 Sous quelle condition la borne déterminée par l'inégalité de Poincaré est elle meilleure que celle de Cheeger ?

Pour terminer ce paragraphe, nous proposons une borne inférieure pour λ_2 .

Théorème 3.4.3 Pour la dynamique de Metropolis, nous avons

$$\lambda_2 \geq 1 - \frac{Z_T}{\deg(G)} \left(\frac{\#E}{2} \right)^2 e^{-h_*/T}.$$

Commentaires. Pour résumer, il existe deux constantes $C_1, C_2 > 0$ telles que

$$C_1 e^{-h^*/T} \leq 1 - \lambda_2 \leq C_2 e^{-h^*/T}.$$

Le comportement de la seconde valeur propre à basse température est donc déterminé par la constante h_* .

Démonstration. La démonstration repose sur un ensemble test S bien choisi et l'inégalité isopérimétrique. Soit i_0 et j_0 tels que $f(i_0) \leq f(j_0)$ et $H(i_0, j_0) - f(i_0) - f(j_0) = h_*$. Posons

$$S = \{j \in E ; H(i_0, j) < H(i_0, j_0)\}$$

alors $i_0 \in S$ et $j_0 \in \bar{S}$. Soit $i \in S$ et $j \notin S$, nous avons (exercice)

- $f(j) \geq H(i_0, j_0)$,
- $f(j) > f(i)$.

Finalement,

$$\begin{aligned} \lambda_2 &\geq 1 - \bar{\phi}(S) \\ &= 1 - \frac{\sum_{i \in S, j \notin S} \pi(j)/\deg(G)}{\pi(S)\pi(\bar{S})} \\ &\geq 1 - \frac{Z_T}{\deg(G)} \sum_{i \in S, j \notin S} e^{-(H(i_0, j_0) - f(i_0) - f(j_0))/T} \end{aligned}$$

■

3.4.3 Borne sur ρ

Pour donner une borne sur ρ , nous admettons le résultat suivant concernant la plus petite des valeurs propres (on peut démontrer ce résultat à l'aide de la seconde inégalité de Poincaré).

Théorème 3.4.4 (*Ingrassia, 1994*). *Nous avons*

$$\lambda_{|E|} \geq -1 + \frac{2}{\deg(G)} \left(A + \frac{B}{e^{\delta/T} - 1} \right)^{-1}$$

où A et B sont des constantes positives et

$$\delta = \min_{i \in E} \min_{j \rightarrow i ; f(i) \neq f(j)} |f(i) - f(j)|.$$

Finalement, nous avons

Théorème 3.4.5 *Pour tout T suffisamment petit,*

$$\rho \leq 1 - \frac{1}{b_\Gamma \gamma_\Gamma \deg(G)} e^{-h^*/T}.$$

Exercice 3.4.3 *Rédiger la démonstration.*

3.5 Un problème

Soit E un ensemble fini et P la matrice de transition d'une chaîne de Markov apériodique et irréductible définie sur E . Selon les notations du cours, π est la loi invariante associée à P . On note P^* l'opérateur adjoint de P dans $\ell_2(\pi)$

$$\forall f, g \in \ell_2(\pi), \quad \langle Pf, g \rangle = \langle f, P^*g \rangle.$$

Toutes les espérances et variances sont prises par rapport à la loi π .

I

- 1) Montrer, pour tout $i, j \in E$, que

$$p^*(i, j) = \frac{\pi(j)}{\pi(i)} p(j, i).$$

- 2) Montrer que $M = PP^*$ est une matrice de transition π -réversible.

- 3) On considère la forme de Dirichlet associée à l'opérateur M

$$\forall f \in \ell_2(\pi), \quad \mathcal{E}_M(f, f) = \langle (I - M)f, f \rangle.$$

Démontrer l'identité suivante (*identité de Mihail*)

$$\forall f \in \ell_2(\pi), \quad \text{Var}(f) = \text{Var}(P^*f) + \mathcal{E}_M(f, f).$$

(Établir en premier lieu le résultat pour une fonction d'espérance nulle.)

- 4) Montrer que les valeurs propres de M sont toutes positives.

II

- 1) Pour tout $i \in E$, on définit

$$\forall t \geq 0, \quad g_t(i) = \frac{\pi_t(i)}{\pi(i)}$$

où $\pi_t = \pi_0 P^t$ désigne la loi à l'instant t . Montrer que

$$\text{Var}(g_t) = \chi_t^2 := \sum_{i \in E} \frac{(\pi_t(i) - \pi(i))^2}{\pi(i)}.$$

- 2) Déduire de la question 1) la relation suivante

$$\forall t \geq 0, \quad \chi_t^2 = \chi_{t+1}^2 + \mathcal{E}_M(g_t, g_t).$$

- 3) On note $\lambda_2(M)$ la seconde plus grande valeur propre de M . Montrer, à l'aide de l'identité de Mihail, que

$$\forall t \geq 0, \quad \lambda_2(M)\chi_t^2 \geq \chi_{t+1}^2.$$

- 4) Soit $\|\nu - \pi\|_{TV}$ la distance en variation totale séparant les mesures de probabilité ν et π . Montrer, à l'aide de l'inégalité de Cauchy-Schwarz, que

$$\forall t \geq 0, \quad 4\|\pi_t - \pi\|_{TV}^2 \leq \chi_t^2.$$

En déduire

$$\forall t \geq 0, \quad 4\|\pi_t - \pi\|_{TV}^2 \leq \lambda_2(M)^t \chi_0^2.$$

III

- 1) On suppose que E est muni d'un graphe (connexe) et on considère la marche aléatoire Q sur ce graphe

$$\forall i, j (j \neq i) \in E, \quad q(i, j) = \begin{cases} 1/d_* & \text{si } i \text{ voisin de } j, \\ 0 & \text{sinon,} \end{cases}$$

et $q(i, i) = 1 - \sum_{j \neq i} q(i, j)$ (on note $d_* = \max\{\deg(i) ; i \in E\}$). Montrer, en utilisant l'inégalité de Poincaré associée à un ensemble de chemins Γ quelconque, que

$$\lambda_2(Q) \leq 1 - \frac{|E|}{d_*^2 \gamma_* b_*}$$

où γ_* est le nombre maximal d'arêtes contenues dans un chemin $\gamma \in \Gamma$ et

$$b_* = \max_e \{\gamma \in \Gamma ; e \in \gamma\}.$$

- 2) Soit n un entier impair. Soit R la matrice de transition de la marche aléatoire symétrique sur le groupe cyclique Z_n (les entiers modulo n)

$$\forall i \in Z_n, \quad r(i, i+1) = r(i, i-1) = \frac{1}{2}.$$

Montrer que

$$\lambda_2(R) \leq 1 - \frac{8n}{(n-1)^2(n+1)}.$$

- 3) On considère la marche aléatoire P sur Z_n allant dans le sens des aiguilles d'une montre

$$\forall i \in Z_n, \quad p(i, i+1) = 1 - p(i, i) = \frac{1}{2}.$$

Donner une estimation de la vitesse de convergence de la chaîne associée vers sa loi invariante.

Algorithmes stochastiques pour l'optimisation
Lundi 26 février 2001

Durée 3h - Tous documents autorisés.

Soit E un ensemble fini et π_T la loi de Gibbs associée à la fonction objectif f , que l'on suppose positive. Soit Q_T la matrice de transition de la dynamique de Metropolis. On modifie cette dynamique de sorte que $q_T(i, i) \geq 1/2$, pour tout $i \in E$. Dans ce problème, on suppose que le paramètre T varie avec le pas d'itération de la chaîne de la manière suivante

$$\forall t \geq 2, \quad T_t = \frac{f_{\max}}{\ln t},$$

où $f_{\max} = \max\{f(i), i \in E\}$. Nous supposons de plus que

$$h_* \leq \frac{f_{\max}}{6}.$$

Nous désignons désormais par P_t la loi de la chaîne non-homogène au pas t et par π_t la mesure de Gibbs π_{T_t} . Nous définissons enfin la norme $\|\cdot\|_t$ de la manière suivante

$$\forall g \in \ell_2(\pi_t), \quad \|g\|_t^2 = \sum_{i \in E} \frac{g^2(i)}{\pi_t(i)}.$$

L'objectif de ce problème est de montrer l'existence d'une constante C_0 et d'un indice t_0 tels que

$$\forall t \geq t_0, \quad \|P_{t+1} - \pi_{t+1}\|_{t+1}^2 \leq \frac{C_0}{\sqrt{t}}. \quad (*)$$

- 1) Soit π_* la loi uniforme sur les minima absolus de f . Montrer, à partir de l'inégalité (*), que

$$\lim_{t \rightarrow \infty} \|P_t - \pi_*\|_2 = 0$$

où $\|g\|_2 = \sqrt{\sum_{i \in E} g^2(i)}$ est la norme usuelle de $\ell_2(1)$.

- 2) Soit $i \in E$. Montrer que $\pi_t(i)/\pi_{t+1}(i) \leq (1 + 1/t)$. En déduire que

$$\|\pi_t - \pi_{t+1}\|_{t+1} \leq \frac{1}{\sqrt{t}}.$$

- 3) Soit π une loi de probabilité sur E et Q une matrice de transition ergodique π -réversible dont les valeurs propres sont positives. Démontrer que

$$\forall g \in \ell_2(\pi), \quad \text{Var}(g) \geq \text{Var}(Qg) + \mathcal{E}(g, g).$$

(Établir en premier lieu le résultat pour une fonction d'espérance nulle.)

- 4) Déduire de la question précédente, l'inégalité suivante

$$\|P_{t+1} - \pi_{t+1}\|_{t+1}^2 \leq \lambda_2(Q_{T_{t+1}}) \|P_t - \pi_{t+1}\|_{t+1}^2$$

où $\lambda_2(Q_{T_{t+1}})$ désigne la seconde valeur propre de la matrice $Q_{T_{t+1}}$. (On pourra s'inspirer du problème traité pendant le cours, partie II.)

- 5) En utilisant l'inégalité de Poincaré, montrer qu'il existe un pas t_1 tel que

$$\forall t \geq t_1, \quad \sqrt{\lambda_2(Q_{T_{t+1}})} \leq 1 - \frac{1}{t^{1/5}}$$

et un pas t_0 tel que

$$\forall t \geq t_0, \quad \sqrt{\lambda_2(Q_{T_{t+1}})} \sqrt{(1 + \frac{1}{t})} \leq 1 - \frac{1}{t^{1/4}}.$$

- 6) Montrer que

$$\|P_t - \pi_{t+1}\|_{t+1} \leq \|P_t - \pi_t\|_{t+1} + \|\pi_t - \pi_{t+1}\|_{t+1},$$

En déduire, pour $t \geq t_0$, que

$$\|P_{t+1} - \pi_{t+1}\|_{t+1} \leq (1 - \frac{1}{t^{1/4}}) \|P_t - \pi_t\|_t + \frac{1}{\sqrt{t}}.$$

- 7) En supposant $\|P_t - \pi_t\|_t \geq \frac{2}{t^{1/4}}$, montrer que

$$\frac{1}{\sqrt{t}} \leq \frac{1}{2t^{1/4}} \|P_t - \pi_t\|_t.$$

- 8) Terminer la démonstration du résultat (*) en raisonnant par récurrence et en discutant les cas

$$\|P_t - \pi_t\|_t \geq \frac{2}{t^{1/4}}$$

et

$$\|P_t - \pi_t\|_t < \frac{2}{t^{1/4}}.$$

Le problème est inspiré de l'article de A. Nolte et R. Schrader, "A note on the finite time behavior of simulated annealing", paru dans Mathematics of Operations Research, 25, 3, 2000, pp. 476-484.

Chapter 4

Chaînes de Markov à transitions rares

4.1 Définition

Nous considérons une famille de chaînes de Markov définies sur un ensemble d'états fini, noté X , dont le comportement est dicté par des événements rares. Formellement, une famille de chaînes de Markov de matrice de transition P_T , $T > 0$ est à transitions rares, s'il existe une collection de coûts positifs $\{V(x, y) , x, y \in X\}$ telle que

$$\forall x, y \in X, \quad c_1 e^{-V(x,y)/T} \leq p_T(x, y) \leq c_2 e^{-V(x,y)/T}$$

pour $c_1, c_2 \geq 0$. On définit parfois les chaînes à transitions rares à l'aide d'une condition plus faible

$$\forall x, y \in X, \quad -\lim_{T \rightarrow 0} \log p_T(x, y) = V(x, y).$$

Les deux conditions signifient que les probabilités de transition sont de l'ordre de $\exp(-V(x, y)/T)$ lorsque T est petit.

Exercice 4.1.1 Montrer que la dynamique de Metropolis définit une famille de chaînes de Markov à transitions rares. Montrer qu'il en est de même pour les algorithmes de population par mutation/sélection et mutation+sélection.

4.2 Lemmes sur les chaînes de Markov

Soit X un ensemble fini et P la matrice de transition d'une chaîne de Markov ergodique, définie sur X . Les lemmes en question sont en fait des résultats importants qui établissent des formules tout à fait générales pour la loi invariante et les temps d'atteinte d'un sous-ensemble donné (Freidlin et Wentzell, chapitre 6).

4.2.1 Sous-graphes

Nous avons besoin de définir des sous-graphes (orienté) du graphe de transition. Soit (X, G) une structure de graphe définie sur l'ensemble X et $A \subset X$.

Définition 4.2.1 Un graphe g appartient à $G(A)$ si

- de tout $x \in X \setminus A$ part exactement une arête,
- de tout $x \in A$ ne part aucune arête,
- g ne contient pas de cycle.

Définition 4.2.2 Pour tout sous-graphe g , on note

$$p(g) = \prod_{e \in g} p(e_-, e_+).$$

4.2.2 Mesure invariante

Théorème 4.2.1 Soit P une matrice de transition ergodique définie sur X et de graphe de transition (X, G) . Alors la loi stationnaire de la chaîne associée est

$$\forall x \in X, \quad \pi(x) = \frac{\sum_{g \in G(x)} p(g)}{Z}$$

où Z est une constante de normalisation.

Exercice 4.2.1 Rédiger la démonstration.

4.2.3 Temps d'atteinte

Soit $A \subset X$. On considère le temps d'atteinte de l'ensemble A par la chaîne (X_t) de matrice de transition P

$$T_A = \inf\{t \geq 1 ; X_t \in A\}.$$

Lemme 4.2.1 Soit $x \in A, y \in A$ et $N_A(y)$ le nombre de visites en y avant de sortir de A . Alors, nous avons

$$E[N_A(y) \mid X_0 = x] = \sum_{t=0}^{\infty} p_A^{(t)}(x, y) = (I_A - P_A)^{-1}(x, y)$$

où I_A désigne la matrice identité dans A et P_A est la restriction de la matrice P à $A \times A$.

Démonstration. Nous avons

$$N_A(y) = \sum_{t=0}^{\infty} \mathbb{1}_{(X_0 \in A; \dots; X_{t-1} \in A; X_t = y)}$$

d'où

$$E[N_A(y) \mid X_0 = x] = \sum_{t=0}^{\infty} P(X_0 \in A; \dots; X_{t-1} \in A; X_t = y \mid X_0 = x)$$

La propriété de Markov permet de conclure. ■

Lemme 4.2.2 Soit $x \in A, y \in A$

$$(I_A - P_A)^{-1}(x, y) = \frac{\sum_{g \in G_{x,y}(\bar{A} \cup y)} p(g)}{\sum_{g \in G(\bar{A})} p(g)}$$

où $G_{x,y}(\bar{A} \cup y)$ est l'ensemble des graphes de $G(\bar{A} \cup y)$ pour lesquels il existe un chemin entre x et y .

Démonstration. Vérification pénible. ■

Théorème 4.2.2 Pour tout $A \subset X$ (propre) et $x \in A$, nous avons

$$E[T_{\bar{A}} \mid X_0 = x] = \sum_{y \in A} \frac{\sum_{g \in G_{x,y}(\bar{A} \cup y)} p(g)}{\sum_{g \in G(\bar{A})} p(g)}$$

Exercice 4.2.2 Rédiger la démonstration.

4.3 Application aux chaînes à transitions rares

4.3.1 Mesure invariante

Théorème 4.3.1 Soit π_T la loi invariante d'une chaîne de Markov à transition rares ergodique. Nous avons

$$\lim_{T \rightarrow 0} -T \log \pi_T(x) = W(x) - W_{\min}$$

où

$$\forall x \in X, \quad W(x) = \min_{g \in G(x)} V(g)$$

avec

$$V(g) = \sum_{(y \rightarrow z) \in g} V(y, z).$$

Exercice 4.3.1 Rédiger la démonstration.

Commentaires.

- Le passage au logarithme et l'asymptotique considérée transforment la somme (d'exponentielles) en min et le produit en somme. On parle parfois de point de vue $(\min, +)$.

- Il faut interpréter ce résultat en pensant que π_T se comporte (grossièrement) comme une mesure de Gibbs lorsque T est petit

$$\pi_T(x) \approx \frac{e^{-W(x)/T}}{Z}.$$

Bien entendu, la mesure invariante se concentre sur l'ensemble

$$\mathcal{W}_* = \{x ; W(x) = W_{\min}\}.$$

Pour utiliser une chaîne de Markov à transition rares à des fins d'optimisation, il faudra pouvoir associer à tout élément de \mathcal{W}_* , un minimum absolu de la fonction objectif f car la chaîne va en toutes circonstances converger vers cet ensemble.

4.3.2 Temps d'atteinte

Théorème 4.3.2 Soit $a_* \in X$ et T_* le temps d'atteinte de a_* par la chaîne de matrice de transition P_T . Nous avons

$$\lim_{T \rightarrow 0} T \log E[T_* | X_0 = x] = W(a_*) - \min_{y \neq a_*} \min_{g \in G_{xy}(\{a_*, y\})} V(g).$$

Supposons que $\mathcal{W}_* = \{a_*\}$, alors

$$\lim_{T \rightarrow 0} T \log E[T_* | X_0 = x] = W_{\min} - W_2$$

où $W_2 = \min_{x \neq y} \min_{g \in G(x,y)} V(g)$.

Exercice 4.3.2 Rédiger la démonstration.

Commentaires. On peut montrer par ailleurs que

$$-\lim_{T \rightarrow 0} T \log(1 - \lambda_2) = -\lim_{T \rightarrow 0} T \log(1 - \rho_T) = W_{\min} - W_2 ; .$$

Ainsi lorsque la chaîne se concentre en un point unique (que l'on aimerait faire correspondre au minimum absolu de f)

$$1 - \lambda_2 \approx \frac{1}{E[T_*]},$$

un résultat que l'on pouvait intuiter dès le chapitre 2. Ce résultat montre directement, pour la dynamique de Metropolis, que le temps d'atteinte de i_* sera de l'ordre de $\exp(h_*/T)$ (le vérifier).

Définition 4.3.1 Soit $A \subset X$ et $x \in A$. On appelle hauteur de sortie de l'ensemble A partant de x , la constante

$$H_A(x) = \lim_{T \rightarrow 0} T \log E[T_A | X_0 = x]$$

et

$$H_A = \max_{x \in A} H_A(x).$$

Exercice 4.3.3 Etablir une formule générale pour $H_A(x)$, dans le cas d'une chaîne à transitions rares.

Pour terminer ce paragraphe, nous supposons que le minimum de W est atteint en un point unique, noté a_* , et nous précisons le comportement du temps d'atteinte de a_* . Posons, pour la suite

$$h_* = W_{\min} - W_2.$$

Nous montrons que T_* est de l'ordre de $e^{h_*/T}$ (c'est à dire concentré autour de son espérance).

Proposition 4.3.1 Soit $\epsilon > 0$ et $x \neq a_*$. Nous avons

$$\lim_{T \rightarrow 0} -T \log P(T_* > e^{(\epsilon+h_*)/T} \mid X_0 = x) = \infty$$

et

$$\liminf_{T \rightarrow 0} -T \log \min_{x \neq a_*} P(T_* < e^{(h_*-\epsilon)/T} \mid X_0 = x) \geq \epsilon$$

Démonstration. Démontrons la première identité. Par la propriété de Markov, nous avons, pour k, ℓ entiers

$$P(T_* > k\ell \mid X_0 = x) \leq \left(\max_{x \neq a_*} P(T_* > k \mid X_0 = x) \right)^\ell.$$

En choisissant $k = \exp((\epsilon/2 + h_*)/T)$ et $\ell = \lfloor \exp(\epsilon/2T) \rfloor$, nous obtenons

$$\begin{aligned} P(T_* > e^{(\epsilon+h_*)/T} \mid X_0 = x) &\leq \left(\max_{x \neq a_*} P(T_* > k \mid X_0 = x) \right)^\ell \\ &\leq \left(\max_{x \neq a_*} E[T_* \mid X_0 = x] e^{-(\epsilon/2+h_*)/T} \right)^\ell. \end{aligned}$$

En passant au logarithme

$$\log(P()) \leq \lfloor \exp(\epsilon/2T) \rfloor \left(\max_{x \neq a_*} \log E[T_* \mid X_0 = x] - (\epsilon/2 + h_*)/T \right)$$

et le premier résultat suit. Pour le second, rappelons que

$$E[Y] \leq \sum_{n \geq 0} P(Y > n)$$

si $Y \geq 0$ (un exercice de théorie de la mesure!). Par conséquent,

$$\begin{aligned} E[T_* e^{-(h_*-\epsilon)/T} \mid X_0 = x] &\leq \sum_{k \geq 0} P(T_* > k e^{-(\epsilon-h_*)/T} \mid X_0 = x) \\ &\leq \sum_{k \geq 0} \left(\max_{x \neq a_*} P(T_* > e^{-(\epsilon-h_*)/T} \mid X_0 = x) \right)^k \end{aligned}$$

En sommant la série géométrique, on conclut

$$\min_{x \neq a_*} P(T_* \leq e^{(h_*-\epsilon)/T} \mid X_0 = x) \leq e^{(h_*-\epsilon)/T} E[T_* \mid X_0 = x]^{-1}$$

■

4.4 Propriétés des chaînes à transitions rares

4.4.1 Altitudes

Définition 4.4.1 On considère une famille de chaîne de Markov à transitions rares associée à la fonction de coût V . Pour tout chemin $\gamma = (x_k)_{k=0,\dots,r}$ de x vers y dans le graphe de transition, on définit l'altitude par

$$\text{alt}(\gamma) = \max_{i=0,\dots,r-1} W(x_k) + V(x_k, x_{k+1}).$$

Exercice 4.4.1 Montrer que cette définition coïncide avec celle donnée précédemment dans le cas de la dynamique de Metropolis.

Définition 4.4.2 Pour tout $x, y \in X$, on pose

$$H(x, y) = \min_{\gamma_{xy}} \text{alt}(\gamma_{xy}).$$

Commentaires. La notion de cycle joue un rôle clé dans le formalisme des chaînes à transitions rares. Par manque d'espace, nous n'y ferons qu'une brève allusion ici (mais nous utiliserons cette notion dans le chapitre suivant concernant l'application aux dynamiques de population). La démonstration du résultat ci-dessous utilise cette notion. Nous l'admettrons. Le lecteur intéressé pourra consulter les notes de Catoni (1997).

Théorème 4.4.1 Pour tout sous-ensemble propre de X , la hauteur de sortie est donnée par

$$H(A) = \max_{x \in A} \min_{y \notin A} \{H(x, y) - W(x)\}.$$

En particulier, si a_* est un point quelconque de X , nous avons un équivalent (logarithmique) du temps d'atteinte de ce point

$$H(X \setminus a_*) = \min_{x \neq a_*} \{H(a_*, x) - W(x)\}.$$

Cette définition rappelle donc (et généralise) celle de la hauteur critique de Hajek, introduite pour la dynamique de Metropolis.

4.4.2 Convergence

Nous cherchons à estimer la probabilité d'être dans l'état y au bout d'un temps long, et à montrer que cette probabilité est proche de la probabilité stationnaire.

Proposition 4.4.1 *Supposons que $\mathcal{W}_* = \{a_*\}$. Choisissons α tel que*

$$h_* = H(X \setminus a_*) < \alpha$$

alors, pour tout $x, y \in X$ et $N = \lfloor e^{\alpha/T} \rfloor$

$$\liminf_{T \rightarrow 0} -T \log P(X_N = y \mid X_0 = x) \geq \min\{\alpha - h_*, W(y) - W_{\min}\}.$$

La démonstration repose sur le lemme suivant.

Lemme 4.4.1 *Soit P une chaîne de Markov ergodique de loi invariante π . Alors, nous avons*

$$\forall x, y \in X, \quad \sup_{t \geq 0} p^t(x, y) \leq \frac{\pi(y)}{\pi(x)}$$

Démonstration. Fixons $x \in X$. On pose $g_t(y) = p^t(x, y) \frac{1}{\pi(y)}$, $y \in X$. Alors, on vérifie facilement que

$$g_{t+1}(y) \leq \max_z g_t(z).$$

Il vient

$$\max_z g_t(z) \leq \max_z g_0(z) = \frac{1}{\pi(x)}$$

et le résultat suit. ■

Nous pouvons démontrer le théorème.

Démonstration. Il s'agit de majorer brutalement. Pour atteindre y , on peut passer par a_* ou non. Ainsi,

$$\begin{aligned} P(X_N = y \mid X_0 = x) &\leq P(X_N = y; T_{a_*} > e^{\alpha/T} \mid X_0 = x) \\ &\quad + P(X_N = y; T_{a_*} \leq e^{\alpha/T} \mid X_0 = x) \\ &\leq P(T_{a_*} > e^{\alpha/T} \mid X_0 = x) + \sup_t p^t(a_*, y) \\ &\leq e^{-\alpha/T} E[T_{a_*} \mid X_0 = x] + \frac{\pi(y)}{\pi(a_*)} \end{aligned}$$

Nous avons donc

$$\liminf_{T \rightarrow 0} -T \log P \geq \min(\alpha - h_*, W(y) - W(a_*)).$$
■

Montrons maintenant une “inégalité” dans le sens inverse. Pour simplifier, on supposera que \mathcal{W}_* est réduit à un point.

Proposition 4.4.2 *Supposons que $\mathcal{W}_* = \{a_*\}$. Choisissons α tel que*

$$h_* = H(X \setminus a_*) < \alpha$$

alors, pour tout $x, y \in X$ et $N = \lfloor e^{\alpha/T} \rfloor$

$$\limsup_{T \rightarrow 0} -T \log P(X_N = y \mid X_0 = x) \leq W(y) - W_{\min}.$$

Commentaires. Les deux propositions que nous venons d'énoncer s'interprètent en disant que sur échelle de temps de l'ordre de $e^{\alpha/T}$ pour $\alpha > h_*$, nous avons grossièrement, si $W(y)$ est proche de W_{\min} ,

$$P(X_N = y \mid X_0 = x) \approx e^{-(W(y)-W_{\min})/T} \approx \pi_T(y).$$

En choisissant $T = \alpha / \log(t)$, on obtient l'estimation suivante

$$P(X_t = y \mid X_0 = x) \approx (1/t)^{(W(y)-W_{\min})/\alpha}.$$

Ainsi, pour la dynamique de Metropolis, cela permet d'obtenir l'estimation d'erreur suivante, si $T_t = \alpha / \log(t)$ ($W = f$ et $f_{\min} = 0$),

$$P(f(X_t) \geq \theta \mid X_0 = x) \approx \left(\frac{1}{t}\right)^{\theta/\alpha}$$

qui tend vers 0 lorsque t grandit.

Démonstration. La démonstration de la proposition reprend l'idée de couplage mentionnée au chapitre 2. On considère la chaîne définie sur $X \times X$ à partir de deux copies indépendantes de la chaîne de départ. La première copie démarre du point x alors que la seconde est initialisée en un point tiré selon la loi invariante. Le couplage est encore une chaîne à transitions rares associée à la fonction de coût

$$V_2((x_1, x_2), (y_1, y_2)) = V(x_1, x_2) + V(y_1, y_2)$$

On montre que $H_2(X \times X \setminus (a_*, a_*)) = h_*$ (exercice). Prenons N comme dans l'énoncé. Alors

$$P_{x,\pi}(X_N^1 = y) \geq P_{x,\pi}(X_N^1 = y, T_{(a_*, a_*)} \leq N)$$

Puisque l'on passe par (a_*, a_*) , on a, d'après la propriété de Markov

$$P_{x,\pi}(X_N^1 = y, T_{(a_*, a_*)} \leq N) = P_{x,\pi}(X_N^2 = y, T_{(a_*, a_*)} \leq N)$$

et

$$P_{x,\pi}(X_N^1 = y) \geq P_{\pi}(X_N^2 = y) - P_{x,\pi}(T_{(a_*, a_*)} > N)$$

On conclut du fait que

$$P_{\pi}(X_N^2 = y) = \pi(y)$$

et grâce à l'inégalité de Tchebishev. ■

Chapter 5

L'algorithme de Mutation/Sélection

5.1 Préambule

Une chaîne de Markov à transitions rares peut définir une méthode de recherche approchée du minimum absolu d'une fonction. Nous avons étudié le cas de l'algorithme de Metropolis, pour lequel les résultats du chapitre précédent s'appliquent immédiatement. L'étude des algorithmes de population, définis dans le premier chapitre, relève aussi du formalisme des chaînes de Markov à transitions rares. L'application des résultats n'est toutefois pas aussi directe que pour l'algorithme de Metropolis. En particulier, la chaîne doit "visiter" très souvent les "bonnes" populations (celles qui contiennent le minimum absolu de f). Mathématiquement, cela signifie que la loi stationnaire doit être concentrée sur cet ensemble de populations. Nous verrons que cela se réalise lorsque la dimension du vecteur de solution (la taille de la population) est assez grande. Dans le formalisme du chapitre 4, cela revient à dire que le minimum de l'énergie virtuelle coïncide avec celui de f (en un sens que nous rendrons clair). La vitesse de convergence de l'algorithme sera précisée à travers l'estimation des temps d'atteinte de la solution optimale.

Nous allons détailler les résultats pour l'algorithme de mutation/sélection. Le lecteur intéressé par l'algorithme génétique de mutation+sélection pourra lire l'article (les articles) de R. Cerf. L'étude faite dans ce chapitre suit assez fidèlement la démarche adoptée par Cerf (1996, 1997). L'avantage de cette présentation est de pouvoir expliciter de manière poussée les différentes constantes d'importance (les indices géométriques). Le chapitre qui vient reproduit un article récent consacré à la dynamique de mutation/sélection (et contient quelques redondances). Il peut être lu indépendamment des autres chapitres.

5.2 Introduction

Les algorithmes de population constituent des méthodes génériques d'optimisation dont l'utilité pratique est largement reconnue, en particulier pour des problèmes multimodaux, chahutés ou discontinus. Ces méthodes reposent sur des opérateurs prob-

abilistes, conçus sur le modèle de l'évolution naturelle. Les algorithmes se fondent sur le comportement collectif d'une *population* d'individus en interaction, chaque individu représentant une solution potentielle au problème d'optimisation.

La population est initialisée de manière arbitraire, puis itérativement modifiée au moyen d'opérateurs stochastiques appelés *mutation*, *sélection* et *recombinaison* (qui est parfois ommise). L'interaction entre individus provient du processus de sélection, qui est en fait le principal opérateur d'optimisation. La mutation est un opérateur d'exploration, dont le rôle est d'introduire de nouveaux concurrents dans le processus de sélection. Les nouvelles solutions créées par cet opérateur sont souvent appelées *descendance*, et certains descendants sont susceptibles de remplacer leurs "parents". Bien que les mérites de l'opérateur de recombinaison soient souvent loués [11], [30], ce dernier n'est pas nécessaire à la convergence vers la solution optimale. Nous l'abandonnerons donc. Trois courants principaux peuvent être identifiés dans la littérature concernant les algorithmes évolutionnaires : les algorithmes génétiques, la programmation évolutionnaire et les stratégies d'évolution. Pour une vue d'ensemble de ces différentes instances, on pourra consulter les références [23, 5].

De nombreux auteurs ont proposé d'adapter les résultats concernant le recuit simulé aux algorithmes génétiques. Dans [17], la "probabilité de mutation" converge vers zero. Cerf [10] utilise le même point de vue, mais en renforçant simultanément la sélection (roulette de Boltzmann). La relation entre ces algorithmes et le recuit simulé a été établie à travers une théorie appelée *recuit généralisé* (GSA), fondée sur les propriétés des chaînes de Markov à transitions rares [53, 7].

Par ailleurs, relativement peu d'efforts ont été consacrés aux stratégies d'évolution [5] dans le contexte du recuit généralisé. La méthode d'*exploration/sélection* (E/S) repose sur un principe algorithmique proche de ces stratégies [24]. Les deux méthodes utilisent un processus de sélection déterministe. Toutefois, il existe des différences méthodologiques importantes. En particulier, le nombre de descendants (obtenus par mutation) sera aléatoire dans E/S, alors qu'il est constant dans les stratégies d'évolution. De plus, le pourcentage de descendants tend vers zéro dans E/S, et joue ainsi le rôle d'une température.

L'analyse de l'algorithme suivra sagement l'approche utilisée par Cerf pour les algorithmes génétiques [10, 11]. Toutefois, tout sera plus simple. Par conséquence, les constantes géométriques critiques seront plus explicites. Résumons les résultats que Cerf a obtenu pour les algorithmes génétiques (sans préciser les hypothèses). Tout d'abord, il existe une taille de population critique, en dessous de laquelle la population peut se trouver piégée dans une solution sous-optimale. La taille critique est sensible au "diamètre" de l'espace de recherche (voir [10]). De plus, la convergence de l'algorithme peut être quantifiée au moyen de deux grandeurs appelées *hauteur critique* et *exposant de convergence*. Ces deux grandeurs mesurent pour l'algorithme la difficulté posée par le problème d'optimisation. Nous rappelerons plus loin la définition de ces grandeurs. Malheureusement, il n'est pas possible en général d'en fournir une expression analytique simple, ni même des bornes suffisamment explicites.

En revanche, nous établirons dans ce chapitre une description assez précise de ces grandeurs clés pour l'algorithme de mutation/sélection. Voici l'organisation du

chapitre. Le paragraphe 5.3 définit l'algorithme de base. Ce dernier utilise un couplage entre mutation et sélection différent de celui des algorithmes génétiques. Les résultats principaux seront énoncés dans le paragraphe 5.4. Le paragraphe 5.5 rappelle les notations utilisées ainsi que le rôle des outils de grandes déviations. Les résultats auxiliaires et les démonstrations sont présentés dans le paragraphe 5.7. Le chapitre se conclut par une discussion des problèmes d'implantation, et donne une comparaison de la méthode avec l'utilisation parallèle du recuit simulé.

5.3 L'algorithme d'Exploration/Sélection

Soit E un ensemble de cardinal fini et f une fonction positive :

$$f : E \rightarrow \mathbb{R}_+. \quad (5.1)$$

Notons A^* le sous-ensemble des minima de f

$$A^* = \operatorname{argmin}_{a \in E} f(a). \quad (5.2)$$

Mis à part le paragraphe 5.7.1, les résultats seront établis sous l'hypothèse d'injectivité

$$f(a) = f(b) \text{ implique } a = b. \quad (5.3)$$

Le sous-ensemble A^* se réduit alors à $A^* = \{(a^*)\}$.

L'espace E est muni d'une structure de graphe (E, G) appelé *graphe d'exploration*. La stratégie d'Exploration/Sélection utilise un vecteur de solutions potentielles au problème de minimisation. Chaque solution est considérée comme un individu qui effectue une marche aléatoire sur le graphe d'exploration. Le processus d'exploration agit indépendamment sur chaque individu, et consiste à choisir un voisin au hasard dans le graphe. Par analogie avec les autres algorithmes évolutionnaires, le processus d'exploration sera aussi appelé *Mutation*. Le voisin résultant du pas de la marche aléatoire sera dit *descendant*. La stratégie comporte deux paramètres : le premier est la *taille de la population*

$$n \geq 2, \quad (5.4)$$

et le second est appelé *probabilité de mutation*

$$0 < p < 1. \quad (5.5)$$

Ce dernier paramètre représente la proportion de descendants à chaque génération. L'espace des populations est défini de la manière suivante

$$X = E^n, \quad (5.6)$$

(n répliques de E). Pour toute population $x = (x_1, \dots, x_n) \in X$, nous notons

$$E_x = \{x_1, \dots, x_n\} \subset E \quad (5.7)$$

l'ensembles des “types” contenus dans x . Le point minimal dans E_x est noté

$$x_* = \operatorname{argmin}_{x_i \in E_x} f(x_i). \quad (5.8)$$

Le graphe d’exploration est supposé symétrique (non-orienté) et connexe. Nous notons $\deg(a)$ le degré du sommet $a \in E$, et $N(a) \subset E$ son voisinage. L’algorithme se décrit informellement comme suit.

- Choisir arbitrairement n individus dans E ;
- Répéter
 - 1) Sélectionner le meilleur individu de la population;
 - 2) Pour $i = 1 \dots n$, créer un descendant de l’individu i et remplacer le parent avec la probabilité p ; avec la probabilité $(1 - p)$, remplacer l’individu i by l’individu sélectionné à l’étape 1);

jusqu’à **arrêt**.

Au pas 2), nous imposons au descendant d’être différent de l’individu sélectionné au pas 1). Tous les choix se font indépendamment. Le critère d’arrêt est un nombre fini (grand) de générations.

Modéliser les algorithmes génétiques à l’aide de chaînes de Markov est un procédé classique [46, 43, 14, 11]. L’état de la population à la génération t is noté X_t , et les probabilités de transition sont données par

$$\forall x, y \in X, \quad \operatorname{Prob}(X_{t+1} = y \mid X_t = x) = q(x, y) \quad (5.9)$$

où

$$q(x, y) = \prod_{i=1}^n (p a(x, y_i) + (1 - p) \delta_{x_*, y_i}), \quad (5.10)$$

et, pour tout $y_i \neq x_i$,

$$a(x, y_i) = \begin{cases} \frac{1}{\deg(x_i)} & \text{si } y_i \in N(x_i) \setminus \{x_*\}, \\ 0 & \text{sinon.} \end{cases} \quad (5.11)$$

Comme d’usage, $\delta_{a,b}$ désigne le symbole de Kroneker, égal à 1 si $a = b$, et 0 sinon. L’opérateur d’exploration a est complètement défini en posant

$$a(x, x_i) = 1 - \sum_{y_i \neq x_i} a(x, y_i). \quad (5.12)$$

Donnons une autre description des probabilités de transition. Pour x, y deux populations, considérons l’ensemble d’indices entiers $i \in \{1, \dots, n\}$ défini par

$$I(x, y) = \{1 \leq i \leq n ; y_i \neq x_*\}. \quad (5.13)$$

le cardinal de cet ensemble est noté

$$V_1(x, y) = \#I(x, y) . \quad (5.14)$$

Nous avons

$$\forall x, y \in X , \quad q(x, y) = \pi(x, y) p^{V_1(x, y)} (1 - p)^{n - V_1(x, y)} \quad (5.15)$$

où

$$\pi(x, y) = \prod_{i \in I(x, y)} a(x, y_i) . \quad (5.16)$$

Pour mettre en évidence la relation avec le recuit, nous introduisons un paramètre positif $T > 0$ et écrivons la probabilité de mutation de la manière suivante :

$$p = e^{-1/T} . \quad (5.17)$$

Le paramètre T est appelé *temperature*. Lorsque T tend vers 0, nous avons

$$\forall x, y \in X , \quad q(x, y) \sim \exp(-V_1(x, y)/T) \quad (5.18)$$

où le symbole \sim signifie

$$-T \log q(x, y) \rightarrow V_1(x, y) , \quad \text{lorsque } T \rightarrow 0 , \quad (5.19)$$

et la définition de $V_1(x, y)$ est étendue de sorte que

$$V_1(x, y) = \infty , \quad \text{si } \pi(x, y) = 0 . \quad (5.20)$$

5.4 Résultats principaux

5.4.1 Définitions

Pour commencer, nous rappelons certaines définitions liées au graphe (E, G) . Un chemin est une suite de sommets

$$\gamma : a^0 \rightarrow a^1 \rightarrow \dots \rightarrow a^r , \quad a^i \in E , i = 0, \dots, r \quad (5.21)$$

reliés par les arêtes $a^i \rightarrow a^{i+1}$. La longueur d'un chemin est le nombre r d'arêtes dans le chemin. La "distance", i.e., la longueur du chemin le plus court, entre deux sommets a et b est notée $d(a, b)$.

Ensuite, nous définissons deux grandeurs géométriques d'importance cruciale vis à vis de la convergence de l'algorithme. Le premier indice concerne la taille de population. Il est défini par

$$n_* = \max_{a \neq a^*} d(a, a^*) \quad (5.22)$$

où a^* est le minimum absolu de f . Cet indice correspond à la distance maximale séparant a^* d'un sommet arbitraire.

Le second indice quantifie la difficulté de l'algorithme à traiter le problème de minimisation. Formellement, il est défini par

$$d_* = \max_{a \neq a^*} \min_{b: f(b) < f(a)} d(a, b). \quad (5.23)$$

Le maximum est atteint en un sommet qui présente un minimum local de f . En fait, cet indice mesure la plus grande distance possible entre un minimum local et un sommet meilleur. Cette constante peut être vue comme une mesure de sortie des minima locaux (non-globaux). Notons que Suzuki (1993) a considéré une grandeur analogue dans le contexte des algorithmes génétiques élitistes.

5.4.2 Énoncés

Théorème 5.4.1 Temps d'atteinte de a^* . Soit $n > n_*$ et

$$\tau = \inf\{t \geq 1, X_t \ni a^*\}.$$

Alors, pour tout $x \not\ni a^*, x \in X$, nous avons

$$T \log E[\tau \mid X_0 = x] \rightarrow d_*, \quad \text{lorsque } T \rightarrow 0.$$

En pratique, la température T peut être modifiée à chaque génération de l'algorithme. Comme pour le recuit, nous cherchons un schéma décroissant convergent (voir le paragraphe 4.4). Notons X_t au lieu de $X_t^{T(t)}$, $t \geq 1$, pour signifier que la température T dépend de t . La chaîne (X_t) devient alors non-homogène.

Théorème 5.4.2 Schéma optimal. Soit $n > n_*$ et $x \in X$. Considérons une suite décroissante $(T(t))_{t \geq 0}$ qui converge vers zéro. Alors, nous avons

$$\text{Prob}(X_t \ni a^* \mid X_0 = x) \rightarrow 1.0 \quad \text{lorsque } t \rightarrow \infty \quad (5.24)$$

si et seulement si

$$\sum_{t=0}^{\infty} e^{-d_*/T(t)} = \infty. \quad (5.25)$$

Théorème 5.4.3 Exposant optimal. Soit $n > n_*$. Il existe deux constantes $R_1 > 0$ et $R_2 > 0$, et une constante α_* telle que, pour tout $t \geq 1$

$$\frac{R_1}{t^{\alpha_*}} \leq \inf_{0 \leq T(t) \leq \dots \leq T(1)} \max_{x \in X} P(X_t \not\ni a^* \mid X_0 = x) \leq \frac{R_2}{t^{\alpha_*}}.$$

De plus, la constante α_* satisfait à

$$\frac{n - n_*}{d_*} \leq \alpha_* \leq \frac{n(n_* + 1 - d_*) - 1}{d_*}.$$

Cette constante est l'exposant de convergence de (X_t) (à définir dans le paragraphe 5.5).

5.5 Notations et rappels

La clé de ce chapitre est que notre modèle relève du le formalisme du recuit généralisé présenté dans [53]. Ce cadre est bien adapté pour étudier des algorithmes de type Metropolis. Plusieurs modèles d'algorithmes génétiques ont déjà été étudiés à l'aide de ce formalisme [10, 11, 18]. Un noyau markovien q définit un algorithme de Metropolis généralisé, ou recuit généralisé s'il existe $\kappa > 0$ t.q.

$$\frac{1}{\kappa} \pi(x, y) e^{-V_1(x, y)/T} \leq q(x, y) \leq \kappa \pi(x, y) e^{-V_1(x, y)/T} \quad (5.26)$$

où V_1 (le coût de communication) satisfait $V_1(x, y) \geq 0$ et $V_1(x, y) = +\infty$ ssi $\pi(x, y) = 0$.

Le *coût de communication en plusieurs pas* de x vers y in X est défini par

$$V(x, y) = \inf \left\{ \sum_{k=0}^{r-1} V_1(x_k, x_{k+1}), x_0 = x, x_k \in X, x_r = y, r \geq 1 \right\}. \quad (5.27)$$

Énergie virtuelle. Rappelons que l'énergie virtuelle W , définie sur X , décrit le comportement asymptotique de la chaîne. Dans [25], un équivalent logarithmique de la loi de probabilité stationnaire π_T est donné

$$\forall x \in X, \lim_{T \rightarrow 0} -T \log \pi_T(x) = W(x) - W_* \quad (5.28)$$

où W_* est la valeur minimale de W . Soit \mathcal{W}^* l'ensemble des populations minimales pour W . L'équation (5.28) dit que la loi π_T se concentre sur \mathcal{W}^* .

Altitudes. Pour $x, y \in X$, $x \neq y$, et une trajectoire γ_{xy} de la chaîne (X_t) (de x vers y)

$$\gamma_{xy} = (x_0 = x \rightarrow x_1 \rightarrow \dots \rightarrow x_r = y), \quad (5.29)$$

l'*altitude* est définie par

$$H(\gamma_{xy}) = \max_{0 \leq k < r} \{W(x_k) + V(x_k, x_{k+1})\} \quad (5.30)$$

où le max est pris sur l'ensemble des sommets de γ_{xy} . Soit $H(x, y)$ la valeur minimale de $H(\gamma_{xy})$ pour toutes les trajectoires γ_{xy} . La grandeur $H(x, y)$ est appelée *altitude de communication* entre x et y , et est symétrique [53].

Définition des cycles. Soit $\lambda \geq 0$ et

$$W_\lambda = \{x \in X ; W(x) \leq \lambda\}. \quad (5.31)$$

Considérons les classes d'équivalence \mathcal{C}_λ de la relation \mathcal{R}_λ définie sur W_λ par

$$\forall x, y \in W_\lambda, x \mathcal{R}_\lambda y \text{ ssi } H(x, y) \leq \lambda. \quad (5.32)$$

et

$$\forall x, \quad x \mathcal{R}_\lambda x. \quad (5.33)$$

Un ensemble $\pi \subset X$ est un cycle si $\pi = \mathcal{C}_\lambda$ pour $\lambda \geq 0$ (π peut se réduire à un singleton).

Hauteur critique et exposant de convergence. La hauteur critique est définie par

$$H_1 = \max\{H(\pi) ; \pi \text{ cycle n'intersectant pas } \mathcal{W}^*\}. \quad (5.34)$$

La définition de la hauteur de sortie $H(\pi)$ est complexe (voir le théorème 4.4.1). Grâce au théorème 5.7.3, nous verrons que H_1 est donné par

$$H_1 = \max_{x \neq (a^*)} H(x, (a^*)) - W(x) \quad (5.35)$$

lorsque $n > n_*$. L'*exposant de convergence* est décrit par

$$\alpha_* = \min\left\{\frac{W(\pi) - W_*}{H(\pi)} ; \pi \text{ cycle n'intersectant pas } \mathcal{W}^*\right\}, \quad (5.36)$$

où $W(\pi)$ est le minimum de W dans π .

5.6 Le rôle des populations uniformes

Tout au long de ce chapitre, la population uniforme (a, \dots, a) sera identifiée à $a \in E$

$$(a) = (a, \dots, a). \quad (5.37)$$

L'ensemble des populations uniformes est noté U . Par définition (équation 5.14), nous avons,

$$\forall x \in X, \quad V_1(x, (x_*)) = 0, \quad (5.38)$$

De plus, pour $x \in X$ et $y \neq (x_*)$,

$$V_1(x, y) > 0. \quad (5.39)$$

Définissons une nouvelle fonctionnelle de coût sur l'ensemble U . Pour tout $x, y \in U$,

$$V_U(x, y) = \inf\left\{\sum_{k=0}^{r-1} V_1(x_k, x_{k+1}), x_0 = x, x_k \notin U \ (1 \leq k < r), x_r = y, r \geq 1\right\}. \quad (5.40)$$

Pour les populations (a) et (b) , les notations $V(a, b)$ et $W(a)$ remplaceront désormais $V((a), (b))$ et $W((a))$.

Théorème 5.6.1 Soient (a) et (b) dans U . Nous avons

$$W(a) = \min_{g \in G_U(a)} \sum_{(b \rightarrow c) \in g} V_U(b, c)$$

où le min parcourt les a -graphes définis sur U . De plus, $H(a, b)$ peut être calculée à partir de V_U (plutôt que V).

Démonstration. Appliquer le Théorème 5.8 de [10] avec $H = U$. ■

Dans les calculs à venir, la chaîne (X_t) pourra être identifiée avec la chaîne qu'elle induit sur U , et associée à V_U . Nous renvoyons le lecteur à [10] pour la définition de la chaîne induite.

5.7 Démonstrations

5.7.1 Concentration

Dans ce paragraphe (seulement), nous levons l'hypothèse d'injectivité faite sur f .

Théorème 5.7.1 Concentration. Soit $n > n_*$. Nous avons, pour tout $x \in X$,

$$\lim_{T \rightarrow 0} \lim_{t \rightarrow \infty} \text{Prob}(X_t \text{ contient un élément de } A^* \mid X_0 = x) = 1. \quad (5.41)$$

Lemme 5.7.1 Soit $a \in E \setminus A^*$. Alors, nous avons

$$\min\{V_U(a, a^*) ; a^* \in A^*\} = \min\{d(a, a^*) ; a^* \in A^*\}. \quad (5.42)$$

où d est la distance sur le graphe (E, G) .

Démonstration. Soit $a \in E \setminus A^*$. Considérons un sommet a^* qui réalise le min dans le membre de gauche de l'équation (5.42). De manière évidente, nous avons $V_U(a, a^*) \geq d(a, a^*)$. Considérons un chemin qui réalise le min dans la définition de $d(a, a^*)$:

$$a_0 = a \rightarrow a_1 \rightarrow \dots \rightarrow a_r = a^*,$$

et la trajectoire suivante

$$\begin{aligned} x_0 &= (a) \\ &\downarrow \\ x_1 &= (a_1, a, \dots, a) \\ &\dots \\ &\downarrow \\ x_k &= (a_k, \tilde{x}_k, \dots, \tilde{x}_k) \\ &\downarrow \\ x_r &= (a^*, \dots) \\ &\downarrow \\ x_{r+1} &= (a^*) \end{aligned} \quad (5.43)$$

où

$$\tilde{x}_1 = a \quad (5.44)$$

et

$$\forall k = 2, \dots, r, \quad \tilde{x}_k = \begin{cases} a_{k-1} & \text{si } f(a_{k-1}) < f(\tilde{x}_{k-1}) \\ \tilde{x}_{k-1} & \text{sinon.} \end{cases} \quad (5.45)$$

Donc, nous avons

$$\forall k = 0, \dots, r-1, \quad V_1(x_k, x_{k+1}) = 1, \quad (5.46)$$

et $V_U(a, a^*) \leq d(a, a^*)$. ■

Démonstration du Théorème 5.7.1. La chaîne de transition q associée à l'algorithme est irréductible

$$\forall (x, y) \in X \times X, \quad \exists r \geq 1 \quad q^{(r)}(x, y) > 0, \quad (5.47)$$

et (X_t) converge vers la loi stationnaire π_T , lorsque t tend vers l'infini. Soit (a) une population uniforme telle que $a \notin A^*$. Nous avons

$$n > \max_{a \notin A^*} \{d(a, A^*)\}. \quad (5.48)$$

D'après le lemme 5.7.1, il existe $a^* \in A^*$ tel que

$$n > V_U(a, a^*). \quad (5.49)$$

D'autre part, au moins n mutations sont nécessaires pour sortir de A^*

$$\forall b \notin A^*, \quad V_U(a^*, b) \geq n. \quad (5.50)$$

Soit g un a -graphe sur U tel que

$$V_U(g) = W(a) = \sum_{(u \rightarrow v) \in g} V_U(u, v). \quad (5.51)$$

Puisque $a \notin A^*$ et g est un arbre recouvrant sur U de racine a , il existe $b \in U$ tel que $(a^* \rightarrow b)$ appartient à g . Nous construisons un a^* -graphe en détruisant l'arête $(a^* \rightarrow b)$ dans g et en introduisant l'arête $(a \rightarrow a^*)$. Ainsi, nous avons

$$W(a^*) \leq W(a) + V_U(a, a^*) - V_U(a^*, b). \quad (5.52)$$

Donc

$$\forall a \notin A^*, \quad W(a) > W(a^*), \quad (5.53)$$

et la chaîne de Markov se concentre sur l'ensemble des populations minimales. ■

Dans le reste de ce manuscrit, nous supposerons f injective, et donc $A^* = \{a^*\}$.

5.7.2 Temps d'atteinte

Nous devons considérer des sous-graphes particuliers de X pour estimer le temps d'atteinte de a^* par la chaîne (X_t) . Soit $\{a, b\}$ une paire de sommets dans E , et $G(a, b)$ l'ensemble des graphes orientés qui "aboutissent" en a et b . Pour tout $g \in G(a, b)$, $c \neq a, c \neq b$ est le sommet de départ d'exactement une arête (aucune arête ne part de a ou b). Soit

$$W_{**} = \min\{V_U(g); g \in G(a, b); a \neq b \in U\}. \quad (5.54)$$

et

$$\tau = \inf\{t \geq 1, X_t \ni a^*\}, \quad (5.55)$$

On note

$$H_0 = \lim_{T \rightarrow 0} T \log E[\tau \mid X_0 = x]. \quad (5.56)$$

Proposition 5.7.1 *Soit $n > n_*$, $x \in X$, $x \not\ni a^*$. Nous avons*

$$H_0 = \min_{a \neq a_*} \min_{g \in G(a^*, a)} V_U(g) - W_* = W_{**} - W_*.$$

Démonstration. Elle découle du théorème 5.7.1 et des lemmes [25], chapitre 6. ■

Théorème 5.7.2 *Soit $n > n_*$. Nous avons*

$$H_0 = d_*$$

Démonstration. Soit g_* le a^* -graphe sur U tel que

$$W_* = W(a^*) = V_U(g_*). \quad (5.57)$$

Dans g_* , une arête partant de (a) va vers la population uniforme (b) telle que

$$V(a, b) = \min_{b: f(b) < f(a)} d(a, b). \quad (5.58)$$

Considérons la population (a_*) pour laquelle

$$V_U(a_*, b) = d_* \quad (5.59)$$

où b vérifie

$$(a_*) \rightarrow (b) \in g_* . \quad (5.60)$$

Construisons un graphe

$$g^* \in G(a^*, a_*) \quad (5.61)$$

en supprimant l'arête $(a_*) \rightarrow (b)$ dans g_* . Par définition, on a

$$W_{**} \leq V_U(g^*) = W_* - d_* . \quad (5.62)$$

Ceci montre que

$$d_* \leq H_0 . \quad (5.63)$$

Par construction, g^* est en fait le graphe pour lequel le minimum W_{**} est atteint, et l'inégalité précédente est une égalité. ■

5.7.3 Cycles

Théorème 5.7.3 Soit $n > n_*$. Sur U , un cycle contient (a^*) , ou se réduit à un singleton.

Démonstration. On démontre le résultat en raisonnant par l'absurde. Soit π un cycle qui ne contient pas (a^*) et $(a) \in \pi$ tel que

$$\forall(b) \in \pi, b \neq a, \quad f(a) < f(b). \quad (5.64)$$

Par définition d'un cycle, nous avons

$$H(a, b) \leq \lambda \quad (5.65)$$

pour un $\lambda > 0$. Au moins n mutations sont nécessaires pour aller de (a) vers (b) ($(b) \in \pi \setminus (a)$), et

$$V_U(a, b) > n. \quad (5.66)$$

D'après le lemme 5.7.1, nous avons $V_U(a, a^*) = d(a, a^*)$, et

$$H(a, a^*) \leq W(a) + V_U(a, a^*) \leq W(a) + d(a, a^*) \leq W(a) + n_*. \quad (5.67)$$

De plus,

$$W(a) + n_* \leq W(a) + V_U(a, b) \leq \lambda \quad (5.68)$$

En assemblant les équations (5.67) et (5.68), on obtient $(a^*) \in \pi$. ■

5.7.4 Schémas optimaux

Théorème 5.7.4 Soit $n > n_*$. Nous avons

$$H_1 = d_*.$$

Nous commençons par un lemme facile.

Lemme 5.7.2 Soit $a \in E$, $a \neq a^*$, et $b \in E$ le plus proche sommet tel que $f(b) < f(a)$. Alors, nous avons

$$V_U(a, b) = d(a, b). \quad (5.69)$$

Démonstration. Clairement, nous avons $V_U(a, b) \geq d(a, b)$. Maintenant, considérons une chemin dans (E, \mathcal{G}) réalisant $d(a, b)$:

$$a_1^0 = a \rightarrow a_1^1 \rightarrow \dots \rightarrow a_1^r = b$$

et la trajectoire suivante dans X

$$\begin{aligned}
 x_0 &= (a) \\
 &\downarrow \\
 x_1 &= (a_1^1, a, \dots, a) \\
 &\downarrow \\
 &\cdots \\
 &\downarrow \\
 x_r &= (b, a, \dots, a) \\
 &\downarrow \\
 x_{r+1} &= (b) .
 \end{aligned} \tag{5.70}$$

Alors le coût de cette trajectoire est exactement

$$\sum_{k=0}^r V_1(x_k, x_{k+1}) = d(a, b), \tag{5.71}$$

d'où $V_U(a, b) = d(a, b)$. ■

Démonstration du théorème 5.7.4. D'après le théorème 5.7.3, nous avons

$$H_1 = \max_{a \neq a^*} \{H(a, a^*) - W(a)\}. \tag{5.72}$$

Montrons que $H_1 \leq d_*$. Fixons $a \in E$, et considérons la trajectoire

$$\gamma : (a^0) = (a) \rightarrow (a^1) \rightarrow \dots \rightarrow (a^r) = (a^*), \tag{5.73}$$

dans U où a^i est le plus proche sommet de a^{i-1} tel que $f(a^i) < f(a^{i-1})$, $i = 1, \dots, r$. D'après le lemme 5.7.2, nous avons

$$H(a, a^*) \leq \max_{i=0, \dots, r-1} \{W(a^i) + d(a^i, a^{i+1})\}, \tag{5.74}$$

et ainsi

$$H_1 \leq d_*. \tag{5.75}$$

Pour l'autre inégalité, la Proposition 14 de [7] peut être appliquée. Ce résultat (classique) dit que

$$H_0 \geq \max_{a \neq a^*} \{H(a, a^*) - W(a)\}. \tag{5.76}$$

Le membre de gauche est égal à d_* d'après le théorème 5.7.2. ■

La démonstration du Théorème 5.25 découle de [53] Théorème 2.22, p. 981.

5.7.5 Exposant de convergence

Théorème 5.7.5 Soit $n > n_*$. Nous avons

$$\frac{n - n_*}{d_*} \leq \alpha_* \leq \frac{n(n_* + 1 - d_*) - 1}{d_*}.$$

Démonstration. D'après le théorème 5.7.3, nous avons

$$\alpha_* = \min_{b \neq a^*} \left\{ \frac{W(b) - W_*}{H(\{b\})} \right\}. \quad (5.77)$$

Par définition de H_1 , nous avons

$$\alpha_* \geq \frac{\min_{b \neq a^*} W(b) - W_*}{H_1} \quad (5.78)$$

Soit $b_* \neq a^*$ tel que $W(b_*) = \min_{b \neq a^*} W(b)$. Considérons le b_* -graphe qui réalise le min dans la dernière définition. A partir de ce graphe, créons un a^* -graphe, disons g , en supprimant l'arête $a^* \rightarrow a$ et en ajoutant $b_* \rightarrow a^*$. D'après le lemme 5.7.1, nous obtenons

$$V_U(g) - W(b_*) = d(b_*, a^*) - V_U(a^*, a) \geq W_* - W(b_*) \quad (5.79)$$

et, ainsi

$$W(b_*) - W_* \geq V_U(a^*, a) - d(b_*, a^*) \geq n - n_*. \quad (5.80)$$

Ceci montre que

$$\alpha_* \geq \frac{n - n_*}{d_*}. \quad (5.81)$$

Pour le membre de droite, remarquons que

$$\alpha_* \leq \frac{W(b^*) - W_*}{H_1} \quad (5.82)$$

où b^* est tel que $H_e(b^*) = H_1$. Créons un b^* -graphe, disons g' , en supprimant $b^* \rightarrow b$ et en ajoutant $a^* \rightarrow b^*$ dans le graphe qui réalise l'argument minimum de W . Nous obtenons

$$V(g') - W_* = V(a^*, b^*) - d_* \geq W(b^*) - W_* \quad (5.83)$$

Pour calculer $V(a^*, b^*)$, considérons le chemin

$$b^* = a^r \rightarrow \dots \rightarrow a^0 = b$$

qui réalise

$$d_* = \max_{b^* \neq a^*} \min_{b'; f(b') < f(b^*)} d(b^*, b'), \quad (5.84)$$

tel que $r = d_*$. Construisons une trajectoire entre (a^*) et (b) , puis la trajectoire γ_{bb^*} entre (b) et (b^*) comme suit

$$\begin{aligned}
 x_0 &= (b, b, \dots, b) \\
 &\quad \downarrow \\
 x_1 &= (a^1, a^1, \dots, a^1) \\
 &\quad \downarrow \\
 &\quad \cdots \\
 &\quad \downarrow \\
 x_k &= (a^k, \tilde{x}_k, \dots, \tilde{x}_k) \\
 &\quad \downarrow \\
 &\quad \cdots \\
 &\quad \downarrow \\
 x_r &= (b^*, a^1, \dots, a^1) \\
 &\quad \downarrow \\
 x_{r+1} &= (b^*)
 \end{aligned} \tag{5.85}$$

où \tilde{x}_k est défini comme dans la démonstration du lemme 5.7.1. Moins de $(n_* + 1 - d_*)$ arêtes sont nécessaires pour aller de (a^*) à (a^1) . Au pire, n mutations simultanées sont nécessaires à chaque pas. Le coût vérifie donc

$$V(a^*, b^*) \leq n(n_* + 1 - d_*) + d_* - 1. \tag{5.86}$$

Ceci montre que

$$\alpha_* \leq \frac{n(n_* + 1 - d_*) - 1}{d_*}. \tag{5.87}$$

A nouveau, la démonstration du Théorème 5.4.3 découle de [53] Théorème 2.22, p. 981. ■

5.8 Implantation et exemples

5.8.1 Implantation

Nous modifions sensiblement l'algorithme présenté dans la section 5.3. La nouvelle version repose sur un tirage binomial du nombre de descendants par mutation. La mutation est appliquée aux individus dans l'ordre des indices dans la population. Voici l'algorithme.

- Initialiser une population x de n individus numérotés.
- Répéter
 1. Tirer un nombre au hasard N selon $bin(n, p)$;
 2. Sélectionner le meilleur individu x_* de la population;

3. Créer des descendants x'_1, \dots, x'_N des N premiers individus, et remplacer les “parents” par leurs enfants. Remplacer les $n - N$ derniers individus par x_* ;
4. Modifier p ;

Jusqu'à **arrêt**.

La probabilité de transition $q_M(x, y)$ entre x et $y \in X$ est donnée par

$$q_M(x, y) = P(N = V_1(x, y))\pi_M(x, y). \quad (5.88)$$

où $\pi_M(x, y) = 0$ si $y_i \notin N(x_i) \cap (E \setminus \{x_*\})$ pour $i \leq V_1(x, y)$ ou $y_i \neq x_*$ pour $i > V_1(x, y)$. Dans les autres cas, $\pi(x, y)$ est strictement positif (indépendant de p) et correspond aux choix de voisins dans le graphe. Cette nouvelle chaîne de Markov satisfait à nouveau des inégalités de grandes déviations. Pour un $\kappa_M > 0$, nous avons

$$\frac{1}{\kappa_M} \pi_M(x, y) e^{-V_1(x, y)/T} \leq q_M(x, y) \leq \kappa_M \pi_M(x, y) e^{-V_1(x, y)/T} \quad (5.89)$$

où V_1 est la même fonctionnelle de coût que pour q . En conséquence, les résultats énoncés dans la section 5.4 sont aussi valides pour cette nouvelle version.

L'avantage de ce procédé de simulation est que la taille effective de population est $N + 1$. Il est possible d'utiliser cette version pour économiser l'espace mémoire. Mieux, la caractéristique de cette méthode est de fournir une forme de recherche hiérarchique. De telles idées ont été considérées dans le passé (e.g., [16]), mais semblent peu exploitées dans le contexte évolutionnaire. Les individus effectuent différents “degrés” de recherche, selon leur position dans la population. Pour des tailles assez grandes, les premiers individus peuvent voyager assez librement dans l'espace de recherche, sans rappel de la sélection. A l'opposé les individus d'indice proche de $\lfloor np \rfloor$ sont assez souvent rappelés par la sélection. Cette version a été expérimentée dans [24], et apparaît plus efficace que celle proposée dans la section 5.3. Nous l'avons utilisée pour obtenir les résultats numériques du paragraphe suivant.

5.8.2 Comparaisons

Étant donné deux algorithmes A et B , et une classe de fonctions tests, une question cruciale est de déterminer si A est “meilleur” que B . Par le théorème “No-Free-Lunch” [58], la réponse peut être difficile à donner lorsque la classe est trop vaste. Notre approche consiste à comparer deux algorithmes sur la base de leur probabilité de succès, que l'on estime grâce à l'exposant de convergence. Cela permet de construire des classes de fonctions pour lesquelles un algorithme dépasse un autre (asymptotiquement). Toutefois, les exposants de convergence doivent être explicités à l'aide de bornes calculables.

Dans ce paragraphe, nous comparons E/S, algorithme A , avec n recuits simulés en parallèle, notre B , n étant la taille de population de A . L'exposant de l'algorithme B est inférieur à $n(\Delta - h_*)/h_*$ où h_* est la hauteur de Hajek [31], et

$$\Delta = \max\{|f(a) - f(b)| ; a, b \in E\}. \quad (5.90)$$

E	0	1	2	3	4	5	6	7
$f(E)$	2.5	4.3	4.6	1.0	2.2	0.0	1.5	12.0

E	8	9	10	11	12	13	14	15
$f(E)$	10.2	11.0	2.0	4.1	2.8	6.3	12.01	9.4

Table 5.1: Fonction objectif pour le premier problème test sur le groupe cyclique à 16 éléments.

Selon le théorème 5.4.3, l'algorithme A peut être préféré à B si

$$h_* > (\Delta - h_*)d_* , \quad (5.91)$$

et n vérifie

$$n\left(1 - \frac{(\Delta - h_*)d_*}{h_*}\right) > n_* . \quad (5.92)$$

Pour évaluer une telle affirmation, des simulations numériques ont été effectuées sur un problème très simple, pour lequel les constantes h_* , n_* et d_* se calculent facilement. Le problème test est défini sur $E = \mathbb{Z}/16\mathbb{Z}$, et l'exploration agit comme une marche aléatoire sur E (un voisinage de z est constitué des entiers $z - 1$ et $z + 1$ modulo 16). Les valeurs de f sont données dans la Table 5.1.

Nous trouvons

$$h_* = 10.0 , \quad n_* = 7 , \quad d_* = 4 , \quad (5.93)$$

et

$$1 - \frac{(\Delta - h_*)d_*}{h_*} = 0.196 . \quad (5.94)$$

Par la suite, n a été pris égal à 40, de sorte que l'équation (5.92) puisse être vérifiée.

La probabilité de succès (atteinte de $a^* = 5$) a été estimée à partir d'un grand nombre de simulations. La population a été initialisée au sommet 10, correspondant à la pire situation. Un première remarque à faire est que les stratégies à température infinie sont assez efficaces, étant donné de peu d'états à énumérer. En modifiant E/S pour devenir élitiste, A et B sont équivalents. Après 10 itérations, nous avons approximativement 0.75 chance d'atteindre a^* . Cette probabilité croît vers 0.99 avec 10 itérations supplémentaires.

Les résultats concernant six schémas de température sont donnés dans les tables 5.3 et 5.4. Tout d'abord, la température a été fixée à $T = 0.83$ ($p = 0.3$). Les résultats montrent que l'algorithme A a 0.96 chance d'atteindre a^* en 25 itérations, alors que le recuit parallèle B n'a pratiquement aucune chance de le faire. Lorsque $T = 0.33$, ($p = 0.05$), A a 0.99 chance de trouver a^* , alors que cette probabilité est presque nulle pour B .

E	0	1	2	3	4	5	6	7
$f(E)$	1.2	1.3	1.4	1.02	1.22	0.0	1.0	2.01

E	8	9	10	11	12	13	14	15
$f(E)$	1.82	1.94	1.01	1.45	1.23	1.64	2.02	1.74

Table 5.2: Fonction objectif pour le second problème test sur le groupe cyclique à 16 éléments.

Nous rapportons des résultats du même type lorsque des schémas décroissants sont utilisés. Considérons

$$\forall t \geq 1, \quad \beta(t) = 1/T(t). \quad (5.95)$$

Les schémas logarithmiques $\beta_1(t) = \log(t)/4 + 1.2$ et $\beta_2(t) = \log(t)/10 + 0.5$ ont d'abord été testés. Avec β_1 , le minimum est atteint en 50 itérations pour A avec la probabilité 0.97, alors que cette probabilité est 0 pour l'algorithme B , même avec 500 itérations. Le second schéma semble plus favorable à B , mais A est encore le plus efficient.

Le phénomène se reproduit avec des schémas linéaires $\beta_3(t) = 0.83 + 0.05t$, et $\beta_4(t) = 0.51 + 0.02t$. Avec β_3 , le minimum est atteint en 100 itérations pour A avec la probabilité 0.91, alors que cette probabilité est 0.0 pour B (500 itérations). Avec β_4 , la probabilité croît vers 0.15 pour l'algorithme B , mais A est encore meilleur.

Notons de plus que le nombre d'évaluation de f est significativement plus petit pour A que pour B .

Des arguments analogues peuvent être appliqués pour décider quand B est meilleur que A . La condition est alors

$$\frac{n(n_* + 1 - d_*) - 1}{d_*} < \frac{n\delta}{h_*}. \quad (5.96)$$

avec

$$\delta = \min\{f(a) - f(a_*); a \neq a_* \in E\}. \quad (5.97)$$

Pour évaluer cette affirmation, des simulations numériques ont été effectuées pour un second problème test (Table 5.2).

Les constantes critiques h_* , n_* est d_* sont

$$h_* = 1.0, \quad n_* = 7, \quad d_* = 4, \quad (5.98)$$

et nous prenons $n = 8$, de sorte que la condition (5.96) puisse être vérifiée. Les mêmes schémas ont été étudiés (β_2 a été changé : $\beta_2(t) = 0.5 + \log(t)$).

Les résultats rapportés dans les tables 5.5 et 5.6 montrent que 8 recuits sont plus efficaces qu'un E/S de taille $n = 8$. Pour les schémas constants, la probabilité de succès est proche de 1.0 pour B . Cette probabilité est plus petite pour A (quasi nulle pour $p = 0.05$). L'algorithme B dépasse l'algorithme A pour tous les autres schémas. Ces résultats démontrent la valeur de la condition (5.96), bien que cette dernière se fonde sur des estimations assez grossières.

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	0.88	—	0.65	0.98	0.66	0.98
25	0.97	0.12	0.92	1	0.78	1
50	1	0.30	0.97	1	0.89	1
100	1	0.61	1	1	0.91	1
500	1	0.99	1	1	1	1

Table 5.3: Probabilité de succès pour l'algorithme exploration/sélection (taille de population $n = 40$), en fonction de la température inverse et du nombre d'itérations pour le premier problème. Les résultats sont moyennés sur 1,000 runs.

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	—	—	—	—	—	—
25	—	—	—	—	—	—
50	—	—	—	—	—	—
100	0.001	—	0.0	0.19	0.0	0.14
500	0.03	0.0	0.0	0.31	0.0	0.15

Table 5.4: Probabilité de succès pour 40 recuits simulés en parallèle en fonction de la température inverse et du nombre d'itérations pour le premier problème. Les résultats sont moyennés sur 1,000 runs.

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	0.68	0.26	0.60	0.49	0.66	0.78
25	0.82	0.41	0.73	0.61	0.81	0.88
50	0.99	0.84	0.99	0.87	0.97	1
100	1	0.99	1	0.97	0.99	1

Table 5.5: Probabilité de succès pour 8 recuits simulés en parallèle en fonction de la température inverse et du nombre d'itérations pour le second problème. Les résultats sont moyennés sur 1,000 runs.

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	—	—	—	—	—	0.36
25	0.24	—	0.05	—	0.17	0.49
50	0.54	—	0.13	—	0.19	0.73
100	0.82	0.01	0.19	0.02	0.23	0.79

Table 5.6: Probabilité de succès pour l'algorithme exploration/sélection (taille de population $n = 8$), en fonction de la température inverse et du nombre d'iterations pour le second problème. Les résultats sont moyennés sur 1,000 runs.

Itération de la procédure MOSES

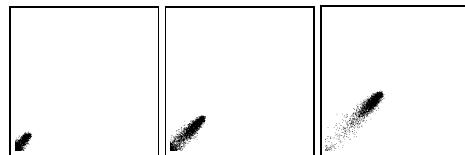


Figure 1: Itérations 100,200,300

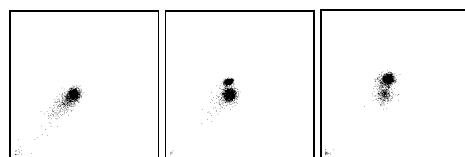


Figure 2: Itérations 400,600,800

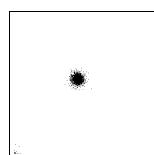


Figure 3: Itération 1000

Taille de la population $n = 5000$. Diamètre du graphe d'exploration
 $D = 252$.

Figure 5.1: L'évolution de l'algorithme pour un problème quadratique perturbé. La population est initialisée dans le coin gauche puis évolue vers le minimum absolu selon une trajectoire privilégiée.

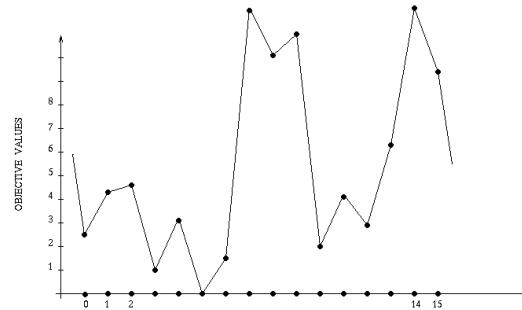


Figure 5.2: Problème 1.

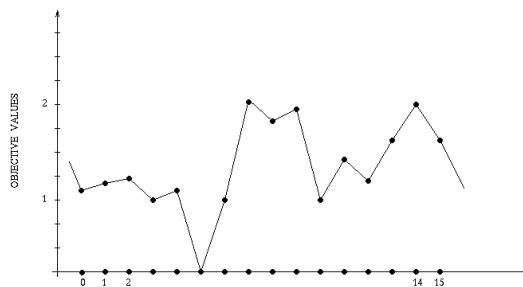


Figure 5.3: Problème 2.

Bibliography

- [1] E.H.L Aarts and J.H.M Korst. Simulated annealing and Boltzmann machines, Wiley, New York, 1988.
- [2] Aldous, J. Fill. Projet de livre, source internet, 1997.
- [3] R. Azencott. Simulated annealing, *Sém. Bourbaki*, 697 (1988), 161-175.
- [4] T. Bäck, H.P. Schwefel. An overview of evolutionary algorithms for parameters optimization, *Evolutionary Computation*, 1, (1993), 1-24.
- [5] T. Bäck. Evolutionary algorithms in theory and practice, Oxford university press, NY, 1996.
- [6] Bertsimas, D., Tsitsiklis, J. (1993). Simulated Annealing. *Statist. Sci.* **8** 10-15.
- [7] O. Catoni. Simulated annealing algorithms and Markov chains with rare transitions, Lectures Notes DEA Univ. Paris XI, 1997.
- [8] R. Cerf. Asymptotic convergence of a genetic algorithm, *C.R. Acad. Sci. Paris Sér. I Math.*, 319 (1994), 271-276.
- [9] R. Cerf. Une théorie asymptotique des algorithmes génétiques, PhD thesis, Montpellier II, 1994.
- [10] R. Cerf. The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Probab. Statist.*, 32 (1996), 455-508.
- [11] R. Cerf. A new genetic algorithm. *Ann. Appl. Probab.* 6 (1996), 778-817.
- [12] Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the Laplacian. In R.C. Gunning (ed.) *Problems in Analysis*: Princeton University Press, pp. 195-199.
- [13] Cinlar. Introduction to stochastic processes, xx, 198x.
- [14] U.K. Chakraborty, D. Kalyanmoy and M. Chakraborty. Analysis of selection algorithms : a Markov chain approach, *Evol. Comput.* 4 (1996), 2, 133-167.
- [15] L. Davis. Handbook of genetic algorithms, Van Nostrand Reinhold, New-York, 1991.
- [16] D.A. Dawson. Stochastic models of parallel systems for global optimization, *IMA Vol. Math. Appl.* 9 (1987), 25-44.

- [17] T.E. Davis and J.C. Principe. A simulated annealing like convergence theory for the simple genetic algorithm. In R.K. Belew and L.B. Booker eds. *Proceedings of the fourth Int. Conf. on Genetic Algorithms*, San Mateo CA, Morgan Kauffman, 174-181, 1991.
- [18] P. Del Moral, L. Miclo. On the convergence and the applications of the generalized simulated annealing, preprint, Université de Toulouse, 1999.
- [19] Diaconis P., Saloff-Coste L. (1993) Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3, pp. 696-730.
- [20] Diaconis, P., Saloff-Coste, L. (1996). What do we know about the Metropolis algorithm? *J. Comp. Sci. Syst.* To appear.
- [21] Diaconis, P., Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36-61.
- [22] W. Feller. An introduction to probability theory and its applications, vol. II, Wiley, New York, 1971.
- [23] D.B. Fogel. Evolutionary computation: toward a new philosophy of machine intelligence, IEEE Press, Piscatawan, NJ, 1995.
- [24] O. François. An evolutionary strategy for global minimization and its Markov chain analysis, *IEEE transactions on Evolutionary Computation*, 2, (1998), 77-90.
- [25] M.I. Freidlin and A.D. Wentzell. Random perturbations of dynamical systems, Springer Verlag, New York, 1984.
- [26] Frigessi, A., Hwang, C.R., Sheu, S.J., Di Stefano, P. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. *J Roy. Statist. Soc. Ser. B* **55** 205-220.
- [27] Frigessi, A., Martinelli, F., Stander, J. (1997). Computational complexity of Markov chain Monte Carlo methods. *Biometrika* **84**, 1-18.
- [28] Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the bayesian restoration of images. *IEEE Trans. Patt. Ana. Mach. Intel.* **6** 721-741.
- [29] Glauber, R.J. (1963). Time dependent statistics of the Ising model. *J. Math. Phys.* **4** 294-307.
- [30] D. Goldberg. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading, MA, 1989.
- [31] B. Hajek. Cooling schedules for optimal annealing, *Math. Oper. Res.* 13 (1988), 311-329.
- [32] Holland, J.H. Adaptation in natural and artificial systems. Univ. Michigan Press. Ann Arbor, 1975.
- [33] Holley, R., Stroock, D. (1988). Simulated Annealing via Sobolev Inequalities. *Comm. Math. Phys.* **115** 553-569.

- [34] L. Ingber and B. Rosen. Genetic algorithm and very fast simulated annealing - a comparison. *Math. Comp. Model.*, 16 (1992), 11, 87-100.
- [35] Ingrassia, S. (1993). Geometric approaches to the estimation of the spectral gap of time reversible Markov chains. *Combin. Probab. Comput.* **2** 301-323.
- [36] Ingrassia, S. (1994). On the rate of convergence of the Metropolis algorithm and the Gibbs sampler by geometric bounds. *Ann. Appl. Probab.* **4** 347-389.
- [37] Kemeny, J.L. Snell. Finite Markov Chains, , 196x.
- [38] Lawler, G.F., Sokal, A.D. (1988). Bounds on the L^2 spectrum for Markov Chains and Markov Processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.* **309** 557-580.
- [39] Y. Leung, Y.Gao and Z.B. Xu. Degree of population diversity - A perspective on premature convergence in genetic algorithms and its Markov chain analysis, *IEEE Trans. Neural Networks*, 8 (1997), 1165-1175.
- [40] Lezaud, P. Chernoff bounds for finite Markov chains, PhD Thesis, Toulouse, 1996
- [41] Martinelli, F., Olivieri, E. (1994). Approach to equilibrium of Glauber dynamics in the one phase region I: the attractive case. *Comm. Math. Phys.* **161** 487-514.
- [42] Metropolis, M., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. (1953). Equation of state calculation by fast computing machines. *J. Chem. Phys.* **21** 1087-1092.
- [43] A. Nix and M. Vose. Modeling genetic algorithms with Markov chains, *Ann. Math. Art. Intell.*, 5 (1992), 1, 79-88.
- [44] S.M. Ross, Stochastic processes, wiley and sons, NY, 1983.
- [45] Rechenberg, I. Evolutionsstrategie: Optimierung technischer systeme nach prinzipien der biologischen evolution. Stuttgart, Frommann-Holzboog Verlag, 1973.
- [46] G. Rudolph. Convergence analysis of canonical genetic algorithms, *IEEE Trans. Neural Networks*, 5 (1994), 96-101.
- [47] Saloff-Coste, L. (1997). *Lectures on Finite Markov Chains*, (Saint-Flour) Lect. Notes Math. 1665, Springer Berlin, 301-413.
- [48] Sinclair, A., Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.* **82** 93-133.
- [49] Sinclair, A. (1991) Improved bounds for mixing rates of Markov chains and multicommodity flows. *Combin. Probab. Comput.* **1** 351-370.
- [50] J. Suzuki. A Markov chain analysis on a genetic algorithm. In S. Forrest ed., *Proceedings of the fifth Int. Conf. on Genetic Algorithms* Urbana-Champaign, Morgan Kauffman, San Mateo, CA, 146-153, 1993.

- [51] A. Trouv . Parallelisation massive du recuit simul . PhD thesis. Universit  Paris XI, 1993.
- [52] A. Trouv . Asymptotical behaviour of several interacting annealing processes. *Probab. Theory Related Fields*, 102 (1995), 123-143.
- [53] A.Trouv . Cycle decomposition and simulated annealing, *SIAM J. Control Optim.*, 34, 3 (1996), 966-986.
- [54] J.N. Tsitsiklis. Markov chains with rare transitions and simulated annealing, *Math. Oper. Res.*, 14 (1989) 70-90.
- [55] P.J.M. Van Laarhoven and E.H.L Aarts. Simulated annealing: theory and applications. Reidel, Dordrecht, 1987.
- [56] A.D. Ventcel. On the asymptotics of eigenvalues of matrices with elements of order $\exp(-V_{ij}/2\epsilon^2)$, *Dokl. Akad. Nauk. SSSR*, 202 (1972), 65-68.
- [57] M. Vose. A closer look at mutation in genetic algorithms, *Ann. Math. Art. Intell.*, 10 (1994), 4, 423-435.
- [58] D.H. Wolpert, W.G. MacReady. No free lunch theorem for Optimization, *IEEE Transactions on Evolutionary Computation*, 1(1), (1997), 67-82.