



CHAMPS DE MARKOV EN TRAITEMENT D'IMAGE

Module C3M

Marc SIGELLE et Florence TUPIN
Département Traitement de Signal et des Images
E-mail : author@tsi.enst.fr
Année 1999

Contents

1 Définition et simulation d'un champ de Markov	9
1 Introduction	9
2 Description et modélisation probabiliste de l'image	10
3 Champs de Markov - Champs de Gibbs	12
4 Echantillonnage de MRF	14
5 Applications : restauration et segmentation	22
6 Le recuit simulé	25
7 Conclusion	28
2 Estimateurs dans un cadre markovien	29
1 Introduction	29
2 Modélisation bayésienne et fonction de coût	29
3 Estimateur MAP	30
4 Estimateur MPM	31
5 Estimateur TPM	32
6 Solutions algorithmiques des estimateurs dans le cas markovien	33
7 Comparaison des estimateurs MAP, MPM, et TPM	36
8 Conclusion	38
3 Estimation des paramètres	41
1 Introduction	41
2 Données complètes	42
3 Données incomplètes	46

4 Quelques applications des champs markoviens	61
1 Applications sur le graphe des pixels	61
2 Applications sur des graphes de primitives	74
3 Conclusion	78
A Statistique exponentielle linéaire d'un paramètre θ	81
1 Introduction - Notations et définitions	81
2 Comportement en fonction du paramètre θ positif	82
3 Vraisemblance pour une statistique linéaire exponentielle	83
B Développement en cumulants des statistiques exponentielles	87
1 Introduction	87
2 Définition	87
3 Cumulants de distributions de Gibbs mono-paramétrées	88
4 Le cas des distributions de Gibbs multi-paramétrées	89
C Échantillonnage des distributions de Gibbs - recuit simulé	91
1 Rappels sur les mesures, noyaux et le coefficient de Dobrushin	91
2 Échantilleurs de Gibbs et de Metropolis	97
3 Échantillonnage homogène des distributions de Gibbs : convergence	100
4 Échantillonnage inhomogène des distributions de Gibbs	101

*Ce document est dédié à la mémoire de Marc Bernard,
qui a introduit avec brio les Champs de Markov au Département Images de l'ENST*

Introduction

Les champs de Markov sont utilisés depuis maintenant une dizaine d'années en traitement d'images et font désormais partie des techniques de base de cette discipline. Nés à l'origine dans le cadre de la physique statistique pour étudier les phénomènes de transition de phase [Ising(1925)], ils sont rapidement appliqués aux réseaux bidimensionnels que constituent les images. Les premiers modèles sont restreints par des contraintes de causalité [Abend *et al.*(1965)] [Kanal(1980)] et donc limités jusqu'à l'article fondateur de Geman et Geman en 84 [Geman and Geman(1984)] qui ouvre réellement leur utilisation en traitement d'images. Cette utilisation ira croissant avec les années et donnera lieu à de nombreuses recherches sur les problèmes que suscite leur utilisation (calcul des paramètres des modèles, techniques de simulation et de recherche de solutions rapides, etc.). A côté de ces travaux toujours d'actualité, se sont ouvertes de nouvelles recherches dédiées à des tâches de plus haut niveau et visant à l'interprétation des images.

Ce polycopié présente d'une façon relativement complète les champs de Markov et les différents domaines qui s'y rattachent. Le lecteur y trouvera plusieurs degrés de détails et de difficultés suivant ses motivations. Il s'organise comme suit:

- le chapitre 1 est un chapitre d'introduction aux techniques markoviennes et établit des notions de base; il donne la définition des champs markoviens et des champs de Gibbs et traite des techniques d'échantillonnage (échantillonneur de Gibbs, de Métropolis, recuit simulé); les applications en restauration et segmentation sont également mentionnées; des résultats concernant les statistiques exponentielles sont donnés dans les annexes A et B de ce chapitre, et les preuves de convergence des échantillonneurs de Gibbs et de Métropolis ainsi que du recuit simulé sont données en annexe C;
- le chapitre 2 présente différents estimateurs d'un champ de Markov caché correspondant à diverses fonctions de coût (estimateurs MAP, MPM, TMP); les algorithmes permettant d'obtenir ces estimateurs sont également décrits;
- le chapitre 3 aborde le problème de l'estimation des paramètres d'un champ markovien, aussi bien dans le cas de données complètes qu'incomplètes; les différentes techniques existant sur ce sujet sont présentées ainsi que leurs limites (limites d'applications -modèle de champ spécifique par exemple, limites théoriques -absence de preuves de convergence, etc.);

- le chapitre 4 présente un certain nombre d'applications pour lesquelles les champs markoviens ont montré leurs potentialités, qu'il s'agisse de problèmes de bas-niveau comme la restauration ou la segmentation d'images, ou de plus haut niveau, comme la détection de réseaux ou l'interprétation d'une scène.

Les approches concernant la préservation des contours dans le cadre de la modélisation markovienne et s'appuyant sur l'utilisation de processus bords explicites ou reposant sur le choix de modèles particuliers ne sont pas abordés ici. Par ailleurs, les problèmes spécifiques de la multi-résolution dans le cadre markovien ne sont pas développés.

Chapter 1

Définition et simulation d'un champ de Markov

1 Introduction

L'image numérique se présente et se manipule sous forme d'un tableau bidimensionnel (ou n-dimensionnel) d'une variable entière quantifiée. L'information véhiculée par cette représentation va en réalité bien au delà de la seule donnée des niveaux de gris en chaque site, et la description d'une image se fait en termes de zones, contours, structures définis par les contrastes, textures, etc. qui peuvent être présents dans l'image. Le niveau de gris en un site n'est donc souvent pas significatif en lui-même, mais dans ses relations et interactions avec les pixels voisins.

Cette propriété des images, à savoir les interactions locales entre niveaux de gris voisins pour définir les différentes régions de l'image, va nous permettre d'utiliser un formalisme markovien dans de nombreux traitements, qu'il s'agisse de restauration, de segmentation et plus tard d'analyse complète des images. Le principe est de définir des énergies locales entre groupes de sites reflétant les interactions entre niveaux de gris. L'énergie globale est alors reliée à la probabilité d'apparition de l'image dans le cadre des champs de Gibbs.

Dans ce chapitre, nous introduisons tout d'abord de façon intuitive la notion d'énergie locale avant de définir plus formellement un champ de Markov et d'énoncer le théorème d'équivalence entre champs de Markov et champs de Gibbs. Les algorithmes d'échantillonnage d'un champ de Markov (échantilleur de Gibbs et algorithme de Métropolis) sont ensuite présentés, ainsi que les modèles markoviens les plus courants. L'utilisation des MRF en traitement d'images dans un cadre bayésien, montre la nécessité de pouvoir accéder aux configurations les plus probables d'un champ markovien et nous amène à la présentation du recuit simulé.

2 Description et modélisation probabiliste de l'image

2.1 Description de l'image

L'image est formée d'un ensemble fini S de sites s_i correspondant aux pixels. S est donc essentiellement un réseau discret fini, partie de \mathbb{Z}^d , si on note d la dimension de l'espace (2 le plus classiquement, 3 pour les volumes, etc.). A chaque site est associé un descripteur, représentant l'état du site et qui peut être son niveau de gris, une étiquette, ou une information plus complexe, et prenant ses valeurs dans E .

La notion d'interactions locales nécessite de structurer les relations spatiales entre les différents sites du réseau. Pour ce faire, on munit S d'un système de voisinage \mathcal{V} défini de la façon suivante:

$$\mathcal{V}_s = \{t\} \text{ tels que } \begin{cases} \cdot & s \notin \mathcal{V}_s \\ \cdot & t \in \mathcal{V}_s \Rightarrow s \in \mathcal{V}_t \end{cases}$$

A partir d'un système de voisinage, un système de cliques peut être déduit : une clique est soit un singleton de S , soit un ensemble de sites tous voisins les uns des autres. En fonction du système de voisinage utilisé, le système de cliques sera différent et fera intervenir plus ou moins de sites comme illustré sur la figure 1.1. On notera \mathcal{C} l'ensemble des cliques relatif à \mathcal{V} , et \mathcal{C}_k l'ensemble des cliques de cardinal k .

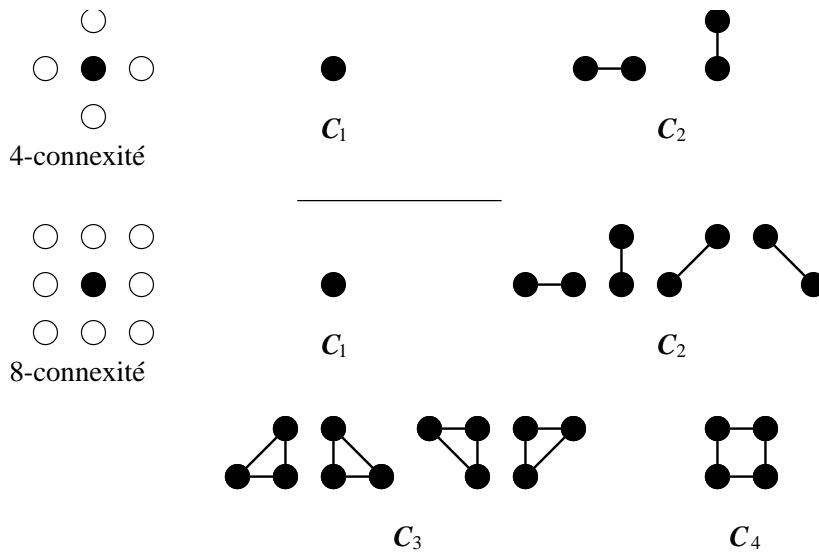


Figure 1.1: Les cliques associées à divers systèmes de voisinage en dimension $d = 2$

Les interactions locales entre niveaux de gris (ou descripteurs) de sites voisins peuvent alors s'exprimer comme un potentiel de clique. Soit c une clique, on lui associe le potentiel U_c dont la valeur dépend des niveaux de gris (ou descripteurs) des pixels constituant la clique. En poursuivant ce raisonnement, on peut définir l'énergie globale de l'image comme la somme des

potentiels de toutes les cliques :

$$U = \sum_{c \in \mathcal{C}} U_c$$

et l'énergie locale en un site comme la somme des potentiels de toutes les cliques auxquelles il appartient :

$$U_s = \sum_{c \in \mathcal{C} / s \in c} U_c$$

Nous avons jusqu'ici considéré le cas d'une image pour illustrer les notions de voisinage, de clique et de potentiel, mais le formalisme markovien se définit très généralement sur tout graphe. Soit un ensemble de sites S dénombrable (sommets du graphe), et une relation de voisinage, les cliques sont alors définies comme les sous-graphes complets du graphe. C'est l'utilisation de graphes plus généraux que ceux définis sur la grille de l'image qui permettent des traitements de plus haut niveau.

2.2 Modélisation probabiliste de l'image

La définition des champs de Markov qui sera donnée dans la section suivante nécessite une modélisation probabiliste de l'image. Ainsi, l'image dont nous disposons va être considérée comme une réalisation d'un champ aléatoire. Soit s un site de l'image, on peut en effet lui associer une variable aléatoire X_s prenant ses valeurs dans E . Le niveau de gris x_s en s n'est ainsi qu'une réalisation¹ de la v.a X_s . On définit alors le champ aléatoire $X = (X_s, X_t, \dots)$ prenant ses valeurs dans $\Omega = E^{|S|}$. On trouvera aussi le terme de processus aléatoire pour X ; en toute rigueur, "processus" devrait être réservé au cas d'un ensemble d'indexation continu, et champ au cas discret.

Dans ce cadre probabiliste, l'image considérée est simplement une réalisation x du champ. La probabilité globale de x , $P(X = x)$, permet d'accéder en quelque sorte à la vraisemblance de l'image, et les probabilités conditionnelles locales d'une valeur en un site permettent de mesurer le lien statistique entre un niveau de gris et le reste de l'image. L'hypothèse markovienne permet d'évaluer ces quantités.

Notons que nous nous plaçons dans le cas où E , l'espace de valeurs des descripteurs, est quantifié, ce qui nous permet de manipuler des probabilités. Dans le cas où cet espace est continu, il faut remplacer P par une densité de probabilité, mais dans ce cas le théorème que nous allons voir ci-dessous n'est plus valable.

¹On notera généralement en lettres majuscules les variables aléatoires (v.a) et en minuscules leurs réalisations.

3 Champs de Markov - Champs de Gibbs

3.1 Définition d'un champ de Markov

Considérons x_s la valeur du descripteur prise au site s et $x^s = (x_t)_{t \neq s}$ la configuration de l'image excepté le site s . La définition d'un champ de Markov est alors la suivante :

X est un champ de Markovssi la probabilité conditionnelle locale en un site n'est fonction que de la configuration du voisinage du site considéré

ce qui s'exprime de façon formelle par :

$$P(X_i = x_i / x^i) = P(X_i = x_i / x_j, j \in \mathcal{V}_i)$$

Ainsi, le niveau de gris en un site ne dépend que des niveaux de gris des pixels voisins de ce site. Cette hypothèse markovienne se justifie bien dans le cas de la plupart des images naturelles constituées de zones homogènes ou texturées ainsi que pour une large gamme d'images de synthèse. Plus généralement, une connaissance locale de l'image suffit souvent à réaliser son interprétation partielle et donc cette hypothèse markovienne sera souvent justifiée sur des graphes plus globaux que le graphe des pixels.

Notons qu'en l'absence de contrainte sur le système de voisinage, tous les champs aléatoires peuvent être considérés comme markoviens à condition de prendre un voisinage suffisamment grand. L'intérêt de cette modélisation réside bien sûr dans le cas où la propriété markovienne est vérifiée pour des voisinages restreints permettant des calculs rapides.

3.2 Equivalence champs de Markov - champs de Gibbs

La modélisation markovienne prend toute sa puissance grâce au théorème que nous allons voir maintenant. En effet, celui-ci permettra d'accéder aux expressions des probabilités conditionnelles locales. Il nous faut au préalable définir un certain nombre de notions relatives aux mesures et champs de Gibbs.

Définition d'une mesure de Gibbs: La mesure de Gibbs de fonction d'énergie (ou d'Hamiltonien) $U : \Omega \rightarrow \mathbb{R}$ est la probabilité P définie sur Ω par :

$$P(X = x) = \frac{1}{Z} \exp(-U(x))$$

avec

$$U(x) = \sum_{c \in \mathcal{C}} U_c(x)$$

où \mathcal{C} est le système de cliques associé au système de voisinage \mathcal{V} de U ². $Z = \sum_{x \in \Omega} \exp(-U(x))$ est une constante de normalisation appelée fonction de partition de Gibbs. En pratique, il est quasi

²Il est toujours possible de trouver un système de voisinage \mathcal{V} permettant de décomposer U ; le cas extrême correspondant à des sites tous voisins les uns des autres.

impossible de calculer cette constante à cause du très grand nombre de configurations possibles. Ne serait-ce que dans le cas d'une image binaire ($\text{Card}(E) = 2$) et de taille $= 512 \times 512$, on a $2^{2^{62144}}$ configurations possibles!

La notation couramment utilisée pour $U(x)$ est abusive car $U_c(x)$ ne dépend pas de l'ensemble de la configuration x mais seulement de x restreinte à la clique c ($U_c(x) = U_c(x_j, j \in c)$).

Nous pouvons maintenant définir le champ de Gibbs de potentiel associé au système de voisinage \mathcal{V} : c'est le champ aléatoire X dont la probabilité est une mesure de Gibbs associée au système de voisinage \mathcal{V} , ce qui implique :

$$P(X = x) = \frac{1}{Z} \exp(-U(x)) = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} U_c(x)\right)$$

L'énergie globale d'un champ de Gibbs possède donc la propriété de se décomposer sous forme d'une somme d'énergies locales, qui comme on le verra par la suite permettront d'accéder aux probabilités conditionnelles locales. Notons ici que plus une configuration d'un champ de Gibbs a une énergie faible, plus elle est probable.

Le théorème de Hammersley-Clifford [Besag(1974)] établit alors le résultat fondamental suivant sous les hypothèses :

- S fini ou dénombrable,
- système de voisinage \mathcal{V} borné,
- espace des états E discret

X est un champ de Markov relativement à \mathcal{V}
 et $P(X = x) > 0 \quad \forall x \in \Omega$
 \Leftrightarrow
X est un champ de Gibbs de potentiel associé à \mathcal{V}

Par exemple, si nous considérons un champ de Markov de voisinage 4-connexe, nous pouvons écrire l'énergie de la configuration x sous la forme :

$$U(x) = \sum_{c=(i) \in \mathcal{C}_1} U_c(x_i) + \sum_{c=(i,j) \in \mathcal{C}_2} U_c(x_i, x_j)$$

Notons que rien n'interdit la non-stationnarité du champ, c'est à dire la variation des potentiels U_c en fonction de la localisation de la clique c dans l'image [Descombes(1993)]. D'autre part, rien n'impose la symétrie des potentiels et on peut avoir $U_{c=(r,s)}(0, 1) \neq U_c(1, 0)$.

Le théorème de Hammersley-Clifford, et la forme bien spécifique de probabilité de X qui en résulte, va permettre de lier les probabilités globales et locales comme nous allons le voir maintenant. En effet si nous cherchons à écrire la probabilité conditionnelle locale $P(x_s / X^s = x^s)$, nous avons grâce au résultat précédent :

$$P(X_s = x_s / X^s = x^s) = \frac{P(X = x)}{P(X^s = x^s)} = \frac{\exp(-U(x_s, x^s))}{\sum_{x_s \in E} \exp(-U(x_s, x^s))}$$

Définissons l'énergie locale U_s par:

$$\begin{aligned} U_s(x_s / x_t, t \in \mathcal{V}_s) &= \sum_{c \in \mathcal{C} / s \in c} U_c(x_s, x_t, t \in \mathcal{V}_s) \\ &= \sum_{c \in \mathcal{C} / s \in c} U_c(x_s, V_s) \end{aligned}$$

en notant $V_s = (x_t, t \in \mathcal{V}_s)$. Cette énergie locale ne fait donc intervenir que les voisins de i . On peut alors écrire l'énergie globale $U(x)$ sous la forme :

$$\begin{aligned} U(x) &= \sum_{c \in \mathcal{C} / s \notin c} U_c(x) + \sum_{c \in \mathcal{C} / s \in c} U_c(x) \\ &= \sum_{c \in \mathcal{C} / s \notin c} U_c(x) + U_s(x_s / V_s) \end{aligned}$$

En simplifiant l'expression de la probabilité conditionnelle locale en supprimant les termes communs qui font intervenir les cliques ne contenant pas le site i au numérateur et au dénominateur, on a:

$$\begin{aligned} P(X_s = x_s | X^s = x^s) &= \frac{\exp\left(-\sum_{c \in \mathcal{C} / s \notin c} U_c(x) - U_s(x_s / V_s)\right)}{\sum_{x_s \in \Lambda} \exp\left(\sum_{c \in \mathcal{C} / s \notin c} U_c(x) - U_s(x_s / V_s)\right)} \quad (1.1) \\ &= \frac{\exp(-U_s(x_s / V_s))}{\sum_{x_s \in \Lambda} \exp(-U_s(x_s / V_s))} \end{aligned}$$

L'expression obtenue, qui ne fait intervenir que les potentiels des cliques contenant le site s (ce qui nous permet de retrouver au passage l'hypothèse markovienne), est très importante. En effet, autant il n'est pas possible partant d'une configuration x d'accéder à sa probabilité à cause de la constante de normalisation, autant il est possible de calculer en chaque site la probabilité conditionnelle locale. Cette expression sera à la base de tous les algorithmes de simulation de champs markoviens que nous verrons dans la section suivante.

Remarque : Le théorème de Hammersley-Clifford n'est valable que lorsqu'aucune configuration n'est interdite (condition $P(X = x) > 0 \ \forall x$). Des solutions seront évoquées dans le chapitre, lorsqu'il est souhaitable de supprimer certaines configurations irréalistes (par exemple avoir une route à l'intérieur d'une zone de mer, ou de l'os dans la matière blanche du cerveau).

4 Echantillonnage de MRF

Si nous résumons les résultats précédents, la définition d'un champ de Markov passe par la définition de sa fonction d'énergie U . Celle-ci nécessite la définition d'un système de voisinage, qui définit alors le système de cliques, et de fonctions de potentiel associées aux cliques. Ces

fonctions de potentiel permettent d'accéder à la probabilité globale d'une configuration, et aux probabilités conditionnelles locales.

Le problème qui se pose alors est, étant défini un champ de Markov, comment pouvons-nous réaliser le tirage d'une configuration (une image ici) en suivant la loi de probabilité de Gibbs caractéristique de ce champ? Deux algorithmes ont été proposés pour synthétiser des réalisations d'un champ de Markov qui sont:

- l'échantillonneur de Gibbs,
- l'algorithme de Métropolis

que nous allons décrire maintenant.

4.1 L'échantillonneur de Gibbs

Cet algorithme, proposé par Geman et Geman [Geman and Geman(1984a)], repose sur la construction itérative d'une suite d'images. A la convergence, i.e après un nombre d'itérations suffisant, les images construites sont des réalisations tirées selon la loi de Gibbs globale.

La méthode de construction de l'image à l'itération n , partant de l'image à l'itération $n - 1$ se fait par mise à jour successive des sites de l'image. A l'étape n :

- choix d'un site s ;
- au site s , selon la configuration des voisins V_s pour l'image $x^{(n-1)}$, calcul de la probabilité conditionnelle locale :

$$P(X_s = x_s | V_s) = \frac{\exp(-U_s(x_s | V_s))}{\sum_{x_s \in \Lambda} (\exp(-U_s(\xi | V_s)))}$$

- mise à jour du site s par tirage aléatoire selon la loi $P(X_s = x_s | V_s)$

On considère que l'algorithme a convergé après un grand nombre d'itérations ou lorsque le nombre de changements est faible. Le choix du site s considéré à l'étape n peut se faire de n'importe quelle façon à condition de balayer tous les sites un très grand nombre de fois (théoriquement un nombre infini de fois). Les méthodes usuelles consistent à tirer un site selon une loi uniforme, ou effectuer un balayage classique, ligne par ligne, de l'image.

Cet algorithme construit en réalité une suite d'images $x^{(n)}$ qui sont les observations d'une suite $X^{(n)}$ de champs aléatoires constituant une chaîne de Markov pour un certain noyau de transition. On peut montrer le théorème suivant, lorsque la séquence balaye chaque site une infinité de fois :

$$\forall x^{(0)} \forall x \in \Omega \quad \lim_{n \rightarrow \infty} P(X^{(n)} = x | X^{(0)} = x^{(0)}) = P(x)$$

où P est la mesure de Gibbs associée au champ de Markov considéré. Ainsi, après un grand nombre d'itérations, les images $x^{(n)}$ générées sont des réalisations de la loi globale $P(x)$, et ceci indépendamment de la configuration initiale $x^{(0)}$.

La preuve de ce théorème, et donc de la convergence de l'algorithme, est donnée en annexe C de ce chapitre, après le rappel d'un certain nombre de résultats nécessaires sur les statistiques exponentielles en annexe A et annexe B.

On parle de l'échantillonneur de Gibbs comme d'un algorithme de "relaxation", car il procède par mise à jour successive des sites, et "probabiliste" car celle-ci est fondée sur un tirage aléatoire.

4.2 L'algorithme de Métropolis

L'échantillonneur de Gibbs est un algorithme très utilisé en traitement d'images pour la synthèse de champs de Markov. Néanmoins, un algorithme antérieur et issu de la physique statistique avait été mis au point dans les années 50 par Métropolis [Metropolis *et al.*(1953)].

Cet algorithme repose sur un principe similaire à l'échantillonneur de Gibbs, et il s'agit également d'un algorithme de relaxation probabiliste. Le principe est là encore de construire une suite d'images qui seront des tirages selon la loi du champ de Markov après un nombre suffisamment grand d'itérations. Mais la mise à jour en un site s'effectue de façon différente. Ainsi à l'étape n :

- choix d'un site s
- tirage aléatoire d'un descripteur λ dans E selon une loi uniforme;
- calcul de la variation d'énergie pour $x_s^{(n-1)} \rightarrow \lambda$:

$$\Delta U = U_s(\lambda \mid V_s^{(n-1)}) - U_s(x_s^{(n-1)} \mid V_s^{(n-1)})$$

- deux cas sont alors possibles:

1. $\Delta U < 0$, le changement est accepté : $x_s^{(n)} = \lambda$;
2. $\Delta U \geq 0$, le changement est accepté ou refusé par tirage selon la probabilité $p = \exp(-\Delta U)$ et $1 - p$.

Le système de balayage des sites et le critère d'arrêt sont similaires à ceux de l'échantillonneur de Gibbs. La différence avec l'échantillonneur de Gibbs réside dans le tirage au sort du nouveau niveau de gris (ou descripteur), au lieu de considérer la loi définie par tous les descripteurs. Comme on ne considère que la variation énergétique entre les 2 configurations, l'algorithme de Métropolis est plus rapide à chaque étape que l'échantillonneur de Gibbs, qui lui nécessite le calcul de la fonction de partition locale. Mais la convergence peut être plus lente car le taux d'acceptation est strictement inférieur à 1 (les transitions ne sont pas toujours acceptées, contrairement au cas de l'échantillonneur de Gibbs).

Là encore, le principe est de construire une chaîne de Markov selon un certain noyau de transition (différent de celui intervenant dans l'échantillonneur de Gibbs). Le théorème précédent est alors encore vérifié pour l'algorithme de Métropolis comme le montre l'annexe C.

4.3 Quelques MRF fondamentaux

Nous présentons ici quelques uns des champs de Markov les plus utilisés. Comme indiqué précédemment, ces champs sont définis par leur voisinage et leurs fonctions de potentiel. Ils sont illustrés par le tirage de réalisations selon l'échantillonneur de Gibbs.

Modèle d'Ising :

Ce modèle est le plus ancien (1925 [Ising(1925)]) et a été développé lors de l'étude du ferromagnétisme en physique statistique. L'espace des descripteurs est celui des états des spins, i.e $E = \{-1, 1\}$ (espace binaire), et le voisinage est constitué par les 4 ou 8 plus proches voisins dans un espace bidimensionnel. Les potentiels sont des potentiels en tout ou rien :

$$\begin{aligned} U_{c=(s,t)}(x_s, x_t) &= -\beta \text{ si } x_s = x_t \\ &= +\beta \text{ si } x_s \neq x_t \end{aligned}$$

Ce qui s'écrit également :

$$U_{c=(s,t)}(x_s, x_t) = -\beta x_s x_t$$

Les potentiels des cliques d'ordre 1 (clique constituée par un seul spin) sont de la forme $-Bx_s$. L'énergie totale s'écrit :

$$U(x) = - \sum_{c=(s,t) \in \mathcal{C}} \beta x_s x_t - \sum_{s \in S} B x_s$$

β est la constante de couplage entre sites voisins et B représente un champ magnétique externe. Lorsque β est positif, les configurations les plus probables (i.e d'énergies plus faibles) sont celles pour lesquelles les spins sont de même signes (ferro-magnétisme), alors que dans le cas de β négatif, au contraire, on favorisera l'alternance de spins de signes opposés (antiferromagnétisme). La valeur (signe et valeur absolue) de β conditionne donc la régularité du modèle d'Ising. Quant au champ magnétique externe relatif au potentiel d'ordre 1, il favorise a priori par son signe un spin ou un autre.

La figure 1.2 montre des réalisations de modèles d'Ising pour différents paramètres (la régularisation appelée “critique” correspond à l'apparition des zones homogènes).

Modèle de Potts :

Il s'agit d'une extension du modèle d'Ising [Wu(1982)] pour un espace m -aire, i.e $E = \{0, m - 1\}$. Il peut s'agir de plusieurs niveaux de gris, mais plus souvent pour ce modèle, d'étiquettes (labels) pouvant représenter une classification de l'image (par exemple les classes *eau*, *forêt*, *champ*, *ville*). Le voisinage considéré est 4- ou 8-connexe et les potentiels sont comme précédemment en tout ou rien mais définis seulement pour les cliques d'ordre 2:

$$\begin{aligned} U_{c=(s,t)}(x_s, x_t) &= -\beta \text{ si } x_s = x_t \\ &= +\beta \text{ si } x_s \neq x_t \end{aligned}$$

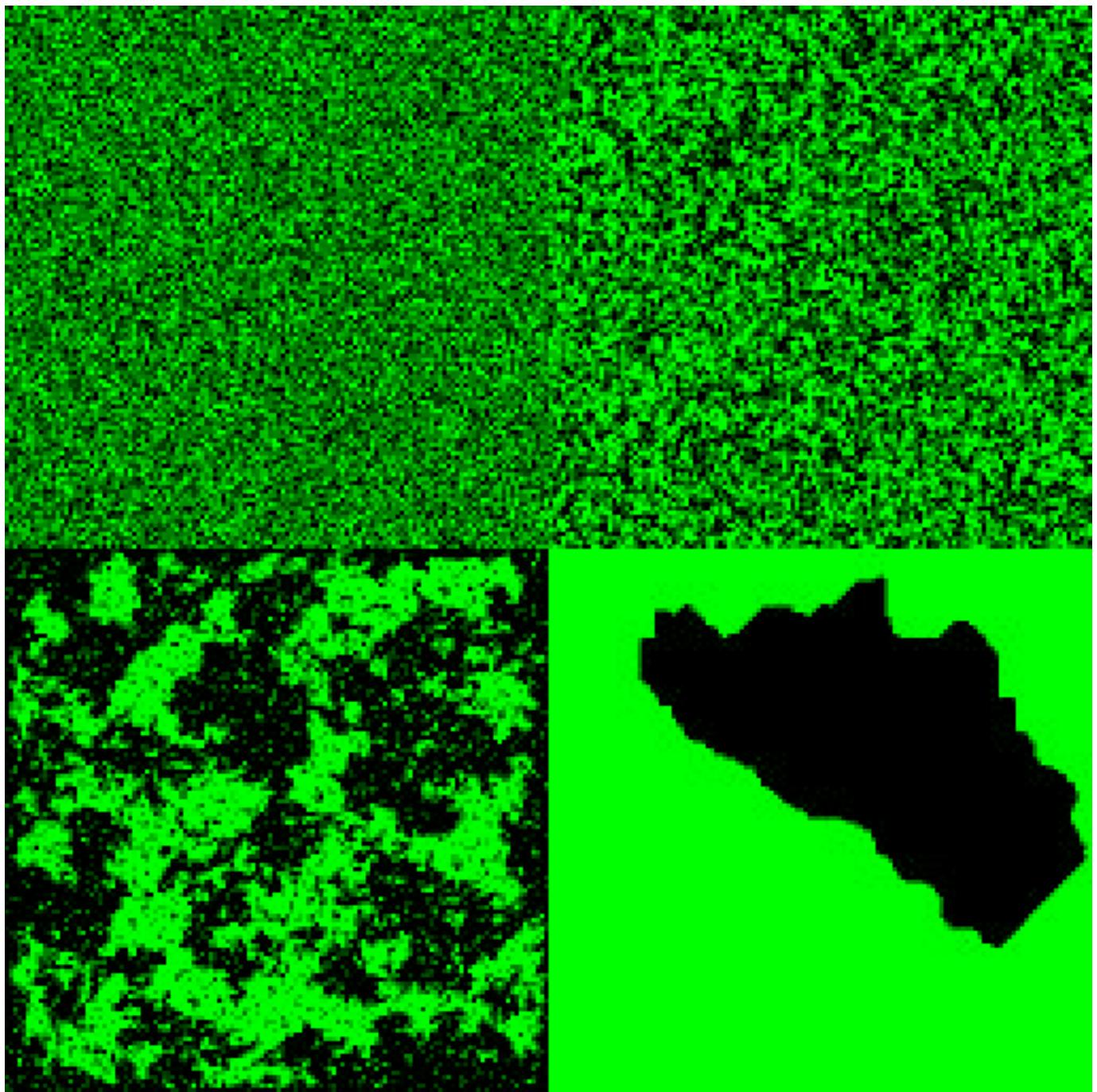


Figure 1.2: Le Modèle d'Ising plan et 4-connexe pour différentes valeurs de paramètres - Simulations en raccordement torique

A	B
C	D

- A : Image aléatoire : $\beta = 0$
- B : régularisation faible : $\beta = 0.2$
- C : régularisation “critique” : $\beta \approx 0.44$
- D : régularisation forte : $\beta = 4.0$

Lorsque β est positif, les configurations les plus probables correspondent à des sites voisins de même niveau de gris ou descripteur, ce qui donne des réalisations constituées par des larges zones homogènes. La taille de ces régions est gouvernée par la valeur de β . Des exemples de réalisations pour différentes valeurs de β sont montrés figure 1.3.

Il est possible de définir des modèles utilisant des pondérations β différentes en fonction des directions des cliques, et de privilégier ainsi certaines directions.

Ce modèle permet également de prendre en compte différentes relations entre les régions (i.e entre différentes valeurs des descripteurs). On peut par exemple définir des pondérations $\beta(e_s, e_t)$ pour $e_i, e_j \in E$. Dans notre exemple de classification en 4 étiquettes *eau*, *forêt*, *champ*, *ville*, une configuration de sites avec les étiquettes *champ* / *forêt* peut être supposée plus probable qu'une configuration *ville* / *forêt*, d'où des valeurs $\beta(\text{champ}, \text{forêt})$ et $\beta(\text{ville}, \text{forêt})$ différentes.

Modèle markovien gaussien :

Ce modèle est réservé aux images en niveaux de gris $E = \{0, \dots, 255\}$ et ne convient pas bien aux images d'étiquettes. Le voisinage est 4 ou 8-connexe et l'énergie est de la forme:

$$U(x) = \beta \sum_{c=(s,t)} (x_s - x_t)^2 + \alpha \sum_{s \in S} (x_s - \mu_s)^2$$

Le premier terme correspondant aux cliques d'ordre 2 est un terme de régularisation, qui favorise les faibles différences de niveaux de gris entre sites voisins pour $\beta > 0$. Le second terme peut correspondre à un terme d'attache aux données dans le cas où on possède une image de données extérieures. Des exemples de synthèse de modèles gaussiens sont montrés figure 1.4. Le rapport $\frac{\alpha}{\beta}$ pondère les influences respectives de l'attache aux données et de la régularisation, et les valeurs absolues des paramètres caractérisent le caractère plus ou moins "piqué" ou équiréparti au contraire de la distribution.

Le modèle gaussien favorise des niveaux de gris proches pour des pixels voisins dans tous les cas. Or si on considère une image naturelle cet aspect est néfaste à proximité des contours car il favorisera la présence d'un dégradé. Aussi, de nombreuses fonctions ϕ ont été proposées pour modéliser les potentiels des cliques d'ordre 2 : $U_{c=(s,t)} = \phi(x_s - x_t)$. En particulier, la fonction suivante permet de respecter les contours de l'image [Geman and McClure(1985)]:

$$\phi(x) = \frac{1}{1 + (\frac{x}{\delta})^2}$$

et est donc très utilisée en restauration. Ces modèles permettent de synthétiser des textures très variées.

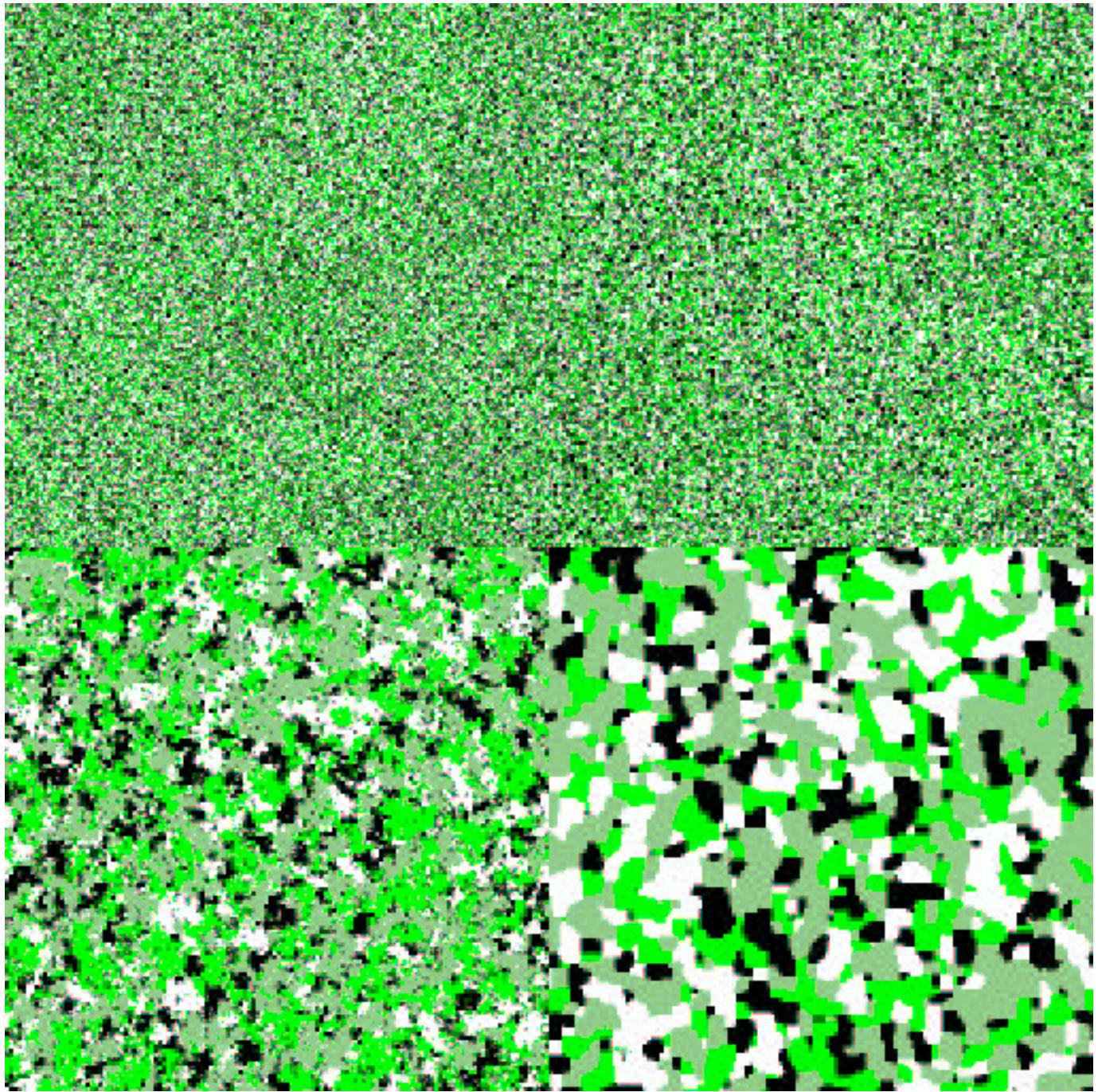


Figure 1.3: Le Modèle de Potts 2D et 4-connexe pour différentes valeurs de paramètres ($m = 4$) - Simulations en raccordement torique

A	B
C	D

- A : Image aléatoire : $\beta = 0$
- B : régularisation faible : $\beta = 0.2$
- C : régularisation “critique” : $\beta \approx 1,099$
- D : régularisation forte : $\beta = 4.0$

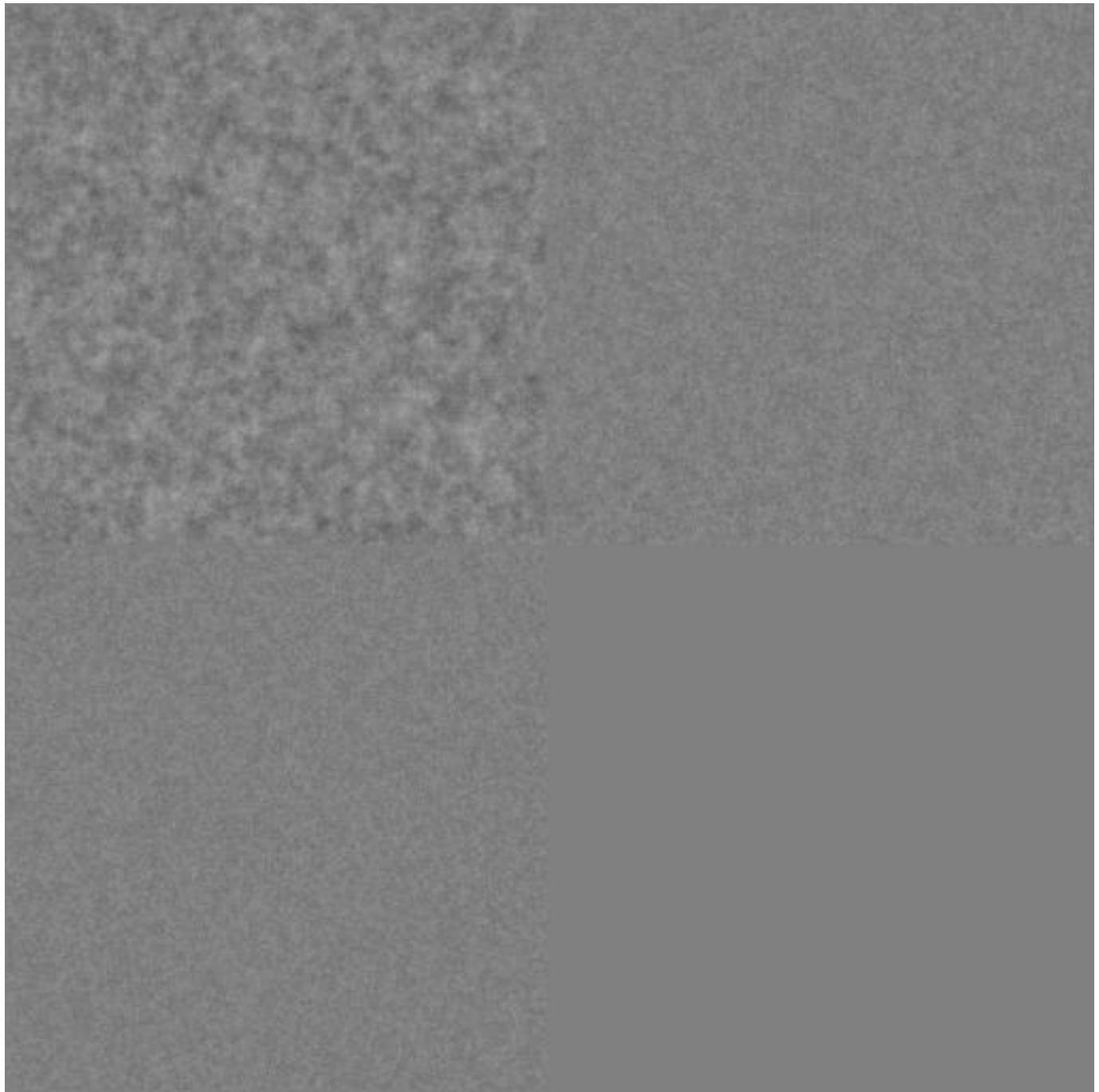


Figure 1.4: Le modèle gaussien 4-connexe

$$U(x) = \sum_{c=(s,t)} (x_s - x_t)^2 + \alpha \sum_{s \in S} (x_s - \mu_s)^2$$

A	B
C	D

- **A** : $\alpha = 5.10^{-4}$
- **B** : $\alpha = 5.10^{-3}$
- **C** : $\alpha = 2.10^{-3}$
- **D** : $\alpha = \infty$ ($\mu = 127$) pour toutes simulations)

5 Applications : restauration et segmentation

Nous avons étudié dans la section précédente des algorithmes de simulation permettant des générer des réalisations d'un champ de Markov donné. Nous abordons ici leur utilisation en tant qu'outils de traitements d'images pour deux grandes applications de bas-niveau : la restauration d'images et la segmentation.

5.1 Cadre bayésien

Pour ces deux applications, on peut mener modéliser le problème dans un cadre bayésien de la façon suivante. Nous disposons d'une certaine donnée (image) que nous noterons y et que nous pouvons considérer comme une réalisation d'un champ aléatoire Y . Nous cherchons une réalisation x de l'image restaurée ou segmentée, que nous pouvons modéliser comme un champ de Markov X . X est le champ des étiquettes (labels) dans le cas de la segmentation, le champ des intensités dans le cas de la restauration. Les espaces de configurations ne sont donc pas nécessairement les mêmes pour X et Y . Ces deux champs sont liés par le processus d'acquisition de l'image, qui conduit du champ idéal X , le processus image originel que nous cherchons, au champ bruité Y que nous observons. La restauration ou la segmentation ont pour objectif d'inverser le processus et donc de remonter à une réalisation de X à partir de l'observation des données bruitées y . On parle dans ce contexte de champ de Markov "caché" pour X , ou de données incomplètes puisque y n'est pas une réalisation de X .

On peut par exemple utiliser le critère du maximum a posteriori et rechercher la configuration \hat{x} maximisant la probabilité de X conditionnellement à la donnée y i.e $P(X = x / Y = y)$. Or la règle de Bayes permet d'écrire:

$$P(X = x / Y = y) = \frac{P(Y = y / X = x)P(X = x)}{P(Y = y)}$$

Expression dans laquelle il s'agit alors d'analyser chacun des termes $P(Y = y | X = x)$ et $P(X = x)$, sachant que $P(Y)$ est une constante (indépendante de la réalisation x). Le premier terme $P(Y = y | X = x)$ décrit justement le processus d'observation et d'acquisition des données. L'hypothèse la plus courante (dont la validité reste à justifier) consiste à supposer l'indépendance conditionnelle des pixels (bruit non corrélé par exemple) :

$$P(Y = y / X = x) = \prod_s P(Y_s = y_s / X_s = x_s)$$

Cette écriture n'est plus valable lorsqu'il y a une convolution par la fonction de transfert du système d'acquisition, mais on peut montrer que le champ a posteriori reste markovien.

Par ailleurs, on fait sur le champ X recherché une hypothèse markovienne selon un voisinage \mathcal{V} et un modèle donné dépendant de l'application. On peut alors écrire :

$$P(X = x) = \frac{\exp(-U(x))}{Z}$$

Si on revient maintenant à la distribution a posteriori, celle-ci s'exprime par:

$$\begin{aligned} p(X = x \mid Y = y) &\propto p(Y \mid X)p(X) \\ &\propto e^{\ln p(Y \mid X) - U(x)} \\ &\propto e^{-\mathcal{U}(x \mid y)} \end{aligned}$$

avec

$$\mathcal{U}(x \mid y) = \sum_{s \in S} -\ln p(y_s \mid x_s) + \sum_{c \in \mathcal{C}} U_c(x) \quad (1.2)$$

Par conséquent, sous les hypothèses précédentes, on constate que la distribution a posteriori est une distribution de Gibbs et que donc le champ X conditionnellement à y est également un champ de Markov (théorème de Hammersley-Clifford). Ainsi, il est possible de simuler des réalisations de ce champ à l'aide de l'échantillonneur de Gibbs ou de l'algorithme de Métropolis. Mais la configuration x qui nous intéresse est celle maximisant la probabilité a posteriori, donc la réalisation la plus probable du champ de Gibbs, ou encore celle qui minimise l'énergie $\mathcal{U}(x \mid y)$. Un algorithme a été proposé pour atteindre cet (ou ces) état(s) d'énergie minimale, il s'agit du recuit simulé qui est décrit dans la section suivante.

5.2 Cas de la restauration

Reprendons la démarche précédente et exprimons plus en détails l'énergie $\mathcal{U}(x \mid y)$ dans un cas particulier de restauration.

Dans le cas où le processus d'acquisition entraîne une dégradation de l'image sous forme d'un bruit blanc gaussien de variance σ , on a la probabilité conditionnelle suivante:

$$P(y_s \mid x_s) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x_s - y_s)^2}{2\sigma^2}$$

En effet, soit B le bruit, on peut écrire :

$$P(y_s \mid x_s) = \frac{P(x_s, y_s)}{P(x_s)}$$

avec $y_s = x_s + b_s$ et X_s et B_s sont des variables aléatoires indépendantes. Donc, on a :

$$P(X_s = x_s, Y_s = y_s) = P(X_s = x_s, B_s = y_s - x_s) = P(X_s = x_s)P(B_s = y_s - x_s)$$

Et finalement:

$$P(y_s \mid x_s) = P(B_s = y_s - x_s)$$

ce qui permet d'obtenir l'expression donnée précédemment lorsque le bruit est gaussien de variance σ^2 .

La probabilité a priori $P(X = x)$ permet d'introduire les contraintes que nous souhaitons imposer à la solution (i.e que nous supposons pour le processus originel). En faisant l'hypothèse X markovien nous nous restreignons à des contraintes locales, le plus souvent de régularité entre sites voisins. On choisit fréquemment un modèle avec des potentiels d'ordre 2 :

$$P(X = x) = \frac{1}{Z} \exp(-\beta \sum_{(s,t) \in \mathcal{C}_2} \phi(x_s, x_t))$$

On a alors l'énergie suivante correspondant à la distribution de Gibbs du champ a posteriori:

$$\mathcal{U}(x / y) = \sum_{s \in S} \frac{(x_s - y_s)^2}{2\sigma^2} + \beta \sum_{(s,t) \in \mathcal{C}_2} \phi(x_s, x_t) \quad (1.3)$$

Le champ X conditionnellement à y est donc un champ de Gibbs pour le même système de voisinage que X . La constante β pondère l'influence entre le terme d'attache aux données (cliques d'ordre 1) qui impose des niveaux de gris x_s de l'image restaurée proches de ceux y_s de la donnée bruitée, et le terme de régularisation (cliques d'ordre 2) qui impose une solution constituée de zones homogènes. Le modèle pour X peut être soit markovien gaussien, soit plus adapté à la restauration des contours avec une fonction ϕ appropriée (cf. section 4.3).

5.3 Cas de la segmentation

Dans ce contexte, le champ markovien X est défini sur un autre espace de configurations que Y car seulement quelques étiquettes sont considérées : $E = \{1, \dots, m - 1\}$. Dans ce cas le processus de passage de X (champ des labels) à Y ne décrit pas tant le processus d'acquisition que l'apparence des classes dans l'image. Le terme $P(Y = y / X = x)$ traduit donc la probabilité de réalisation d'une configuration donnée connaissant son étiquetage (i.e connaissant la classe de chaque pixel). En supposant l'indépendance des sites les uns par rapport aux autres, et en supposant que le niveau de gris y_s en un site s ne dépend que de l'étiquette x_s en ce site, on a :

$$P(Y = y / X = x) = \prod_s P(y_s / x_s)$$

Les valeurs des probabilités conditionnelles sont données par l'histogramme conditionnel des niveaux de gris pour une classe donnée. Par exemple, si on suppose que chaque classe i a une distribution gaussienne de moyenne μ_i et d'écart-type σ_i , on a :

$$P(y_s / x_s = i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_s - \mu_i)^2}{2\sigma_i^2}\right)$$

Si comme précédemment on fait une hypothèse markovienne sur X et qu'on se limite aux cliques d'ordre 2, on a :

$$P(X = x) = \frac{1}{Z} \exp(-\beta \sum_{(s,t) \in \mathcal{C}_2} \phi(x_s, x_t))$$

D'où l'énergie a posteriori:

$$\mathcal{U}(x / y) = \sum_s \frac{(y_s - \mu_{x_s})^2}{2\sigma_{x_s}^2} + \log \sqrt{2\pi} \sigma_{x_s} + \beta \sum_{(s,t) \in \mathcal{C}_2} \phi(x_s, x_t) \quad (1.4)$$

Le champ des labels conditionnellement à y est markovien et d'énergie de Gibbs $\mathcal{U}(x / y)$. Là encore, comme pour la restauration, le terme d'ordre 1 exprime le respect des données (le niveau de gris doit correspondre à la classe), et le terme d'ordre 2 la contrainte de régularisation introduite. On choisit souvent un modèle de Potts pour X , ce qui donne une image segmentée avec de larges zones homogènes.

6 Le recuit simulé

Nous avons vu dans les applications précédentes qu'il était utile de pouvoir déterminer le ou les états d'énergie minimale qui correspondent au maximum de la probabilité d'un champ markovien. L'algorithme du recuit simulé permet de trouver ces configurations.

Avant de présenter cet algorithme, nous avons besoin de quelques résultats sur les distributions de Gibbs avec paramètre de température que nous présentons maintenant.

6.1 Distribution de Gibbs avec température

Une distribution de Gibbs avec paramètre de température est une probabilité qui s'écrit :

$$P_T(X = x) = \frac{1}{Z(T)} \exp - \frac{U(x)}{T}$$

avec $Z(T) = \sum_x \exp - \frac{U(x)}{T}$ et $T > 0$. Le terme de température provient de l'analogie avec la physique statistique.

Il est intéressant d'étudier le comportement de cette distribution pour des valeurs extrêmes du paramètre de température.

- $T \rightarrow \infty$:

On a $\exp - \frac{U(x)}{T} \rightarrow 1$ et comme $\sum_x P_T(X = x) = 1$, on obtient

$$P_T(X = x) \rightarrow \frac{1}{\text{Card } \Omega}$$

Donc P_T converge vers la probabilité uniforme sur Ω , i.e pour une température infinie tous les états sont équiprobables.

- $T \rightarrow 0$:

Notons U^* l'énergie minimale et Ω^* l'ensemble des configurations atteignant l'énergie minimale $\Omega^* = \{x_1, \dots, x_k\}$ (x_1, \dots, x_k sont les minima globaux de l'énergie). On peut écrire:

$$\begin{aligned} P_T(X = x) &= \frac{\exp - \frac{U(x)}{T}}{\sum_y \exp - \frac{U(y)}{T}} \\ &= \frac{\exp - \frac{U(x) - U^*}{T}}{\sum_y \exp - \frac{U(y) - U^*}{T}} \\ &= \frac{\exp - \frac{U(x) - U^*}{T}}{\sum_{y \notin \Omega^*} \exp - \frac{U(y) - U^*}{T} + \sum_{y \in \Omega^*} 1} \end{aligned}$$

◊ Si $x \notin \Omega^*$, on a $U(x) - U^* > 0$ et $\exp(\frac{U(x)-U^*}{T}) \rightarrow 0$ pour $T \rightarrow 0$. Donc $P_T(x) \rightarrow 0$ si x n'est pas un minimum global de l'énergie.

◊ Si $x \in \Omega^*$, on a :

$$P_T(x_1) = P_T(x_2) = \dots = P_T(x_k) = \frac{1}{k}$$

(il y a une somme finie de termes qui tendent vers 0 au dénominateur).

Ce qui signifie que lorsque la température est nulle P_T est uniformément distribuée sur les minima globaux de l'énergie, i.e sur les configurations les plus probables. C'est ce résultat qui est à la base de l'algorithme de recuit simulé.

On retrouvera ces résultats ainsi qu'une étude complète du comportement de P_T et son espérance en fonction de T dans l'annexe A sur l'étude des statistiques exponentielles.

6.2 Le recuit simulé

Cet algorithme est dédié à la recherche d'une configuration d'énergie minimale d'un champ de Gibbs. L'idée d'intégrer une paramètre de température et de simuler un recuit a été initialement proposée par Kirkpatrick [Kirkpatrick *et al.*(1982)] et reprise par Geman et Geman [Geman and Geman(1984a)] qui ont proposé l'algorithme suivant.

Comme les algorithmes de simulation, c'est un algorithme itératif qui construit la solution au fur et à mesure. Le déroulement de l'algorithme est le suivant (en notant n le numéro de l'itération):

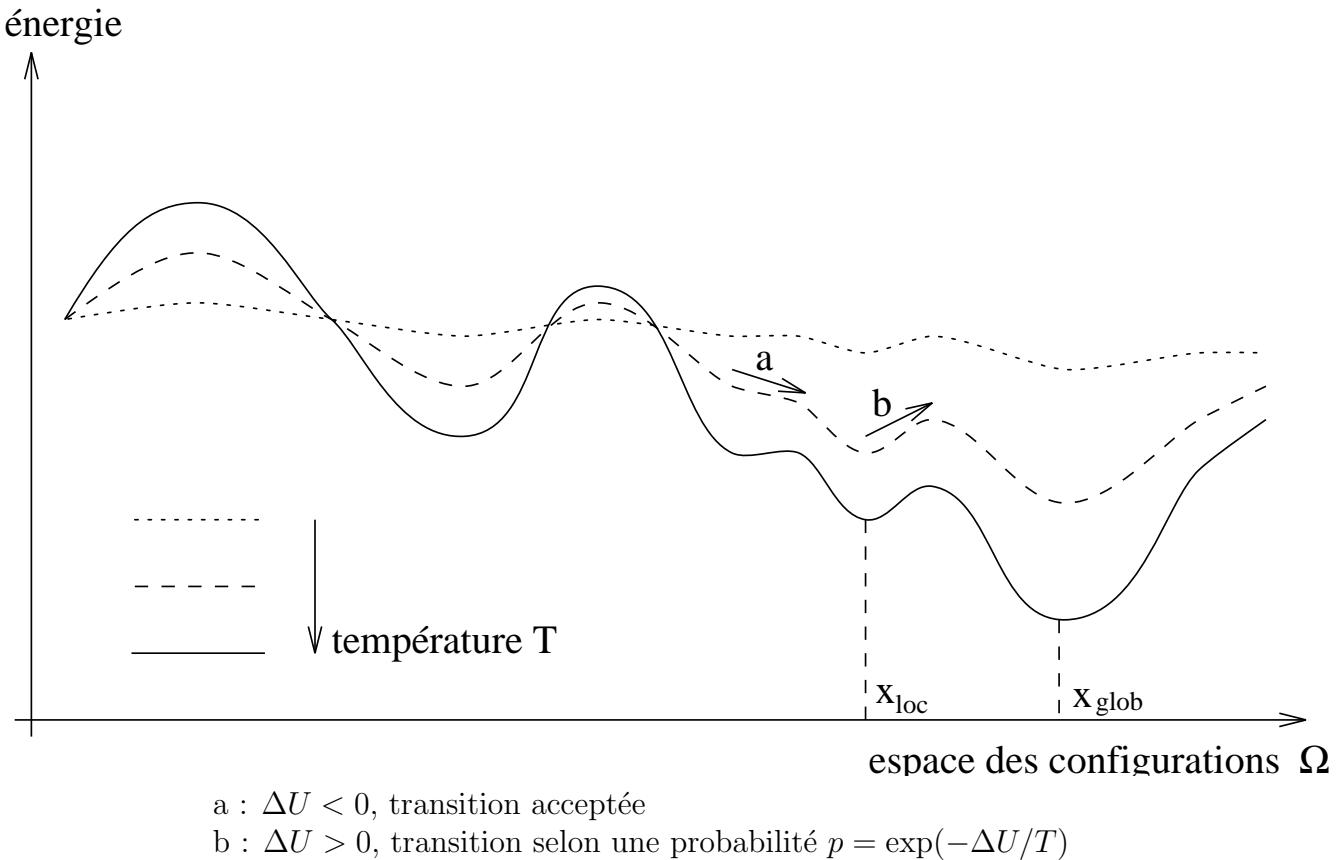
- choix d'une température initiale $T^{(0)}$ "grande"
- choix d'une configuration initiale quelconque $x^{(0)}$
- à l'étape n

- simulation d'une configuration $x^{(n)}$ pour la loi de Gibbs d'énergie $\frac{U(x)}{T^{(n)}}$ à partir de la configuration $x^{(n-1)}$; la simulation peut se faire par l'échantillonneur de Gibbs ou l'algorithme de Métropolis; on réalise en général un balayage complet de l'image à la température $T^{(n)}$;
- diminution lente de la température : $T^{(n)} > \frac{c}{\log(1+n)}$
- arrêt lorsque le taux de changement est faible.

La décroissance logarithmique de la température est un rythme très lent; en pratique des décroissances géométriques sont utilisées, souvent sans dégradation notable des résultats obtenus. La constante c intervenant dans la décroissance dépend de la variation énergétique globale maximale sur l'espace des configurations.

La figure 6.2 montre l'évolution du paysage énergétique représenté en 1 dimension au fur et à mesure de la décroissance en température. Au départ, toutes les configurations sont équiprobables puis les minima énergétiques apparaissent et s'accentuent.

Notons que contrairement aux algorithmes de l'échantillonneur de Gibbs et de Métropolis qui échantillonnent selon la loi de Gibbs et qui sont en mesure de donner toutes les configurations possibles, les images obtenues par recuit simulé sont uniques et doivent en théorie correspondre aux minima globaux de l'énergie.



Il existe une preuve de convergence de cet algorithme, qui repose à nouveau sur la construction d'une chaîne de Markov, mais qui est inhomogène cette fois-ci à cause de la variation du paramètre de température. On trouvera cette preuve en annexe B. Intuitivement, le recuit simulé permet d'atteindre un optimum global, car il accepte des remontées en énergie. Avec la décroissance de la température, ces sauts énergétiques sont progressivement supprimés au fur et à mesure qu'on se rapproche de l'optimum global. La descente en température doit donc se faire suffisamment lentement pour que l'algorithme ne reste pas “piégé” dans un minimum local de l'énergie.

7 Conclusion

Nous avons abordé dans ce chapitre les notions fondamentales du formalisme markovien. Nous avons vu comment se définissait un champ de Markov et comment l'équivalence champ de Markov - champ de Gibbs permettait de définir sa loi de probabilité en terme d'énergie. Nous avons abordé les algorithmes de synthèse permettant d'échantillonner selon une loi de Gibbs, et le recuit simulé permettant de trouver une configuration d'énergie minimale.

Les annexes A, B, C de ce chapitre étudient les statistiques exponentielles et exposent les preuves théoriques de convergence des algorithmes d'échantillonage et du recuit simulé.

Le chapitre suivant (chap.2) revient sur la recherche d'une solution dans le cadre d'un champ de Markov caché. En effet, le critère du maximum de la probabilité a posteriori que nous avons utilisé pour présenter les applications de restauration et de segmentation n'est pas toujours le plus adapté. D'autres critères correspondant à d'autres fonctions de coût peuvent être utilisés.

Le chapitre 3 aborde ensuite le problème de l'estimation des paramètres des champs markoviens. Par exemple le choix de la valeur β qui définit un modèle de Potts ou qui pondère le terme d'attache aux données et celui de régularisation.

Chapter 2

Estimateurs dans un cadre markovien

1 Introduction

Nous avons vu au chapitre précédent comment il était possible d'utiliser le formalisme markovien à des fins de restauration et de segmentation. On se situe alors dans le cadre de données incomplètes (on parle aussi de champs de Markov cachés) car la réalisation dont on dispose est une réalisation bruitée (ou plus généralement vue à travers le système d'acquisition) du champ de Markov originel. En notant Y le champ dont on observe une réalisation, et X le champ initial, l'objectif est alors d'obtenir la “meilleure” réalisation \hat{x} de X connaissant l’observation y , autrement dit, reconstruire x de manière optimale vis à vis d'un certain critère. Dans le précédent chapitre, nous nous étions intéressés à la réalisation maximisant la probabilité a posteriori $P(X = x / Y = y)$, et nous avions vu un algorithme permettant d'obtenir cette réalisation: le recuit simulé. En réalité d'autres choix sont possibles, auxquels correspondent d'autres méthodes de résolution, et que nous allons aborder dans ce chapitre.

2 Modélisation bayésienne et fonction de coût

Si nous reprenons rapidement le raisonnement effectué au chapitre précédent, on peut écrire, en appliquant la règle de Bayes:

$$P(X = x | Y = y) = \frac{P(Y = y / X = x)P(X = x)}{P(Y = y)}$$

On montre alors que sous certaines hypothèses (indépendance des sites dans la probabilité conditionnelle $P(Y / X)$ et hypothèse markovienne pour le champ X , cf. chap. 1), la distribution a posteriori est une distribution de Gibbs et donc que le champ X conditionnellement à la donnée y est markovien. Cette propriété n'est pas nécessaire pour les notions suivantes, mais elle sera primordiale pour les algorithmes de résolution.

Le problème est alors de déterminer une estimation \hat{x} optimisant un certain critère, où \hat{x}

est une fonction déterministe ϕ de la donnée y :

$$\hat{x} = \phi(y) \text{ avec } \phi : \Omega \rightarrow \Omega$$

L'estimation bayésienne procède alors comme suit. On se donne une fonction de coût, L définie de $\Omega \times \Omega$ dans \mathbb{R}^+ , qui représente le coût de remplacer x par $\phi(y)$, et qui possède les propriétés suivantes :

$$\begin{aligned} \forall x, x' \in \Omega \times \Omega : \\ \bullet L(x, x') &\geq 0 \\ \bullet L(x, x') &= 0 \Leftrightarrow x = x' \end{aligned}$$

L'estimateur optimal, i.e la fonction ϕ optimale est alors la fonction minimisant l'espérance (notée $E[\cdot]$) du coût, c'est à dire :

$$E[L(X, \phi(y)) / Y = y] = \sum_{x \in \Omega} L(x, \phi(y)) P(x / y)$$

La fonction ϕ^{opt} minimise donc l'erreur moyenne conditionnellement à y , et l'estimateur “optimal” est $\hat{x} = \phi^{\text{opt}}(y)$.

Suivant les fonctions de coût envisagées, on obtient différents estimateurs et différentes méthodes de résolution associées. Nous présentons d'abord les estimateurs MAP, MPM et TPM avant d'aborder les solutions algorithmiques dans la section 6.

3 Estimateur MAP

Considérons la fonction de coût suivante :

$$\begin{aligned} L(x, x') &= 1 \text{ si } x \neq x' \\ L(x, x') &= 0 \text{ sinon} \end{aligned}$$

Cette fonction consiste donc à pénaliser toute différence entre deux configurations, et ce, quelque soit le nombre de sites en lesquels elles diffèrent. Nous pouvons alors écrire :

$$\begin{aligned} E[L(X, \phi(y)) / y] &= \sum_{x \in \Omega} L(x, \phi(y)) P(x / y) \\ &= \sum_{x \neq \phi(y)} P(x / y) \\ &= 1 - \sum_{x = \phi(y)} P(x / y) \\ &= 1 - P(X = \phi(y) / y) \end{aligned}$$

Par conséquent, la fonction ϕ^{opt} minimisant l'espérance pour cette fonction de coût, est celle qui maximise la probabilité a posteriori :

$$\hat{x} = \phi^{\text{opt}}(y) = \text{Argmax}_{\phi}[P(X = \phi(y) / y)]$$

Il nous faut donc trouver la réalisation \hat{x} , fonction de y , maximisant la probabilité a posteriori $P(X / y)$. On parle de l'estimateur MAP (maximum a posteriori) ou de maximum de vraisemblance a posteriori.

On retrouve donc avec cette fonction de coût en tout ou rien, la démarche intuitive que nous avions présentée au chapitre 1, à savoir chercher la configuration maximisant la probabilité conditionnellement à la donnée disponible.

4 Estimateur MPM

Considérons maintenant la fonction de coût définie par :

$$L(x, x') = \sum_{s \in S} L(x_s, x'_s) = \sum_{s \in S} \mathbf{1}_{x_s \neq x'_s}$$

La fonction de coût pénalise cette fois-ci proportionnellement au nombre de différences entre deux configurations. Elle paraît donc plus “naturelle” que la fonction de coût en tout ou rien précédente.

Dans le cas d'une fonction de coût définie comme somme de coûts en chaque site, on peut faire le raisonnement suivant:

$$\begin{aligned} E[L(X, \phi(y)) / Y = y] &= \sum_{x \in \Omega} L(x, \phi(y)) P(x / y) \\ &= \sum_{x \in \Omega} \sum_{s \in S} L(x_s, \phi(y)_s) P(x / y) \\ &= \sum_{s \in S} \sum_{x \in \Omega} L(x_s, \phi(y)_s) P(x / y) \\ &= \sum_{s \in S} \sum_{x_s} \sum_{x^s} L(x_s, \phi(y)_s) P(x^s, x_s / y) \\ &= \sum_{s \in S} \sum_{x_s} L(x_s, \phi(y)_s) \sum_{x^s} P(x^s, x_s / y) \end{aligned}$$

Or $\sum_{x^s} P(x^s, x_s / y) = P(X_s = x_s / y)$, donc on peut faire apparaître les probabilités conditionnelles et espérances en chaque site s :

$$\begin{aligned} E[L(X, \phi(y)) / Y = y] &= \sum_{s \in S} \sum_{x_s} L(x_s, \phi(y)_s) P(X_s = x_s / y) \\ &= \sum_{s \in S} E[L(X_s, \phi(y)_s) / y] \end{aligned}$$

On passe donc de la probabilité conditionnelle globale d'une configuration à la probabilité conditionnelle en un site. Il s'agit d'une somme de termes positifs, et par conséquent la fonction ϕ optimale minimise en chaque site l'espérance conditionnelle du coût local $E[L(X_s, \phi(y)_s) / y]$. Ce résultat est valable pour toutes les fonctions de coût définies par une somme de coûts en chaque site.

Dans le cas de la fonction définie ci-dessus, nous pouvons écrire:

$$\begin{aligned} E[L(X_s, \phi(y)_s) / y] &= \sum_{x_s} L(x_s, \phi(y)_s) P(X_s = x_s / y) \\ &= \sum_{x_s \neq \phi(y)_s} P(X_s = x_s / y) \\ &= 1 - P(X_s = \phi(y)_s / y) \end{aligned}$$

Ainsi, la valeur optimale de $\phi(y)$ ou de \hat{x} en chaque site est telle que :

$$\hat{x}_s = \phi^{\text{opt}}(y)_s = \text{Argmax}_{\phi}[P(X_s = \phi(y)_s / y)]$$

i.e on maximise en chaque site la marginale a posteriori $P(X_s / y)$.

On obtient donc des estimateurs du maximum a posteriori locaux, à calculer en chaque pixel contrairement à la recherche précédente qui était globale. L'estimateur est appelé “maximum de vraisemblance a posteriori local” ou maximum posterior marginal (pour maximum a posteriori de la marginale) abrégé en MPM.

Nous verrons dans la section 6 les solutions algorithmiques pour calculer cet estimateur.

5 Estimateur TPM

Considérons maintenant la fonction de coût définie par :

$$L(x, x') = \|x - x'\| = \sum_{s \in S} (x_s - x'_s)^2$$

Il s'agit de l'erreur quadratique et elle pénalise cette fois-ci directement la somme des différences entre les deux configurations. Elle peut donc être plus adaptée dans certains cas que les précédentes, puisqu'elle tient compte non seulement du nombre de différences comme le MPM, mais aussi de leurs valeurs.

Dans ce cas, on a en utilisant le résultat établi précédemment :

$$\begin{aligned} E[L(X, \phi(y)) / Y = y] &= \sum_{x \in \Omega} L(x, \phi(y)) P(x / y) \\ &= \sum_{s \in S} E[(X_s - \phi(y)_s)^2 / y] \end{aligned}$$

On cherche donc ϕ , telle que :

$$E[(X_s - \phi(y)_s)^2 / y] \text{ maximum}$$

Nous allons écrire cette espérance sous une nouvelle forme en utilisant la moyenne conditionnelle au site s , $\bar{x}_s = E[X_s / y] = \sum_{x_s \in E} x_s P(x_s / y)$:

$$\begin{aligned} E[(X_s - \phi(y)_s)^2 / y] &= \sum_{x_s \in E} (x_s - \phi(y)_s)^2 P(x_s / y) \\ &= \sum_{x_s \in E} (x_s - \bar{x}_s + \bar{x}_s - \phi(y)_s)^2 P(x_s / y) \\ &= \sum_{x_s \in E} (x_s - \bar{x}_s)^2 P(x_s / y) + \sum_{x_s \in E} (\bar{x}_s - \phi(y)_s)^2 P(x_s / y) \\ &= K + (\bar{x}_s - \phi(y)_s)^2 \end{aligned}$$

où K est une constante ne dépendant pas de ϕ donc n'intervenant pas dans la minimisation. Par conséquent, le minimum de l'erreur est atteint pour la fonction ϕ telle que :

$$\phi^{\text{opt}}(y)_s = E[X_s / y]$$

Cet estimateur consiste à prendre en chaque site la moyenne conditionnelle locale donnée par la loi a posteriori, d'où le nom de TPM (“Thresholded Posterior Mean”).

La section suivante traite des solutions algorithmiques pour calculer cet estimateur et les estimateurs MAP et MPM que nous avons vu précédemment.

6 Solutions algorithmiques des estimateurs dans le cas markovien

Chacun des estimateurs, MAP, MPM, ou TPM nécessite l'évaluation de quantités différentes.

◊ Pour le MAP, on cherche \hat{x} telle que :

$$\hat{x} = \text{Argmax}_x P(X = x / y)$$

◊ Pour le MPM, on cherche en chaque site \hat{x}_s telle que :

$$\hat{x}_s = \text{Argmax}_{x_s} P(X_s = x_s / y)$$

◊ Pour le TPM, on cherche en chaque site \hat{x}_s telle que :

$$\hat{x}_s = E[X_s / y]$$

Or la taille de l'espace des configurations Ω ne permet pas un calcul direct des quantités $P(x / y)$ et $P(x_s / y)$. Aussi réalise-t-on en pratique des approximations de type Monte-Carlo. En effet, supposons que l'on soit capable de tirer des réalisations de X selon sa loi conditionnelle à y , et notons les $x(1), \dots, x(N)$. Il est alors possible de calculer des approximations des estimateurs

MPM et TPM commençons allons le voir ci-dessous. Le tirage des réalisations ne pose quant à lui pas de problème particulier car nous avons vu au chapitre 1 comment tirer des réalisations d'un champ de Gibbs avec l'échantillonneur de Gibbs et l'algorithme de Métropolis. Or sous les hypothèses rappelées en début de la section 1, la probabilité a posteriori $P(X / y)$ est une distribution de Gibbs.

6.1 Estimateur MPM

Nous supposons ici que nous disposons de N échantillons de X tirés selon la loi a posteriori et nous cherchons ici à estimer la distribution conditionnelle en chaque site $P(X_s = \lambda / y) \forall \lambda \in E$. Nous allons estimer cette quantité par la fréquence empirique de λ au site s dans les échantillons $x(k)$ de X , i.e :

$$\hat{P}(X_s = \lambda / y) = \frac{1}{N} \sum_{k=1}^N \mathbf{1}_{x(k)_s = \lambda}$$

L'estimation au sens du MPM est alors donnée en chaque site en choisissant la valeur de x_s dans Ω maximisant $P(X_s / y)$.

6.2 Estimateur TPM

La démarche est similaire à celle effectuée dans le paragraphe précédent. On approxime l'espérance conditionnelle en chaque site par la moyenne empirique en ce site des N échantillons tirés selon la loi a posteriori :

$$\hat{E}(X_s / y) = \frac{1}{N} \sum_{k=1}^N x(k)_s$$

L'estimation au sens du TPM est alors donnée en chaque site par sa moyenne empirique.

Remarquons que cet estimateur est mal adapté à une problématique de segmentation car la moyenne des étiquettes n'a alors aucun sens.

6.3 Estimateur MAP

Dans ce cas, la connaissance de N échantillons tirés selon la loi a posteriori n'est pas suffisante pour obtenir une réalisation correspondant au minimum global de l'énergie.

Nous avons vu au chapitre 1 comment l'algorithme du recuit simulé permettait d'obtenir la (les) configuration(s) la (les) plus probable(s) d'une distribution de Gibbs, i.e minimisant l'énergie. Il est donc possible d'obtenir l'estimateur MAP par un recuit simulé utilisant la distribution a posteriori avec paramètre de température.

Malheureusement, cet algorithme est très lourd en temps de calcul puisqu'il demande la génération d'un grand nombre de configurations au fur et à mesure que la température décroît.

Des algorithmes sous-optimaux sont donc souvent utilisés en pratique. Besag [Besag(1986)] a ainsi proposé un autre algorithme, beaucoup plus rapide, mais pour lequel nous n'avons pas de preuve de convergence vers un minimum global. Il s'agit de l'ICM, Iterated Conditional Mode, que nous allons présenter ici.

Cet algorithme, comme les échantillonneurs vus au chapitre 1, est un algorithme itératif modifiant à chaque étape les valeurs x_s de l'ensemble des sites de l'image. Mais à la différence de ces algorithmes qui étaient stochastiques par essence, la modification d'une valeur se fait ici de façon déterministe.

On construit donc, partant d'une configuration initiale $x(0)$, une suite d'images $x(n)$, convergeant vers une approximation du MAP \hat{x} recherché. Soit T un tour (visite de tous les sites de l'image), on parlera dans la suite d'itérations à chaque mise à jour d'un site et d'étape à chaque mise à jour de toute l'image (i.e accomplissement d'un tour).

Le déroulement de l'étape n s'effectue de la façon suivante : on parcourt tous les sites et en chaque site, on effectue les deux opérations suivantes :

1. calcul des probabilités conditionnelles locales, pour toutes les valeurs possibles de λ dans E du site:

$$P(X_s = \lambda / y, \hat{x}_r(k), r \in \mathcal{V}_s)$$

(en pratique, calcul plus simplement des énergies conditionnelles locales)

2. mise à jour de la valeur par le λ maximisant la probabilité conditionnelle locale:

$$\hat{x}_s(k+1) = \text{Argmax}_{\lambda} P(X_s = \lambda / y, \hat{x}_r(k), r \in \mathcal{V}_s)$$

(ou de façon équivalente, minimisant l'énergie conditionnelle locale)

Le processus s'arrête lorsque le nombre de changements d'une étape à l'autre devient suffisamment faible.

On peut montrer que l'énergie globale de la configuration \hat{x} diminue à chaque itération. En effet, en appelant s le pixel qui est mis à jour à l'itération k :

$$\begin{aligned} P(X = \hat{x}(k+1) / y) &= P(\hat{x}^s(k+1), \hat{x}_s(k+1) / y) \\ &= P(\hat{x}_s(k+1) / y, \hat{x}^s(k+1))P(\hat{x}^s(k+1) / y) \end{aligned}$$

Or comme seule x_s est modifiée de l'itération k à l'itération $k+1$, on a : $x^s(k+1) = x^s(k)$. Donc :

$$P(X = \hat{x}(k+1) / y) = P(\hat{x}_s(k+1) / y, \hat{x}^s(k))P(\hat{x}^s(k) / y)$$

Et par construction de l'algorithme :

$$P(\hat{x}_s(k+1) / y, \hat{x}^s(k)) \geq P(\hat{x}_s(k) / y, \hat{x}^s(k)) \Rightarrow P(\hat{x}(k+1) / y) \geq P(\hat{x}(k) / y)$$

Cet algorithme, contrairement au recuit simulé, est très rapide (une dizaine de balayages permettent d'arriver à convergence) et peu coûteux en temps de calcul puisqu'il ne nécessite que le calcul des énergies conditionnelles locales. En contrepartie, ses performances dépendent très fortement de l'initialisation (par rapport à la forme du paysage énergétique) puisqu'il converge vers un minimum local. L'ICM s'apparente à une descente en gradient (on fait baisser l'énergie à chaque itération) ou à un recuit simulé gelé à température nulle, et peut donc rester bloquer dans le minimum énergétique local le plus proche de l'initialisation. Le recuit simulé, au contraire, grâce au paramètre de température et aux remontées en énergie qu'il autorise permet d'accéder au minimum global.

Notons qu'il a également été proposé d'utiliser la programmation dynamique pour estimer le MAP [Derin and Elliott(1987)]. Mais il est alors nécessaire d'être dans une configuration simple de segmentation (peu d'étiquettes, dimensions raisonnables) et seule une approximation peut être obtenue.

7 Comparaison des estimateurs MAP, MPM, et TPM

Nous comparons dans cette section les trois estimateurs dans le cadre de la restauration. Dans le cas du MAP, les résultats sont obtenus par ICM et par recuit simulé. L'image à restaurer (figure 2.2.a) est une image bruitée par un bruit blanc gaussien. L'énergie a posteriori utilisée s'écrit :

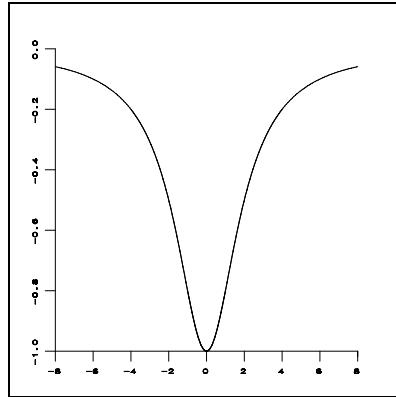
$$\mathcal{U}(x / y) = \sum_s \frac{(y_s - x_s)^2}{\sigma^2} + \sum_{c \in \mathcal{C}} \frac{-\beta}{\phi}(x_r - x_s) \text{ avec } \beta > 0$$

$$\text{et } \phi(x) = \frac{1}{1 + (\frac{x}{\delta})^2}$$

En toute rigueur, il ne faudrait pas définir d'énergie négative (terme $-\beta\phi(x)$), mais les potentiels sont définis à une constante près et cet artifice permet de mettre en évidence l'évolution du terme de régularisation de l'énergie en fonction des écarts des niveaux de gris dans l'image restaurée. La fonction ϕ utilisée a la forme indiquée figure 2.1. Cette fonction est bien sûr minimale lorsque $x = 0$, mais ses valeurs se stabilisent au dessus d'un certain seuil (contrôlé par la valeur de δ). Ceci permet de seuiller les pénalités imposées par le terme de régularisation en présence de contours dans l'image.

Les paramètres utilisés sont fixés aux valeurs suivantes : $\sigma = 28$ (écart-type du bruit), $\delta = 10$ (saut en amplitude à partir duquel on considère qu'il y a un contour), $\beta = 0.5$ pondération de l'influence relative des deux termes). L'initialisation est donnée par l'image à restaurer. Pour tous les estimateurs, on réalise 600 itérations. Dans le cas du recuit simulé, la température initiale est de 6. Les résultats sont montrés sur la figure 2.2.

On constate visuellement que le meilleur résultat est obtenu par le MAP du recuit simulé. Comme l'image originale est une assez bonne initialisation, il n'y a pas de grandes différences

Figure 2.1: Allure de la ϕ fonction avec $\delta = 2$.

entre les algorithmes de recuit simulé et d'ICM pour l'estimateur du MAP. Ce n'est pas vrai dès que l'initialisation s'écarte du résultat à obtenir et les différences avec le recuit simulé peuvent être très importantes. Par ailleurs, on constate que l'estimateur TPM, qui est par définition plus local, donne un résultat plus bruité et moins régularisé. Cette analyse visuelle est confirmée par l'étude statistique qui peut être effectuée sur des zones homogènes de l'image. Le tableau ci-dessous donne les statistiques d'une zone sombre et d'une zone claire de l'image originale et pour les différents résultats de restauration. Les écart-types les plus faibles sont obtenus pour l'estimateur MAP.

Table 2.1: Statistiques sur des zones homogènes pour les différents résultats de restauration

Statistiques sur l'image originale			
zone 1		135.744	27.799
zone 2		90.969	27.118
Estimateur MAP ICM			
zone 1		136.704	11.546
zone 2		91.653	6.919
Estimateur TPM			
zone 1		136.232	11.341
zone 2		91.636	6.947
Estimateur MAP recuit simulé			
zone 1		136.175	10.272
zone 2		91.963	6.118

On notera également que des points isolés de faible ou fort niveau de gris subsistent dans l'image restaurée. Ceci est lié à l'utilisation de la ϕ fonction qui ne "régularise plus" au delà d'un certain seuil contrôlé par la valeur de δ .

En ce qui concerne les temps de calcul, les méthodes se répartissent comme suit : l'algorithme le plus rapide pour converger est sans conteste l'ICM, les algorithmes de recuit simulé et de TPM (ou MPM) étant à peu près équivalents. En effet, plus le nombre d'itérations est grand, meilleure est l'estimation de la moyenne a posteriori.

Les conclusions qui sont données ici ne sont pas nécessairement valables pour une application en segmentation. L'estimateur du TPM peut en effet dans certains cas donner de meilleurs résultats que le MAP. Par ailleurs, l'ICM peut s'avérer très utile lorsqu'on connaît une configuration proche de la configuration optimale.

8 Conclusion

Nous avons abordé dans ce chapitre les différents estimateurs bayésiens (MAP, MPM, TPM) et les algorithmes qui leur sont associés dans un cadre markovien. Il n'existe pas de conclusion universelle sur l'utilisation a priori de l'un ou l'autre de ces algorithmes pour une application précise. Il s'agit le plus souvent pour l'utilisateur de réaliser le meilleur compromis en termes de coût (temps, place mémoire, etc) et de qualité du résultat obtenu.

Le chapitre suivant aborde le problème de l'estimation des paramètres qui interviennent dans les différents termes de l'énergie. De très nombreuses approches statistiques ont été proposées pour résoudre ce problème complexe, notamment dans le cas de données incomplètes, dont on trouvera une synthèse.

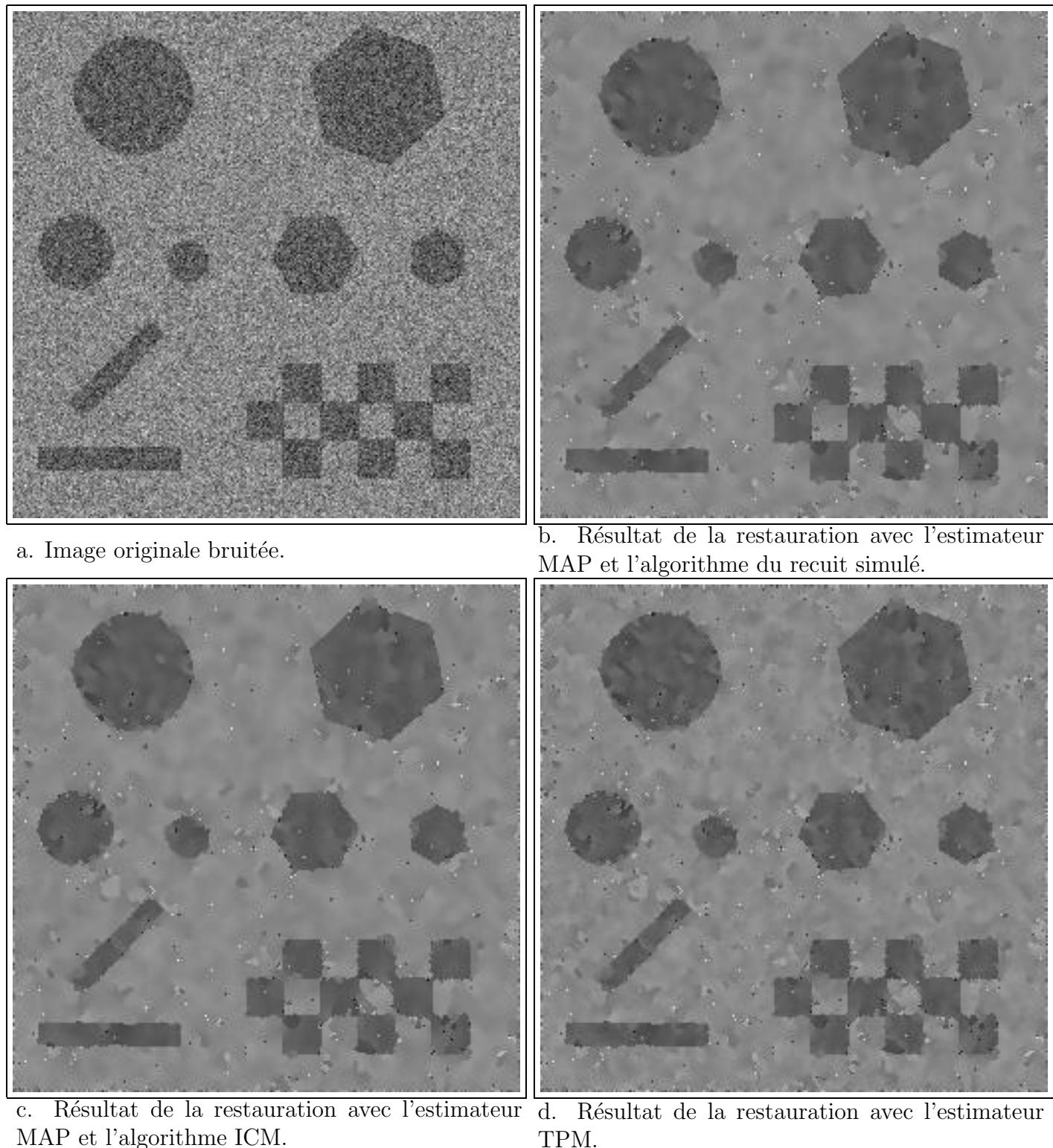


Figure 2.2: Comparaison des algorithmes ICM, Recuit Simulé et TPM en restauration.

Chapter 3

Estimation des paramètres

1 Introduction

Le problème de l'estimation de paramètres (encore appelés hyperparamètres dans la littérature), revient très fréquemment en traitement d'image par champs de Markov. Donnons-en plusieurs exemples :

1. On se donne une réalisation d'un champ de Markov associé au modèle d'Ising, du type de la Fig. 1.2.D, mais on ne connaît pas ses paramètres. Quels sont-ils ?
2. On veut généraliser ceci à une image de texture donnée dont on connaît le modèle sous-jacent (par exemple un modèle gaussien en 4-connexité du type de 1.4.A) mais pas les paramètres, qui sont du type : moyenne locale, “variance locale”, poids de la “régularisation locale”. Quels sont-ils ? Leur connaissance pourrait bien en effet servir à la classification d'images composée de zones texturées en se basant sur l'estimation locale de tels paramètres. On classifierait alors selon les valeurs de ces attributs locaux.
3. On veut segmenter une image, et pour cela apprendre les paramètres de chaque classe, ainsi que le coefficient optimal du modèle de régularisation adapté cette tâche. On sait en effet que le résultat d'une segmentation par estimateur MAP par exemple dépend fondamentalement du poids respectif de la régularisation par rapport à celui de l'attache aux données. Il faut donc là aussi estimer ce poids d'une façon optimale dans un sens à définir.

L'ensemble forme un problème réputé difficile. Nous ne démentirons pas ici la difficulté de ces problèmes, accentuée par le fait que de très nombreuses variantes ont été mises au point dans la littérature pour le résoudre. Certaines tentatives de comparaison de ces approches ont déjà été effectuées [Pieczynski(1994)]. Nous exposerons ici les variantes les plus fréquemment utilisées. Elles se décomposent en deux classes fondamentales:

- le cas des données dites **complètes**, correspondant aux deux premiers problèmes cités plus haut : un échantillon d'une distribution de Gibbs est connu. Il s'agit de "remonter" aux paramètres de cette distribution.
- le cas des données dites **incomplètes**, correspondant au dernier des problèmes cités plus haut Chap. 2 Section 1. Là non seulement le résultat de traitement est inconnu, mais les paramètres sont également à estimer.

2 Données complètes

Notons dans ce qui va suivre x une configuration observée relativement à une distribution de Gibbs donnée P_θ dont l'énergie associée puisse s'écrire sous la forme d'une fonction linéaire d'un paramètre θ , par exemple $U(x) = \theta \Phi(x)$, où Φ est un potentiel donné (voir annexe A). Un principe naturel en vue de la recherche de θ semblerait d'écrire la vraisemblance de la donnée x :

$$L(\theta) = P_\theta(x) = \frac{\exp -\theta \Phi(x)}{Z_\theta}$$

et de chercher par exemple la valeur de l'hyperparamètre $\hat{\theta}$ maximisant cette vraisemblance $L(\theta)$. Le problème essentiel est que l'on ne sait en général pas calculer exactement la fonction de partition Z_θ . Même pour des modèles aussi simples et fondamentaux que ceux d'Ising et de Potts, le résultat (analytique) est obtenu après des calculs excessivement compliqués [Onsager(1944), Landau and Lifschitz(1961)]. Dans les autres cas on sera donc amené :

- soit à effectuer des approximations de la fonction de partition globale au moyen des fonctions de partition conditionnelles locales (codages paragraphe 2.1, pseudo-vraisemblance paragraphe 2.2)
- soit à employer des algorithmes itératifs (gradient stochastique paragraphe 2.3) à partir de la vraisemblance exacte, mais dont il s'agit alors de prouver la convergence ainsi que le type d'optimum trouvé (local, global).

2.1 Méthode des codages

Le principe de la méthode des codages [Besag(1974)] est le suivant. Une fois défini un système de voisinage pour un champ de Markov, nous sommes capables de définir un certain nombre de sous réseaux, chacun formé de sites/pixels indépendants les uns des autres : chacun de ces sous réseaux est appelé un codage. Par exemple avec un voisinage en 4 connecté il existe deux codages comme le montre la figure ci dessous:

2	1	2	1	2
1	2	1	2	1
2	1	2	1	2

Dans le cas de la huit connexité nous aurions pu définir 4 codages différents :

3	4	3	4	3
2	1	2	1	2
3	4	3	4	3
2	1	2	1	2
3	4	3	4	3

Nous allons poser le problème d'estimation dans le cadre de chaque codage. Pour un codage donné les différents sites/pixels le constituant sont **indépendants** les uns des autres puisqu'ils ne sont pas des voisins pour le champ de Markov de départ. La probabilité globale d'un codage se trouvera donc être le produit des probabilités individuelles de chacun des sites du codage. Or du fait de la structure de Markov du champ de départ cette probabilité individuelle se trouve être la probabilité conditionnelle locale du site/pixel dans le champ de Markov. Pour un codage Cod_n donné nous pouvons donc écrire :

$$P_\theta(\{X_s = x_s\}_{s \in \text{Cod}_n} / \{X_r = x_r\}_{r \notin \text{Cod}_n}) = \prod_{s \in \text{Cod}_n} P_\theta(X_s = x_s / V_s) \quad (3.1)$$

Etant donné la structure des probabilités locales telles qu'elles ont été décrites dans les chapitres précédents la fonction de vraisemblance devient calculable, puisque les fonctions de partition conditionnelles locales le sont. Dans le cas où la dépendance des énergies locales est linéaire vis-à vis des paramètres, la **log-vraisemblance** associée,

$$\log P_\theta(\{X_s = x_s\}_{s \in \text{Cod}_n} / \{X_r = x_r\}_{r \notin \text{Cod}_n}) = -\theta \sum_{c \in \mathcal{C}} U_c(x) - \sum_{s \in \text{Cod}_n} \log(Z_s)$$

est une fonction **concave** du (des) paramètre(s), car somme de fonctions concaves en vertu des résultats de l'annexe A. Elle se prête donc bien à la recherche d'un optimum par une méthode classique de type gradient etc. . On peut également montrer qu'il s'agit dans ce cas d'un simple problème de moindres carrés [Derin *et al.*(1985)].

2.2 Pseudo-vraisemblance

Il apparaît en fait expérimentalement que la méthode des codages n'est pas fiable. La méthode du maximum de vraisemblance vrai paraît quant à elle incalculable. Des algorithmes ont cependant été étudiés pour tenter de résoudre ce problème [Younes(1988)], voir paragraphe 2.3. En fait nous allons utiliser une méthode intermédiaire qui elle aura de bonnes propriétés : la méthode du pseudo-maximum de vraisemblance [Graffigne(1987)]. Du maximum de vraisemblance vrai nous allons prendre l'idée de travailler sur l'ensemble de l'image et non séparément sur des réseaux indépendants. De la méthode de codage nous conservons l'idée de manipuler une fonction de vraisemblance produit des probabilités locales de chacun des sites/pixels. Cette fonction sera appelée pseudo-maximum de vraisemblance, et elle s'écrira :

$$PL_\theta(X = x) = \prod_{s \in S} P(X_s = x_s / V_s) \quad (3.2)$$

ou, encore, en considérant le logarithme de cette fonction, et l'expression de la probabilité locale de chaque site/pixel :

$$\log PL_\theta(X = x) = -\theta \sum_{c \in C} U_c(x) - \sum_{s \in S} \log(Z_s) \quad (3.3)$$

Maintenant l'expression $-\log(Z_s)$ devient calculable, puisque reliée à la fonction de normalisation de la probabilité conditionnelle locale telle qu'elle a été décrite au Chap. 1. Donc un raisonnement identique à celui du paragraphe précédent conduit au fait que la log-pseudo-vraisemblance est une fonction concave des paramètres lorsque l'énergie en dépend de façon linéaire. Les algorithmes usuels de type gradient ou gradient conjugué s'appliquent donc aussi ici naturellement à la recherche de l'optimum (qui est unique comme précédemment).

Il s'agit alors de qualifier la valeur des paramètres obtenus par cette méthode par rapport à la valeur vraie. Des résultats théoriques importants ont été obtenus à ce sujet [Graffigne(1987), Guyon(1992)] : la méthode de la pseudo-vraisemblance est **consistante** et **convergente**.

2.3 Algorithme du gradient stochastique

Partons de la vraisemblance **exacte** du paramètre. Elle s'écrit bien sûr :

$$L(\theta) = P_\theta(x) = \frac{\exp -\theta \Phi(x)}{Z_\theta}$$

La valeur de l'hyperparamètre satisfaisant au principe du maximum de vraisemblance

$$\hat{\theta} = \arg \max_{\theta} P_\theta(x)$$

doit donc vérifier l'équation :

$$\left(\frac{\partial \log P_\theta(x)}{\partial \theta} \right)_{\hat{\theta}} = -\Phi(x) - \left(\frac{\partial \log Z_\theta}{\partial \theta} \right)_{\hat{\theta}} = 0 \quad (3.4)$$

D'après les résultats de l' annexe A (ou également annexe B), cette valeur optimale est donc unique et doit satisfaire l'équation suivante

$$\mathbb{E}_\theta[\Phi] = \Phi(x) \quad (3.5)$$

Il s'agit là de ce que l'on appelle une **équation stochastique**. Comme remarqué en introduction de la Section 2 cette équation ne peut être résolue exactement, pour la raison que $\mathbb{E}_\theta[\Phi]$, qui dérive de la fonction de partition Z_θ , ne peut en général être calculé de façon analytique exacte. On est donc amené à employer des algorithmes basés sur un schéma itératif de type Newton-Raphson, mais adapté à ce cadre stochastique. Un schéma rigoureux conduirait à chaque étape (n) à :

$$\theta_{n+1} = \theta_n - \frac{\mathbb{E}_{\theta_n}[\Phi] - \Phi(x)}{\left(\frac{\partial(\mathbb{E}_\theta[\Phi])}{\partial\theta} \right)_{\theta_n}}$$

c'est-à-dire d'après les résultats des annexes A et B, au schéma itératif :

$$\theta_{n+1} = \theta_n + \frac{\mathbb{E}_{\theta_n}[\Phi] - \Phi(x)}{\text{var}_{\theta_n}(\Phi)}$$

L'idée est alors de remplacer les grandeurs statistiques mises en jeu par leurs valeurs empiriques approchées. Ainsi pour l'espérance du potentiel de régularisation Φ , on prendra sa moyenne empirique au cours d'une seule itération (c'est-à-dire la valeur effective obtenue !) d'un échantillonneur de Gibbs (ou de Metropolis) mené avec la valeur courante du paramètre. Quand à la variance de ce potentiel, on l'estime encore plus crûment par une grandeur positive fixée V ! On peut montrer que le prix à payer pour cette approximation est l'introduction d'un terme correctif supplémentaire en $\frac{1}{n+1}$ dans le schéma itératif à l'itération (n). C'est le principe de l'algorithme de gradient stochastique [Younes(1988)] :

$$\begin{cases} \theta_0 & \text{arbitraire} \\ x^{(0)} & \text{tiré au hasard} \end{cases}, \quad \theta_{n+1} = \theta_n + \frac{\Phi(x^{(n)}) - \Phi(x)}{(n+1) V} \quad \text{pour } n \geq 1 \quad (3.6)$$

Il est très important de noter ici que $x^{(n)}$ ($n \geq 1$), échantillon de la distribution P_{θ_n} obtenu par une dynamique de Gibbs (ou de Metropolis) associée à la valeur **courante** du paramètre θ_n , est **généré à partir** de l'échantillon $x^{(n-1)}$ obtenu à l'itération précédente (c'est-à-dire lors de la valeur précédente du paramètre) . On peut alors montrer que cet algorithme stochastique converge presque sûrement, en termes de probabilité, vers la valeur optimale $\hat{\theta}$ lorsque le coefficient V est choisi suffisamment grand.

3 Données incomplètes

Dans cette section, nous abordons le problème de l'estimation des paramètres dans le cas des données incomplètes dites encore manquantes¹. Dans ce cas nous connaissons une observation y , échantillon de la v.a. Y , mais elle est appelée incomplète (c'est-à-dire dégradée), car reliée à une scène originale x , non-dégradée, dont le champ aléatoire correspondant sera noté X . La relation entre y et x s'effectue via une loi de probabilité conditionnelle représentant l'attache aux données (cf. Chap. 2) dont nous explicitons la dépendance p.r. à un paramètre λ positif :

$$\Pr(Y = y / X = x, \Lambda = \lambda) = \frac{\exp - \lambda U(y / x)}{Z_\lambda}$$

Il est très important dans la suite de cette présentation de supposer que la fonction de partition $Z_\lambda = \sum_{y \in \Omega} \exp - \lambda U(y / x)$ est indépendante de x . Ainsi, dans le cas d'un bruit blanc gaussien **additif** en restauration ou en déconvolution et l'approximation discrète finie, on a :

$$\Pr(Y = y / X = x, \Lambda = \lambda) = \frac{\exp - \lambda \|y - x\|^2}{Z_\lambda}$$

où R est la matrice associée à la réponse impulsionnelle de la fonction de flou (l'identité en restauration) et $Z_\lambda = \left(\sqrt{\frac{\pi}{2\lambda}}\right)^{|S|}$.

On suppose que l'on dispose également d'une connaissance *a priori* sur la scène à retrouver x , que ce soit en segmentation ou restauration, via la distribution de Gibbs suivante:

$$P_\theta(x) = \frac{\exp - \theta \Phi(x)}{Z_\theta}$$

Nous commencerons par généraliser, dans un cadre de Maximum de Vraisemblance, la méthode de gradient stochastique vue au paragraphe précédent lorsque la loi d'observation (attache aux données) est complètement connue [Sigelle(1997)], c'est-à-dire que nous nous focaliserons sur l'estimation du meilleur "paramètre de régularisation" θ . Puis nous comparerons cette méthodes à des variantes importantes similaires répertoriées dans la littérature.

Nous aborderons ensuite l'estimation des paramètres d'attache aux données, en particulier dans le cadre de la segmentation d'images. Cela nous permettra de décrire ensuite la seconde grande classe de méthodes adaptée à l'estimation des paramètres : l'EM (Expectation-Maximisation).

3.1 Gradient stochastique généralisé [Younes(1989)]

On va prouver dans cette partie que l'estimation du paramètre de régularisation connaissant la forme du potentiel *a priori* ne peut être dissociée dans la plupart des cas de la forme de

¹On supposera pour simplifier que l'espace des configurations Ω est **fini**. Il faut toutefois savoir que la méthodologie présentée ici se généralise correctement dans le cas où Ω est infini, à condition que les quantités rencontrées (qui sont souvent des fonctions de partition, c'est-à-dire ici des intégrales de fonctions prises sur Ω) existent et soient finies.

l'attache aux données, supposée connue ici par simplicité, **c'est-à-dire que λ est connu**. On suppose accéder par exemple d'une façon ou d'une autre à la variance d'un bruit gaussien en restauration d'image bruitée (et à la réponse impulsionale du flou si l'on est en déconvolution). Dans le cas supposé où aucune information *a priori* sur le paramètre de régularisation n'est disponible, c'est-à-dire lorsque θ suit la distribution uniforme sur \mathbb{R} , la vraisemblance de ce paramètre, qui est la grandeur adéquate à étudier connaissant l'observation incomplète y et le paramètre λ , s'écrit :

$$\begin{aligned} L(\theta) &= \Pr(Y = y, \Lambda = \lambda / \Theta = \theta) \\ &= \sum_{x \in \Omega} \Pr(X = x, Y = y, \Lambda = \lambda / \Theta = \theta) \\ &= \sum_{x \in \Omega} \Pr(Y = y / X = x, \Lambda = \lambda, \Theta = \theta) \cdot \Pr(X = x, \Lambda = \lambda / \Theta = \theta) \end{aligned}$$

En remarquant que l'attache aux données est indépendante de l'hyperparamètre θ (resp. la régularisation est indépendante du paramètre λ), il s'ensuit

$$L(\theta) = \sum_{x \in \Omega} \Pr(Y = y / X = x, \Lambda = \lambda) \Pr(X = x / \Theta = \theta) \quad (3.7)$$

On reconnaît sous le signe somme une expression qui rappelle la probabilité a posteriori d'une configuration x . Explicitons-le en fonction des énergies de Gibbs associées :

$$\begin{aligned} L(\theta) &= \sum_{x \in \Omega} \frac{\exp - \lambda U(y / x)}{Z_\lambda} \cdot \frac{\exp - \theta \Phi(x)}{Z_\theta} \\ &= \frac{1}{Z_\lambda \cdot Z_\theta} \sum_{x \in \Omega} \exp - \lambda U(y / x) - \theta \Phi(x) \end{aligned}$$

Il apparaît donc sous le signe somme la fonction de partition :

$$Z_{\theta, \lambda} = \sum_{x \in \Omega} \exp - \lambda U(y / x) - \theta \Phi(x)$$

associée à la distribution de Gibbs *a posteriori* $P_{\theta, \lambda}(X = x)$ d'énergie

$$\mathcal{U}(x / y) = \lambda U(y / x) + \theta \Phi(x)$$

d'où l'on tire en définitive :

$$L(\theta) = \frac{Z_{\theta, \lambda}}{Z_\lambda \cdot Z_\theta} \quad (3.8)$$

La valeur optimale de l'hyperparamètre de régularisation $\hat{\theta}$ doit donc satisfaire l'équation suivante :

$$\left(\frac{\partial \log L(\theta)}{\partial \theta} \right)_{\hat{\theta}} = \mathbb{E}_{\hat{\theta}}[\Phi] - \mathbb{E}_{\theta, \lambda}[\Phi] = 0 \Rightarrow \mathbb{E}_{\hat{\theta}}[\Phi] = \mathbb{E}_{\theta, \lambda}[\Phi] \quad (3.9)$$

où l'on rappelle que $\mathbb{E}_\theta[\cdot]$ est l'espérance d'une v.a. sous la distribution de Gibbs **a priori** d'énergie $\theta \Phi(x)$, tandis que $\mathbb{E}_{\theta, \lambda}[\cdot]$ signifie l'espérance statistique sous la distribution de Gibbs **a posteriori**, dont la fonction énergie est $\mathcal{U}(x / y) = \lambda U(y / x) + \theta \Phi(x)$. Notons également d'après une remarque effectuée en annexe B, que les deux grandeurs statistiques $\mathbb{E}_{\theta, \lambda}[\Phi]$ et $\mathbb{E}_\theta[\Phi]$ sont des fonctions monotones décroissantes de θ , ce qui implique que *plusieurs* valeurs optimales de l'hyperparamètre peuvent exister (Voir Fig. 3.1).

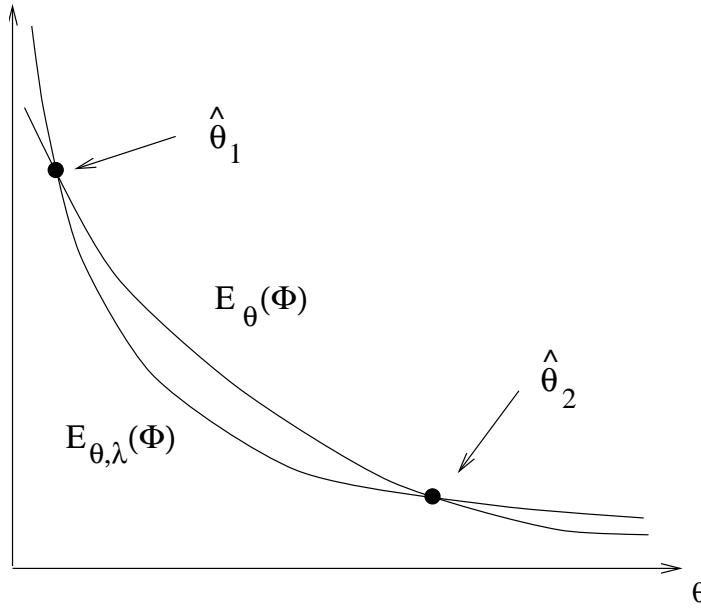


Figure 3.1: Valeur(s) optimale(s) $\hat{\theta}$ du paramètre de régularisation en données incomplètes.

A la limite $\lambda \rightarrow +\infty$, ou comment retrouver le cas des données complètes : il est très important de noter le fait suivant : supposons que nous soyons en restauration par commodité, c'est-à-dire que la matrice $R = \mathbf{1}$, et qu'à la limite $\lambda \rightarrow +\infty$, on ait :

$$\Pr(Y = y / X = x, \Lambda = \lambda) \rightarrow \delta_y(x)$$

Cela est en phase avec les résultats du Chap. 1 et de l'annexe A appliqués à l'attache aux données considérée comme distribution de Gibbs de paramètre λ tendant vers $+\infty$. En effet on peut supposer que l'énergie associée $U(y / x)$ est minimale en $x = y$ et que ce minimum est unique, ce que l'on attend effectivement d'un terme d'attache aux données ² . Cela signifie alors que l'observation y devient *complète*, car elle est la seule configuration de probabilité non nulle. Nous allons retrouver la cohérence des résultats des paragraphes précédents. En effet il vient :

$$\mathbb{E}_{\hat{\theta},\lambda}[\Phi] \rightarrow \frac{\sum_{x \in \Omega} \delta_y(x) \Phi(x) \exp -\hat{\theta} \Phi(x)}{\sum_{x \in \Omega} \delta_y(x) \exp -\hat{\theta} \Phi(x)} = \Phi(y)$$

L'équation du gradient stochastique généralisé devient donc :

$$\mathbb{E}_{\hat{\theta}}[\Phi] = \Phi(y) \tag{3.10}$$

On est donc ramené à l'estimateur du Maximum de Vraisemblance de l'hyperparamètre θ pour le cas de la donnée complète y et de la fonction énergie Φ (eq. 3.5).

²Il faut être prudent dans le cas continu : ainsi une loi gaussienne “tend vers la distribution de Dirac en y ” lorsque sa variance tend vers 0.

Implémentation : si l'on applique un schéma de Newton-Raphson à l'équation stochastique eq. 3.9 on obtient directement :

$$\theta_{n+1} = \theta_n - \frac{\mathbb{E}_{\theta_n}[\Phi] - \mathbb{E}_{\theta_n,\lambda}[\Phi]}{\left(\frac{\partial(\mathbb{E}_{\theta_n}[\Phi] - \mathbb{E}_{\theta_n,\lambda}[\Phi])}{\partial \theta_n} \right)} = \theta_n + \frac{\mathbb{E}_{\theta_n}[\Phi] - \mathbb{E}_{\theta_n,\lambda}[\Phi]}{\text{var}_{\theta_n}(\Phi) - \text{var}_{\theta_n,\lambda}(\Phi)}. \quad (3.11)$$

Il faut d'abord noter que dans ce schéma le dénominateur du dernier terme peut être de signe quelconque contrairement au cas des données complètes eq. 3.6, ce qui est relié à l'existence de plusieurs solutions possibles en données incomplètes. Maintenant, en s'inspirant du raisonnement effectué pour le gradient stochastique pour les données complètes, on s'aperçoit qu'il est nécessaire d'approximer des quantités statistiques (moyenne, variance) reliées aux distributions a posteriori et a priori. On a donc besoin d'échantillonner **deux** champs de Markov en général : celui lié à la régularisation pure (champ a priori) et le champ de Markov postérieur comprenant l'attachement aux données³. Le schéma empirique qui en résulte naturellement est donc le suivant [Younes(1989)] :

$$\begin{cases} \text{un échantillon } x^{(n)} \text{ généré par } P_{\theta_n} \text{ (distribution a priori)} \\ \text{un échantillon } \tilde{x}^{(n)} \text{ généré par } P_{\theta_n,\lambda} \text{ (distribution a posteriori)} \end{cases}$$

avec la procédure d'évolution de l'hyperparamètre :

$$\theta_{n+1} = \theta_n + \frac{1}{n} \cdot \frac{\Phi(x^{(n)}) - \Phi(\tilde{x}^{(n)})}{(< \text{var}_{\theta_n}(\Phi) > - < \text{var}_{\theta_n,\lambda}(\Phi) >)}$$

Il faut noter en particulier la présence des estimateurs empiriques des variances $< \text{var}_{\theta_n}(\Phi) >$ et $< \text{var}_{\theta_n,\lambda}(\Phi) >$ (qui nécessitent donc en fait au moins deux échantillons pour chacune des distributions et à chaque étape du schéma d'évolution de l'hyperparamètre). Ainsi en segmentation, on doit échantillonner aussi bien la distribution a posteriori, ce qui sera par ailleurs utile en vue de la segmentation à convergence des paramètres par l'un des estimateurs MPM ou TPM (Chap. 2), que la distribution a priori qui correspond au modèle de régularisation pur : modèles d'Ising, de Potts, généralisation, et ceci pour la valeur courante du (des) hyperparamètre(s). La convergence de cette méthode est à montrer dans le cadre général des algorithmes stochastiques [Métivier and Priouret(1987)], qui dépasse largement le cadre de ce document de cours ainsi que la compétence de ses auteurs !

³Il faut noter que dans un certain nombre de circonstances usuelles l'espérance et la variance a priori du potentiel Φ peuvent en fait être calculées analytiquement dans l'approximation continue en fonction du paramètre de regularization θ [Bouman and Sauer(1993), Saquib *et al.*(1998)] Le schéma de descente de gradient s'écrit alors comme suit :

$$\theta_{n+1} = \theta_n + \frac{\mathbb{E}_{\theta_n}[\Phi] - \Phi(Y^{(n)})}{\text{var}_{\theta_n}(\Phi) - < \text{var}_{\theta_n,\lambda}(\Phi) >}. \quad (3.12)$$

3.2 Comparaison avec d'autres variantes d'estimation

Dans ce paragraphe on présente d'autres méthodes usuelles d'estimation des hyperparamètres en données incomplètes, qui sont en fait reliées à d'autres estimateurs que le maximum de vraisemblance. Toute la subtilité réside ici dans le choix de la fonction de type vraisemblance à maximiser. On va voir en effet que de très légères différences conduisent à des formes fort différentes d'équations stochastiques, et donc des résultats d'estimation a priori fort différents. On suppose dans les deux premiers cas qu'un **résultat “optimal” de restauration** (ou segmentation) x^* est connu à l'étape d'estimation où l'on se situe.

3.2.1 Loi jointe de l'observation et du résultat connaissant la valeur courante des hyperparamètres [Lakshmanan and Derin(1989), Descombes *et al.*(1996)]

On peut écrire que le paramètre optimal doit satisfaire :

$$\hat{\theta} = \arg \max_{\theta} \Pr_{\theta}(X = x^*, Y = y / \Theta = \theta, \Lambda = \lambda) \quad .$$

En utilisant comme précédemment le fait que l'attache aux données (resp. la régularisation) est indépendante de l'hyperparamètre θ (resp. de λ), on a :

$$\begin{aligned} & \Pr(X = x^*, Y = y / \Theta = \theta, \Lambda = \lambda) \\ &= \Pr(Y = y / X = x^*, \Theta = \theta, \Lambda = \lambda) \Pr(X = x^* / \Theta = \theta, \Lambda = \lambda) \\ &= \Pr(Y = y / X = x^*, \Lambda = \lambda) \Pr(X = x^* / \Theta = \theta) \end{aligned}$$

Et comme le premier terme (attache aux données) ne dépend pas de θ ,

$$\hat{\theta} = \arg \max_{\theta} \Pr_{\theta}(X = x^* / \Theta = \theta, \Lambda = \lambda) = \frac{\exp -\theta \Phi(x^*)}{[Z_{\theta} = \sum_{x \in \Omega} \exp -\theta \Phi(x)]}$$

On “retombe” alors sur l'estimateur au sens du **maximum de vraisemblance** pour le champ de Markov a priori d'énergie $\theta \Phi(x)$ et pour la donnée ici **complète** x^* , c'est-à-dire :

$$\mathbf{E}_{\hat{\theta}}[\Phi] = \Phi(x^*) \quad .$$

On peut donc utiliser une technique de gradient stochastique classique pour estimer ce paramètre (cf. paragraphe 2.3). On peut ensuite itérer en effectuant le traitement désiré (restauration, segmentation) avec les nouvelles valeurs des hyperparamètres obtenues. On obtient donc une nouvelle configuration optimale à partir de laquelle on peut re-estimer les paramètres, et ainsi de suite [Khounri(1997), Zerubia and Blanc-Féraud(1998)]. La méthode est théoriquement convergente vers les valeurs optimales des paramètres et de la configuration.

3.2.2 Probabilité du résultat conditionnellement à l'observation et aux hyperparamètres

Dans ce cas, le paramètre optimal doit vérifier :

$$\hat{\theta} = \arg \max_{\theta} \Pr_{\theta}(X = x^* / Y = y, \Theta = \theta, \Lambda = \lambda) \quad .$$

En appliquant la seconde formule de Bayes et en utilisant le même argument d'indépendence des différentes lois de probabilités vis-à-vis des différents hyperparamètres mis en jeu, on obtient :

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \frac{\Pr(Y = y / X = x^*, \Theta = \theta, \Lambda = \lambda)}{\sum_{x \in \Omega} \Pr(Y = y / X = x, \Theta = \theta, \Lambda = \lambda)} \frac{\Pr(X = x^* / \Theta = \theta, \Lambda = \lambda)}{\Pr(X = x / \Theta = \theta, \Lambda = \lambda)} \\ &= \arg \max_{\theta} \frac{\exp - \lambda U(y / x^*) - \theta \Phi(x^*)}{[Z_{\theta, \lambda} = \sum_{x \in \Omega} \exp - \lambda U(y / x) - \theta \Phi(x)]} \end{aligned}$$

On tombe cette fois sur l'estimateur au sens du **maximum de vraisemblance** pour la distribution de Gibbs a posteriori d'énergie

$$\theta \Phi(x) + \lambda U(y / x)$$

et pour la donnée **complète** x^* , c'est-à-dire :

$$\mathbb{E}_{\hat{\theta}, \lambda}[\Phi] = \Phi(x^*) \quad .$$

Les deux méthodes aboutissent donc à l'estimation d'un paramètre unique, et peuvent donc être implémentées avec une technique de gradient stochastique classique [Younes(1988), Khoumri(1997)]. Voir la comparaison au paragraphe suivant. Nous citerons également l'alternative des méthodes MCMCML (Monte Carlo Markov Chains Maximum Likelihood) [Descombes *et al.*(1996)].

3.2.3 Comparaison avec l'estimateur du gradient stochastique généralisé

Il est très intéressant de comparer les résultats précédents avec le “véritable” gradient stochastique généralisé vu précédemment. On rappelle que celui-ci permet de maximiser la probabilité du paramètre de régularisation conditionnellement à l'**observation uniquement** :

$$\hat{\theta} = \arg \max_{\theta} \Pr_{\theta}(Y = y / \Theta = \theta, \Lambda = \lambda)$$

Voir aussi [Younes(1989)]. En toute rigueur, il est l'estimateur exact au sens du MV du paramètre de régularisation. Les deux autres cités précédemment ont essentiellement l'avantage d'être plus rapides, car ne nécessitant qu'un échantillonnage à chaque étape, et permettant de faire alterner le traitement d'image désiré avec l'estimation des paramètres.

En pratique, les trois estimateurs cités aboutissent à des valeurs du paramètres de régularisation semblables, ce qui provient du fait que les modèles adoptés pour l'attache aux données et pour la régularisation sont souvent choisis en cohérence avec l'observation fournie.

3.3 Estimation des hyperparamètres d'attache aux données : le cas de la segmentation

Supposons que l'on veuille estimer également **les paramètres intervenant dans l'attache aux données**, que nous regrouperons sous une variable λ . Ainsi, dans le cas important de la **segmentation** où le niveau de gris de chacune des m régions d'une image est supposé suivre une loi gaussienne de moyenne μ_i et de variance σ_i^2 , λ est l'ensemble $\{\sigma_i, \mu_i\}_{i=1,m}$. Nous étudierons ici l'estimation de ces paramètres dans le cas **gaussien** par raison de commodité. La vraisemblance en données incomplètes des paramètres de type λ s'écrit alors (cf. eq. 3.7) :

$$\begin{aligned} L(\lambda) &= \Pr(Y = y / \Lambda = \lambda, \Theta = \theta) \\ &= \sum_{x \in \Omega} \Pr(Y = y / X = x, \Lambda = \lambda) \cdot \Pr(X = x / \Theta = \theta) \\ &= \sum_{x \in \Omega} \frac{\exp - \lambda U(y / x)}{Z_\lambda} \cdot \frac{\exp - \theta \Phi(x)}{Z_\theta} \\ &= \frac{1}{Z_\theta} \sum_{x \in \Omega} \exp \{ - \lambda U(y / x) - \log Z_\lambda - \theta \Phi(x) \} \end{aligned}$$

Il nous faut maintenant inclure dans l'énergie a posteriori toute la dépendance vis-à-vis des paramètres d'attache aux données, c'est-à-dire en fait la quantité $\log Z_\lambda$ (calculable ici). En effet, toute décision de segmentation a posteriori doit prendre en compte cette dépendance (ce qui n'est pas le cas en restauration). Définissons pour cela la notation :

$$\begin{aligned} \Psi_{\theta,\lambda}(x) &= \mathcal{U}(x / y, \lambda, \theta) = \lambda U(y / x) + \log Z_\lambda + \theta \Phi(x) \\ &= \sum_{s \in S} \left(\frac{1}{2 \sigma_{x_s}^2} (y_s - \mu_{x_s})^2 + \log \sigma_{x_s} \right) + \theta \Phi(x) \quad \text{dans le cas gaussien} \end{aligned}$$

La fonction de partition a posteriori associée s'écrit :

$$Z_{\theta,\lambda} = \sum_{x \in \Omega} \exp - \mathcal{U}(x / y, \lambda, \theta) = \sum_{x \in \Omega} \exp - \Psi_{\theta,\lambda}(x)$$

et la log-vraisemblance des paramètres d'attache aux données est donc

$$\log L(\lambda) = \log \left\{ \frac{1}{Z_\theta} \sum_{x \in \Omega} \exp - \Psi_{\theta,\lambda}(x) \right\} = \log Z_{\theta,\lambda} - \log Z_\theta$$

Intéressons nous à maximiser cette quantité vis-à-vis d'un des paramètres de λ_i . On obtient

$$\frac{\partial \log L(\lambda)}{\partial \lambda_i} = \frac{\partial \log Z_{\theta,\lambda}}{\partial \lambda_i} = 0$$

Nous allons pour cela généraliser ici une formule de base vue en annexe A et annexe B. Considérons une distribution de Gibbs $P_{\theta,\lambda}(X = x)$ associée à une fonction énergie $\Psi_{\theta,\lambda}(x)$ dépendant de façon **non-linéaire** d'un ensemble de paramètres $\lambda = \{\lambda_i\}$. La fonction de partition associée s'écrit bien sûr :

$$Z_{\theta,\lambda} = \sum_{x \in \Omega} \exp - \Psi_{\theta,\lambda}(x)$$

D'où l'on déduit pour chacun des paramètres λ_i :

$$\frac{\partial \log Z_{\theta,\lambda}}{\partial \lambda_i} = \frac{\sum_{x \in \Omega} -\frac{\partial \Psi_{\theta,\lambda}}{\partial \lambda_i}(x) \exp -\Psi_{\theta,\lambda}(x)}{Z_{\theta,\lambda}} = -\mathbb{E}_{\theta,\lambda}\left[\frac{\partial \Psi_{\theta,\lambda}}{\partial \lambda_i}\right] \quad (3.13)$$

On doit donc avoir

$$\forall \lambda_i \in \lambda, \quad \mathbb{E}_{\theta,\lambda}\left[\frac{\partial \Psi_{\theta,\lambda}}{\partial \lambda_i}\right] = 0 \quad (3.14)$$

Précisons qu'il s'agit bien de l'espérance **a posteriori**. Il nous reste donc à analyser la dépendance précise de l'énergie a posteriori vis-a-vis de tous les paramètres $\{\sigma_i, \mu_i\}_{i=1,m}$. Il est alors commode de re-écrire l'énergie a posteriori en employant les fonctions caractéristiques de chaque classe :

$$\begin{aligned} \Psi_{\theta,\lambda}(x) &= \sum_{s \in S} \left(\frac{1}{2 \sigma_{x_s}^2} (y_s - \mu_{x_s})^2 + \log \sigma_{x_s} \right) + \theta \Phi(x) \\ &= \sum_{i=1}^m \left[\sum_{s \in S} \left(\frac{1}{2 \sigma_i^2} (y_s - \mu_i)^2 + \log \sigma_i \right) \cdot \mathbf{1}_{x_s=i} \right] + \theta \Phi(x) \end{aligned}$$

Il en résulte des formules importantes pour la suite:

$$\left\{ \begin{array}{lcl} \frac{\partial \Psi_{\theta,\lambda}(x)}{\partial \mu_i} &=& \frac{1}{\sigma_i^2} \sum_{s \in S} (\mu_i - y_s) \mathbf{1}_{x_s=i} \\ \frac{\partial \Psi_{\theta,\lambda}(x)}{\partial \sigma_i} &=& \frac{1}{\sigma_i} \sum_{s \in S} \left(\frac{-(y_s - \mu_i)^2}{\sigma_i^2} + 1 \right) \mathbf{1}_{x_s=i} \end{array} \right. \quad (3.15)$$

- Examinons d'abord le cas d'un paramètre de moyenne μ_i donné. On doit avoir :

$$\mathbb{E}_{\theta,\lambda}\left[\frac{\partial \Psi_{\theta,\lambda}}{\partial \mu_i}\right] = \frac{1}{\sigma_i^2} \mathbb{E}_{\theta,\lambda}\left[\sum_{s \in S} (\mu_i - y_s) \mathbf{1}_{x_s=i}\right] = 0$$

En utilisant la formule bien commode

$$\mathbb{E}_a[\mathbf{1}_{X=u} f(X)] = P_a(X=u) f(u) \quad (= \sum_{\xi \in \Omega} \mathbf{1}_{\xi=u} P_a(X=\xi) f(\xi))$$

il en résulte (si σ_i est fini) :

$$\sum_{s \in S} (\mu_i - y_s) P_{\theta,\lambda}(X_s = i) = 0$$

c'est-à-dire

$$\forall i \in [1..m], \quad \mu_i = \frac{\sum_{s \in S} y_s P_{\theta,\lambda}(X_s = i)}{\sum_{s \in S} P_{\theta,\lambda}(X_s = i)} \quad (3.16)$$

L'interprétation physique en est claire : on obtient le barycentre des observations en chaque site de l'image pondérées par la probabilité **a posteriori** d'avoir une classe déterminée.

- De la même façon , on obtient pour les variances de chaque classe :

$$\mathbb{E}_{\theta,\lambda}\left[\frac{\partial \Psi_{\theta,\lambda}}{\partial \sigma_i}\right] = \frac{1}{\sigma_i} \mathbb{E}_{\theta,\lambda}\left[\left(\frac{-(y_s - \mu_i)^2}{\sigma_i^2} + 1\right) \mathbf{1}_{x_s=i}\right] = 0$$

ce qui équivaut, d'après la remarque précédente, à

$$\sum_{s \in S} \left(\frac{-(\mu_i - y_s)^2}{\sigma_i^2} + 1 \right) P_{\theta, \lambda}(X_s = i) = 0$$

c'est-à-dire

$$\forall i \in [1..m] , \quad \sigma_i^2 = \frac{\sum_{s \in S} (y_s - \mu_i)^2 P_{\theta, \lambda}(X_s = i)}{\sum_{s \in S} P_{\theta, \lambda}(X_s = i)} \quad (3.17)$$

les μ_i pouvant être calculés par la formule (3.16). Le résultat est similaire à une variance empirique, mais avec pondération en chaque site par la probabilité a posteriori que ce site ait le label étudié.

En définitive on aboutit donc à des équations 3.16 et 3.17 auto-cohérentes, de forme bien plus compliquée (car sous forme de fraction) que celles des gradients stochastiques simple et généralisé. On aimeraient pouvoir les remplacer par des formes itératives simples : c'est ce qui va justifier l'emploi des méthodes de type EM, à adapter dans le cadre gibbsien défini ici.

Remarque Appliquons la méthode de Laksmanan-Derin (sous-section 3.2.1) à l'attache aux données définie ici, lorsque l'on dispose en plus d'une segmentation courante x^* . Du fait que

$$\Pr(X = x^*, Y = y / \Theta = \theta, \Lambda = \lambda) = \Pr(Y = y / X = x^*, \Lambda = \lambda) \Pr(X = x^* / \Theta = \theta)$$

on obtient :

$$\hat{\lambda} = \arg \max_{\lambda} \Pr(Y = y / X = x^*, \Lambda = \lambda)$$

c'est-à-dire puisque la composante de régularisation $\theta \Phi$ ne dépend pas de λ

$$\frac{\partial \log \Pr(Y = y / X = x^*, \Lambda = \lambda)}{\partial \lambda} = \frac{\partial \Psi_{\theta, \lambda}(x^*)}{\partial \lambda} = 0$$

donc immédiatement

- pour les paramètres de moyennes,

$$\frac{1}{\sigma_i^2} \sum_{s \in S} (\mu_i - y_s) \mathbf{1}_{x_s^* = i} = 0 \Rightarrow \forall i \in [1..m] , \quad \mu_i = \frac{\sum_{s \in S} y_s \mathbf{1}_{x_s^* = i}}{\sum_{s \in S} \mathbf{1}_{x_s^* = i}}$$

- pour les paramètres de variance,

$$\frac{1}{\sigma_i^2} \sum_{s \in S} \left(\frac{-(\mu_i - y_s)^2}{\sigma_i^2} + 1 \right) \mathbf{1}_{x_s^* = i} \Rightarrow \forall i \in [1..m] , \quad \sigma_i^2 = \frac{\sum_{s \in S} (y_s - \mu_i)^2 \mathbf{1}_{x_s^* = i}}{\sum_{s \in S} \mathbf{1}_{x_s^* = i}}$$

En définitive, ce sont les moyennes et variances empiriques calculées sur la segmentation courante que l'on retrouve, ce qui confirme la cohérence de cette méthode particulière.

3.4 Expectation-maximization (EM)

Récapitulons les résultats précédents obtenus concernant l'estimation au Maximum de Vraisemblance des hyperparamètres dans le cas des données incomplètes :

- pour l'hyperparamètre de régularisation nous sommes arrivés à l'équation stochastique :

$$\mathbb{E}_{\hat{\theta}}[\Phi] = \mathbb{E}_{\hat{\theta}, \lambda}[\Phi]$$

On peut remarquer qu'à la limite $\lambda \rightarrow +\infty$ (données complètes)

$$\hat{\theta} = \arg \max_{\theta} \log P_{\theta}(Y = y) = \arg \max_{\theta} \mathbb{E}_{\theta, \lambda \rightarrow +\infty} [\log P_{\theta}(Y = y)]$$

que l'on peut aussi écrire :

$$\hat{\theta} = \arg \max_{\theta} \mathbb{E}_{\hat{\theta}, \lambda \rightarrow +\infty} [\log \Pr(Y = y, X = x / \Theta = \theta, \Lambda = \lambda)]$$

où les espérances sont calculées selon la loi **a posteriori**. En effet, à la limite indiquée, les lois jointe et a posteriori se confondent avec la loi d'observation, c'est-à-dire la mesure de masse 1 placée en la configuration observée y , car pour toute valeur finie de θ , on a :

$$\Pr(Y = y, X = \xi / \Theta = \theta, \Lambda = \lambda) \rightarrow \frac{\delta_y(\xi) \exp -\theta \Phi(\xi)}{\sum_{\eta \in \Omega} \delta_y(\eta) \exp -\theta \Phi(\eta)} = \delta_y(\xi)$$

ce qui implique $\mathbb{E}_{\theta, \lambda}[f] \rightarrow f(y)$ pour toute fonction $f : \Omega \mapsto \mathbb{R}$ (prendre $f(\xi) = \log P_{\theta}(X = \xi)$). On aimerait généraliser au cas incomplet “véritable”, c'est-à-dire avec données manquantes.

- pour les paramètres d'attache aux données en segmentation avec m classes nous sommes partis de l'ensemble d'équations (3.14)

$$\mathbb{E}_{\theta, \lambda} \left[\frac{\partial \Psi_{\theta, \lambda}}{\partial \lambda_i} \right] = 0$$

qui peut, étant donné la forme adoptée pour l'énergie a posteriori, se mettre sous la forme

$$\mathbb{E}_{\theta, \lambda} \left[\frac{\partial \log \Pr(Y = y / X = x, \Lambda = \lambda)}{\partial \lambda_i} \right] = 0$$

Cela nous a permis d'arriver au système d'équations :

$$\forall i \in [1..m] , \quad \begin{cases} \mu_i = \frac{\sum_{s \in S} y_s P_{\theta, \lambda}(X_s = i)}{\sum_{s \in S} P_{\theta, \lambda}(X_s = i)} \\ \sigma_i^2 = \frac{\sum_{s \in S} (y_s - \mu_i)^2 P_{\theta, \lambda}(X_s = i)}{\sum_{s \in S} P_{\theta, \lambda}(X_s = i)} \end{cases}$$

Nous en avons conclu qu'il serait désirable de résoudre ces équations de manière itérative. Par exemple si les probabilités a posteriori $P_{\theta, \lambda}(\cdot)$ étaient connues (ou apprises) à une étape

donnée on pourrait les injecter dans la première équation à condition de pouvoir estimer pour toute valeur de θ l'espérance a priori du potentiel de régularisation ainsi que dans le deuxième système.

Ces deux arguments vont nous fournir les principes de l'EM, avec un certain nombre de résultats théoriques très puissants à la clé [Baum *et al.*(1970), Redner and Walker(1984)].

3.4.1 Le principe EM sous forme exacte

De façon à essayer d'optimiser “séparation” les hyperparamètres d'attache aux données et celui de régularisation, il s'impose de considérer la probabilité jointe de l'observation et d'une donnée incomplète (un étiquetage de l'image par exemple). En effet, cette loi est le produit de la loi d'observation par la probabilité a priori. Le principe de l'EM s'écrit alors sous la forme suivante :

$$\max_{\theta, \lambda} \mathbb{E}_{\theta, \lambda} [\log \Pr(X = x, Y = y / \Lambda = \lambda, \Theta = \theta)] \quad (3.18)$$

où nous précisons bien que l'espérance est bien prise **selon la loi a posteriori**.

La recherche associée des paramètres optimaux exacts est manifestement insoluble en pratique. C'est pourquoi on lui substitue donc une forme itérative plus simple à résoudre, et pour laquelle des résultats théoriques existent.

3.4.2 Le principe EM sous forme itérative

Supposons que les paramètres θ_n et λ_n soient connus à une étape n . On définit la notation :

$$Q(\theta, \lambda, \theta_n, \lambda_n) = \mathbb{E}_{\theta_n, \lambda_n} [\log \Pr(X = x, Y = y / \Lambda = \lambda, \Theta = \theta)]$$

et on recherche alors :

$$\begin{aligned} (\theta_{n+1}, \lambda_{n+1}) &= \arg \max_{\theta, \lambda} Q(\theta, \lambda, \theta_n, \lambda_n) \\ &= \arg \max_{\theta, \lambda} \mathbb{E}_{\theta_n, \lambda_n} [\log \Pr(X = x, Y = y / \Lambda = \lambda, \Theta = \theta)] \end{aligned}$$

Du fait de la “séparabilité” de la loi jointe, la fonction objectif $Q(\theta, \lambda, \theta_n, \lambda_n)$ s'écrit :

$$Q(\theta, \lambda, \theta_n, \lambda_n) = \mathbb{E}_{\theta_n, \lambda_n} [\log \Pr(Y = y / X = x, \Lambda = \lambda) + \log \Pr(X = x / \Theta = \theta)]$$

Cela correspond donc à l'optimisation séparée (partie optimisation de l'EM) :

$$\left\{ \begin{array}{l} \theta_{n+1} = \arg \max_{\theta} \mathbb{E}_{\theta_n, \lambda_n} [\log \Pr(X = x / \Theta = \theta)] \\ \lambda_{n+1} = \arg \max_{\lambda} \mathbb{E}_{\theta_n, \lambda_n} [\log \Pr(Y = y / X = x, \Lambda = \lambda)] \end{array} \right.$$

On peut montrer théoriquement que la fonction objectif $Q(\theta_{n+1}, \lambda_{n+1}, \theta_n, \lambda_n)$ croît en fonction de l'itération n [Baum *et al.*(1970), Redner and Walker(1984)], ce qui aboutit à terme à un optimum local des valeurs des paramètres.

Examinons plus précisément ce qui se passe pour chacune des catégories de paramètres :

- en ce qui concerne le paramètre de régularisation, cela correspondra à :

$$\mathbb{E}_{\theta_n, \lambda_n} \left[\frac{\partial \log \Pr(X = x / \Theta = \theta)}{\partial \theta} \right] = \mathbb{E}_{\theta_n, \lambda_n} \left[-\Phi - \frac{\partial \log Z_\theta}{\partial \theta} \right] = 0$$

c'est-à-dire en vertu de résultats supposés maintenant acquis :

$$\mathbb{E}_{\theta_{n+1}}[\Phi] = \mathbb{E}_{\theta_n, \lambda_n}[\Phi] \quad (3.19)$$

Supposons que l'on sache d'une manière ou d'une autre calculer ou estimer l'espérance a posteriori du potentiel de régularisation pour la valeur courante des paramètres. Ainsi, en pratique, on approxime cette quantité par sa moyenne empirique au cours d'un échantillonneur de Gibbs (ou de Metropolis) pris pour la valeur courante des paramètres⁴. On est alors ramené à un problème d'estimation du paramètre de régularisation pour la donnée complète $\mathbb{E}_{\theta_n, \lambda_n}[\Phi]$! On peut donc appliquer la technique du pseudo-maximum de vraisemblance [Chalmond(1989)] ou bien encore celle du gradient stochastique [Younes(1991)] pour re-estimer la nouvelle valeur du paramètre de régularisation.

On peut aussi remarquer que la valeur $\hat{\theta}$ trouvée à la convergence doit vérifier :

$$\mathbb{E}_\theta[\Phi] = \mathbb{E}_{\hat{\theta}, \hat{\lambda}}[\Phi]$$

c'est-à-dire le principe du Maximum de Vraisemblance en données incomplètes avec comme valeur du paramètre d'attache aux données sa valeur optimale $\hat{\lambda}$! On observe donc bien ici une cohérence entre les deux approches fondamentales de l'estimation en données incomplètes décrites dans ce chapitre (MV, EM).

- pour les paramètres d'attache aux données, on a vu plus haut que

$$\frac{\partial \log \Pr(Y = y / X = x, \Lambda = \lambda)}{\partial \lambda_i} = \frac{\partial \Psi_{\theta, \lambda}(x)}{\partial \lambda_i}$$

Le principe EM s'écrit donc

- pour les paramètres de moyenne :

$$\begin{aligned} \mathbb{E}_{\theta_n, \lambda_n} \left[\frac{\partial \Psi_{\theta, \lambda}}{\partial \mu_i} \right] &= \frac{1}{\sigma_i^2} \mathbb{E}_{\theta_n, \lambda_n} \left[\sum_{s \in S} (\mu_i - y_s) \mathbf{1}_{x_s=i} \right] \\ &= \frac{1}{\sigma_i^2} \sum_{s \in S} (\mu_i - y_s) P_{\theta_n, \lambda_n}(X_s = i) = 0 \\ \Rightarrow \forall i \in [1..m], \quad \mu_i(n+1) &= \frac{\sum_{s \in S} y_s P_{\theta_n, \lambda_n}(X_s = i)}{\sum_{s \in S} P_{\theta_n, \lambda_n}(X_s = i)} \end{aligned} \quad (3.20)$$

⁴Ceci correspond à l'étape Estimation de l'algorithme EM.

- pour les paramètres de variances :

$$\begin{aligned}
 \mathbb{E}_{\theta_n, \lambda_n} \left[\frac{\partial \Psi_{\theta, \lambda}}{\partial \sigma_i} \right] &= \frac{1}{\sigma_i} \mathbb{E}_{\theta_n, \lambda_n} \left[\left(\frac{-(y_s - \mu_i)^2}{\sigma_i^2} + 1 \right) \mathbf{1}_{x_s=i} \right] \\
 &= \sum_{s \in S} \left(\frac{-(y_s - \mu_i)^2}{\sigma_i^2} + 1 \right) P_{\theta_n, \lambda_n}(X_s = i) = 0 \\
 \Rightarrow \forall i \in [1..m] \ , \ \sigma_i(n+1)^2 &= \frac{\sum_{s \in S} (y_s - \mu_i(n+1))^2 P_{\theta_n, \lambda_n}(X_s = i)}{\sum_{s \in S} P_{\theta_n, \lambda_n}(X_s = i)} \tag{3.21}
 \end{aligned}$$

Deux remarques insistantes :

1. Ce sont bien les distributions **a posteriori** qui sont mises en jeu dans l'estimation itérative (eqs. 3.20 et 3.21). Elles sont donc à re-estimer à chaque itération n . Un cas extrêmement important (qui sort du propos de ce document) est celui des chaînes de Markov cachées (HMM). Il s'agit de modèles mono-dimensionnels très utilisés en Traitement de la Parole et Traitement de l'Écriture, donc de chaînes de Markov, pour lesquelles on sait calculer explicitement toutes ces probabilités a posteriori [Rabiner(1989)]. Il nous faut par contre en traitement d'image estimer d'une façon ou d'une autre les probabilités a posteriori apparaissant dans les expressions (3.20) et (3.21), et qui sont également nécessaires pour estimer l'espérance a posteriori du potentiel de régularisation dans (3.19). Une manière bien naturelle est de remplacer ces probabilités a l'étape courante du processus EM par la fréquence empirique d'apparition des labels lors d'une série d'**échantillonnages de la distribution de Gibbs a posteriori courante** menés à l'aide d'un échantillonneur de type Gibbs ou Metropolis [Chalmond(1989)]. Notons pour cela N le nombre d'itérations effectué avec l'échantillonneur ainsi sélectionné :

$$\bullet \quad - - - - - \quad \bullet \\ (n) \qquad \qquad N \qquad \qquad (n+1)$$

Notons également $N_s(i)$ le nombre de fois que le label i a été tiré en un site s au cours de l'ensemble de ces N échantillonnages. On peut donc écrire : $P_{\theta_n, \lambda_n}(X_s = i) \approx \frac{N_s(i)}{N}$, et il en résulte les estimations empiriques des paramètres de moyennes et de variance de chaque classe :

$$\forall i \in [1..m] \ , \ \left\{ \begin{array}{lcl} \mu_i(n+1) & = & \frac{\sum_{s \in S} y_s N_s(i)}{\sum_{s \in S} N_s(i)} \\ \sigma_i(n+1)^2 & = & \frac{\sum_{s \in S} (y_s - \mu_i(n+1))^2 N_s(i)}{\sum_{s \in S} N_s(i)} \end{array} \right.$$

De la même façon, notons $x^{(n)}(k)$ la série d'images échantillons ainsi obtenue pour $k = 1..N$. On obtient pour l'espérance a posteriori du potentiel de régularisation

$$\mathbb{E}_{\theta_n, \lambda_n} [\Phi] \approx \frac{1}{N} \left[\sum_{k=1}^N \Phi(x^{(n)}(k)) \right]$$

C'est donc ensuite que l'on procéde à l'étape d'estimation EM comme indiqué plus haut, à partir de ces valeurs empiriques a posteriori.

2. On retrouve le caractère itératif que nous voulions anticiper à propos des équations plus haut. Là aussi, à convergence, on doit satisfaire à la forme exacte des équations (3.16) et (3.17) obtenues par le Maximum de Vraisemblance.

3.4.3 Variantes

Une variante importante, appelée ICE ou Iterative Conditional Estimation est la suivante [Pieczynski(1994)] :

considérons les N échantillons $x^{(n)}(k)$ obtenus précédemment à une étape courante n de l'EM. On pourrait songer à estimer moyenne et variance des classes pour chacun d'eux considéré comme segmentation courante, puis prendre la moyenne empirique (arithmétique) des valeurs ainsi obtenues

$$\forall i \in [1..m] , \quad \begin{cases} \mu_i(n+1) &= \frac{1}{N} \sum_{k=1}^N \left(\frac{\sum_{s \in S} y_s \mathbf{1}_{x_s^{(n)}(k)=i}}{\sum_{s \in S} \mathbf{1}_{x_s^{(n)}(k)=i}} \right) \\ \sigma_i(n+1)^2 &= \frac{1}{N} \sum_{k=1}^N \left(\frac{\sum_{s \in S} (y_s - \mu_i(n+1))^2 \mathbf{1}_{x_s^{(n)}(k)=i}}{\sum_{s \in S} \mathbf{1}_{x_s^{(n)}(k)=i}} \right) \end{cases}$$

Cette estimation est différente de l'EM. On peut montrer qu'elle correspond en fait à l'estimateur de la moyenne a posteriori des paramètres.

3.5 Conclusion pour l'estimation en données incomplètes

On voit donc au terme de cette partie que l'estimation en données incomplètes se prête à un nombre très riche de variantes. On pourrait ainsi parfaitement “estimer” préférable (!) d'estimer certains types de paramètres comme moyenne et variance par l'EM, tandis que l'on adopterait la méthode de Lakhshmanan-Derin (paragraphe 3.2.1) pour le paramètre de régularisation, ou réciproquement ! Ces variantes méritent encore d'être examinées et comparées entre elles de façon exhaustive, dans la lignée de l'approche suivie dans [Pieczynski(1994)].

Il faut aussi préciser que les méthodes utilisant un échantillonnage de la distribution a posteriori (comme l'EM) permettent en même temps de fournir une série de configurations échantillons se prêtant favorablement à la segmentation ou à la restauration dans le cadre des estimateurs de type TPM ou MPM (Chap 2), lorsque la convergence des paramètres vers leur valeur optimale est supposée atteinte.

Si ce travail est susceptible d'intéresser le lecteur, nous aurons alors rempli notre but !

Chapter 4

Quelques applications des champs markoviens

Ce chapitre présente quelques applications des champs markoviens en traitement d'images pour illustrer les potentialités de ce domaine. On s'intéressera dans un premier temps à des applications de bas niveau (analyse de textures et segmentation) avant de présenter des applications manipulant des graphes construits à partir de primitives plus complexes (régions, segments).

1 Applications sur le graphe des pixels

Cette première partie présente trois applications de bas-niveau, i.e qui travaillent sur le graphe des pixels :

- La première est une analyse de textures s'appuyant sur un modèle markovien ; elle permet de discriminer différentes textures de l'image à l'aide des paramètres du champ extraits; un exemple d'applications pour la détection des zones urbaines en imagerie satellitaire est donné;
- La seconde est un simple exemple de segmentation appliquée en imagerie radar mettant en évidence la souplesse de ce modèle pour prendre en compte des statistiques très variées dans les images et détaillant quelques aspects pratiques de segmentation; on se placera dans un cadre supervisé pour le paramètre de régularisation, c'est à dire que celui-ci sera fixé de façon empirique;
- La troisième application présentée est un schéma de fusion markovien, relativement général qui permet de combiner plusieurs sources d'informations ; un exemple est donné dans le cas de l'analyse d'images satellitaires NOAA dans plusieurs bandes spectrales.

1.1 Analyse de textures

Nous nous intéressons dans cette partie à l'utilisation des modèles markoviens pour discriminer différents types de textures dans les images. L'idée est de se placer dans le cas de données complètes, i.e de faire l'hypothèse que l'image dont nous disposons est la réalisation d'un champ markovien, et d'extraire les paramètres de ce champ. La variabilité des paramètres en fonction du type de textures devant alors permettre de réaliser une classification de l'image.

Nous avons déjà vu au paragraphe 2 du chapitre 3 un certain nombre de techniques (méthode des codages, maximum de pseudo-vraisemblance, gradient stochastique) pour calculer les paramètres d'un champ markovien. Toutes ces méthodes, qui ne presupposent pas de modèle pour le champ, sont assez lourdes à mettre en œuvre. Nous allons supposer ici que nous nous situons dans le cadre d'un champ markovien gaussien pour lequel nous pouvons obtenir une expression exacte des paramètres.

Dans le cas d'un champ markovien gaussien, nous pouvons écrire l'énergie $U(x)$ sous la forme :

$$U = \frac{1}{T} \left(\lambda \sum_{s \in S} (x_s - \mu)^2 + \sum_{c=(s,t) \in C} (x_s - x_t)^2 \right)$$

Avec cette écriture, les paramètres caractérisant le champ et que nous cherchons à estimer sont donc la température T (correspondant à une sorte de "variance généralisée"), la moyenne μ et la pondération du terme d'attache aux données de l'énergie, λ .

1.1.1 La méthode des queues de comètes

La méthode que nous allons voir a été proposée dans [Descombes(1993)]. Considérons la probabilité conditionnelle en un site :

$$P(X_s = x_s / x_t, t \in \mathcal{V}_s) = \frac{1}{Z(V_s)} \exp\left\{-\frac{1}{T} \left(\lambda(x_s - \mu)^2 + \sum_{t \in \mathcal{V}_s} (x_s - x_t)^2 \right)\right\}$$

En notant n le nombre de sites voisins de s , et m_s la moyenne locale des niveaux de gris des voisins du site s , $m_s = \frac{\sum_{t \in \mathcal{V}_s} x_t}{n}$, on peut montrer qu'on a :

$$\begin{aligned} P(X_s = x_s / x_t, t \in \mathcal{V}_s) &= \frac{1}{Z(m_s)} \exp\left\{-\frac{n+\lambda}{T} \left(x_s - \frac{1}{n+\lambda} (nm_s + \lambda\mu) \right)^2\right\} \\ &= P(X_s = x_s / m_s) \end{aligned}$$

On a donc remplacé le conditionnement local par l'ensemble des voisins de s , par une seule variable conditionante m_s , ce qui améliorera la robustesse des estimateurs. En effet, le nombre de sites concernés par un conditionnement par m_s sera bien plus grand que celui des sites concernés par une configuration $V_s = (x_t, t \in \mathcal{V}_s)$ du voisinage.

Nous constatons que $P(X_s / m_s)$ est une loi gaussienne, de moments (moyenne et variance) suivants :

$$\begin{aligned}\mathbb{E}(X_s / m_s) &= \frac{nm_s + \lambda\mu}{n + \lambda} \\ \text{var}(X_s / m_s) &= \frac{T}{2(n + \lambda)}\end{aligned}$$

L'espérance et la variance de X_s conditionnellement à m_s étant fonction des paramètres λ , T et μ que nous recherchons, il nous suffit d'estimer ces moments de façon empirique. La variance ne dépendant pas de la valeur conditionnante m_s , on notera le moment théorique par σ^2 . Pour faire ces estimations, on construit l'image dite des “queues de comètes” en raison de son aspect, qui est simplement l'image des fréquences normalisées des niveaux de gris x_s conditionnellement à la moyenne m_s du voisinage. Si on met les valeurs de m_s en colonne et celles de x_s en ligne, chaque ligne de l'image des queues de comètes représente $P(X_s / m_s)$, i.e une gaussienne de moyenne $\frac{nm_s + \lambda\mu}{n + \lambda}$ et de variance $\frac{T}{2(n + \lambda)}$ (figure 4.1).

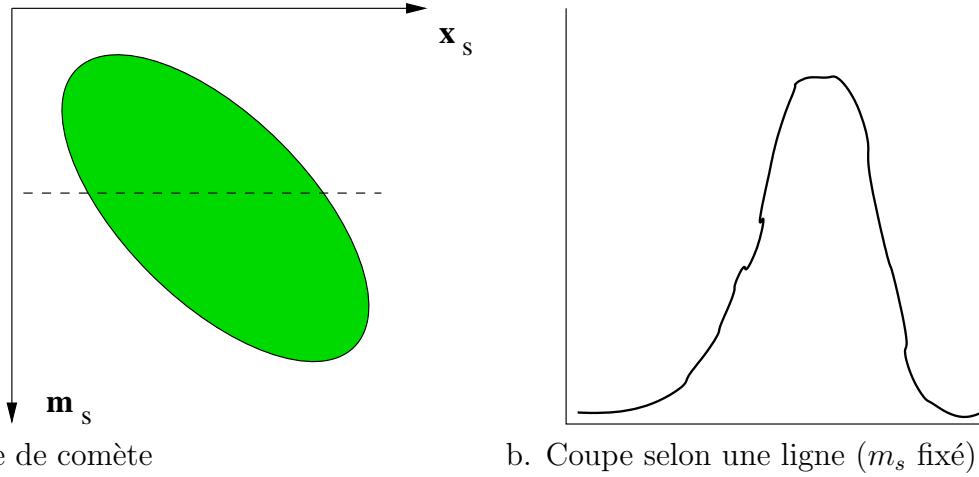


Figure 4.1: Queue de comète d'un modèle gaussien 4-connexe

On considère alors pour la variance, l'estimateur suivant :

$$\widehat{\sigma}^2 = \sum_{m_s} P(m_s) \widehat{\text{var}}(X_s / m_s)$$

en notant $\widehat{\text{var}}(X_s / m_s)$ l'estimation empirique de la variance faite selon une ligne des queues de comètes. Cet estimateur permet d'accorder à chaque probabilité conditionnelle locale une importance proportionnelle au nombre d'échantillons qui la constituent et entraîne une plus grande robustesse de l'estimation.

En ce qui concerne l'espérance, nous avons la relation suivante :

$$\begin{aligned}\mathbb{E}(X_s / m_s) &= \frac{n}{n + \lambda} m_s + \frac{\mu}{n + \lambda} \\ &= \alpha m_s + \beta\end{aligned}$$

Par conséquent, les espérances des probabilités conditionnelles se situent sur une droite. L'estimation empirique des moyennes $\hat{\mathbb{E}}(X_s / m_s)$ permet de faire une estimation aux moindres carrés des paramètres α et β de la droite. Une fois estimés empiriquement $\hat{\sigma}^2$, $\hat{\alpha}$, $\hat{\beta}$, on peut déduire λ , μ , et T par les relations :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{T}}{2(n_s + \hat{\lambda})} \\ \hat{\alpha} &= \frac{n_s}{n_s + \lambda} \\ \hat{\beta} &= \frac{\hat{\lambda}\hat{\mu}}{n_s + \hat{\lambda}}\end{aligned}$$

En explicitant l'estimation aux moindres carrés des paramètres de la droite, on peut exprimer directement λ , μ , et T en fonction des moments d'ordre 1 et 2, conditionnés ou non. On retrouve dans ce cas pour μ , la moyenne empirique des x_s .

1.1.2 Application à la détection des zones urbaines

Cet exemple d'application est tiré de la thèse de X. Descombes [Descombes(1993)].

On peut utiliser les résultats précédents pour analyser les textures présentes sur une image satellitaire SPOT. L'image étant par essence non stationnaire (sinon l'analyse aurait peu d'intérêt!), le calcul des paramètres se fait localement, sur une fenêtre glissante centrée en chaque pixel. La fiabilité des estimateurs s'accroît avec la taille de la fenêtre, en même temps, et de façon antagoniste, que le risque de considérer des mélanges de textures différentes à l'intérieur de la fenêtre d'étude. Une solution pour remédier à ce problème peut être de ne considérer que les échantillons les plus représentés. En cas de mélange, les échantillons appartiendront à la texture la plus présente dans la fenêtre.

Le paramètre de température, est un bon indicateur du milieu urbain qui se présente sur une image SPOT sous une forme assez texturée, type “poivre et sel” avec alternance de niveaux de gris faibles et forts. En effet, ce paramètre qui mesure en quelque sorte le chahut de la zone, a des valeurs plus élevées dans les régions urbaines. Il permet d'obtenir une bonne discrimination du milieu urbain comme indiqué sur les figures ci-dessous (fig. 4.2, 4.3, 4.4).



a. Extrait 1



b. Extrait 2

Figure 4.2: Extraits de l'image SPOT panchromatique originale à 10m.

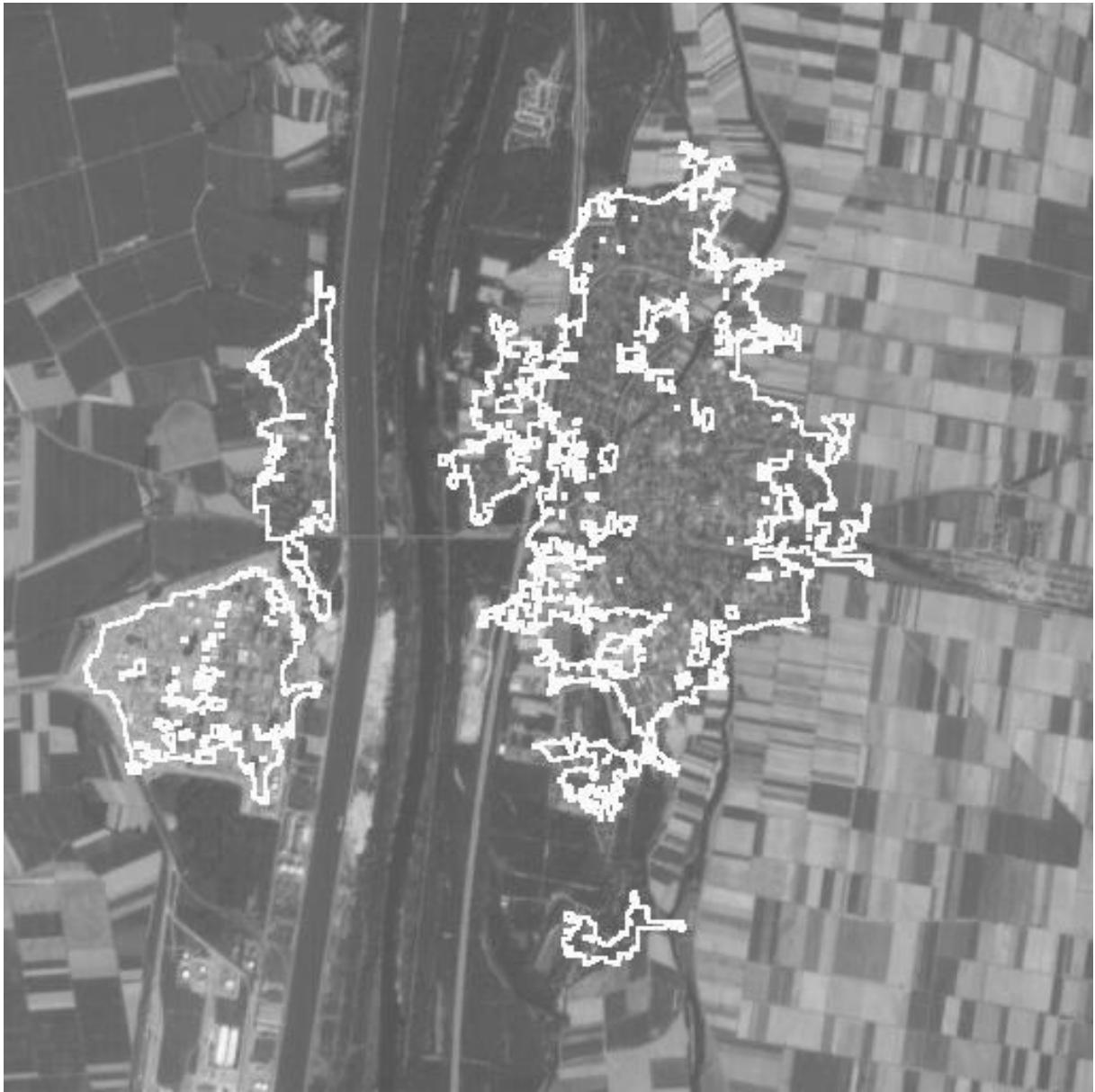


Figure 4.3: Résultat de la détection des zones urbaines sur une image SPOT (extrait 1)

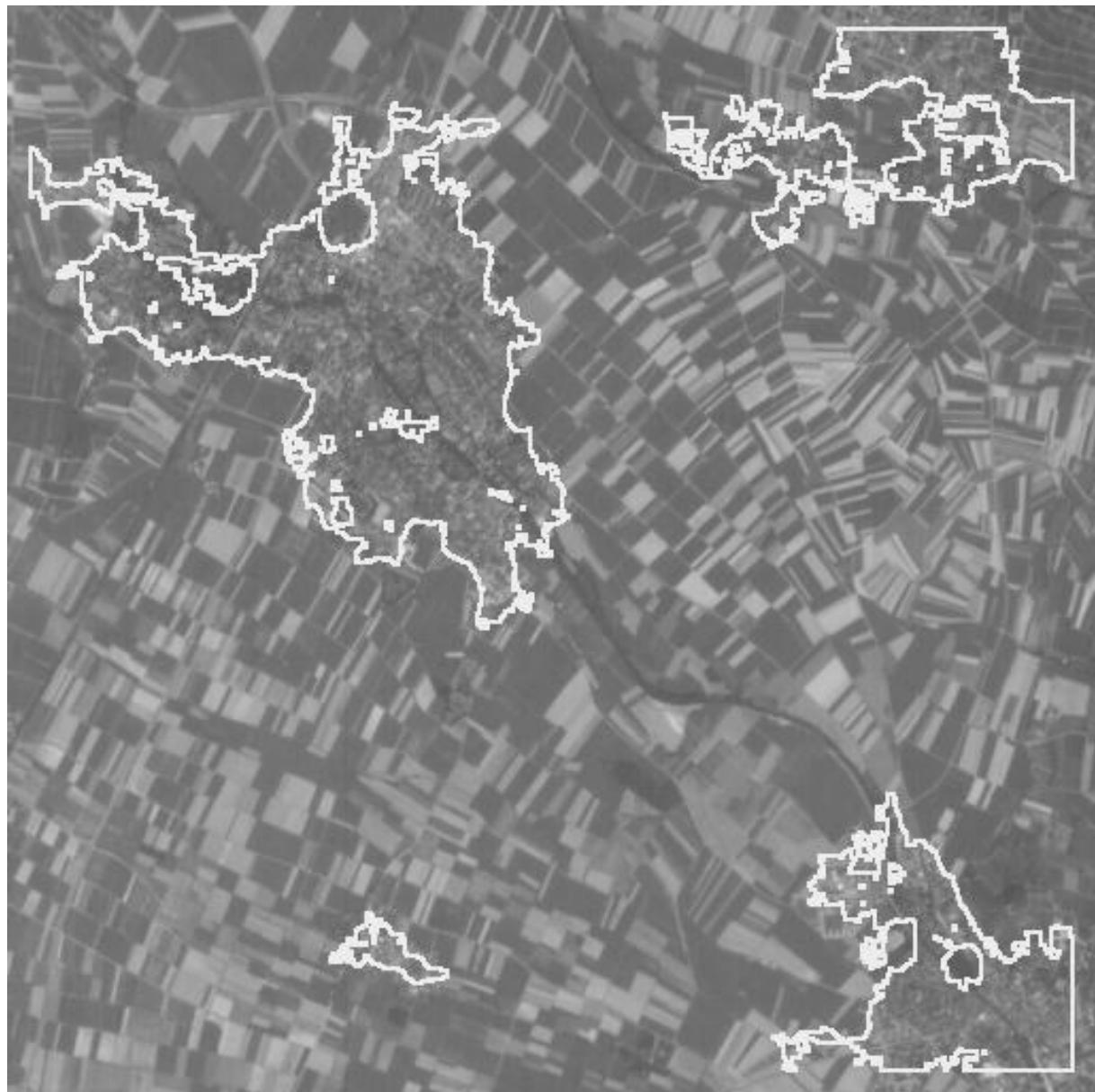


Figure 4.4: Résultat de la détection des zones urbaines sur une image SPOT (extrait 2)

1.2 Segmentation

Nous allons aborder dans cette partie une application très classique des champs markoviens qui est la segmentation. Nous commençons par rappeler le principe de la segmentation markovienne en prenant l'exemple des images radar [Tupin(1997)], puis présentons une méthode de fusion dans un cadre markovien [Descombes(1993)].

1.2.1 Segmentation d'une image radar

Cet exemple d'application est tiré de la thèse de F. Tupin [Tupin(1997)].

La modélisation est similaire à celle qui a été présentée dans le paragraphe 5.3. Ecrivons à nouveau les deux termes intervenant dans la probabilité a posteriori (en gardant les notations du 5.3). Pour la probabilité du champ des observations conditionnellement au champ des étiquettes, en supposant l'indépendance des pixels, on a :

$$P(Y / X = x) = \prod_s P(Y_s = y_s / X_s = x_s)$$

Les images radar sont des images très bruitées par le phénomène de speckle. En revanche, le processus d'acquisition est bien modélisé statistiquement et on a l'expression suivante pour une image radar en amplitude :

$$p(Y_s = y_s / X_s = i) = \frac{2L^L}{\mu_i^L \Gamma(L)} y_s^{(2L-1)} \exp(-\frac{Ly_s^2}{\mu_i})$$

avec L un paramètre du système connu¹ appelé nombre de vues, Γ la fonction Gamma, et μ_i les moyennes en intensité (carré de l'amplitude) des différentes classes i considérées.

Le champ des étiquettes est supposé markovien avec un modèle de Potts qui vise à obtenir des zones homogènes compactes sur l'image segmentée :

$$P(X = x) = \frac{1}{Z} \exp(-\beta \sum_{(s,t) \in \mathcal{C}_2} \phi(x_s - x_t))$$

avec $\beta > 0$, $\phi(0) = 1$ et $\phi(x) = 1 \quad \forall x \neq 0$.

Le champ a posteriori résultant est donc markovien et son énergie s'écrit :

$$\mathcal{U}(x / y) = L \sum_s \left(\frac{y_s^2}{\mu_{x_s}} + \ln \mu_{x_s} \right) + \sum_{c=(s,t)} \phi(x_s - x_t)$$

En choisissant l'estimateur MAP, la solution est obtenue par recuit simulé avec une décroissance géométrique en température, et une température initiale fixée arbitrairement. Le choix des classes se fait de la façon suivante. On se fixe le nombre de classes (15 dans les illustrations

¹Pour les produits PRI du satellite ERS1 $L = 3$.

ci-dessous) et on applique un algorithme de k-moyennes dont le résultat² sert à calculer les valeurs des moyennes en intensité μ_i des différentes classes. Notons que ces classes n'ont pas de contenu sémantique et que la segmentation correspond ici à un “découpage” de l'image.

Le choix de β qui pondère l'influence entre attache aux données et régularisation se fait de façon ad hoc après quelques essais. L'augmentation de β entraîne une augmentation de la taille des zones obtenues par la segmentation. Il serait bien sûr possible d'estimer ce paramètre à l'aide d'une des méthodes décrites au chapitre 3. Notons que ce modèle n'est pas adapté à la préservation des cibles ponctuelles et des lignes qui sont des configurations de forte énergie pour le champ des étiquettes (il faut donc une forte attache aux données pour que ces configurations subsistent dans le résultat final). Il est bien sûr possible de prendre en compte un champ externe pour mieux respecter les lignes par exemple [Tupin *et al.*(1996)] ou d'utiliser un modèle plus approprié [Descombes *et al.*(1995)].

Le tableau ci-dessous 4.1 résume les paramètres utilisés et les figures 4.5 et 4.6 montrent le résultat de la segmentation.

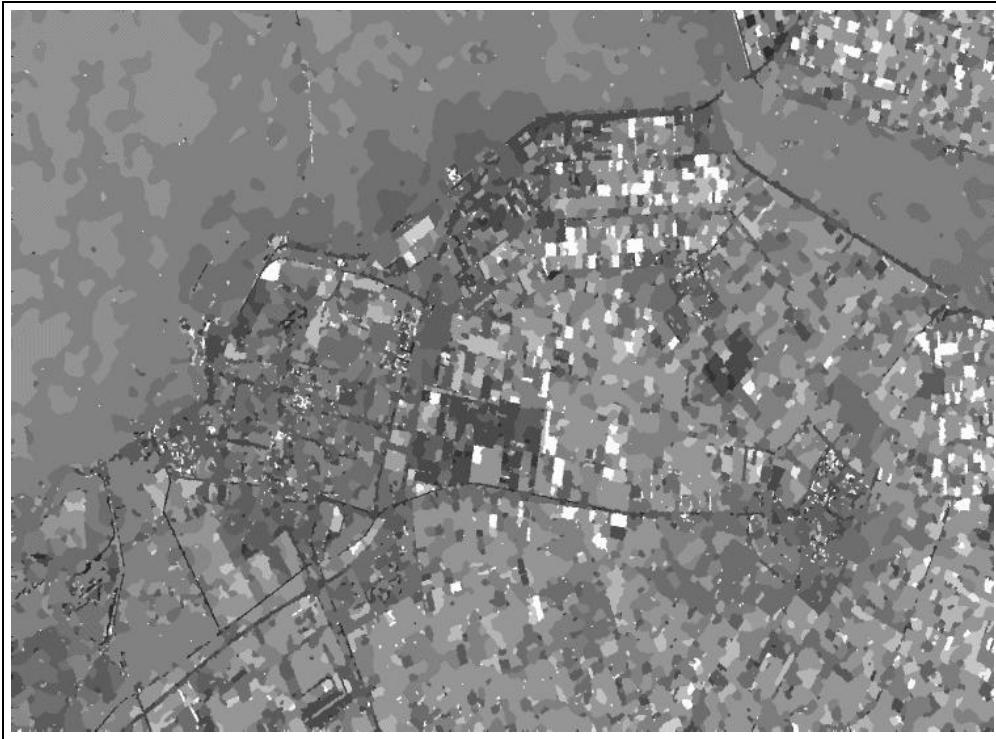
Table 4.1: Valeurs des paramètres de la segmentation

Température initiale	5
Facteur de décroissance géométrique	0,95
Paramètre de régularisation β	0,4
Nombre de classes	15
Nombre d'itérations pour les k-moyennes	20



Figure 4.5: Image radar ERS-1 de la région du Flevoland ©ESA

²Le résultat des k-moyennes est très bruité à cause du bruit multiplicatif présent sur les images radars et l'absence de modes dans l'histogramme. Il ne peut donc pas dans ce cas être utilisé directement comme résultat de segmentation.



a. Régions



b. Contours

Figure 4.6: Résultat des deux étapes de la segmentation (k -moyennes et recuit simulé) sur l'image du Flevoland

1.2.2 Schéma de fusion dans un cadre markovien

- **Principe du schéma de fusion**

Nous nous plaçons maintenant dans le cas où plusieurs sources d'information sont disponibles pour réaliser la classification de l'image. Ces différentes sources peuvent provenir de l'extraction de différents paramètres à partir d'une même image ou directement de plusieurs capteurs.

Notons y le vecteur d'attributs correspondant aux différentes sources d'informations $y = (y^1, \dots, y^K)$ avec K le nombre de données (ou canaux) et M le nombre de classes $E = \{\lambda_1, \dots, \lambda_M\}$. La probabilité a posteriori s'écrit :

$$p(X / Y = y) \propto p(Y / X)p(X)$$

Si nous faisons l'hypothèse que les sources sont **indépendantes** entre elles, et les pixels de chaque source entre eux, on a:

$$\begin{aligned} p(Y / X = x) &= \prod_{s \in S} P(Y_s / X_s = x_s) \\ &= \prod_{s \in S} P(\{Y_s^1, Y_s^2, \dots, Y_s^K\} / X_s = x_s) \\ &= \prod_{s \in S} P(Y_s^1 / X_s = x_s) \dots P(Y_s^K / X_s = x_s) \\ &= \prod_{s \in S} \prod_{k=1}^K P(Y_s^k / X_s = x_s) \end{aligned}$$

Le potentiel d'attache aux données résultant s'exprime alors sous forme d'une somme des potentiels individuels de chaque source :

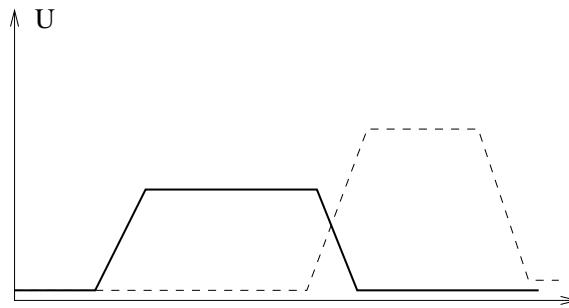
$$U_s(y_s = (y_s^k)_k / x_s = \lambda) = \sum_k U_s(y_s^k / \lambda)$$

Les différentes sources ne renseignant pas de la même façon sur toutes les classes, il est possible d'introduire des coefficients de pondération exprimant la confiance (la fiabilité) qu'on veut accorder à chacune des sources par rapport à une classe. On notera $\gamma_{(k,\lambda)}$ la "confiance" accordée à la source k pour la classe λ . Il faut par ailleurs ne favoriser aucune classe, donc les coefficients doivent vérifier $\sum_k \gamma_{(k,\lambda)} = 1 \quad \forall \lambda$. L'expression de l'attache aux données s'écrit alors :

$$U_s(y_s = (y_s^k)_k / x_s = \lambda) = \sum_k \gamma_{(k,\lambda)} U_s(y_s^k / \lambda)$$

En pratique, il n'est pas toujours nécessaire de faire des modélisations statistiques compliquées pour calculer les potentiels d'attache aux données. Souvent des potentiels très simples, linéaires par morceaux, permettent d'obtenir de bons résultats. On attache un potentiel faible (typiquement 0) à la plage de niveaux de gris qui correspond à la classe considérée et un

potentiel élevé (typiquement 1) ailleurs (voir figure 4.7). Cette définition peut se faire de façon supervisée en analysant l'histogramme ou de façon automatique par une recherche des modes de l'histogramme de l'image (analyse multi-échelle [Aurdal(1997)], recuit simulé sur l'histogramme [Bloch *et al.*(1997)], etc.).



La définition des coefficients de fiabilité des sources est souvent plus problématique. On se limite en général à des remarques de bon sens en affectant $\gamma_{k,\lambda} = 0$ lorsque la source k n'est pas significative pour la classe λ , 0.5 si l'information délivrée est approximative, et 1 si la source est pertinente (avant normalisation).

• Application à l'analyse des images SPOT

*Cet exemple d'application est tiré de la thèse de X. Descombes [Descombes(1993)]. On trouvera d'autres exemples dans [Tupin *et al.*(1996)] [Aurdal(1997)].*

Nous décrivons ici l'application de ce schéma à l'analyse de plusieurs canaux délivrés par le satellite NOAA. Il s'agit de 5 canaux de basse résolution (1.1km) correspondant aux domaines visible, proche infra-rouge, moyen infra-rouge et 2 canaux thermiques, montrés sur la figure 4.8. Les classes qu'on cherche à discriminer sont les suivantes : mer, nuages, continent sans relief, continent avec relief et icebergs.

Une première étape consiste à définir les “sources” que nous allons utiliser. La méthodes des queues de comète que nous avons présentée plus haut permet en effet de déduire de chaque image, 3 images de paramètres (moyenne locale, température et paramètre d'attache aux données λ). Au total 20 images sont donc disponibles, dont les plus significatives pour notre objectif, sont sélectionnées: l'image 1 (visible), l'image de température et l'image de moyenne associées, l'image 3 (moyen infra-rouge) et l'image de température associée (soit 5 en tout).

Pour chacune de ces images, une analyse supervisée de l'histogramme est effectuée, permettant de définir les potentiels (linéaires par morceaux) et la pertinence du canal pour chaque classe. Le terme d'attache aux données de l'énergie a posteriori est alors défini comme mentionné précédemment. Quant au terme de régularisation, les classes recherchées étant relativement compactes, on utilise un modèle de Potts. Le résultat de la segmentation est montré sur la figure 4.9.

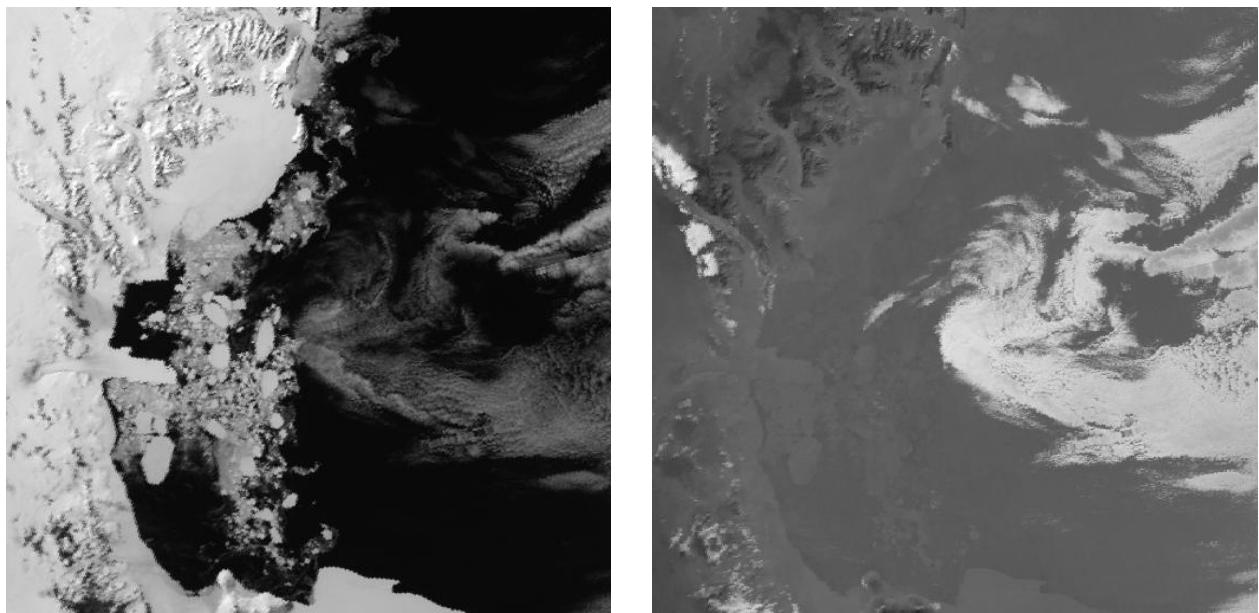


Figure 4.8: Deux des cinq canaux originaux délivrés par le satellite NOAA

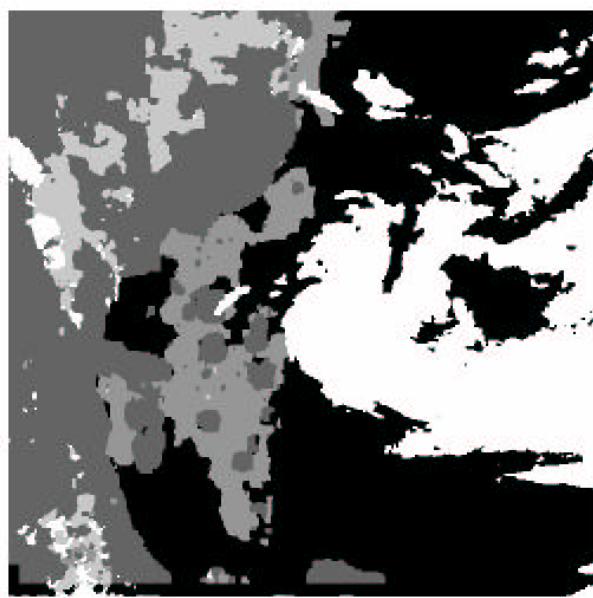


Figure 4.9: Résultat de la segmentation par fusion de canaux

2 Applications sur des graphes de primitives

Comme nous l'avons mentionné au chapitre 1, le formalisme markovien est défini sur tout graphe et son champ d'applications est donc bien plus vaste que la simple grille des pixels. De nombreux problèmes se prêtent à la manipulation de primitives plus complexes, soit pour des raisons de rapidité comme dans le premier exemple que nous décrivons ci-dessous, soit parce qu'il s'agit d'un problème plus proche de l'interprétation d'image, difficile à traiter au niveau du pixel et nécessitant l'introduction d'informations de haut niveau comme dans le second exemple décrit.

2.1 Graphes de régions : application à la segmentation d'une image d'IRM cérébrale

Cet exemple d'application est tiré de la thèse de T. Géraud [Géraud et al.(1995)].

Pour accélérer la segmentation des images, plusieurs applications partent d'une sur-segmentation de l'image qui est ensuite améliorée en fusionnant les régions. Cette fusion, qui procure la segmentation finale, peut se faire dans un cadre markovien. Le graphe est construit à partir des régions de la sur-segmentation, chaque région correspondant à un sommet du graphe et la relation de voisinage étant définie par la relation d'adjacence entre les régions. Le terme d'attaché aux données dépend alors des attributs de la région (niveau de gris moyen des pixels la constituant, moments d'ordre supérieur, etc.) et le terme de régularisation dépend de l'application, un potentiel de Potts pouvant être utilisé lorsqu'on essaye de trouver des zones relativement compactes.

L'objectif de l'exemple décrit ici est de réaliser une segmentation d'images IRM cérébrales. Les classes considérées sont la matière grise, blanche, le liquide céphalo-rachidien, les ventricules et une classe ASI représentant les autres structures internes (noyaux caudés, thalamus, putamen) qui sont difficiles à segmenter et auxquelles on s'intéresse particulièrement dans cette application. Le nombre de pixels de ces images volumiques ($256 \times 256 \times 128$) limite l'utilisation de méthodes markoviennes en raison du temps de calcul. Par contre, l'utilisation d'un graphe de régions construit à partir d'une sur-segmentation, en réduisant drastiquement le nombre de sites permet de réaliser un recuit simulé à un coût raisonnable.

- **Sur-segmentation :** L'étape de sur-segmentation 3D est réalisée par un algorithme calculant la ligne de partage des eaux sur l'image du gradient après fermeture morphologique pour réduire le nombre de bassins. L'image résultat est constituée de zones de niveaux de gris homogènes, auxquelles on associe les attributs suivants : volume (noté vol), niveau de gris moyen (qui sera l'observation y du champ des données), coordonnées du centre du bassin. Le résultat de cette méthode appliquée à l'image 4.10.a est montré sur la figure 4.10.b.

- **Relaxation markovienne :** Un graphe est construit comme indiqué précédemment à partir

de la sursegmentation (fig. 4.10.c). On associe aux arcs du graphe la surface d'adjascence entre les deux régions (notée surf). Les potentiels du champ markovien sont alors définis comme suit :

- Terme d'attache aux données :

$$P(Y_s / x_s) = \exp \left(-\text{vol}_s \sum_s \frac{1}{2\sigma_{x_s}^2} (y_s - \mu_{x_s})^2 \right)$$

(déduit d'une étude statistique des classes [Géraud *et al.*(1995)])

- Terme de régularisation :

$$U_{c=(s,t)}(x) = \text{surf}_{s,t} Q_{(x_s, x_t)}$$

Seuls les potentiels des cliques d'ordre 2 sont choisis non nuls. La matrice Q est une matrice d'adjascence permettant de pondérer les voisinages entre classes suivant qu'ils sont favorisés ou non [Sigelle(1993)].

Le critère utilisé est un critère MAP et la solution est obtenue par un recuit simulé. Le temps de calcul est de moins d'une minute pour un graphe de 32 000 sites. Les résultats pour les ASI sont montrés sur la figure 4.10.c.

2.2 Graphes de segments : application à la détection du réseau routier

Cet exemple d'application est tiré de la thèse de F. Tupin [Tupin(1997)].

Nous nous intéressons dans cette application à la détection du réseau routier et hydrographique dans le cas des images radar. Le bruit de speckle présent sur ces images ne permet pas d'obtenir de très bons résultats de bas niveau et les détecteurs de lignes même adaptés à ce type d'imagerie ont des taux de fausses alarmes élevés si on veut obtenir des taux de détection suffisants.

L'idée est donc de faire suivre l'étape de bas niveau de détection des lignes par une étape de plus haut niveau, dans laquelle on injectera des informations a priori sur la forme des routes. Le cadre markovien, par l'intermédiaire du terme a priori se prête bien à l'introduction de connaissances sur les objets recherchés, à condition que celles-ci puissent s'exprimer de façon locale à l'échelle du graphe. Dans le cas du réseau cette hypothèse (qui assure que le champ soit markovien) est vérifiée, puisque la pratique montre qu'il nous suffit d'informations locales au niveau des segments pour prendre une décision (présence ou absence de réseau routier).

La démarche adoptée pour la détection du réseau est donc la suivante. L'étape de bas niveau permet de détecter des segments candidats. Parmi ceux-ci, certains appartiennent aux objets à détecter, quand d'autres sont de fausses détections. On fait alors l'hypothèse que les segments

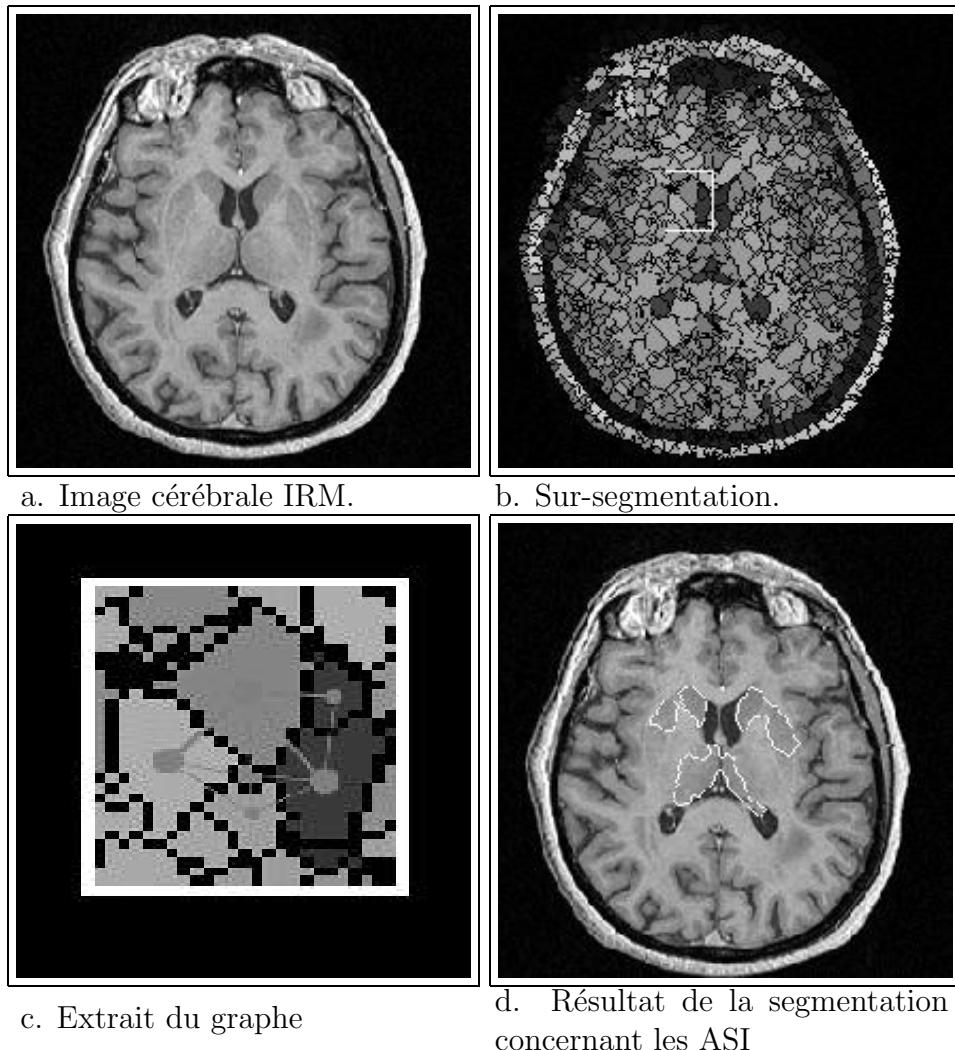


Figure 4.10: Etapes de la segmentation sur le graphe des régions

détectés et toutes les connexions possibles entre ces segments contiennent le réseau routier. Les connexions “possibles” ne sont pas toutes les connexions, mais les connexions raisonnables : entre des segments suffisamment proches, et à peu près “alignés” par exemple. L’ensemble de ces segments (ceux détectés et les connexions) constituent les sommets du graphe (voir figure 4.11). La relation de voisinage entre deux segments est définie par le partage d’une extrémité par ces 2 segments. Les figures ci-dessous illustrent les étapes de la construction du graphe sur un extrait d’image radar (fig. 4.14).

Une fois ce graphe construit, on définit un champ d’observation associé Y et un champ d’étiquettes X (les étiquettes étant simplement 0 pour “non-route” et 1 pour “route”) dont on cherche la configuration optimale au sens du critère MAP.

Les termes de l’énergie sont définis de la façon suivante :

- Terme d’attache aux données : l’observation en un site est définie comme la réponse moyenne du détecteur de lignes le long de ce segment ; les potentiels $U_s(y_s / x_s)$ pour

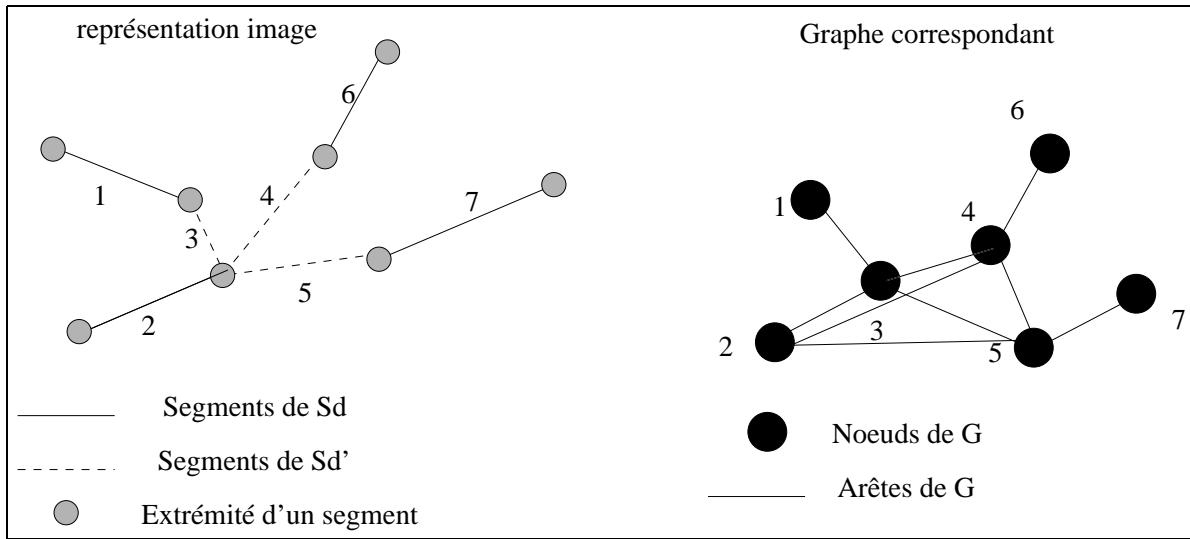


Figure 4.11: Construction du graphe de segments

les labels 0 et 1 sont alors obtenus par une étape d'apprentissage le long de quelques segments appartenant à une vraie route et de quelques segments de "fausse alarme"; la figure 4.12 montre les fréquences des observations y (approximant les $P(y_s / x_s)$) et les potentiels linéaires par morceaux qui en sont déduits (fig. 4.13);

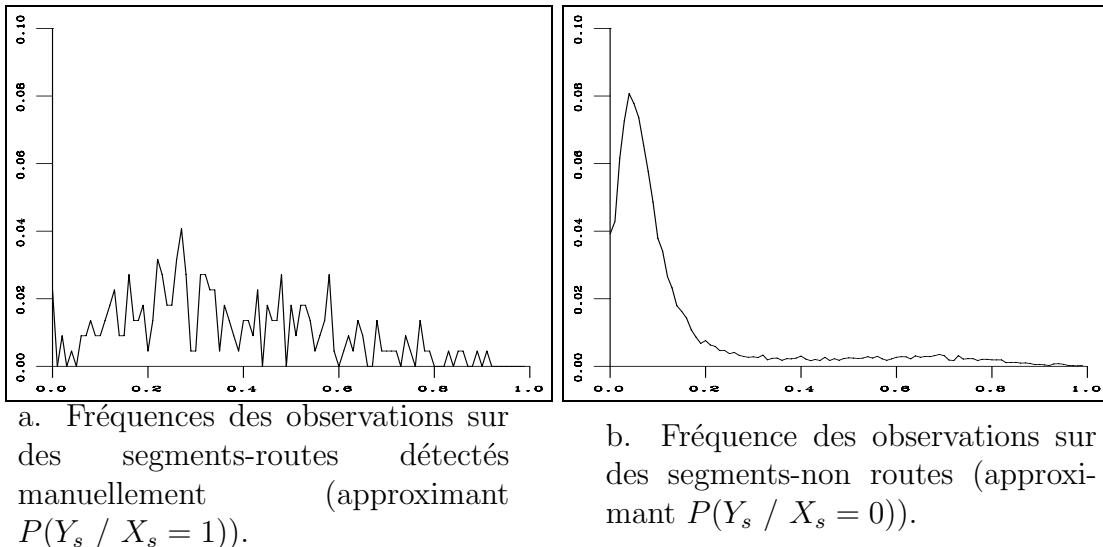


Figure 4.12: Fréquences conditionnelles des observations sur une zone test.

- Énergie a priori : elle permet justement d'intégrer toutes les connaissances a priori qu'on veut prendre en compte pour la détection du réseau, par exemple du type:
 - (i) les routes sont longues et, dans l'absolu, elles ne s'arrêtent pas;
 - (ii) elles ont une courbure relativement faible;
 - (iii) un segment de route est plus souvent connecté en une extrémité à un unique segment de route qu'à plusieurs.

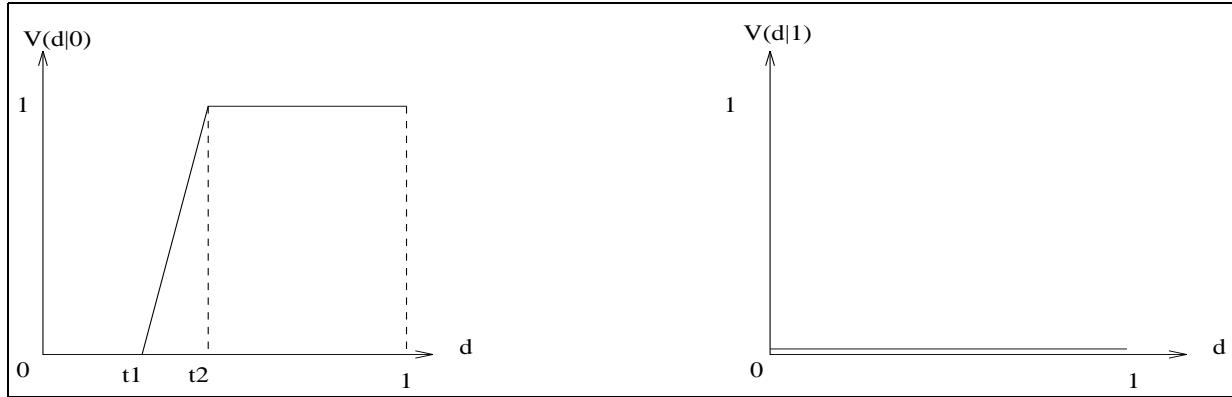


Figure 4.13: Potentiels linéaires par morceaux utilisés

Ces a priori peuvent s'exprimer en définissant les potentiels des cliques d'ordre maximal (rappelons qu'une clique est un ensemble de sites tous voisins les uns des autres, donc dans notre cas, un ensemble de segments qui se rejoignent en un même point); quatre paramètres suffisent pour les contraintes mentionnées précédemment : un paramètre contrôlant les extrémités (qui vise à défavoriser la configuration d'une clique où un seul segment a le label “route”); deux paramètres contrôlant la longueur des routes (qui visent à favoriser des configurations où il y deux segments routes qui se joignent en une extrémité et qui sont “alignés”); un paramètre contrôlant les carrefours (qui vise à défavoriser les configurations où une multitude de segments sont “route”).

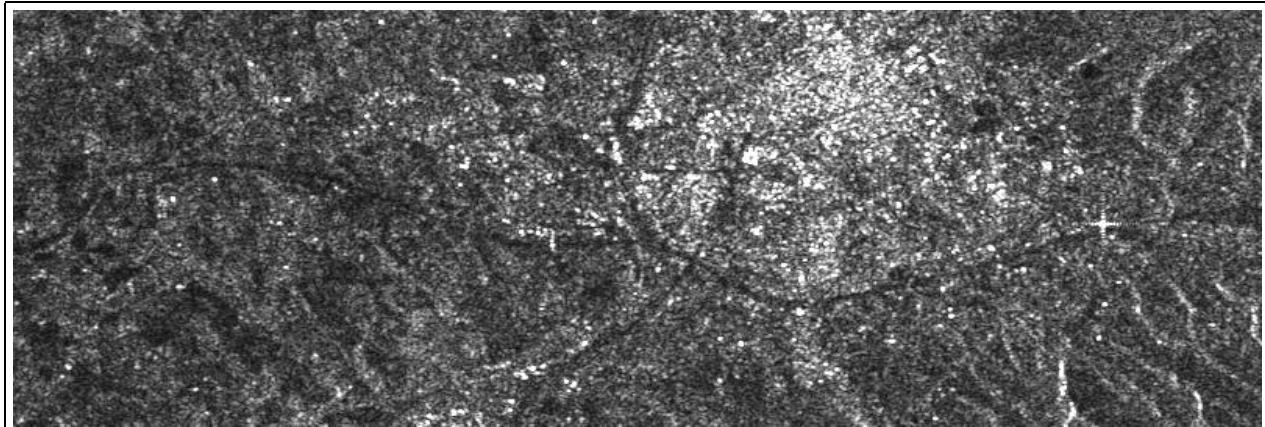
L'apprentissage de ces paramètres se fait difficilement par les méthodes du chapitre 3. En revanche l'étude de configurations extrêmes (chaîne de segments tous à “route”, etc.), plus connue sous le nom de “Boîtes qualitatives d’Azencott” [Azencott(1992)], permet de fixer des intervalles de valeurs pour ces paramètres.

Un exemple de résultat obtenu par recuit simulé est montré sur la figure 4.15.

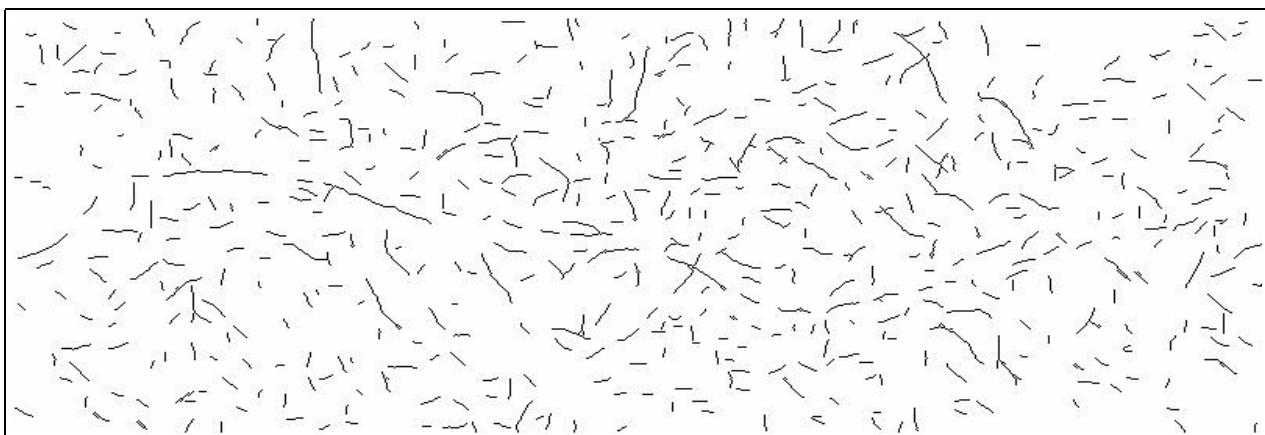
3 Conclusion

Ce chapitre ne se veut pas une présentation exhaustive des applications possibles des champs markoviens, mais présente quelques unes de leurs potentialités dans des domaines et à des niveaux de traitement d'image variés. Leur grande souplesse permet en effet d'introduire toutes sortes d'informations, que ce soit pour le terme d'attache aux données, que ce soit pour les a priori possibles, ou même encore sur la forme du graphe à utiliser.

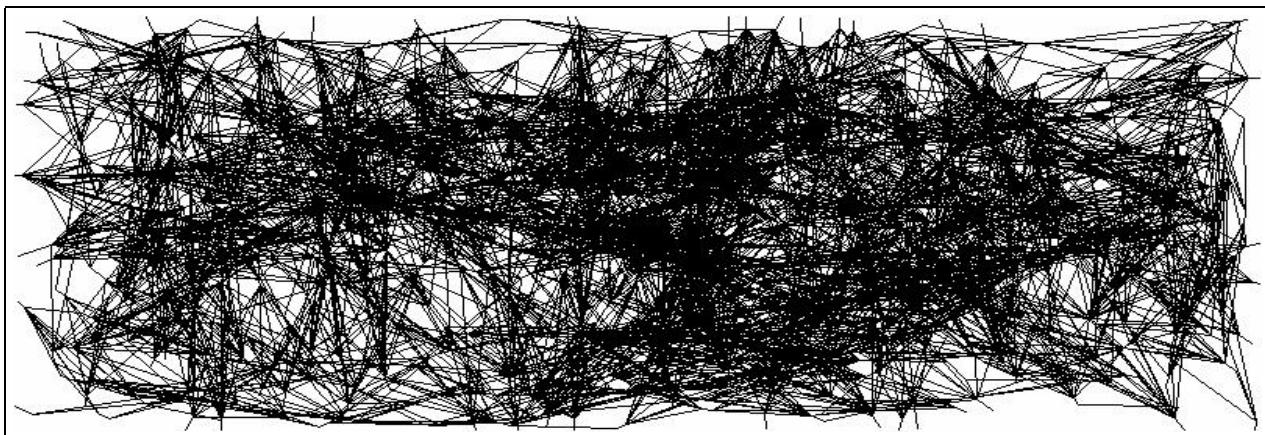
A côté de la multitude d'applications qui peuvent trouver une solution dans un cadre markovien, s'ouvrent des champs de recherche plus théoriques sur l'estimation des paramètres et l'accélération des techniques de recherche de solution.



a. Image originale centrée sur Aix en Provence ©ESA .



b. Ensemble des 839 segments détectés.



c. Ensemble des 8891 segments constituant les sites du graphe.

Figure 4.14: Les étapes de la construction du graphe

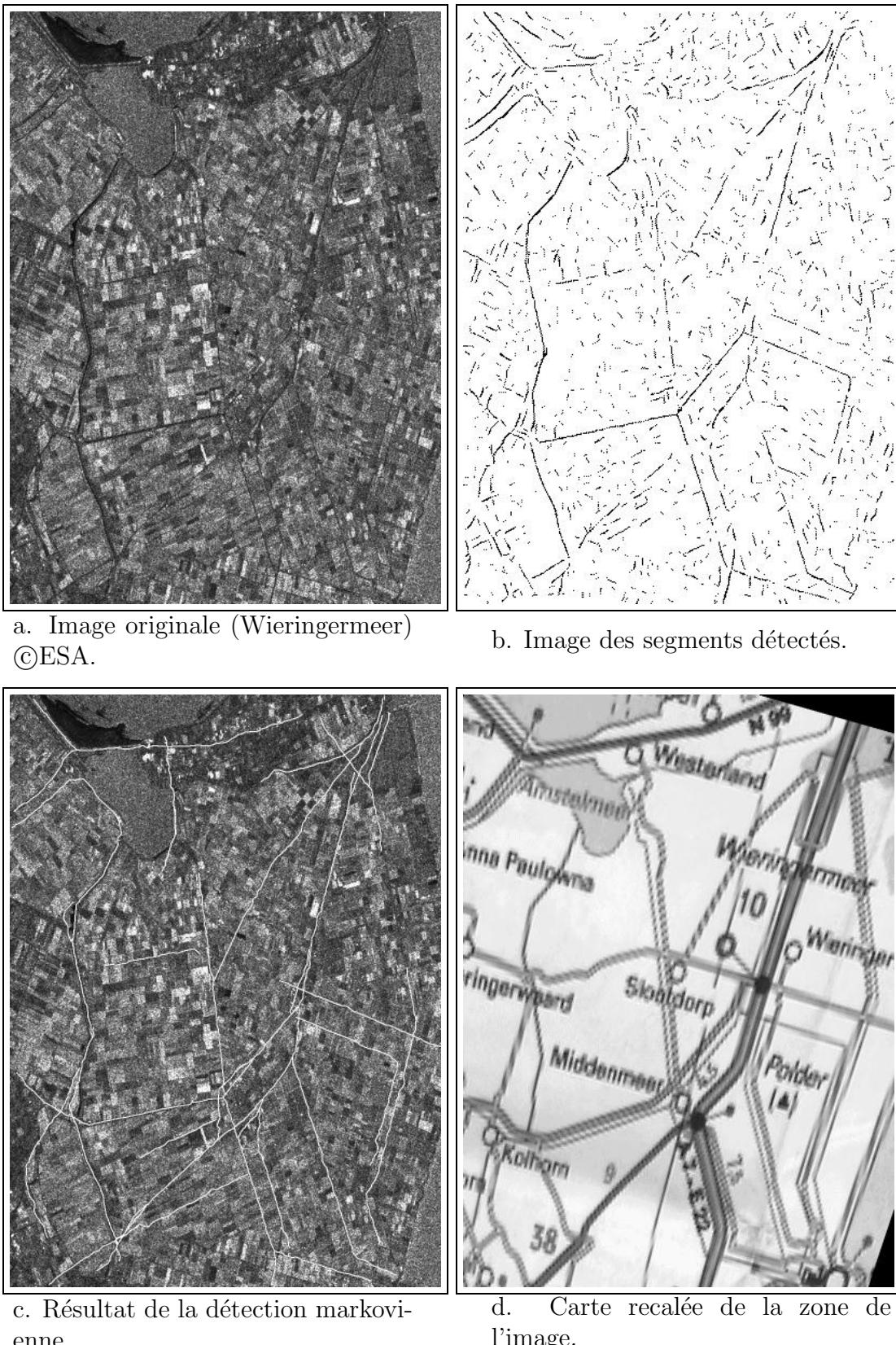


Figure 4.15: Détection des routes sur le Wieringermeer

Appendix A

Statistique exponentielle linéaire d'un paramètre θ

1 Introduction - Notations et définitions

On rappelle brièvement ici les propriétés fondamentales des familles de lois de probabilité exponentielles dépendant linéairement d'un paramètre. Les champs de Markov-Gibbs peuvent en effet être vus comme cas particulier de statistiques (distributions) exponentielles.

On se donne :

- un espace d'états **fini** : Ω . Tout élément de Ω , aussi appelé configuration, sera noté : x .
- une famille de lois de probabilités sur Ω dépendant d'un paramètre réel θ et de la forme :

$$P_\theta(X = x) = P_\theta(x) = \frac{\exp -\theta U(x)}{Z_\theta}$$

- l'énergie de la configuration x est la fonction : $U(x)$. Nous ne faisons pas d'hypothèses sur la forme de l'énergie : elle peut se décomposer en somme d'énergies de cliques ou non.
- la fonction de partition pour la loi P_θ est : $Z_\theta = \sum_{y \in \Omega} \exp -\theta U(y)$.

Remarquons que par rapport au Chap. 1 nous avons “sorti” explicitement le paramètre (supposé scalaire) θ de la fonction énergie $U(x)$, de façon à étudier la dépendance de la distribution de Gibbs associée $P_\theta()$ vis-à-vis de ce paramètre, en particulier lorsque celui-ci peut être considéré comme inverse d'une température T (voir section suivante).

2 Comportement en fonction du paramètre θ positif

Lorsque le paramètre θ est positif (ou nul), il peut être interprété en tant que température inverse : $\theta = \frac{1}{T}$. Le comportement de la distribution $P_\theta()$ en fonction de θ dans le paysage d'énergie des configurations est alors résumé dans la Fig. A.1 :

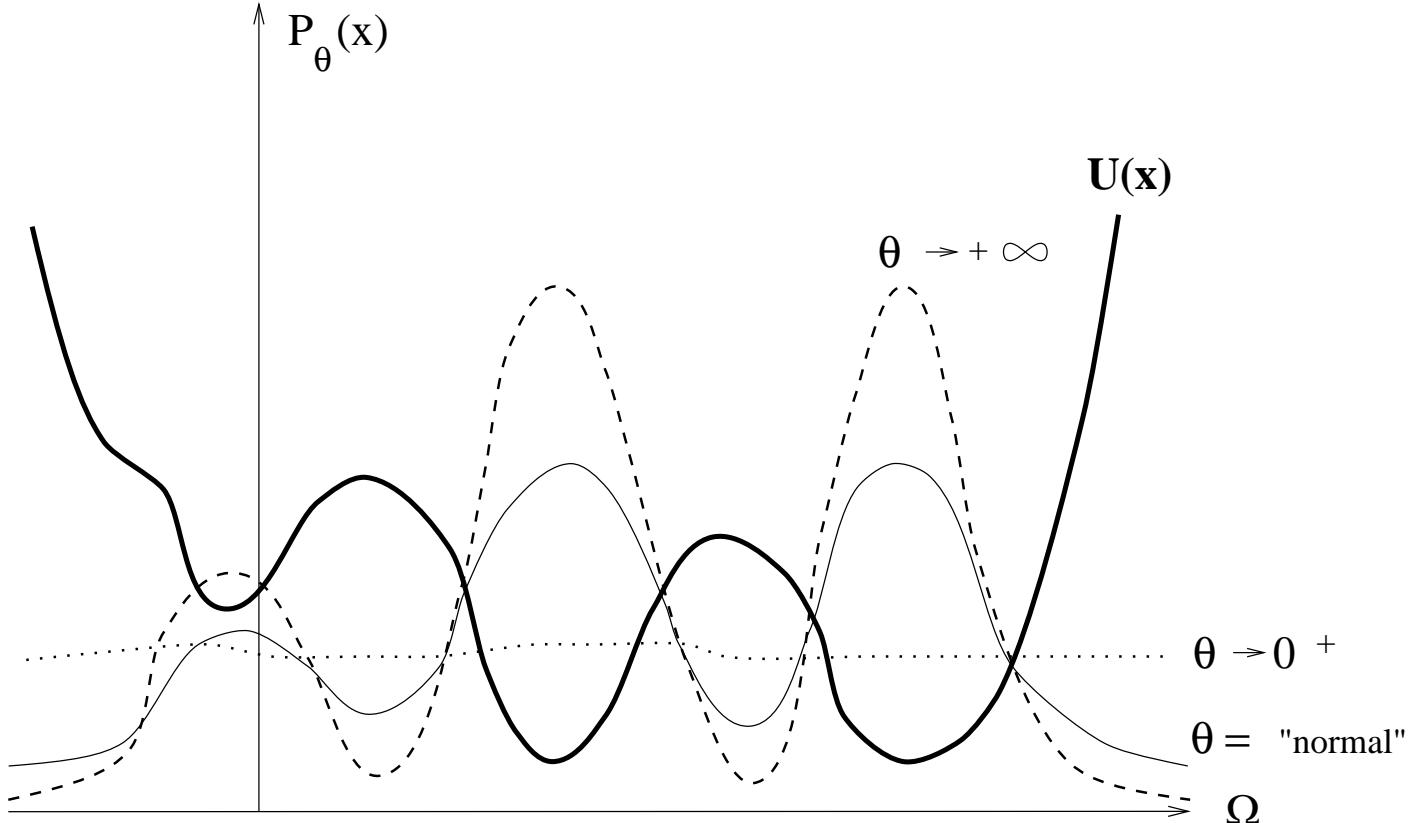


Figure A.1: Allure des distributions de Gibbs pour différentes valeurs du paramètre **θ positif** dans le paysage d'énergie supposé mono-dimensionnel

Il s'explique aux valeurs extrêmes de la façon suivante :

- a) $\theta \rightarrow 0^+$ i.e. $T \rightarrow +\infty$: équidistribution sur Ω .
- b) $\theta \rightarrow +\infty$ i.e. $T \rightarrow 0^+$: équidistribution sur $\Omega^* = \{ x \in \Omega \text{ t.q. } U(x) = \min_{z \in \Omega} U(z) \}$

2.0.1 Démonstration

On note ici la distribution de Gibbs en fonction de la température $P_T() = P_\theta()$.

On utilise alors la relation très utile dans la pratique :

$$\frac{P_T(y)}{P_T(x)} = \exp - \left(\frac{U(y) - U(x)}{T} \right) \quad \forall x, y \in \Omega$$

et le fait que Ω soit fini.

a) lorsque $T \rightarrow \infty$, $\frac{P_T(Y = y)}{P_T(X = x)} \rightarrow 1 \quad \forall x, y \in \Omega$.

On comprend donc bien que la loi $P(\cdot)$ tende vers la loi équidistribuée sur Ω . Plus précisément :

$$P_T(X = x) = \frac{\exp -\frac{U(x)}{T}}{\sum_{y \in \Omega} \exp -\frac{U(y)}{T}} = \frac{1}{\sum_{y \in \Omega} \exp -\left(\frac{U(y) - U(x)}{T}\right)} \quad (\text{A.1})$$

La somme figurant au dénominateur étant *finie* par hypothèse et en vertu de la remarque précédente, on a donc :

$$\lim_{T \rightarrow \infty} P(X = x) = \frac{1}{\text{Card } \Omega} \quad \forall x \in \Omega \text{ CQFD}$$

b) lorsque $T \rightarrow 0^+$, $\frac{P_T(Y = y)}{P_T(X = x)} = \exp -\left(\frac{U(y) - U(x)}{T}\right) \rightarrow 0$ si $U(y) > U(x)$

Notons alors Ω^* l'ensemble des configurations de Ω qui atteignent le minimum *global* de l'énergie, que nous noterons U^* . On réécrit alors (A.1) de la façon suivante :

$$\begin{aligned} P_T(X = x) &= \frac{\exp -\left(\frac{U(x) - U^*}{T}\right)}{\sum_{y \in \Omega} \exp -\left(\frac{U(y) - U^*}{T}\right)} \\ &= \frac{\exp -\left(\frac{U(x) - U^*}{T}\right)}{\text{Card } \Omega^* + \sum_{y \in \Omega, y \notin \Omega^*} \exp -\left(\frac{U(y) - U^*}{T}\right)} \quad (\text{A.2}) \end{aligned}$$

En vertu de la remarque précédente et du fait que la somme figurant au dénominateur est *finie* par hypothèse, on a :

$$\lim_{T \rightarrow \infty} P_T(X = x) = \begin{cases} \frac{1}{\text{Card } \Omega^*} & \text{si } x \in \Omega^* \\ 0 & \text{si } x \notin \Omega^* \end{cases} \quad \text{CQFD}$$

3 Vraisemblance pour une statistique linéaire exponentielle

3.1 notations

Nos notations sont les suivantes pour les moments d'une variable aléatoire X par rapport à une distribution de Gibbs $P_\theta(\cdot)$ donnée :

- l'espérance de la v.a. X est : $\mathbb{E}_\theta[X] = \frac{\sum_{y \in \Omega} X(y) \exp -\theta U(y)}{\sum_{y \in \Omega} \exp -\theta U(y)}$
- la variance de la v.a. X est : $\text{var}_\theta(X) = \mathbb{E}_\theta[X^2] - (\mathbb{E}_\theta[X])^2$

3.2 formules fondamentales

Il faut avoir effectué au moins une fois dans sa vie les calculs suivants :

le fait que $P_\theta(x) = \frac{\exp -\theta U(x)}{Z_\theta}$ entraîne

$$\begin{aligned} \frac{\partial \log P_\theta(x)}{\partial \theta} &= -U(x) - \frac{\partial \log Z_\theta}{\partial \theta} = -U(x) + \frac{\sum_{y \in \Omega} U(y) \exp -\theta U(y)}{\sum_{y \in \Omega} \exp -\theta U(y)} \\ &= -U(x) + \mathbb{E}_\theta[U] \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} \frac{\partial \mathbb{E}_\theta[U]}{\partial \theta} &= \left(\frac{\sum_{y \in \Omega} -U^2(y) \exp -\theta U(y)}{\sum_{y \in \Omega} \exp -\theta U(y)} \right) + \left(\frac{\sum_{y \in \Omega} U(y) \exp -\theta U(y)}{\sum_{y \in \Omega} \exp -\theta U(y)} \right)^2 \\ &= -\text{var}_\theta(U) \leq 0 \end{aligned} \quad (\text{A.4})$$

Remarques

1. Cette dernière équation constitue la formule de la chaleur spécifique en thermodynamique.
2. Ces propriétés peuvent se retrouver très simplement par l'étude du développement en cumulants des distributions de Gibbs. Pour ceux qui sont intéressés, voir annexe B.

3.3 Propriétés de la vraisemblance et de la log-vraisemblance

a. Notations nous noterons les valeurs caractéristiques suivantes importantes pour la suite :

- la moyenne empirique de l'énergie totale est : $\bar{U} = \frac{1}{|\Omega|} \sum_{x \in \Omega} U(x)$
- sa valeur maximale est : $U^{max} = \max_{x \in \Omega} U(x)$ et sa valeur minimale est : $U^* = \min_{x \in \Omega} U(x)$

Nous étudions alors progressivement les quantités suivantes :

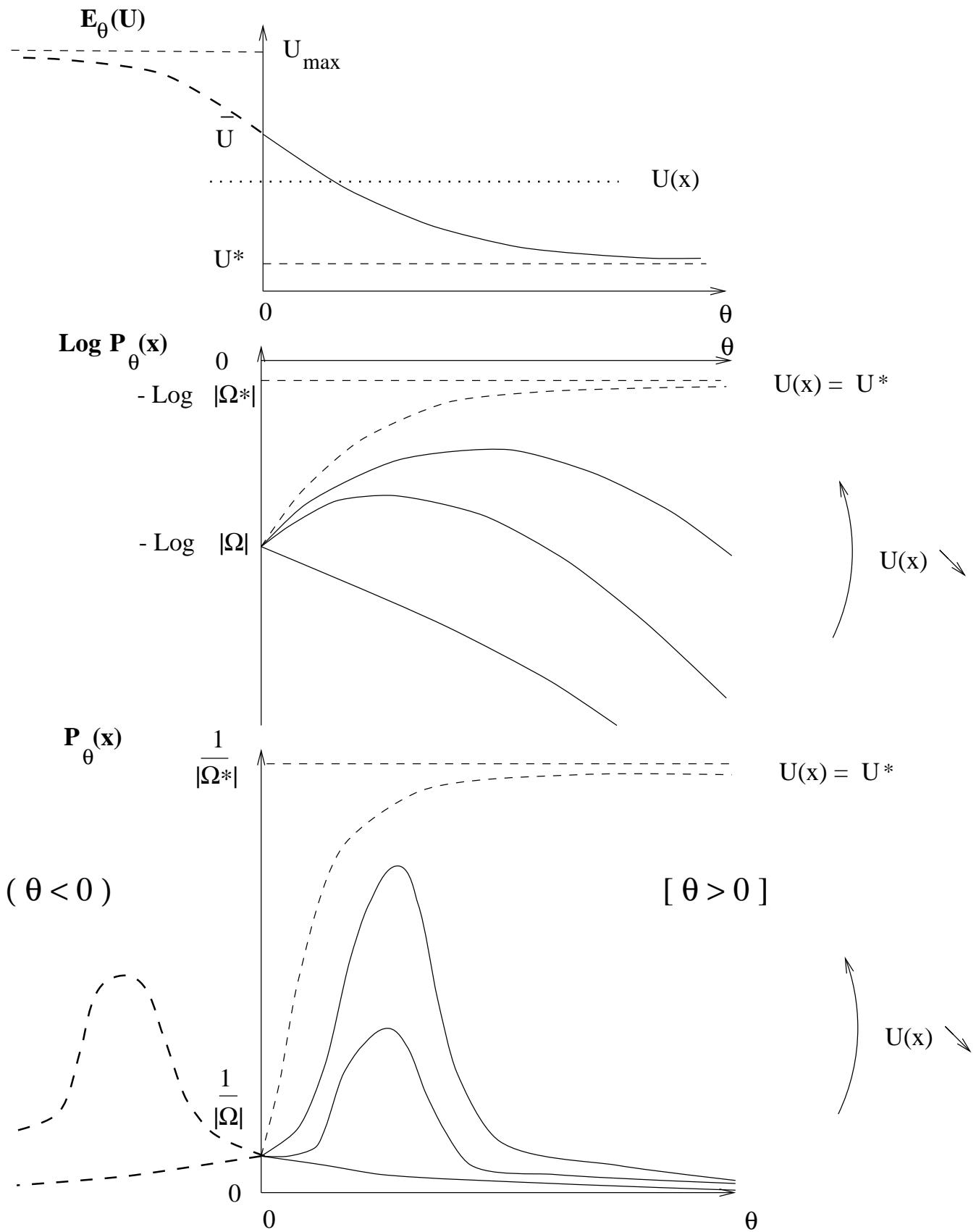


Figure A.2: Comportement de l'énergie statistique moyenne (a), de la log-vraisemblance (b) et de la vraisemblance elle-même (c), *i.e.* distribution de Gibbs, en fonction du paramètre θ .

b. La dérivée de la log-vraisemblance par rapport à θ d'après les équations fondamentales A.3 et A.4 , la dérivée $p.r.$ θ de la log-vraisemblance : $\frac{\partial \log P_\theta(x)}{\partial \theta} = -U(x) + \mathbb{E}_\theta[U]$ est une fonction monotone **décroissante** de θ , quelle que soit la configuration x adoptée . Son domaine de variation est donc le suivant :

entre $-U(x) + \bar{U}$ ($\theta = 0$) et $-U(x) + U^*$ ($\theta = +\infty$) si θ est une température inverse	entre $-U(x) + U^{max}$ ($\theta = -\infty$) et $-U(x) + U^*$ ($\theta = +\infty$) si θ est un paramètre
--------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------

Les propriétés de la vraisemblance et de la log-vraisemblance sont résumées sur la Fig. A.2 .

c. La log-vraisemblance $\log P_\theta(x)$ est donc une fonction **concave** de θ (Fig. A.2 b) .

d. La vraisemblance $P_\theta(x)$ est donc une fonction **unimodale** de θ . Elle admet donc pour chaque configuration x un point maximum qui sera noté θ_x , et qui correspond à la valeur du maximum de vraisemblance (MV) de la distribution $P_\theta()$ pour la configuration x donnée ou observée (Fig. A.2 c) . Plus spécifiquement, cette valeur du maximum de vraisemblance θ_x vérifie en vertu de l'eq. A.3 :

a) $\theta_x = 0$ si $U(x) \geq \bar{U}$ b) $\theta_x > 0$ si $U^* < U(x) < \bar{U}$ c) $\theta_x = +\infty$ si $U(x) = U^*$ si θ est une température inverse	$\theta_x = 0$ si $U(x) = \bar{U}$ b) ou $\theta_x < 0$ si $\bar{U} < U(x) < U^{max}$ c) ou $\theta_x = -\infty$ si $U(x) = U^{max}$ si θ est un paramètre
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

e. Comportement de la vraisemblance lorsque θ croît l'équation fondamentale A.3 entraîne aussi que lorsque θ augmente :

- si $U(x) \geq \bar{U}$ alors $P_\theta(x) \searrow 0$ quand θ croît
- si $U^* < U(x) < \bar{U}$ alors $P_\theta(x) \searrow 0$ dès que $\theta > \theta_x$
- si $U(x) = U^*$ alors $P_\theta(x) \nearrow \frac{1}{|\Omega^*|}$ quand θ croît

Cela sera important dans le cas du recuit simulé : $\frac{1}{\theta} = T \rightarrow 0^+$ (voir annexe C Section 4) .

Appendix B

Développement en cumulants des statistiques exponentielles

1 Introduction

Les résultats de la Section 3 concernant les propriétés fondamentales des distributions de Gibbs vues en tant que statistiques linéaires exponentielles peuvent être rapidement et facilement retrouvés par une analyse en cumulants. Le développement en cumulant est un outil puissant bien connu en Traitement du Signal statistique [Nikias and Petropulu(1993), Cardoso and Moulines(1995)] , ainsi qu'en Physique Statistique pour l'étude des propriétés de corrélation-fluctuation de la matière condensée [Ma(1985)].

On développe d'abord le cas des distributions à un seul paramètre pour des raisons de simplicité. Le cas des distributions multi-paramétrées est abordé ensuite.

2 Définition

On fait l'hypothèse que pour une variable aléatoire (v.a.) X donnée et pour une large classe de distributions on a :

$$\mathbb{E}[\exp \lambda X] = \exp \left\{ \sum_{n=1}^{\infty} \lambda^n \frac{C^n(X)}{n!} \right\} \quad |\lambda| \leq R(X). \quad (\text{B.1})$$

où les $C^n(X)$, $n = 1, 2 \dots$ sont les cumulants respectifs de la v.a. X à chaque ordre n . La formule précédente résulte du développement en série en fonction du paramètre λ , et à l'intérieur du rayon de convergence $R(X)$, de

$$\log \mathbb{E}[\exp \lambda X] = \log \left(\mathbb{E}[1 + \lambda X + \frac{1}{2} \lambda^2 X^2 + \dots] \right)$$

qui s'obtient lui-même facilement à partir du développement de Taylor connu :

$$\log(1 + \epsilon) = \epsilon - \frac{1}{2} \epsilon^2 \dots + (-1)^{n-1} \frac{\epsilon^n}{n} \dots$$

Cela donne immédiatement aux premier et second ordre :

$$C^1(X) = \mathbb{E}[X] \quad \text{and} \quad C^2(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{var}(X) \quad (\text{B.2})$$

On peut aussi écrire par des considérations d'homogénéité par rapport à λ , que l'on a pour une v.a. ξ :

$$\mathbb{E}[\exp \xi] = \exp \sum_{n=1}^{\infty} \frac{C^n(\xi)}{n!} , \quad \text{où } C^1(\xi) = \mathbb{E}[\xi] \text{ et } C^2(\xi) = \text{var}(\xi) \dots \quad (\text{B.3})$$

3 Cumulants de distributions de Gibbs mono-paramétrées

Une des applications principales de l'analyse en cumulants aux distributions de Gibbs est en fait le développement en série de Taylor du logarithme la fonction de partition Z_θ en fonction de θ . Ainsi, la motivation principale de l'estimation des hyperparamètres est de prédire à chaque étape d'un processus (souvent itératif), comment varie, dans un sens à définir, la distribution de Gibbs $P_\theta = \frac{\exp -\theta U(x)}{Z_\theta}$ lorsque l'hyperparamètre dévie "légèrement" de sa valeur courante θ . Pour le montrer, nous partons de la formule suivante :

$$Z_{\theta+\delta\theta} = \sum_{y \in \Omega} \exp -\theta U(y) \exp -\delta\theta U(y) = Z_\theta \mathbb{E}_\theta[\exp -\delta\theta U]$$

qui peut être écrite d'après le paragraphe précédent sous la forme :

$$Z_{\theta+\delta\theta} = Z_\theta \exp \left\{ -\delta\theta C^1(U) + \frac{\delta\theta^2}{2} C^2(U) + \dots \right\} \quad |\delta\theta| \leq R(U)$$

c'est-à-dire : $\log Z_{\theta+\delta\theta} - \log Z_\theta = -\delta\theta \mathbb{E}_\theta[U] + \frac{\delta\theta^2}{2} \text{var}_\theta(U) + \dots \quad |\delta\theta| \leq R(U)$.

En identifiant terme à terme avec un développement en série de Taylor, on obtient facilement aux premiers ordres :

$$\frac{\partial \log Z_\theta}{\partial \theta} = -\mathbb{E}_\theta[U] \quad \text{and} \quad \frac{\partial^2 \log Z_\theta}{\partial \theta^2} = \text{var}_\theta(U) , \quad (\text{B.4})$$

ce qui montre en particulier que : $\frac{\partial \mathbb{E}_\theta[U]}{\partial \theta} = -\text{var}_\theta(U) \leq 0$. On retrouve donc immédiatement que $\mathbb{E}_\theta[U]$ est une fonction monotone décroissante de l'hyperparamètre θ , (cf. annexe A Section 3) ce qui a une grande importance dans l'estimation au maximum de vraisemblance (MV) des champs de Gibbs avec données complètes [Younes(1988)], comme vu au Chap. 3 Section 2. La relation précédente peut aussi être interprétée en tant que formule de la chaleur spécifique en thermodynamique lorsque l'on considère maintenant θ comme l'inverse d'une température.

4 Le cas des distributions de Gibbs multi-paramétrées

Considérons par exemple le cas très fréquent de distributions de Gibbs à deux paramètres :

$$P_{\theta,\lambda}(x) = \frac{1}{Z_{\theta,\lambda}} \exp \{-\theta U(x) - \lambda \Psi(x)\} \text{ où } Z_{\theta,\lambda} = \sum_{y \in \Omega} \exp -\theta U(y) - \lambda \Psi(y) \quad (\text{B.5})$$

On peut immédiatement écrire la formule suivante pour $Z_{\theta+\delta\theta, \lambda+\delta\lambda}$ a partir de l'eq. B.5 :

$$Z_{\theta+\delta\theta, \lambda+\delta\lambda} = \sum_{y \in \Omega} \exp \{-\theta U(y) - \lambda \Psi(y)\} \exp \{-\delta\theta U(y) - \delta\lambda \Psi(y)\} = Z_{\theta,\lambda} \mathbb{E}_{\theta,\lambda}[\exp \{-\delta\theta U - \delta\lambda \Psi\}].$$

Définissons alors une nouvelle variable aléatoire : $\xi = -\delta\theta U - \delta\lambda \Psi$.

On peut utiliser l'expression ci-dessus ainsi que les résultats des paragraphes précédents pour en déduire :

$$\begin{aligned} \log Z_{\theta+\delta\theta, \lambda+\delta\lambda} - \log Z_{\theta,\lambda} &= \log \mathbb{E}_{\theta,\lambda}[\exp \xi] = \mathbb{E}_{\theta,\lambda}[\xi] + \frac{1}{2} \text{var}_{\theta,\lambda}(\xi) + \dots \\ &= -\delta\theta \mathbb{E}_{\theta,\lambda}[U] - \delta\lambda \mathbb{E}_{\theta,\lambda}[\Psi] + \frac{1}{2} \{ \delta\theta^2 \text{var}_{\theta,\lambda}(U) + \delta\lambda^2 \text{var}_{\theta,\lambda}(\Psi) + 2 \delta\theta \delta\lambda \text{cov}_{\theta,\lambda}(U, \Psi) \} + \dots \end{aligned}$$

dont l'identification à un développement en série de Taylor donne aux premiers ordres:

$$\left\{ \begin{array}{l} \frac{\partial \log Z_{\theta,\lambda}}{\partial \theta} = -\mathbb{E}_{\theta,\lambda}[U] \quad \text{and} \quad \frac{\partial \log Z_{\theta,\lambda}}{\partial \lambda} = -\mathbb{E}_{\theta,\lambda}[\Psi] \\ \frac{\partial^2 \log Z_{\theta,\lambda}}{\partial \theta^2} = \text{var}_{\theta,\lambda}(U), \quad \frac{\partial^2 \log Z_{\theta,\lambda}}{\partial \lambda^2} = \text{var}_{\theta,\lambda}(\Psi) \quad \text{and} \quad \frac{\partial^2 \log Z_{\theta,\lambda}}{\partial \theta \partial \lambda} = \text{cov}_{\theta,\lambda}(U, \Psi) \end{array} \right. . \quad (\text{B.6})$$

Cela implique en particulier que :

- $\mathbb{E}_{\theta,\lambda}[U]$ reste une fonction monotone décroissante de θ , quelle que soit la valeur de l'hyperparamètre λ .
 - A la limite $\lambda \rightarrow 0$, $\frac{\partial \mathbb{E}_{\theta,\lambda=0}[\Psi]}{\partial \theta} = -\left(\frac{\partial^2 \log Z_{\theta,\lambda}}{\partial \theta \partial \lambda}\right)_{\theta,\lambda=0} = -\text{cov}_{\theta,\lambda=0}(U, \Psi)$, i.e. :
- $$\frac{\partial \mathbb{E}_{\theta}[\Psi]}{\partial \theta} = -\text{cov}_{\theta}(U, \Psi) \quad \forall \text{ le potentiel } \Psi. \quad (\text{B.7})$$

Cela a également de nombreuses implications en Mécanique Statistique lorsque l'on étudie la réponse linéaire statistique d'un système à une perturbation externe en fonction de quantités décrivant la corrélation entre cette perturbation et l'énergie (interne) initiale du système [Ma(1985)].

On remarque également qu'en écrivant les équations précédentes à sous forme vectorielle et matricielle (c'est-à-dire le gradient et le hessien de $\log Z_{\theta,\lambda}$) :

$$\nabla \log Z_{\theta,\lambda} = - \begin{bmatrix} \mathbb{E}_{\theta,\lambda}[U] \\ \mathbb{E}_{\theta,\lambda}[\Psi] \end{bmatrix} \text{ et } \mathcal{H} = \begin{bmatrix} \text{var}_{\theta,\lambda}(U) & \text{cov}_{\theta,\lambda}(U, \Psi) \\ \text{cov}_{\theta,\lambda}(U, \Psi) & \text{var}_{\theta,\lambda}(\Psi) \end{bmatrix}$$

le hessien est une matrice définie positive, car de variance-covariance. La conséquence en est l'unimodalité de la vraisemblance, des vraisemblances locales (même raisonnement), et donc de la pseudo-vraisemblance quel que soit le nombre (fini) d'hyperparamètres.

Appendix C

Échantillonnage des distributions de Gibbs - recuit simulé

Dans cette partie on s'intéresse aux propriétés de convergence d'une suite d'échantilleurs de distributions de Gibbs associées à une suite $\{\theta_n\}$ de valeurs du paramètre. Nous suivons ici la présentation faite dans [Winkler(1995)].

1 Rappels sur les mesures, noyaux et le coefficient de Dobrushin

1.1 Distance entre mesures

Rappelons qu'une mesure de probabilité (que l'on appellera par commodité mesure dans la suite) sur Ω est une application $\mu : \Omega \mapsto \mathbb{R}$ telle que

- $\mu(x) \geq 0 \quad \forall x \in \Omega$
- $\sum_{x \in \Omega} \mu(x) = 1$

Maintenant la distance en variation entre deux mesures est définie par la relation :

$$\|\mu_1 - \mu_2\| = \sum_{x \in \Omega} |\mu_1(x) - \mu_2(x)| \tag{C.1}$$

On a (voir Fig. C.1):

- $0 \leq \|\mu_1 - \mu_2\| \leq 2$
- $\|\mu_1 - \mu_2\| = 0 \Leftrightarrow \mu_1 = \mu_2$
- $\|\mu_1 - \mu_2\| = 2 \Leftrightarrow \mu_1$ et μ_2 ont des supports disjoints

- $\|\mu_1 - \mu_2\| = 2 - 2 \sum_{x \in \Omega} \min(\mu_1(x), \mu_2(x))$
en effet : $\forall a, b \geq 0 \Rightarrow |a - b| = a + b - 2 \min(a, b)$

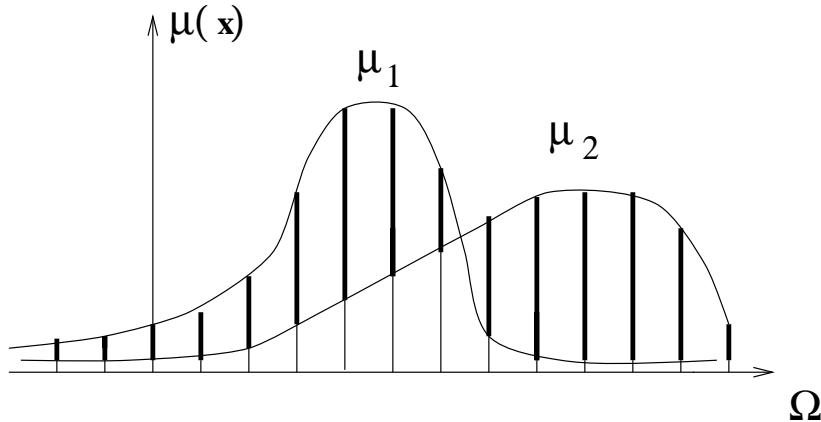


Figure C.1: Distance en variation entre deux mesures

1.2 Noyaux

Un noyau Q est une application $\Omega \times \Omega \mapsto \mathbb{R}$ telle que $Q(x, .)$ est une mesure $\forall x \in \Omega$. On peut donc aussi dire qu'il s'agit d'une famille de mesures sur Ω indiquée par un élément de Ω lui-même ! (voir Fig. C.2).

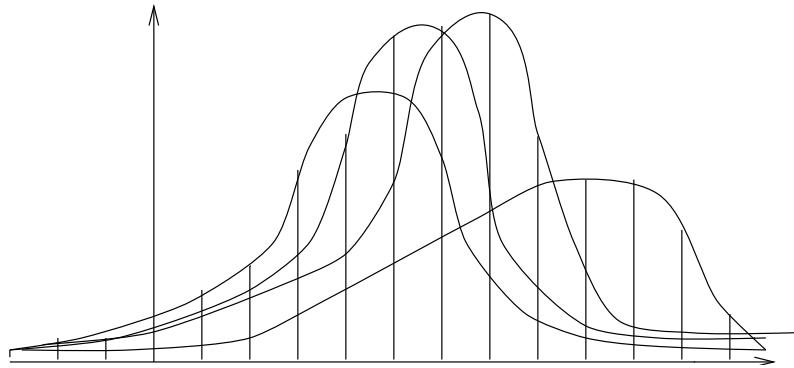


Figure C.2: Un noyau de probabilité

On peut alors définir maintenant :

- le produit d'une mesure par un noyau : c'est la mesure définie par :

$$(\mu Q)(x) = \sum_{y \in \Omega} \mu(y) Q(y, x) \quad \forall x \in \Omega \quad (\text{C.2})$$

- produit de deux noyaux : c'est le noyau défini par :

$$(Q R)(x, y) = \sum_{z \in \Omega} Q(x, z) R(z, y) \quad \forall x, y \in \Omega \quad (\text{C.3})$$

L'interprétation usuelle des noyaux est se fait en termes de probabilités conditionnelles, c'est-à-dire de probabilités de transition en théorie des chaînes de Markov :

$$Q(x, y) = \pi(Y = y / X = x) = \pi(x \rightarrow y)$$

où $\pi(\cdot)$ est une mesure sur Ω . Il résulte aussi des définitions précédentes que l'on a, en tant que mesures :

$$(Q \ R)(x, \cdot) = Q(x, \cdot)R \quad \forall x \in \Omega \quad (\text{C.4})$$

ce qui se révèlera utile pour la suite. On aurait pu d'ailleurs définir le produit de deux noyaux en partant de cette formule.

1.3 Coefficient de contraction de Dobrushin

Ce concept puissant résulte de la définition suivante :

$$c(Q) = \frac{1}{2} \max_{x,y \in \Omega} \|Q(x, \cdot) - Q(y, \cdot)\|$$

Les propriétés immédiates résultant de sa définition sont :

$$\begin{cases} 0 \leq c(Q) \leq 1 \\ c(Q) = 1 - \min_{x,y \in \Omega} \left(\sum_{z \in \Omega} \min(Q(x, z), Q(y, z)) \right) \leq 1 - |\Omega| \min_{a,b \in \Omega} Q(a, b) \end{cases}$$

En particulier, supposons qu'il existe deux valeurs x et $y \in \Omega$ telles que les mesures $Q(x, \cdot)$ et $Q(y, \cdot)$ ont des supports disjoints : considérons-les par exemple comme des probabilités de transition $\pi(x \rightarrow \cdot)$ et $\pi(y \rightarrow \cdot)$. Alors $c(Q) = 1$. Au contraire si toutes les mesures $Q(x, \cdot)$ sont très "semblables", $c(Q) \approx 0$. Dans le cas usuel, $0 < c(Q) < 1$ (voir Fig. C.3).

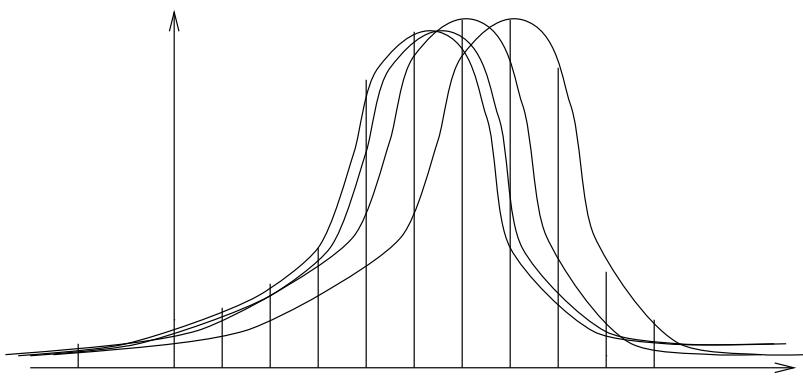
Le coefficient de Dobrushin mesure donc la "dispersion" (resp. l'homogénéité) des lois de probabilités d'un noyau. Tout le problème de l'échantillonnage des distributions de Gibbs va en fait être lié à la dispersion des différentes lois de transitions associées à un noyau donné..

Maintenant les propriétés fondamentales de ce coefficient de contraction de Dobrushin, qui sont à la base des propriétés de convergence des échantillonneurs de Gibbs et de Metropolis ainsi que du recuit simulé, sont les suivantes :

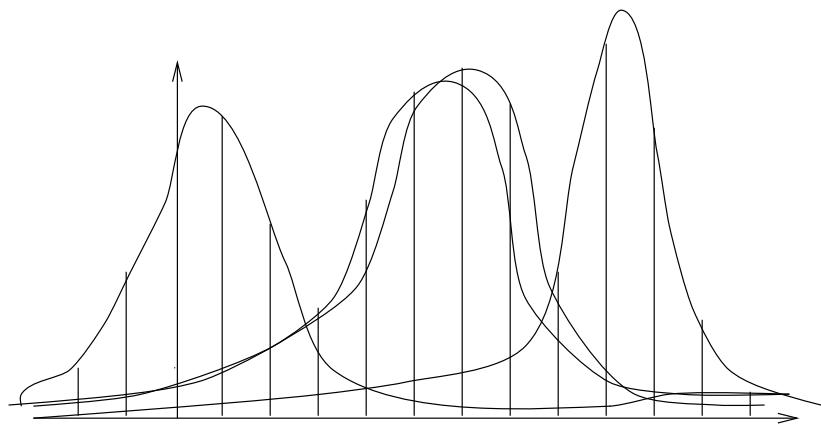
- a) $\|\mu_1 Q - \mu_2 Q\| \leq c(Q) \|\mu_1 - \mu_2\|$
- b) $c(Q \ R) \leq c(Q) c(R)$

On peut voir que la première inégalité doit effectivement être prise au sens large. Prenons en effet à titre d'exemple : $\mu_1 = \delta_a$ et $\mu_2 = \delta_b$, mesures de masse 1 aux configurations respectives **distinguishées** a et b . On a alors

$$\begin{aligned} \mu_1 Q(x) &= \sum_{y \in \Omega} \mu(y) Q(y, x) = Q(a, x) , \quad \text{c'est-à-dire } \mu_1 Q = Q(a, \cdot) \\ &\quad (\text{resp. } \mu_2 Q(x) = Q(b, x) , \quad \text{c'est-à-dire } \mu_2 Q = Q(b, \cdot)) . \end{aligned}$$



Mesures "voisines" : $c(Q) \approx 0$



Deux mesures ou plus a supports disjoints : $c(Q) \approx 1$



Cas "normal" : $0 < c(Q) < 1$

Figure C.3: Coefficient de contraction de Dobrushin : trois cas

Il en résulte

$$\begin{aligned} \|\mu_1 Q - \mu_2 Q\| &= \sum_{x \in \Omega} |Q(a, x) - Q(b, x)| = \|Q(a, \cdot) - Q(b, \cdot)\| \\ &\leq \max_{x, y \in \Omega} \|Q(x, \cdot) - Q(y, \cdot)\| = 2 c(Q) = \|\mu_1 - \mu_2\| c(Q) \end{aligned}$$

L'égalité est atteinte lorsque a et b sont les configurations parmi celles qui réalisent le maximum de $\|Q(a, \cdot) - Q(b, \cdot)\|$ (elles existent puisque Ω est fini). De la même façon on peut généraliser au cas où les deux mesures μ_1 et μ_2 ne diffèrent qu'en les points a et b , c'est-à-dire :

$$\mu_1(x) = A_1 \delta_a(x) + B_1 \delta_b(x) + \nu(x), \quad \mu_2(x) = A_2 \delta_a(x) + B_2 \delta_b(x) + \nu(x) \text{ avec } \nu(x) = \sum_{c \in \Omega, c \neq a, b} \lambda_c \delta_c(x)$$

On a donc $A_1 + B_1 = A_2 + B_2$ ($= 1 - \sum_{c \in \Omega, c \neq a, b} \lambda_c$), d'où il résulte :

- $\mu_1 - \mu_2 = (A_1 - A_2) \delta_a + (B_1 - B_2) \delta_b = (A_1 - A_2) (\delta_a - \delta_b)$
 $\Rightarrow \|\mu_1 - \mu_2\| = \sum_{x \in \Omega} |\mu_1(x) - \mu_2(x)| = 2 |A_1 - A_2| = 2 |B_1 - B_2|$
- $\mu_1 Q - \mu_2 Q = (A_1 - A_2) Q(a, \cdot) + (B_1 - B_2) Q(b, \cdot) = (A_1 - A_2) (Q(a, \cdot) - Q(b, \cdot))$
 $\Rightarrow \|\mu_1 Q - \mu_2 Q\| = |A_1 - A_2| \|Q(a, \cdot) - Q(b, \cdot)\|$

D'où bien sûr aussi

$$\|\mu_1 Q - \mu_2 Q\| \leq \|\mu_1 - \mu_2\| c(Q),$$

et ceci **quels que soient** les poids A_1, A_2, B_1, B_2 compris entre 0 et 1 vérifiant $A_1 + B_1 = A_2 + B_2 \leq 1$. L'égalité entre normes en variation précédente est alors également atteinte lorsque a et b sont les configurations qui réalisent le maximum de $\|Q(a, \cdot) - Q(b, \cdot)\|$. Nous renvoyons le lecteur intéressé par une élégante démonstration dans le cas général de la propriété **a)** à [Winkler(1995)]. On peut aussi remarquer que la deuxième propriété **b)** est une conséquence presque immédiate de la première **a)**, car d'après l'eq. C.4 :

$$\forall x, y \in \Omega \quad (QR)(x, \cdot) - (QR)(y, \cdot) = Q(x, \cdot) R - Q(y, \cdot) R$$

$$\begin{aligned} \text{D'où : } \forall x, y \in \Omega \quad &\|(QR)(x, \cdot) - (QR)(y, \cdot)\| \leq c(R) \|Q(x, \cdot) - Q(y, \cdot)\| \\ \Rightarrow \frac{1}{2} \max_{x, y \in \Omega} &\|(QR)(x, \cdot) - (QR)(y, \cdot)\| \leq \frac{1}{2} \max_{x, y \in \Omega} \|Q(x, \cdot) - Q(y, \cdot)\| c(R) \quad \text{CQFD} \end{aligned}$$

Une fois ceci démontré (ou admis), la plupart des propriétés de l'échantillonnage et du recuit en résultent facilement comme nous allons essayer de le montrer ...

1.4 Mesures invariantes

Ce sont les mesures qui vérifient la propriété suivante

$$(\mu Q)(x) = \mu(x) \quad \forall x \in \Omega$$

pour un noyau Q donné. Elles peuvent être considérées comme des “vecteurs propres” du noyau Q (vu en tant que matrice stochastique) associés à la valeur propre 1.

1.5 Réversibilité

Une mesure est réversible si elle vérifie la propriété suivante

$$\mu(x) Q(x, y) = \mu(y) Q(y, x) \quad \forall x, y \in \Omega$$

pour un noyau Q donné. En sommant cette formule sur y et en se rappelant la définition précédente du produit mesure-noyau, on en déduit que μ est mesure invariante pour Q . On peut interpréter la réversibilité grâce aux probabilités de transition. Le théorème de Bayes implique que pour toute distribution π :

$$\pi(X = x) \pi(Y = y / X = x) = \pi(Y = y) \pi(X = x / Y = y) \quad \forall x, y \in \Omega$$

i.e. $\pi(X = x) \pi(x \rightarrow y) = \pi(Y = y) \pi(y \rightarrow x)$. Cela nous montre et (nous réconforte !) que la mesure π est réversible, et donc invariante, p.r. à sa propre probabilité de transition $\pi(\cdot \rightarrow \cdot)$.

1.6 Irréductibilité

Cette propriété s'écrit :

$$Q(x, y) > 0 \quad \forall x, y \in \Omega, \quad \text{i.e. } c(Q) < 1$$

Le théorème de Perron-Frobenius assure alors **l'existence** et **l'unicité** d'une mesure invariante associée à un noyau Q **irréductible**, c'est-à-dire dont tous les coefficients sont **strictement positifs** (cf. [Winkler(1995)]). Nous allons alors construire une **chaîne de Markov** “d’images” dont la probabilité de transition à chaque étape est régie par un noyau Q irréductible et réversible par rapport à une mesure donnée (voir Fig. C.4). Nous allons montrer la convergence dans un sens à préciser de la probabilité d'état à chaque étape (qui est le produit successif de ces noyaux) vers la mesure invariante associée donnée au départ. Réciproquement, une distribution étant donnée, il s’agit de “trouver” un ou plusieurs noyaux irréductibles l’admettant comme mesure invariante (et si possible réversibles par rapport à celle-ci). La chaîne de Markov associée à ce(s) noyau(x) converge dans ce cas vers la distribution donnée. Pour une distribution de Gibbs, ce seront donc les échantillonneurs basés sur la dynamique de Gibbs ou de Metropolis qui auront les propriétés désirables requises, et qui convergeront donc vers cette distribution lorsque le nombre d’étapes tend vers l’infini.

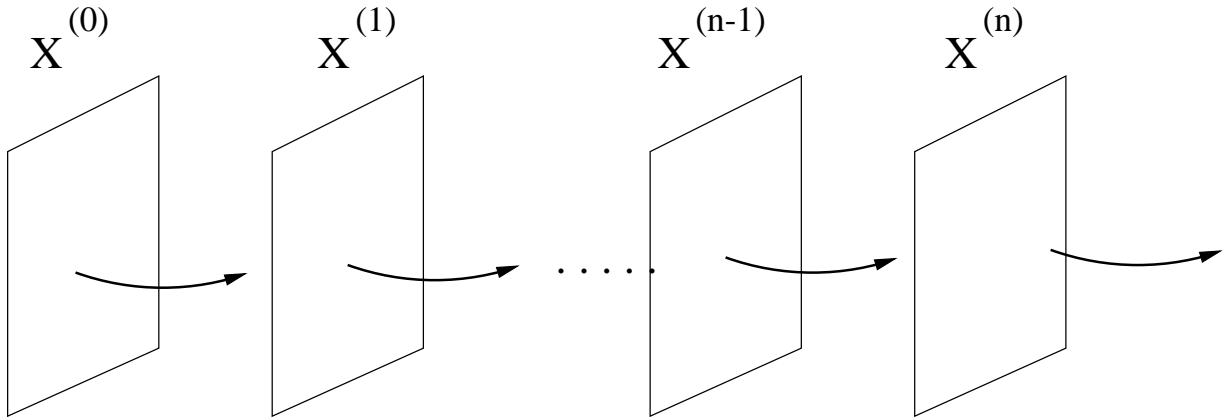


Figure C.4: Une chaîne de Markov d'images

2 Échantillonneurs de Gibbs et de Metropolis

Les deux types fondamentaux de noyaux de transition pour les distributions de Gibbs qui sont des champs de Markov, c'est-à-dire les échantillonneurs de Gibbs et de Metropolis, possèdent tous les deux la propriété désirable de réversibilité *p.r.* à cette distribution, comme nous allons le montrer.

2.1 Échantillonneur de Gibbs

Changeons la valeur de descripteur en un site s : $\xi \rightarrow \eta$. Il s'ensuit que la variation d'énergie totale est égale à la variation d'énergie conditionnelle locale, c'est-à-dire :

$\Delta U = \Delta U(\cdot / V_s)$ d'où il résulte :

$$U(x_s = \xi, x^s) + U(x_s = \eta / V_s) = U(x_s = \eta, x^s) + U(x_s = \xi / V_s)$$

et donc, pour toute valeur de θ :

$$\frac{\exp -\theta U(x_s = \xi, x^s)}{Z_\theta} \cdot \frac{\exp -\theta U(x_s = \eta / V_s)}{Z_\theta^s} = \frac{\exp -\theta U(x_s = \eta, x^s)}{Z_\theta} \cdot \frac{\exp -\theta U(x_s = \xi / V_s)}{Z_\theta^s}$$

Cela montre que le noyau suivant, aussi appelé échantillonneur de Gibbs au site s pour la valeur du paramètre θ :

$$Q_\theta^s(x, y) = \mathbf{1}_{x_r=y_r, r \neq s} \cdot P_\theta(y_s / V_s) = \mathbf{1}_{x_r=y_r, r \neq s} \cdot \frac{\exp -\theta U(y_s / V_s)}{Z_\theta^s}$$

est réversible *p.r.* la distribution de Gibbs

$$P_\theta(x) = \frac{\exp -\theta U(x)}{Z_\theta} ,$$

mais non irréductible, car la seule transition autorisée concerne deux configurations x et y ne différant qu'en un seul site. La propriété d'irréductibilité va apparaître lorsque l'on va

maintenant définir une procédure de balayage de l'ensemble du réseau de sites S .

Soit alors T un arrangement (c'est-à-dire une permutation) de S , que l'on appelle aussi tour. Etant donné deux configurations x et y connues, l'unique façon de "passer" de x à y en suivant l'ordre du tour T est alors :

$$\begin{aligned} Q_\theta(x, y) &= P(Y = y / X = x) \\ &= \prod_{s \in T} P_\theta(y_s / V_s) = \prod_{s \in T} Q_\theta^s(x, y) \end{aligned}$$

ce qui correspond au produit site après site des probabilités conditionnelles locales sachant l'état instantané de chacun des voisinages locaux (voir Fig. C.5).

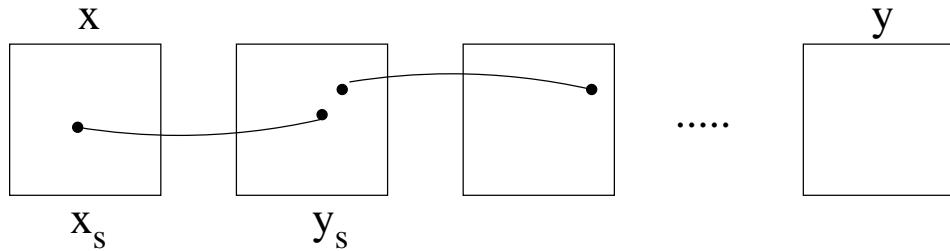


Figure C.5: un tour

Remarquons que la transition entre les configurations x et y données est alors certaine et non pas correspondant au produit "matriciel" des noyaux de transition locaux, car un seul "chemin" menant de x à y a été pris en compte. De plus, vu l'unicité de ce chemin, les fonctions indicatrices de la forme $\mathbf{1}_{\dots r \neq s}$ ont maintenant disparu dans le produit total. Le noyau de transition global $Q_\theta(x, y)$ est donc bien strictement positif, c'est-à-dire **irréductible** et **réversible** car produit simple de termes réversibles.

Dans la suite, nous aurons besoin de majorer le coefficient de contraction de Dobrushin associé à ce noyau. Cela se fait simplement en remarquant que

$$\forall s \in S \quad P_\theta(y_s / V_s) = \frac{1}{\sum_{\xi \in E} \exp \theta [U(y_s / V_s) - U(\xi / V_s)]} \geq \frac{1}{|E|} \exp -\theta \delta_s$$

où

$$\delta_s = \max_{V_s} (\max_{\eta \in E} U(\eta / V_s) - \min_{\xi \in E} U(\xi / V_s))$$

est l'**oscillation** de l'énergie conditionnelle locale $U(\cdot / \cdot)$. On a donc :

$$Q_\theta(x, y) \geq \prod_{s \in T} \left(\frac{1}{|E|} \exp -\theta \delta_s \right) \quad \forall x, y \in \Omega$$

d'où il résulte :

$$\begin{aligned} c(Q_\theta) &\leq 1 - |\Omega| \frac{1}{|E|^{|S|}} \exp (-\theta \cdot \sum_{s \in S} \delta_s) \quad \text{c'est-à-dire} \\ c(Q_\theta) &\leq 1 - \exp (-\Gamma \theta) \quad \text{avec } \Gamma = \sum_{s \in S} \delta_s \end{aligned} \tag{C.5}$$

2.2 Échantillonneur de Metropolis

Supposons pour des raisons physiques que l'on propose

$$(\alpha) \quad Q_\theta(x, y) = 1 \text{ si } U(y) < U(x)$$

c'est-à-dire que la transition de x vers y est certaine lorsque énergie totale décroît. Cela semble raisonnable dans la mesure où les configurations d'énergie faible sont plus probables que celles d'énergie élevée. Il est alors possible de trouver la nature de $Q_\theta(x, y)$ respectant la propriété de réversibilité quels que soient x et y ,

$$P_\theta(x) Q_\theta(x, y) = P_\theta(y) Q_\theta(y, x)$$

Ainsi, pour $U(y) > U(x)$, on a par spécification $Q_\theta(y, x) = 1$, et il s'ensuit

$$(\beta) \quad Q_\theta(x, y) = \frac{P_\theta(y)}{P_\theta(x)} = \exp -\theta (U(y) - U(x)) \text{ si } U(y) > U(x)$$

Toutefois, on s'aperçoit que nombre de transitions y donnent lieu à la valeur $Q_\theta(x, y) = 1$, ce semble peu cohérent avec la théorie des probabilités ! On va en fait pour cela revenir, comme précédemment, au niveau local. On se donne pour cela une loi de choix de la nouvelle valeur de descripteur en un site donné, appelée aussi probabilité d'acceptance :

$$R^s(x, y) = R(x_s, y_s)$$

On suppose que c'est un noyau, et qui est de plus une fonction symétrique de ses arguments : on prend ainsi usuellement $R(x_s, y_s) = \frac{1}{|E|}$, c'est-à-dire la loi équirépartie sur les valeurs de descripteurs possibles. On teste ensuite s'il y a diminution de l'énergie ou non pour cette nouvelle valeur, de sorte que le véritable noyau de transition local s'écrit :

$$S_\theta^s(x, y) = \mathbf{1}_{x_r=y_r, r \neq s} \cdot R^s(x, y) Q_\theta(x, y)$$

Il est bien réversible puisque R est supposé symétrique (voir Fig. C.6).

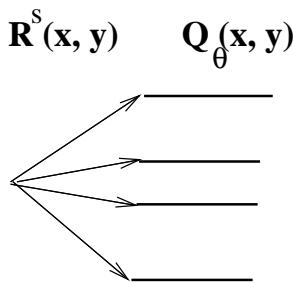


Figure C.6: diagramme de transition pour l'échantillonneur de Metropolis

On définit comme précédemment un tour T du réseau de sites. Le noyau de transition total s'écrit en définitive :

$$S_\theta(x, y) = \prod_{s \in T} S_\theta^s(x, y) = \prod_{s \in T} [R^s(x, y) Q_\theta(x, y)]$$

Il possède donc comme l'échantillonneur de Gibbs les propriétés de **réversibilité p.r.** la distribution de Gibbs $P_\theta(x)$ en vertu de la propriété de symétrie de la fonction R (vérifier), et d'**irréductibilité p.r.** la distribution de Gibbs $P_\theta(x)$.

Nous aurons besoin comme précédemment de minorer les valeurs du noyau de transition ainsi trouvé. Dans le cas où la loi d'acceptance est la loi équidistribuée sur E ,

$$\begin{aligned} S_\theta(x, y) &\geq \frac{1}{|\Omega|} \exp -\theta \Delta \quad \text{avec } \Delta = \max_{y \in \Omega} U(y) - \min_{x \in \Omega} U(x) \\ \Rightarrow c(S_\theta) &\leq 1 - \exp(-\Delta \theta) \end{aligned} \tag{C.6}$$

En résumé pour les deux échantillonneurs, on obtient une majoration du type :

$$c(Q_\theta) \leq 1 - \exp(-\Gamma \theta)$$

Seule la constante Γ dépend de l'échantillonneur étudié.

3 Échantillonnage homogène des distributions de Gibbs : convergence

On se fixe une valeur du paramètre θ et l'on se donne un noyau Q_θ connu (par exemple l'un des deux précédents). Rappelons qu'avec l'hypothèse d'irréductibilité,

$$c(Q_\theta) \leq 1 - \exp(-\Gamma \theta) \quad \text{où } \Gamma > 0$$

Donnons-nous également une mesure initiale μ_0 . On construit comme indiqué précédemment en paragraphe 1.6 une chaîne de Markov de configurations $X^{(n)}$ basée sur le mode de transition entre étapes consécutives. On peut alors calculer la probabilité d'obtenir la configuration x à l'étape n , notée $\pi_n(x)$ en fonction des probabilités de transition intermédiaires, c'est-à-dire du noyau Q_θ :

$$\begin{aligned} \pi_n(x) = P(X^{(n)} = x) &= \sum_{y \in \Omega} P(X^{(n)} = x / X^{(n-1)} = y) P(X^{(n-1)} = y) \\ &= \sum_{y \in \Omega} \pi_{n-1}(y) Q_\theta(y, x) \\ &= \pi_{n-1} Q_\theta(x) = \pi_{n-2} Q_\theta(x)^2 \\ &= \mu_0 Q_\theta^n(x) \text{ par récurrence} \\ \text{c'est-à-dire } \pi_n &= \mu_0 Q_\theta^n \quad \forall n \geq 1 \end{aligned}$$

Rappelons que la mesure invariante (unique) P_θ est telle que $P_\theta = P_\theta Q_\theta^n \quad \forall n \geq 1$.

Cela entraîne que

$$\|\pi_n - P_\theta\| = \|\mu_0 Q_\theta^n - P_\theta Q_\theta^n\| \leq \|\mu_0 - P_\theta\| c(Q_\theta)^n$$

Comme $c(Q_\theta) < 1$ et $\theta < +\infty$, la convergence d'après la norme en variation (appelée aussi convergence en variation) de π_n vers P_θ résulte immédiatement, **quelle que soit** la distribution initiale μ_0 . On peut par exemple prendre $\mu_0 = \delta_{x_0}$, mesure “certaine” en une configuration x_0 donnée.

4 Échantillonnage inhomogène des distributions de Gibbs

Nos hypothèses et notations sont les suivantes :

- une suite $\theta_n \rightarrow \theta$
- une loi de probabilité initiale sur Ω , notée μ_0
- une suite de mesures invariantes notées $P_n = P_{\theta_n}(\cdot)$
- la génération d'échantillons avec des noyaux de transition notés $Q_n = Q_{\theta_n}(\cdot, \cdot)$

A chaque étape n

$$P(X^{(n)} = x) = \pi_n(x) = \mu_0 Q_1 Q_2 \dots Q_n(x)$$

Comme $P_n = P_n Q_n$ pour ≥ 1 puisque mesure invariante associée au noyau Q_n , on peut calculer la différence suivante, (il s'agit du lemme d'Abel voir aussi [Winkler(1995)]) :

$$\begin{aligned} P_n - \pi_n &= P_n - \mu_0 Q_1 Q_2 \dots Q_n = P_n Q_n - \mu_0 Q_1 Q_2 \dots Q_{n-1} Q_n \\ &= (P_n - P_{n-1}) Q_n + (P_{n-1} - \pi_{n-1}) Q_n \\ &\quad \text{et par récurrence} \\ &= (P_n - P_{n-1}) Q_{n-1} + (P_{n-1} - P_{n-2}) Q_{n-1} Q_n \\ &\quad + (P_{n-2} - P_{n-3}) Q_{n-2} Q_{n-1} \dots Q_n \dots \\ &\quad + (P_{n-p+1} - P_{n-p}) Q_{n-p+1} \dots Q_n + (P_{n-p} - \pi_{n-p}) Q_{n-p+1} \dots Q_n \end{aligned} \quad (\text{C.7})$$

On va distinguer dans cette somme entre les p premiers termes et le dernier terme. Comme chacun des coefficients $c(Q_{n-m}) \leq 1$ (p premiers termes), et comme $P_{n-p} - \pi_{n-p}$ est borné (dernier terme), on voit qu'un ensemble de conditions suffisantes de convergence de la chaîne de Markov inhomogène ainsi formée vers la mesure invariante P_θ de paramètre θ est :

1. $\sum_k \|P_k - P_{k-1}\| < +\infty$
2. $\forall p \geq 1, c(Q_{n-p} Q_{n-p+1} \dots Q_n) \rightarrow 0$ quand $n \rightarrow +\infty$

La condition 1 assure d'une part la convergence des p premiers termes vers 0 pour $n - p$ et n assez grands puisque la série $\sum_k \|P_k - P_{k-1}\|$ suit le critère de Cauchy :

$$\sum_{k=q}^m \|P_k - P_{k-1}\| < \epsilon \text{ dès que } q, m > N$$

De plus, comme $\|P_m - P_q\| \leq \sum_{k=q}^m \|P_k - P_{k-1}\|$, la suite $\{P_n\}$ est elle-même de Cauchy, et possède donc une limite qui est nécessairement P_θ . En définitive, $\|P_\theta - \pi_n\| \leq \|P_\theta - P_n\| + \|P_n - \pi_n\| \rightarrow 0$ quand $n \rightarrow +\infty$.

Il reste à examiner la signification physique de la condition 2 ainsi qu'à trouver son domaine de validité.

4.1 Cas du recuit simulé : $\theta_n \rightarrow \theta = +\infty$

Le deuxième ensemble de conditions est plus simple à montrer dans ce cas car, comme on l'a vu précédemment (annexe A) $\forall x \in \Omega$, $P_n(x)$ devient ou bien décroissante ou bien croissante à partir d'un certain rang lorsque θ_n augmente. La première condition est plus compliquée à satisfaire, car $c(Q_n) \rightarrow 1$ lorsque $n \rightarrow +\infty$. Une condition suffisante est que le produit infini associé à sa majoration diverge¹, c'est-à-dire $\lim_{n \rightarrow +\infty} \prod_{k=1}^n (1 - \exp(-\Gamma \theta_k)) = 0$, ce qui conduit immédiatement aux conditions **suffisantes** de Geman et Geman [Geman and Geman(1984b)] :

$$\sum_{n=1}^{+\infty} \exp -\Gamma \theta_n = +\infty \quad (\text{C.8})$$

On obtient donc la condition suffisante sur le paramètre θ (resp. T) :

$$\Rightarrow \Gamma \theta_n \leq \frac{1}{\log n} \text{ ou en termes de température } T_n \geq \frac{\Gamma}{\log n}$$

On voit donc que le recuit simulé doit être conduit de façon lente pour converger vers l'optimum énergétique désiré. On pourrait se demander pourquoi ne pas effectuer un échantillonnage de la distribution de GIbbs à température basse dès le départ. C'est ce qu'on appelle la "trempe" : descente aléatoire directe dans le cas de la dynamique de Metropolis ou ICM dans le cas de la dynamique de Gibbs. En fait cf. paragraphe suivant la condition de convergence 2 devient beaucoup plus longue à atteindre.

4.2 Cas “normal” : $\theta_n \rightarrow \theta < +\infty$

Maintenant, la seconde condition devient triviale car

$$c(Q_m Q_{m+1} \dots Q_n) \leq \prod_{k=m}^n (1 - \exp(-\Gamma \theta_k)) \quad \Gamma > 0$$

Il reste donc à prouver la première condition, ce qui n'est pas si simple : il faut aller voir de plus près par exemple quand θ “oscille” autour de sa valeur limite.

¹On appelle paradoxalement divergent un produit infini dont la limite est 0. En effet il est alors associé à une série divergente:

$$u_n \in [0, 1[\text{ diverge} \Rightarrow \lim_{n \rightarrow +\infty} \prod_{k=1}^n (1 - u_k) = 0$$

Bibliography

- [Abend *et al.*(1965)] Abend, K., Harley, T. J., and Kanal, L. N. (1965). Classification of binary random pattern. *IEEE Transactions on Information Theory*, **11**, 538–544.
- [Aurdal(1997)] Aurdal, L. (1997). *Analyse d’images IRM 3D multi-échos pour la détection et la quantification de pathologies cérébrales*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications.
- [Azencott(1992)] Azencott, R. (1992). Markov field approach : parameter estimation by qualitative boxes. *Cours : Les Houches*.
- [Baum *et al.*(1970)] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, **41**, 164–171.
- [Besag(1974)] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Soc. (series B)*, **36**, 192–326.
- [Besag(1986)] Besag, J. (1986). On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B*, **48**(3), 259–302.
- [Bloch *et al.*(1997)] Bloch, I., Aurdal, L., Bijno, D., and Muller, J. (1997). Estimation of class membership functions for grey-level based image fusion. *ICIP’97 (Santa Barbara)*.
- [Bouman and Sauer(1993)] Bouman, C. and Sauer, K. (1993). A generalized gaussian image model for edge-preserving map estimation. *IEEE Transactions on Image Processing*, **2**(3), 296–310.
- [Cardoso and Moulines(1995)] Cardoso, J.-F. and Moulines, E. (1995). Asymptotic performance analysis of direction finding algorithms based on fourth-order cumulants. *IEEE Transactions on Signal Processing*, **43**(1), 214–224.
- [Chalmond(1989)] Chalmond, B. (1989). An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, **22**(6), 747–761.

- [Derin and Elliott(1987)] Derin, H. and Elliott, H. (1987). Modeling and segmentation of noisy and textured images using gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**(1).
- [Derin *et al.*(1985)] Derin, H., Elliott, H., and Kuang, J. (1985). A new approach to parameter estimation for gibbs random field. In *Int. Conf. on ASSP*.
- [Descombes(1993)] Descombes, X. (1993). *Champs Markoviens en analyse d'images*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications (ENST 93 E 026).
- [Descombes *et al.*(1995)] Descombes, X., Mangin, J. F., Pechersky, E., and Sigelle, M. (1995). Fine structures preserving Markov model for image processing. *The 9th Scandinavian Conference on Image Analysis (Uppsala, Sweden)*, **2**, 349–356.
- [Descombes *et al.*(1996)] Descombes, X., Morris, R., Zerubia, J., and Berthod, M. (1996). Estimation of markov random field prior parameters using markov chain monte carlo likelihood - accepted for publication in ieee transactions on image processing. Technical Report 3015, INRIA.
- [Geman and Geman(1984a)] Geman, S. and Geman, D. (1984a). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.
- [Geman and Geman(1984b)] Geman, S. and Geman, D. (1984b). Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741.
- [Geman and McClure(1985)] Geman, S. and McClure, D. E. (1985). Bayesian image analysis : an application to single photon emission tomography. *Proc. Statist. Comput. sect. (Amer. Statist. Assoc. Washington DC)*, pages 12–18.
- [Géraud *et al.*(1995)] Géraud, T., Mangin, J. F., Bloch, I., and H.Maître (1995). Segmenting internal structures in 3D MR images of the brain by Markovian relaxation on a watershed based adjacency graph. *IEEE ICIP (Austin)*, **III**, 548–552.
- [Graffigne(1987)] Graffigne, C. (1987). *Experiments in Texture Analysis and Segmentation*. Ph.D. thesis, Division of Applied Mathematics - Brown University.
- [Guyon(1992)] Guyon, X. (1992). *Champs aléatoires sur un réseau - modélisations, statistique et applications*. Collection Techniques Stochastiques, Masson.
- [Ising(1925)] Ising, E. (1925). Beitrag zur theorie des ferromagnetisms. *Zeitschrift fur Physik*, **31**, 253–258.
- [Kanal(1980)] Kanal, L. N. (1980). Markov mesh models. *Image Modeling, New York, Academic Press*.

- [Khounri(1997)] Khounri, M. (1997). Estimation d'hyperparamètres pour la déconvolution d'images satellitaires - rapport de stage dea. Technical report, INRIA Sophia.
- [Kirkpatrick *et al.*(1982)] Kirkpatrick, S., Gellatt, C. D., and Vecchi, M. P. (1982). Optimization by simulated annealing. *IBM Thomas J. Watson research Center, Yorktown Heights, NY*.
- [Lakshmanan and Derin(1989)] Lakshmanan, S. and Derin, H. (1989). Simultaneous parameter estimation and segmentation of gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(8), 799–813.
- [Landau and Lifschitz(1961)] Landau, L. and Lifschitz, E. (1961). *Cours de Physique Tome 5 - Physique Statistique*. Editions Mir.
- [Ma(1985)] Ma, S. (1985). *Statistical Mechanics*. World Scientific.
- [Métivier and Priouret(1987)] Métivier, M. and Priouret, P. (1987). Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and Related Fields*, **74**, 403–428.
- [Metropolis *et al.*(1953)] Metropolis, N., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1091.
- [Nikias and Petropulu(1993)] Nikias, C. L. and Petropulu, A. P. (1993). *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. PTR Prentice-Hall.
- [Onsager(1944)] Onsager, L. (1944). *Physical Review*, **65**, 117.
- [Pieczynski(1994)] Pieczynski, W. (1994). Hidden markov fields and iterative conditional estimation. *Traitement du Signal*, **11**(2), 141–153.
- [Rabiner(1989)] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.
- [Redner and Walker(1984)] Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, **26**, 195–239.
- [Saquib *et al.*(1998)] Saquib, S., Bouman, C., and Sauer, K. (1998). Ml parameter estimation for markov random fields with applications to bayesian tomography. *IEEE Transactions on Image Processing*, **7**(7), 1029–1044.
- [Sigelle(1993)] Sigelle, M. (1993). *Champs de Markov en traitement d'images et modèles de la physique statistique : applications en relaxation d'images de classification*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications.

- [Sigelle(1997)] Sigelle, M. (1997). Simultaneous image restoration and hyperparameter estimation for incomplete data by a cumulant analysis. Technical report, INRIA Sophia Antipolis.
- [Tupin(1997)] Tupin, F. (1997). *Reconnaissance des formes et analyse de scénes en imagerie radar à ouverture synthétique*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications.
- [Tupin *et al.*(1996)] Tupin, F., Trouvé, E., Descombes, X., Nicolas, J.-M., and Maître, H. (1996). Improving IFSAR phase unwrapping by early detection of non-interferometric features. *European Symposium on Satellite Remote Sensing III (Taormina, Italy)*.
- [Winkler(1995)] Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods. A Mathematical Introduction*. Applications of mathematics. Springer-Verlag.
- [Wu(1982)] Wu, F. Y. (1982). The potts model. *Reviews of Modern Physics*, **54**(1), 235–267.
- [Younes(1988)] Younes, L. (1988). Estimation and annealing for gibbsian fields. *A. Inst Henri Poincaré*, **24**(2), 269–294.
- [Younes(1989)] Younes, L. (1989). Parametric inference for imperfectly observed gibbsian fields. *Probability Theory and Related Fields*, **82**, 625–645.
- [Younes(1991)] Younes, L. (1991). Parameter estimation for imperfectly observed gibbs fields and some comments on chalmond’s em gibbsian algorithm. In *stochastic models, statistical methods and algorithms in image analysis*. P. Barone and A. Frigessi, Lecture Notes in Statistics, Springer.
- [Zerubia and Blanc-Féraud(1998)] Zerubia, J. and Blanc-Féraud, L. (1998). Hyperparameter estimation of a variational model using a stochastic gradient method. In *SPIE Bayesian Inference for Inverse Problems Proceedings*, volume 3459.