Optimisation et Méthodes de Monte Carlo : entrelacements
Apprentissage Statistique Computationnel

Gersende Fort

IMT & CNRS

Petit Séminaire "Statistique et Optimisation", Avril 2020.

Thèmes de recherche (1/3)

$$\text{argmin}_{\theta \in \Theta} \left(-\ell_n(\theta) + \mathrm{R}(\theta) \right)$$

- dimension finie sauf exceptions (collab. sur quantification d'incertitude) : $\Theta \subseteq \mathbb{R}^d$
- optimisation non convexe
- optimisation lisse ou non lisse
- Ex. : $-\ell_n(\theta)$ fonction de coût, traduisant l'(in)adéquation d'un modèle indexé par θ pour expliquer n observations exemples, données
- Ex. : $\mathrm{R}(\theta)$ traduit une régulatisation / une contrainte / un a priori (MAP en bayésien).

Thèmes de recherche (2/3) - spécificité 1

$$\text{argmin}_{\theta\in\Theta}\left(-\ell_n(\theta)+\mathsf{R}(\theta)\right)$$

- Quantités non explicites, ordre 0 comme ordre 1

$$\ell_n(\theta)=\log\int_{\mathsf{Z}}\exp\left(G(z,\theta)\right)\ \mu(\mathrm{d}z)$$

et si régulière

$$\nabla\ell_n(\theta)=\int_{\mathsf{Z}}\partial_\theta G(z,\theta)\ \underbrace{\frac{\exp\left(G(z,\theta)\right)}{\int_{\mathsf{Z}}\exp(G(u,\theta))\,\mathrm{d}\mu(u)}}_{\text{densité de probabilité}}\ \mu(\mathrm{d}z)$$

- Méthodes d'optimisation privilégiées

∗ Gradient stochastique et ses nombreuses variantes (accélérations, réduction de variance)

∗ Approximation stochastique: zeros du *champ moyen* $h(\theta)\overset{\text{def}}{=}\mathbb{E}\left[H(Z,\theta)\right]$

∗ Algorithmes Majoration-Minoration avec surrogate de type "espérance"

$$\text{Jensen}\qquad -\ell_n(\theta)\leq C(\theta_0)-\int_{\mathsf{Z}}G(z,\theta)\,\frac{\exp\left(G(z,\theta_0)\right)}{\int_{\mathsf{Z}}\exp(G(u,\theta_0))\,\mathrm{d}\mu(u)}\,\mathrm{d}\mu(z)$$

Thèmes de recherche (3/3) - spécificité 2

- Procédures de Monte Carlo dont "par Chaînes de Markov"

$*$ Comment échantillonnner *efficacement* "sous" une loi $\pi_\theta(\mathrm{d}z)$ connue à constante multiplicative près ?

$$\int H(z,\theta)\ \pi_\theta(\mathrm{d}z) \approx \frac{1}{M}\sum_{j=1}^{M} H(Z_j,\theta)$$

$*$ Efficacité : quantification explicite

$$\sup_{\theta\in\Theta}\ \mathbb{E}\left[\left|\frac{1}{M}\sum_{j=1}^{M} H(Z_j,\theta) - \int H(z,\theta)\ \pi_\theta(\mathrm{d}z)\right|^p\right] \leq \frac{C}{M^{p/2}}$$

$*$ Théorie des Chaînes de Markov : critères d'ergodicité $\rightarrow$ obtention de telles bornes

$*$ Stabilité des algorithmes stochastiques

$$\forall \mathcal{K}\subseteq\Theta,\quad \mathbb{E}\left[\left|\frac{1}{M}\sum_{j=1}^{M} H(Z_j,\theta) - \int H(z,\theta)\ \pi_\theta(\mathrm{d}z)\right|^p\ 1_{\theta\in\mathcal{K}}\right] \leq \frac{C}{M^{p/2}}$$

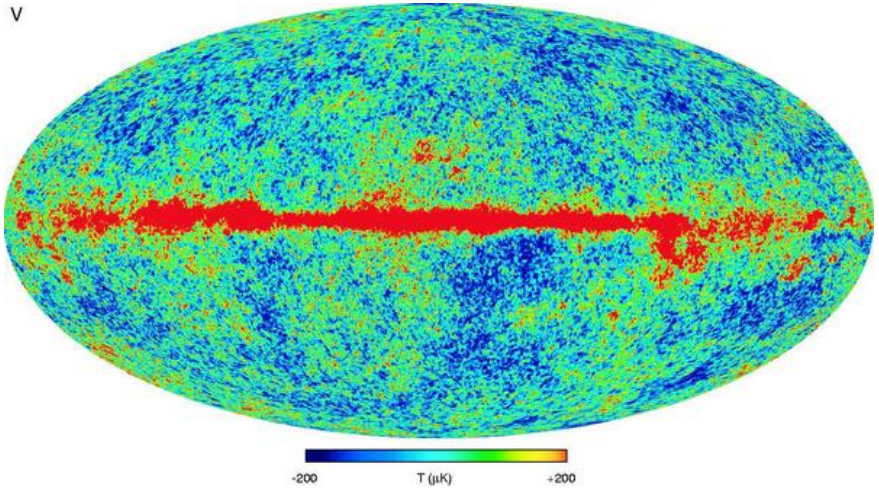
Application 1 - Estimation de paramètres cosmologiques

densité de matière noire, densité de matière baryonique, constante de Hubble, courbure spatiale de l'Univers

---

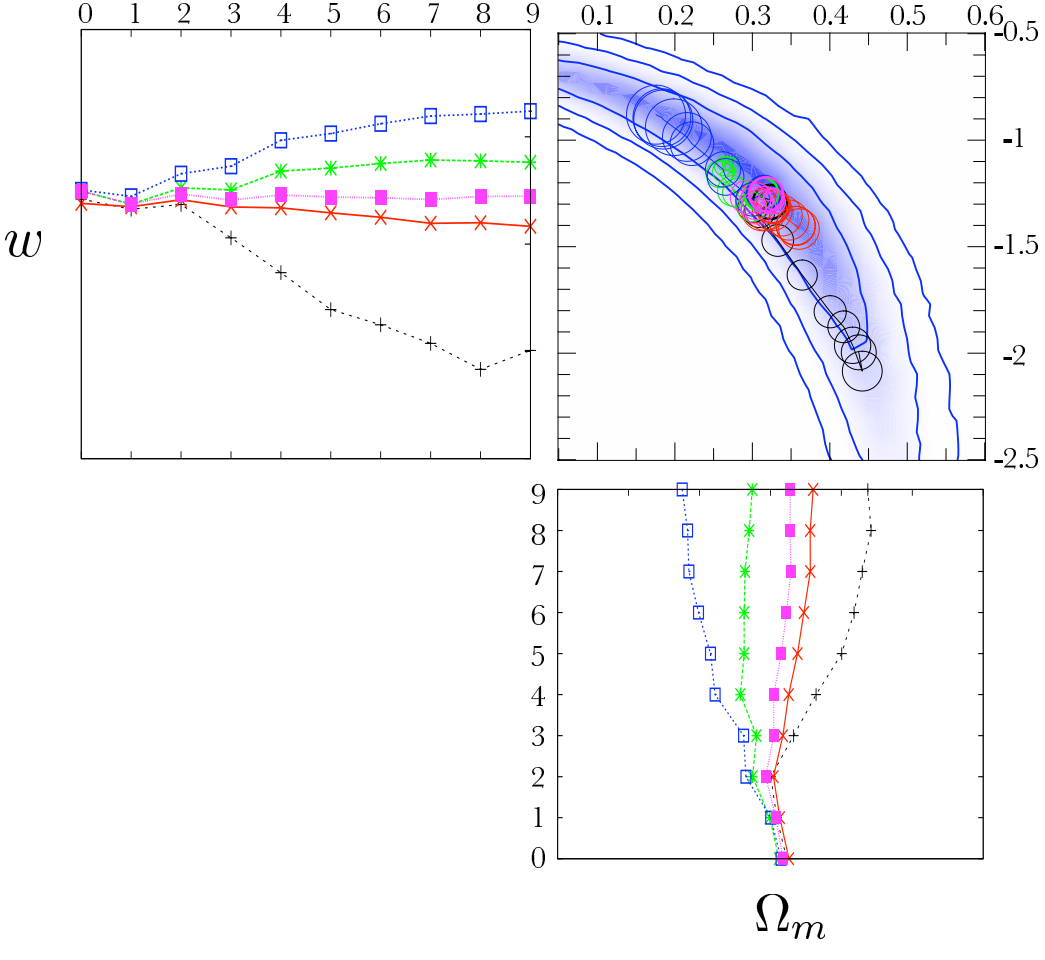
- Vraisemblance des observations : probabilité d'observer une carte de CMB → boîte noire numérique, un appel = plusieurs secondes

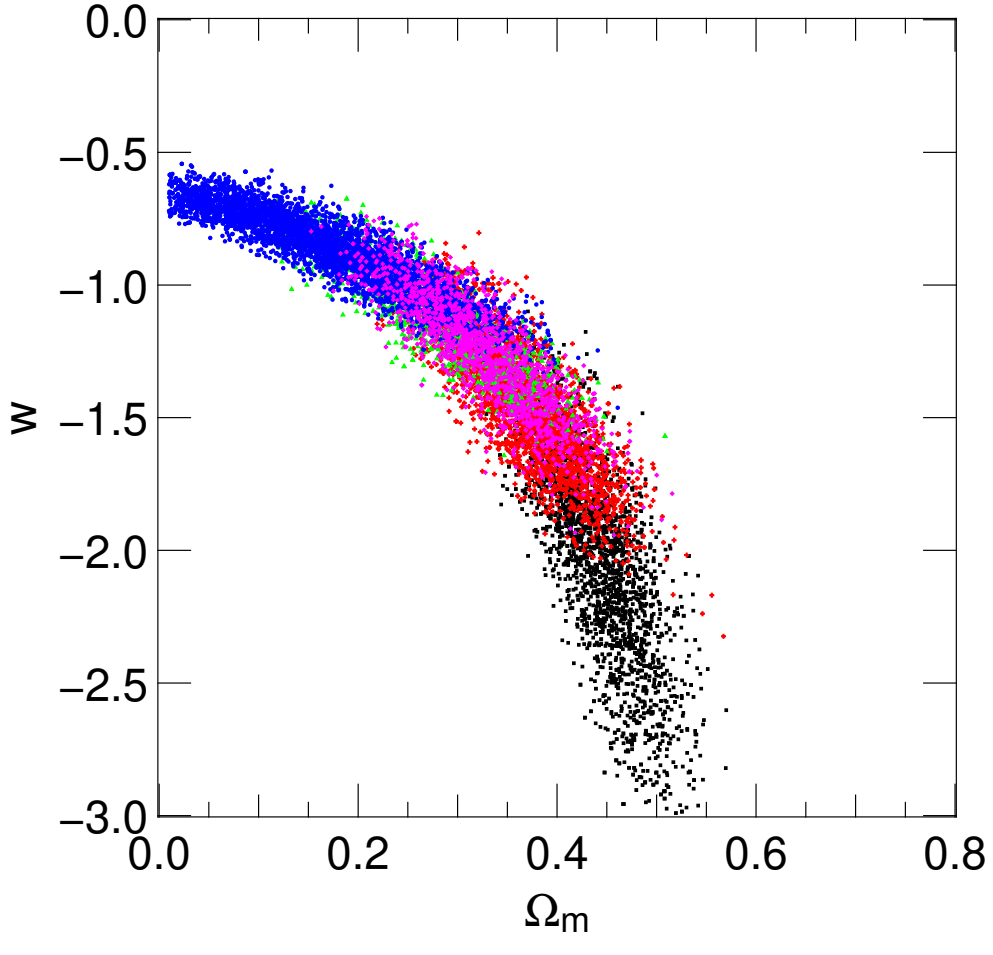
- Exploration de la loi a posteriori par Monte Carlo pour estimation → Procédure d'échantillonnage d'importance adaptatif



Example of survey: WMAP for the Cosmic Microwave Background (CMB) radiations = temperature variations are related to fluctuations in the density of matter in the early universe and thus carry out information about the initial conditions for the formation of cosmic structures such as galaxies, clusters and voids for example.

Ajuster un mélange de cinq gaussiennes à la loi a posteriori ($\rightarrow$ KL $\rightarrow$ EM)



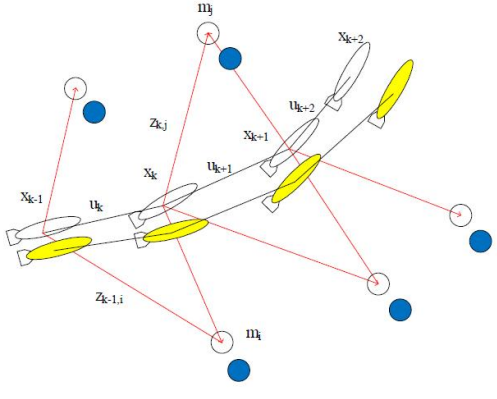


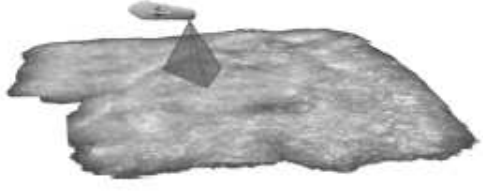
Application 2 - SLAM (cartographie et localisation simultanées)

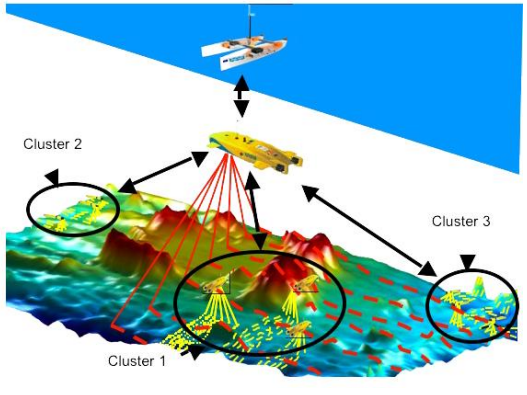
- Traitement bayésien séquentiel dans les modèles de chaînes de Markov cachées :

∗ Identifier la position d'amers (cartographie) et la position d'un robot mobile (localisation) à partir d'une odométrie de qualité moyenne → paramètres + chaîne de Markov cachée

∗ Estimation dans un modèle à **données cachées**, Traitement **en ligne** (et distribué, pas ici) de l'information





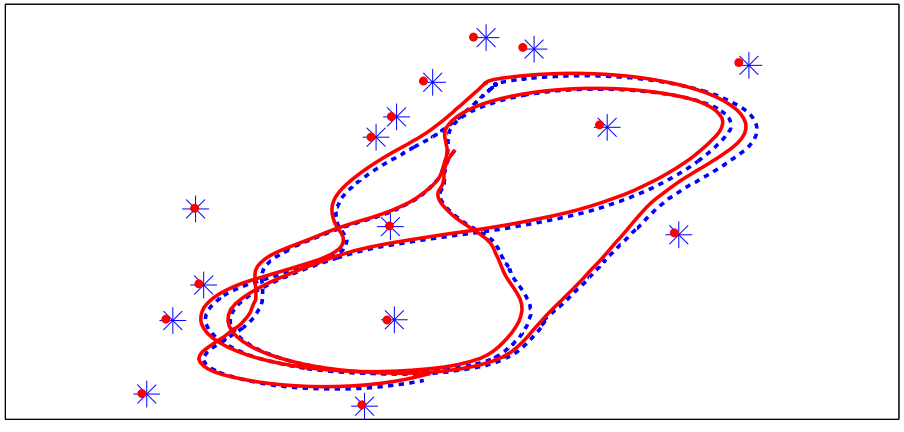


- Traitement en mini-batch des observations (flux) $\rightarrow$ proposition d'un algorithme.

- Algorithme EM en ligne : algorithme majoration-minoration pour **famille de fonctions objectif**.

- doublé d'une approximation stochastique par **filtrage particulaire**

- Preuve de convergence de l'algorithme d'optimisation i.e. convergence vers l'estimateur MV.



Et en ce moment ? (1/4)

- Apprentissage "grande échelle"

$$\text{argmin}_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\theta) + R(\theta) \right),$$

lorsque

$$\mathcal{L}_i(\theta) = -\log \int h_i(z) \ \exp(\langle s_i(z), \phi(\theta) \rangle) \ \mathrm{d}\mu(z)$$

- Optimisation non convexe, lisse (gradient Lipschitz), "somme finie" i.e. ne **pas calculer la somme sur n termes**

- Objectifs: (i) algorithme d'optimisation, (ii) contrôle de convergence **explicite** et **non asymptotique**

- Exemple : classification / régression logistique

$$-\mathcal{L}_i(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \int_{\mathbb{R}^d} \frac{\exp(y_i \langle x_i, z \rangle)}{1 + \exp(\langle x_i, z \rangle)} \ \exp(-0.5 \|z - \theta\|^2) \ \mathrm{d}z$$

Et en ce moment ? (2/4) quel algorithme

$$\text{argmin}_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\theta)+R(\theta)\right), \qquad \mathcal{L}_i(\theta)=-\log\int_{\mathsf{Z}} h_i(z)\ \exp(\langle s_i(z),\phi(\theta)\rangle)\ \mathrm{d}\mu(z)$$

- Algorithme de type Majoration-Minoration via inégalité de Jensen (KL)

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\theta)\leq C(\theta^k)-\frac{1}{n}\sum_{i=1}^{n}\int_{\mathsf{Z}}\langle s_i(z),\phi(\theta)\rangle\ \ \pi_{\theta^k}(\mathrm{d}z)$$

- Doublé d'une procédure qui évite le calcul de **somme** à chaque itération, en ne sélectionnant qu'un exemple par itération du MM

- et redoublé d'une procédure Monte Carlo pour approcher la fonction surrogate.

Et en ce moment ? (3/4) quelles bornes de convergence

$$\text{argmin}_{\theta\in\Theta}\left(F(\theta)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\theta)+R(\theta)\right),\qquad \mathcal{L}_i(\theta)=-\log\int_{\mathsf{Z}}h_i(z)\ \exp(\langle s_i(z),\phi(\theta)\rangle)\ \mathrm{d}\mu(z)$$

- Problème non convexe: comment définir la convergence ?

$$\mathbb{E}\left[\|\nabla F(\theta^{K})\|^2\right]\leq\frac{n^{\mathsf{a}}}{K_{\mathsf{max}}^{\mathsf{b}}}\ \mathcal{B}\ \left(\mathbb{E}\left[F(\theta^{0})\right]-\min F\right)$$

pour un **‘"termination rule"** K aléatoire, entre 0 et K_{max}.

- Bornes de complexité pour un ϵ-*stationary point* :

$$K_{\mathsf{max}}=O\left(n^{\mathsf{a}/\mathsf{b}}\epsilon^{-1/\mathsf{b}}\right)$$

Et en ce moment ? (4/4)

- Dans le détail: algorithmes, état de l'art, résultats théoriques, applications
- Un de ces jours en séminaire $\cdots$
- Travail avec P. Gach (IMT) et E. Moulines (CMAP).