Géométrie Algorithmique
Données, Modèles, Programmes

8. Géométrie des données

Jean-Daniel Boissonnat

Collège de France
31 mai 2017

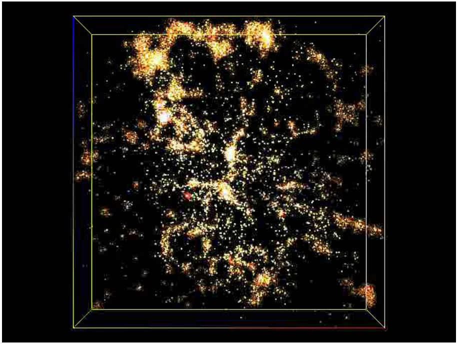
Géométrie algorithmique

Données, modèles, programmes

1. Modèles géométriques discrets
   F. Cazals : Modèles géométriques pour la prédiction des interactions macro-moléculaires
2. La puissance de l'aléa : algorithmes randomisés
   P. Calka : Probabilités géométriques
3. Le calcul géométrique
   S. Pion : La bibliothèque logicielle CGAL
4. Génération de maillages
   J-M. Mirebeau : Les deux réductions de Voronoï et leur application aux équations aux dérivées partielles
5. Courbes et surfaces
   P. Alliez : Reconstruction de surfaces
6. Espaces de configurations
   A. de Mesmay : Dessin de graphes
7. Structures de données géométriques
   D. Feldman : Core sets
8. Géométrie des données
   F. Chazal : Analyse topologique des données

Analyse géométrique des données

Images, texte, voix, signaux neuronaux, traces GPS,...



Géométrisation : Données = points + distances entre points

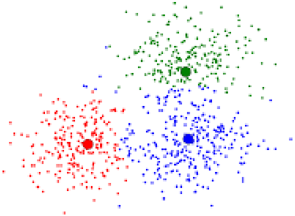
Hypothèse : Les données ont une structure géométrique de "petite" dimension intrinsèque

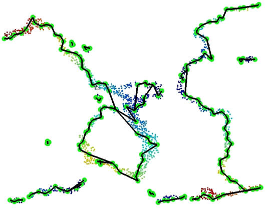
Problème : Inférer la structure à partir des données

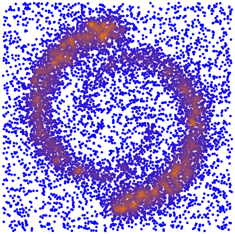
1. Modèles géométriques pour les données
   - Clusters
   - Variétés
2. Robustesse au bruit et analyse multi-échelle
   - Distance à la mesure
   - Analyse multi-échelle, filtration et persistance
3. Géométrie de l'information

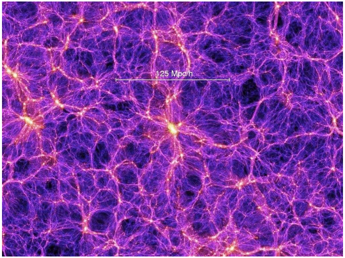
Modèles géométriques pour les données

Clusters et partitions des données









Clusters centrés

Définition : une partition des données en k clusters $C = \{C_1, ..., C_k\}$ représentés par k « centres » $c_1, ..., c_k$

On peut chercher à optimiser différents critères :

- k-centres : $\Phi(C) = \max_{j=1}^{k} \max_{p_i \in C_j} \|p_i - c_j\|$
- k-médianes : $\Phi(C) = \sum_{j=1}^{k} \sum_{p_i \in C_j} \|p_i - c_j\|$
- k-moyennes : $\Phi(C) = \sum_{j=1}^{k} \sum_{p_i \in C_j} \|p_i - c_j\|^2$

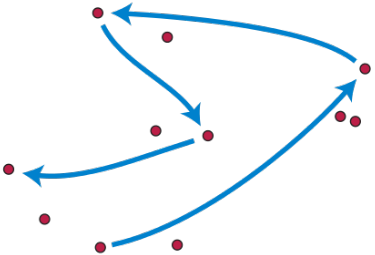
Le problème des k-centres

Problème : minimiser $\Phi(C) = \max_{j=1}^{k} \max_{p_i \in C_j} \|p_i - c_j\|$

Algorithme glouton (farthest traversal)

1. $c_1 :=$ un point de donnée quelconque
2. pour $i = 2, 3, ..., k$,

   $c_i :=$ le point de donnée le plus loin des centres $c_1, ..., c_{i-1}$



Rayon de la partition C : $\Phi(C) = \max_{j=1}^{k} \max_{p_i \in C_j} \|p_i - c_j\|$

Lemme :

S'il existe une solution C de rayon $\Phi(C)$, alors la solution $\tilde{C}$ fournie par l'algorithme glouton est de rayon $\Phi(\tilde{C}) \leq 2\Phi(C)$

Démonstration par l'absurde : $c_1, ..., c_k$ les centres fournis par l'algorithme

- Hyp. : $\exists$ solution de rayon $r/2$ et $\exists p \in \mathcal{P},\ \forall i \in \{1, ..., k\},\ \|p - \tilde{c}_i\| > r$

  $\Rightarrow\ \forall i \in \{1, ..., k\},\ \forall j \in \{1, ..., i-1\},\ \|\tilde{c}_i - \tilde{c}_j\| > r$ (sinon prendre p)

- Les points $\tilde{c}_1, ..., \tilde{c}_k, p$ sont à distance $> r$ les uns des autres

  $\Rightarrow$ aucun de ces $k+1$ pts ne peut appartenir au même cluster de rayon $\leq r/2$

- Il n'existe pas de solution de rayon $r/2$ avec k clusters

Le problème des k-centres

L'algorithme glouton fournit une 2-approximation

Rayon de la partition C : $\Phi(C) = \max_{j=1}^{k} \max_{p_i \in C_j} \|p_i - c_j\|$

Lemme :

S'il existe une solution C de rayon $\Phi(C)$, alors la solution $\tilde{C}$ fournie par l'algorithme glouton est de rayon $\Phi(\tilde{C}) \leq 2\Phi(C)$

Démonstration par l'absurde : $c_1, \ldots, c_k$ les centres fournis par l'algorithme

- Hyp. : $\exists$ solution de rayon $r/2$ et $\exists p \in \mathcal{P},\ \forall i \in \{1, \ldots, k\},\ \|p - \tilde{c}_i\| > r$

  $\Rightarrow\ \forall i \in \{1, \ldots, k\},\ \forall j \in \{1, \ldots, i-1\},\ \|\tilde{c}_i - \tilde{c}_j\| > r$ (sinon prendre p)

- Les points $\tilde{c}_1, \ldots, \tilde{c}_k, p$ sont à distance $> r$ les uns des autres

  $\Rightarrow$ aucun de ces $k+1$ pts ne peut appartenir au même cluster de rayon $\leq r/2$

- Il n'existe pas de solution de rayon $r/2$ avec k clusters

Algorithme glouton et nets

Représentation multi-résolution d'un nuage de points

Notations : $P_i = \{p_1, ..., p_i\}, \quad r_i = d(p_i, P_{i-1})$

Comme P_i grossit avec i : $\quad j \geq i \;\Rightarrow\; r_j \leq r_i$

Lemme A chaque itération $i > 0$, $\;P_i$ est un r_i-net de $\mathcal{P}$.

Démonstration

1. P_i est r_i-dense dans $\mathcal{P}$

2. P_i est λ_i-séparé : $\quad p_a p_b$ plus proche paire dans P_i, $\|p_a - p_b\| = r_b \geq r_i$

Notations : $P_i = \{p_1, ..., p_i\}, \quad r_i = d(p_i, P_{i-1})$

Comme P_i grossit avec i : $\quad j \geq i \;\; \Rightarrow \;\; r_j \leq r_i$

Lemme A chaque itération $i > 0$, $\;P_i$ est un r_i-net de $\mathcal{P}$.

Démonstration

1. P_i est r_i-dense dans $\mathcal{P}$
2. P_i est λ_i-séparé : $\quad p_a p_b$ plus proche paire dans P_i, $\|p_a - p_b\| = r_b \geq r_i$

Le problème des k-moyennes

Problème : minimiser $\Phi(C) = \sum_{j=1}^{k} \sum_{p_i \in C_j} \|p_i - c_j\|^2$

Lemme : Soit un ensemble de points $p_1, ..., p_n$ de $\mathbb{R}^d$. Le point x qui minimise $\sum_{i=1}^{k} \|x - p_i\|^2$ est le barycentre des $p_i = \frac{1}{n} \sum_{i=1}^{k} p_i$

Algorithme de Lloyd

1. Initialisation : prendre k centres
2. Répéter jusqu'à convergence les 2 étapes suivantes
   - 2.1 Partition : Associer chaque point de $\mathcal{P}$ au centre le plus proche
   - 2.2 Relaxation : remplacer le centre d'un cluster par le barycentre correspondant

Le problème des k-moyennes

Problème : minimiser $\Phi(C) = \sum_{j=1}^{k} \sum_{p_i \in C_j} \|p_i - c_j\|^2$

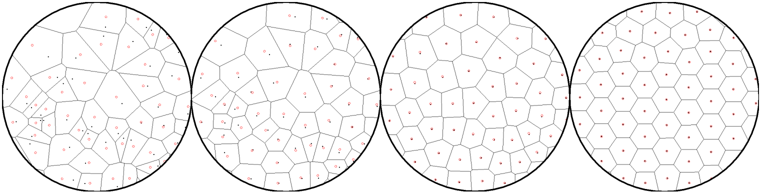
Lemme : Soit un ensemble de points $p_1, ..., p_n$ de $\mathbb{R}^d$. Le point x qui minimise $\sum_{i=1}^{k} \|x - p_i\|^2$ est le barycentre des $p_i = \frac{1}{n} \sum_{i=1}^{k} p_i$

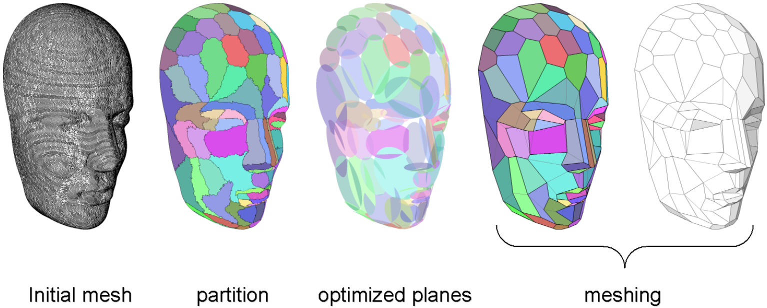
Algorithme de Lloyd

1. Initialisation : prendre k centres
2. Répéter jusqu'à convergence les 2 étapes suivantes
   2.1 Partition : Associer chaque point de $\mathcal{P}$ au centre le plus proche
   2.2 Relaxation : remplacer le centre d'un cluster par le barycentre correspondant

Diagrammes de Voronoï centroïdaux

Application à la modélisation géométrique [Du et al. 1999], [Alliez et col.]

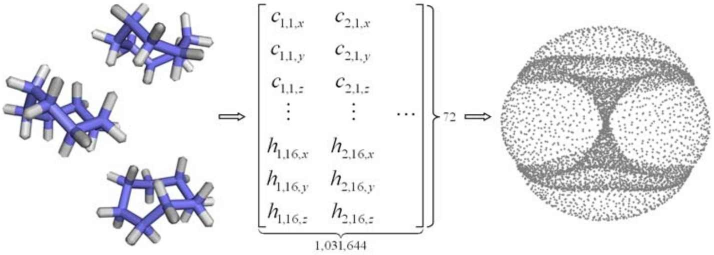




Complexité et qualité des approximations

- k-centres, k-médianes et k-moyennes sont NP-difficiles
- Approcher k-centres à moins d'un facteur 1.86 est NP-difficile
- Des algorithmes approchés polynomiaux sont connus pour les 3 problèmes
- Une bonne initialisation permet de garantir à l'algorithme de Lloyd un facteur d'approximation $\log k$ (k++mean)
- Les clusters ne sont pas toujours centrés !

1. Modèles géométriques pour les données
   - Clusters
   - Variétés
2. Robustesse au bruit et analyse multi-échelle
   - Distance à la mesure
   - Analyse multi-échelle, filtration et persistance
3. Géométrie de l'information



$c_{1,1,x}$ $c_{2,1,x}$
$c_{1,1,y}$ $c_{2,1,y}$
$c_{1,1,z}$ $c_{2,1,z}$
$h_{1,16,x}$ $h_{2,16,x}$
$h_{1,16,y}$ $h_{2,16,y}$
$h_{1,16,z}$ $h_{2,16,z}$
72
1,031,644

antics

Un système dynamique : $\phi : \mathbb{R} \times X \to X$

On s'intéresse aux ensembles invariants $A \subset X$: $\phi(\mathbb{R}, A) = A$

Typiquement $\dim(A) \ll \dim(X)$

On n'a pas accès directement à A mais on observe le système à intervalles réguliers avec une fonction scalaire $\alpha : A \to \mathbb{R}$ et on définit

$$\psi(x) = (\alpha(x), \alpha(\phi(x)), ..., \alpha(\phi^{k-1}(x)) \in \mathbb{R}^k$$

Théorème de Takens : si $k > 2\dim(A)$, alors ψ est (génériquement) un plongement de A dans $\mathbb{R}^k$

Un système dynamique : $\phi : \mathbb{R} \times X \to X$

On s'intéresse aux ensembles invariants $A \subset X$: $\phi(\mathbb{R}, A) = A$

Typiquement $\dim(A) \ll \dim(X)$

On n'a pas accès directement à A mais on observe le système à intervalles réguliers avec une fonction scalaire $\alpha : A \to \mathbb{R}$ et on définit

$$\psi(x) = (\alpha(x), \alpha(\phi(x)), ..., \alpha(\phi^{k-1}(x)) \in \mathbb{R}^k$$

Théorème de Takens : si $k > 2\dim(A)$, alors ψ est (génériquement) un plongement de A dans $\mathbb{R}^k$

Un système dynamique : $\phi : \mathbb{R} \times X \ \rightarrow \ X$

On s'intéresse aux ensembles invariants $A \subset X$: $\phi(\mathbb{R}, A) = A$

Typiquement $\dim(A) \ll \dim(X)$

On n'a pas accès directement à A mais on observe le système à intervalles réguliers avec une fonction scalaire $\alpha : A \rightarrow \mathbb{R}$ et on définit

$$\psi(x) = (\alpha(x), \alpha(\phi(x)), ..., \alpha(\phi^{k-1}(x)) \in \mathbb{R}^k$$

Théorème de Takens : si $k > 2\dim(A)$, alors ψ est (génériquement) un plongement de A dans $\mathbb{R}^k$

Théorème de plongement de Whitney

Toute variété de dimension (intrinsèque) k peut être plongée dans $\mathbb{R}^{2k+1}$

Lemme à la Johnson Lindenstrauss pour les variétés

[Baraniuk & Wakin 2007] [Clarkson 2007]

Soit $\mathbb{M}$ une k-sous-variété de $\mathbb{R}^d$ de portée τ positive et $\epsilon \in (0, 1)$. Si on projette $\mathbb{M}$ sur un plan aléatoire H de dimension

$$k = \Omega(\frac{d}{\varepsilon^2} \log \frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta}),$$

alors, avec probabilité $1 - \delta$,

$$\forall p, q \in \mathbb{M}, \quad (1-\epsilon) \|p - q\|^2 \leq \|f(p) - f(q)\|^2 \leq (1+\epsilon) \|p - q\|^2$$

où $f(p) = \sqrt{\frac{d}{k}}\, \pi_H(p)$

Théorème de plongement de Whitney

Toute variété de dimension (intrinsèque) k peut être plongée dans $\mathbb{R}^{2k+1}$

Lemme à la Johnson Lindenstrauss pour les variétés

[Baraniuk & Wakin 2007] [Clarkson 2007]

Soit $\mathbb{M}$ une k-sous-variété de $\mathbb{R}^d$ de portée τ positive et $\epsilon \in (0, 1)$. Si on projette $\mathbb{M}$ sur un plan aléatoire H de dimension

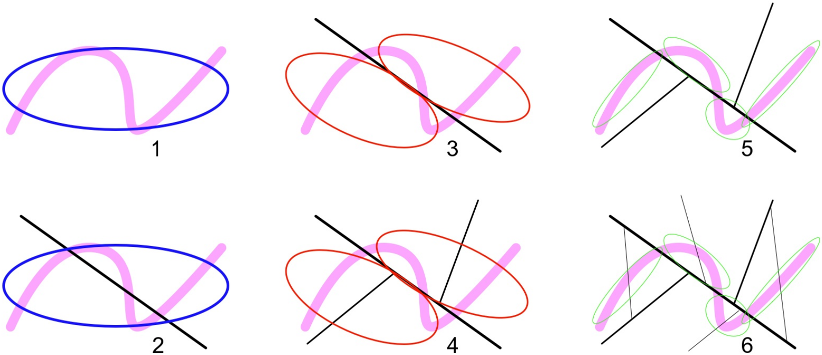
$$k = \Omega(\tfrac{d}{\varepsilon^2} \log \tfrac{1}{\varepsilon} + \tfrac{1}{\varepsilon^2} \log \tfrac{1}{\delta}),$$

alors, avec probabilité $1 - \delta$,

$$\forall p, q \in \mathbb{M},\ \ (1-\epsilon)\left\|p-q\right\|^2 \leq \left\|f(p)-f(q)\right\|^2 \leq (1+\epsilon)\left\|p-q\right\|^2$$

où $f(p) = \sqrt{\frac{d}{k}}\,\pi_H(p)$

Arbres kd ou arbres RP $+$ approximation locale dans chaque cellule



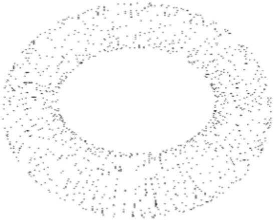
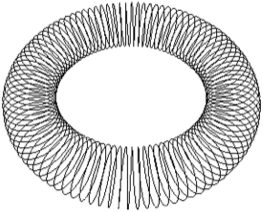
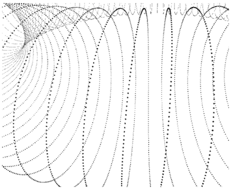
[Freund et al. 2007]

Approximation des variétés

Complexes simpliciaux

Reconstruction de complexes géométriquement et topologiquement fidèles

Quelle est la dimension de la variété ? quelle est sa portée ?

Deux difficultés :

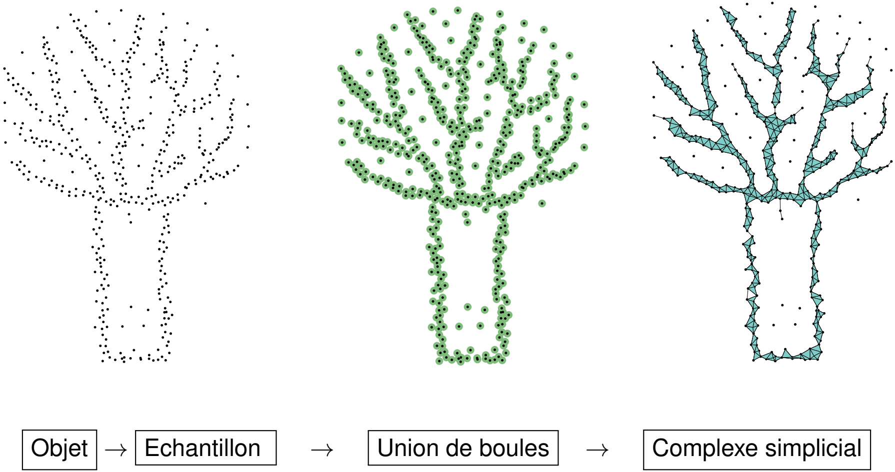
- Choisir une échelle
- Rendre les méthodes robustes au bruit

1. Modèles géométriques pour les données
   - Clusters
   - Variétés
2. Robustesse au bruit et analyse multi-échelle
   - Distance à la mesure
   - Analyse multi-échelle, filtration et persistance
3. Géométrie de l'information

Retour sur la reconstruction homotopique

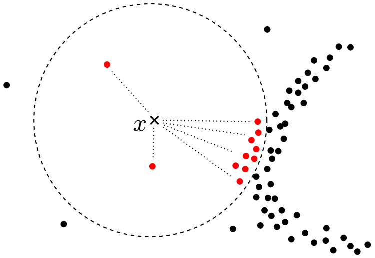
Robustesse vis à vis des points aberrants

Hypothèse du cours 6 : bruit de faible ampleur vis à vis de la portée



Distance à une mesure [Chazal, Cohen-Steiner, Mérigot 2011]

A la recherche d'une distance robuste en présence de points aberrants



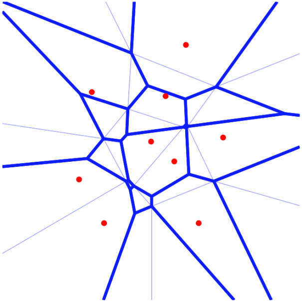
Distance à la mesure

$$d_{\mathcal{P},k}(x) = \left(\frac{1}{k} \sum_{i=1}^{k} \|x - p_i(x)\|^2 \right)^{\frac{1}{2}} \qquad (p_i(x) \ = \ \text{ième ppv}(x) \in \mathcal{P})$$

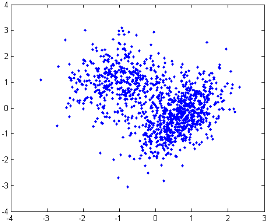
Distance à la mesure et diagrammes de Voronoï

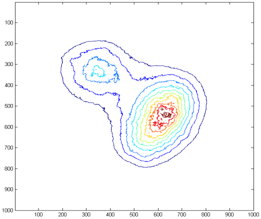
$$d^2_{\mathcal{P},k}(x) = \frac{1}{k} \sum_{i=1}^{k} \|x - p_i(x)\|^2 \qquad (p_i(x) \;=\; \text{ième ppv}(x) \in \mathcal{P})$$

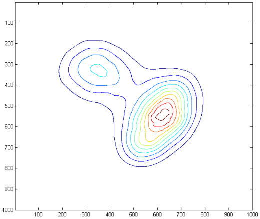
La distance à la mesure a une expression constante sur chaque cellule du diagramme de Voronoï d'ordre k de $\mathcal{P}$

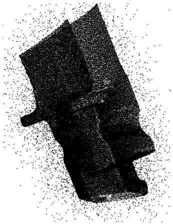


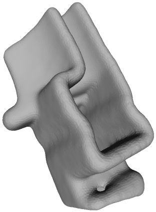
Distance à la mesure et distance au k-ième plus proche voisin











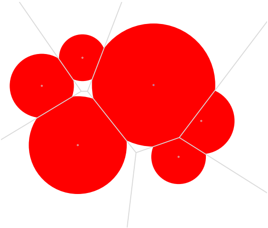
Niveaux de la fonction distance à la mesure

$S_1, S_2, \ldots$ les sous-ensembles de k points de $\mathcal{P}$

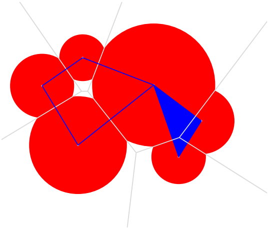
$$\delta(x, S_i) = \frac{1}{k} \sum_{p \in S_i} (x-p)^2 \quad = \quad \text{puissance de } x \text{ par rapport à } B_i = B(c_i, r_i)$$

où $\quad c_i = \frac{1}{k} \sum_{p \in S_i} p \quad$ et $\quad r_i^2 = c_i^2 - \frac{1}{k} \sum_{p \in S_i} p^2$

Les niveaux de la fonction distance à la mesure sont des unions de boules



Union de boules



α-complexe pondéré

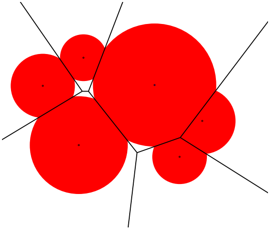
Niveaux de la fonction distance à la mesure

$S_1, S_2, \ldots$ les sous-ensembles de k points de $\mathcal{P}$

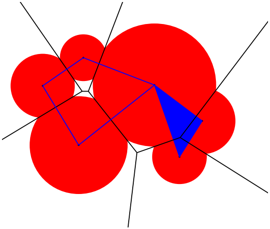
$$\delta(x, S_i) = \frac{1}{k} \sum_{p \in S_i} (x-p)^2 \quad = \quad \text{puissance de } x \text{ par rapport à } B_i = B(c_i, r_i)$$

où $\quad c_i = \frac{1}{k} \sum_{p \in S_i} p \quad$ et $\quad r_i^2 = c_i^2 - \frac{1}{k} \sum_{p \in S_i} p^2$

Les niveaux de la fonction distance à la mesure sont des unions de boules



Union de boules



α-complexe pondéré

Distance à une mesure de probabilité

Mesure uniforme sur un ens. fini de points $\mathcal{P}$: $\mu_{\mathcal{P}}$
(somme de n masses de Dirac $= \frac{1}{n}$))

μ une mesure de probabilité sur $\mathbb{R}^d$

m_0 un paramètre

$$\forall x \in \mathbb{R}^d, \quad \delta_{\mu,m}(x) = \inf\{r \geq 0,\ \mu(B(x,r) \geq m\}$$

Distance à la mesure μ : $d_{\mu,m_0}(x) = \left(\frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2\, dm\right)^{\frac{1}{2}}$

Support de μ : ensemble K t.q. $\mu(\mathbb{R}^d \setminus K) = 0$

Dimension de μ : $\dim(\mu) =$ le plus grand entier κ
$\exists C > 0,\ \forall x \in K,\ \forall r \leq \operatorname{diam}(K) : \quad \mu(B(x,r) \geq C r^{\kappa}$

Distance à une mesure de probabilité

Mesure uniforme sur un ens. fini de points $\mathcal{P}$: $\mu_{\mathcal{P}}$

(somme de n masses de Dirac $= \frac{1}{n}$))

μ une mesure de probabilité sur $\mathbb{R}^d$

m_0 un paramètre

$$\forall x \in \mathbb{R}^d, \quad \delta_{\mu,m}(x) = \inf\{r \geq 0,\ \mu(B(x,r) \geq m\}$$

Distance à la mesure μ : $d_{\mu,m_0}(x) = \left(\frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2\, dm\right)^{\frac{1}{2}}$

Support de μ : ensemble K t.q. $\mu(\mathbb{R}^d \setminus K) = 0$

Dimension de μ : $\dim(\mu) =$ le plus grand entier κ

$\exists C > 0,\ \forall x \in K,\ \forall r \leq \mathrm{diam}(K) : \quad \mu(B(x,r) \geq Cr^\kappa$

Distance à une mesure de probabilité

Mesure uniforme sur un ens. fini de points $\mathcal{P}$: $\mu_{\mathcal{P}}$
(somme de n masses de Dirac $= \frac{1}{n}$))

μ une mesure de probabilité sur $\mathbb{R}^d$

m_0 un paramètre

$$\forall x \in \mathbb{R}^d, \quad \delta_{\mu,m}(x) = \inf\{r \geq 0,\ \mu(B(x,r) \geq m\}$$

Distance à la mesure μ : $d_{\mu,m_0}(x) = \left(\frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2\, dm\right)^{\frac{1}{2}}$

Support de μ : ensemble K t.q. $\mu(\mathbb{R}^d \setminus K) = 0$

Dimension de μ : $\dim(\mu) =$ le plus grand entier κ
$\exists C > 0,\ \forall x \in K,\ \forall r \leq \mathrm{diam}(K)$: $\mu(B(x,r) \geq Cr^{\kappa}$

Distance à une mesure de probabilité

Mesure uniforme sur un ens. fini de points $\mathcal{P}$: $\mu_{\mathcal{P}}$
(somme de n masses de Dirac $= \frac{1}{n}$))

μ une mesure de probabilité sur $\mathbb{R}^d$

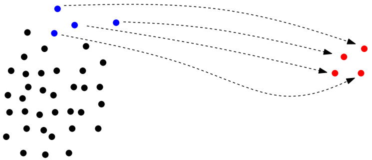
m_0 un paramètre

$$\forall x \in \mathbb{R}^d, \quad \delta_{\mu,m}(x) = \inf\{r \geq 0,\ \mu(B(x,r) \geq m\}$$

Distance à la mesure μ : $d_{\mu,m_0}(x) = \left(\frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2 \, dm \right)^{\frac{1}{2}}$

Support de μ : ensemble K t.q. $\mu(\mathbb{R}^d \setminus K) = 0$

Dimension de μ : $\dim(\mu) =$ le plus grand entier κ
$\exists C > 0,\ \forall x \in K,\ \forall r \leq \operatorname{diam}(K) :\quad \mu(B(x,r) \geq C r^{\kappa}$



Plan de transport π

Ensemble $\{\pi_{i_1 i_2}, i_1 \leq n_1, i_2 \leq n_2\}$ t. q.

1. $\pi_{i_1 i_2} \geq 0$
2. $\sum_{i_1=1}^{n_1} \pi_{i_1 i_2} = n_2$
3. $\sum_{i_2=1}^{n_2} \pi_{i_1 i_2} = n_1$

Coût d'un plan de transport : $C(\pi) = \left(\sum_{i_1,i_2} \|p_{i_1} - p_{i_2}\|^2 \pi_{i_1 i_2}\right)^{1/2}$

Distance de Wasserstein : $d_W(\mathcal{P}_1, \mathcal{P}_2) = \inf_\pi C(\pi)$

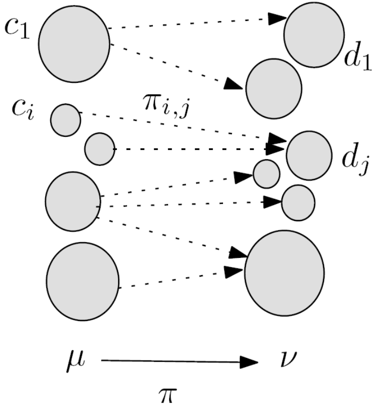
Remarque : Si $\mathcal{P}_1 = \{p_1, ..., p_n\}$ et $\mathcal{P}_2 = \{p_1, ..., p_{n-k-1}, q_1, ..., q_k\}$ avec $d(q_i, \mathcal{P}_1) = R$, alors

$$d_H(\mathcal{P}_1, \mathcal{P}_2) \geq R$$

$$d_W(\mathcal{P}_1, \mathcal{P}_2) \leq \sqrt{\frac{k}{n}}\,(R + \mathrm{diam}(C))$$

Distance de Wasserstein

Cas de deux mesures de probabilité μ et ν sur $\mathbb{R}^d$



Plan de transport entre μ et ν

Une mesure de probabilité π sur $\mathbb{R}^d \times \mathbb{R}^d$ t.q. $\forall X, Y \subset \mathbb{R}^d$

1. $\pi(X \times \mathbb{R}^d) = \mu(X)$
2. $\pi(\mathbb{R}^d \times Y) = \mu(Y)$

Coût d'un plan de transport : $C(\pi) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\pi(x, y) \right)^{1/2}$

Distance de Wasserstein : $d_W(\mu, \nu) = \inf_\pi \; C(\pi)$

Lemme de stabilité

Si deux mesures sont proches au sens de Wasserstein, les distances à ces mesures sont proches au sens de la norme L_∞

Corollaire

Soit μ une mesure de probabilité de dimension $\dim(\mu)$ et de support K, et soit $\mu_{\mathcal{P}}$ la mesure uniforme sur un échantillon $\mathcal{P}$. Alors

$$\|d_{\mathcal{P},k} - d_K\|_\infty \leq \sqrt{\frac{n}{k}}\, d_W(\mu, \mu_{\mathcal{P}}) + C \left(\frac{k}{n}\right)^{\frac{1}{\dim(\mu)}}$$

Reconstruction

Si $d_W(\mu, \mu_{\mathcal{P}})$ est suffisamment petit, les α-niveaux de $d_{\mathcal{P},k}$ sont homotopiquement équivalents aux β-offsets de K (pour certaines plages de valeurs de α et β)

Lemme de stabilité

Si deux mesures sont proches au sens de Wasserstein, les distances à ces mesures sont proches au sens de la norme L_∞

Corollaire

Soit μ une mesure de probabilité de dimension $\dim(\mu)$ et de support K, et soit $\mu_{\mathcal{P}}$ la mesure uniforme sur un échantillon $\mathcal{P}$. Alors

$$\|d_{\mathcal{P},k} - d_K\|_\infty \leq \sqrt{\frac{n}{k}}\, d_W(\mu, \mu_{\mathcal{P}}) + C \left(\frac{k}{n}\right)^{\frac{1}{\dim(\mu)}}$$

Reconstruction

Si $d_W(\mu, \mu_{\mathcal{P}})$ est suffisamment petit, les α-niveaux de $d_{\mathcal{P},k}$ sont homotopiquement équivalents aux β-offsets de K (pour certaines plages de valeurs de α et β)

Distance de Wasserstein

Lemme de stabilité et reconstruction [Chazal, Cohen-Steiner, Mérigot 2011]

Lemme de stabilité

Si deux mesures sont proches au sens de Wasserstein, les distances à ces mesures sont proches au sens de la norme L_∞

Corollaire

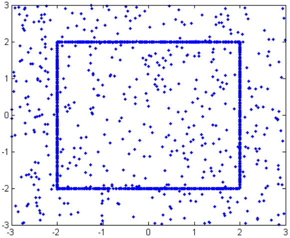
Soit μ une mesure de probabilité de dimension $\dim(\mu)$ et de support K, et soit $\mu_{\mathcal{P}}$ la mesure uniforme sur un échantillon $\mathcal{P}$. Alors

$$\|d_{\mathcal{P},k} - d_K\|_\infty \leq \sqrt{\frac{n}{k}}\, d_W(\mu, \mu_{\mathcal{P}}) + C \left(\frac{k}{n}\right)^{\frac{1}{\dim(\mu)}}$$

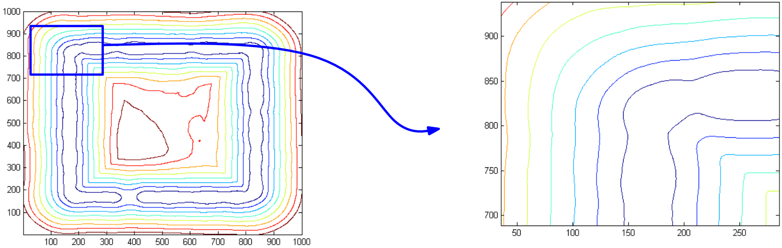
Reconstruction

Si $d_W(\mu, \mu_{\mathcal{P}})$ est suffisamment petit, les α-niveaux de $d_{\mathcal{P},k}$ sont homotopiquement équivalents aux β-offsets de K (pour certaines plages de valeurs de α et β)

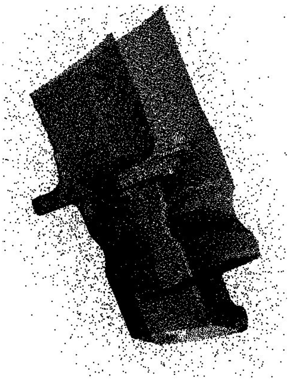
Niveaux de la distance à une mesure

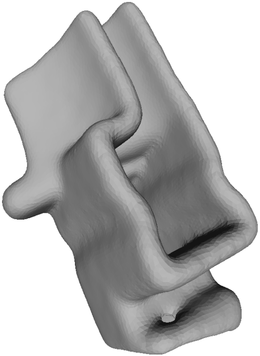


2300 points, 20% outliers



Reconstruction d'un objet 3d bruité





10% de points aberrants

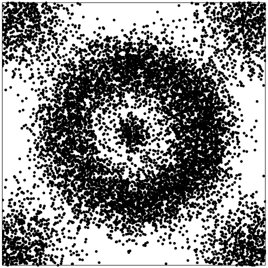
Conclusion sur la distance à la mesure

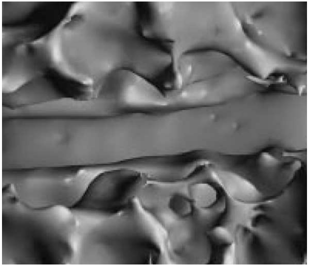
- Pas d'hypothèse sur la nature du bruit
- $d_{k,\mathcal{P}}$ se comporte de façon analogue à une distance
- On peut évaluer son gradient
- Autres applications : estimation de densité, détection de clusters
- Algorithmes d'approximation rapides

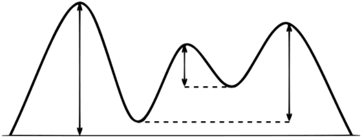
1. Modèles géométriques pour les données
   - Clusters
   - Variétés
2. Robustesse au bruit et analyse multi-échelle
   - Distance à la mesure
   - Analyse multi-échelle, filtration et persistance
3. Géométrie de l’information

Analyse multi-échelle

Clusters et culminance

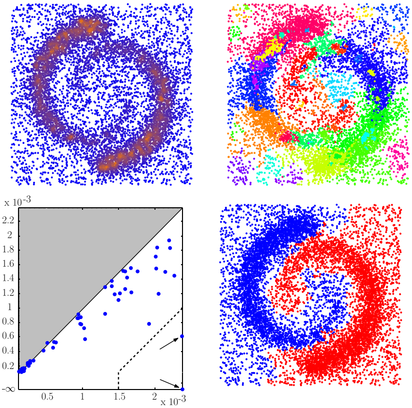






La détection de clusters

L’algorithme Tomato (Topological Mode Analysis Tool) [Chazal et al. 2013]



1. Construire un graphe de voisinage G
2. Estimer la densité du nuage de points en chaque sommet de G
3. Identifier les maxima de densité et les bassins versants associés
4. Rechercher les modes culminants (persistants) et fusionner les bassins versants associés

Le diagramme de persistance associe un point à chaque composante connexe C des surniveaux de la fonction densité. L’abscisse (ordonnée) du point est la valeur de la densité quand C est créée (supprimée)

Filtration d'un complexe simplicial

1. Une filtration de K est une suite de sous-complexes de K

$$\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$$

   t.q. : $K^{i+1} = K^i \cup \sigma^{i+1}$, où σ^{i+1} est un simplexe de K

2. De manière équivalente, une filtration de K est uns suite ordonnée $\sigma_1, \ldots \sigma_m$ des simplexes de K t.q. l'ensemble K^i des premiers i simplexes est un sous-complexe de K

   Cet ordre range les simplexes par dimensions croissantes

Les filtrations jouent un rôle central dans la théorie de la persistance

Filtration d'un complexe simplicial

1. Une filtration de K est une suite de sous-complexes de K

$$\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$$

t.q. : $K^{i+1} = K^i \cup \sigma^{i+1}$, où σ^{i+1} est un simplexe de K

2. De manière équivalente, une filtration de K est uns suite ordonnée $\sigma_1, \ldots \sigma_m$ des simplexes de K t.q. l'ensemble K^i des premiers i simplexes est un sous-complexe de K

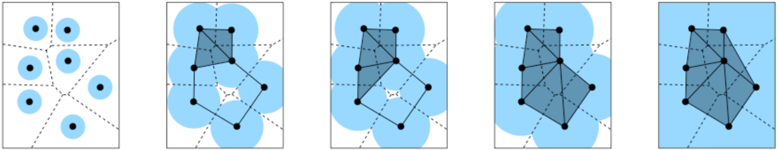
   Cet ordre range les simplexes par dimensions croissantes

Les filtrations jouent un rôle central dans la théorie de la persistance

Filtration de Delaunay

P un ensemble fini de points de $\mathbb{R}^d$ $\qquad U(\alpha) = \bigcup_{p \in P} B(p, \alpha)$

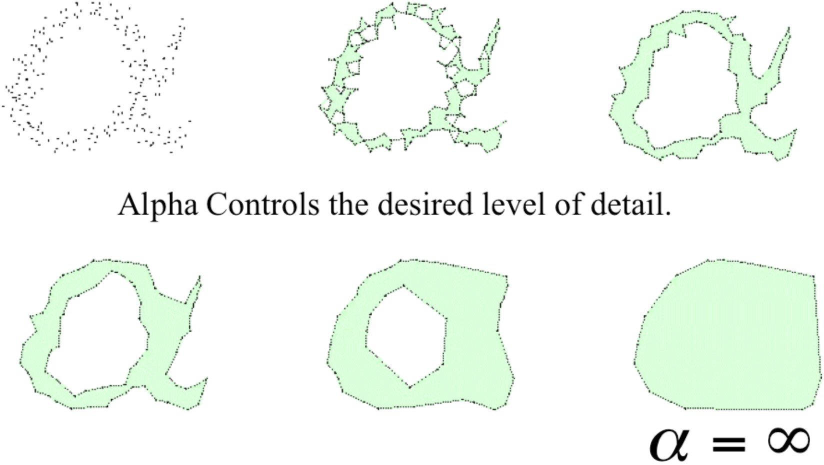
$$\alpha\text{-complexe} = \quad \mathrm{Del}_{|U(\alpha)}(P) \quad \simeq \quad U(\alpha)$$



Filtration de Delaunay de $\mathcal{P}$: $\quad \{\mathrm{Del}_{|U(\alpha)}(\mathcal{P}), \quad \alpha \in \mathbb{R}^+\}$

- Un nombre fini de valeurs α critiques $\quad \alpha_0 = 0, \alpha_1, ..., \alpha_m = +\infty$
- Une suite de complexes emboîtés $\quad \mathrm{Del}_{|U(\alpha_0)}(\mathcal{P}) \subset ... \subset \mathrm{Del}_{|U(\alpha_\infty)}(\mathcal{P})$
- Représentation : associer à chaque $\sigma \in \mathrm{Del}(\mathcal{P})$ une valeur de filtration

min_cc_empty_ball (σ)


Alpha Controls the desired level of detail.
$\alpha = \infty$

Pour en savoir plus sur la persistance topologique

- Séminaire de F. Chazal
- Les colloques des 6 et 8 juin au Collège de France
- Le projet Gudhi (Geometry Understanding in Higher Dimensions) et la bibliothèque Gudhi
  `https://project.inria.fr/gudhi/`

Un point représente une loi de probabilité, par exemple la gaussienne isotrope définie dans $\mathbb{R}^d$

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\|x - \mu\|}{2\sigma^2}\right)$$

peut être représentée par le point (μ, σ) dans l'espace

$$H = \{(\mu, \sigma) \in \mathbb{R}^{d+1}, \sigma > 0\}$$

Quelle distance dans ces espaces ?

- La métrique de Fischer : fournit une structure riemanienne
- Divergences de Kullback-Leibler, Itakura-Saito, etc.

Un point représente une loi de probabilité, par exemple la gaussienne isotrope définie dans $\mathbb{R}^d$

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\|x-\mu\|}{2\sigma^2}\right)$$

peut être représentée par le point (μ, σ) dans l'espace

$$H = \{(\mu, \sigma) \in \mathbb{R}^{d+1}, \sigma > 0\}$$

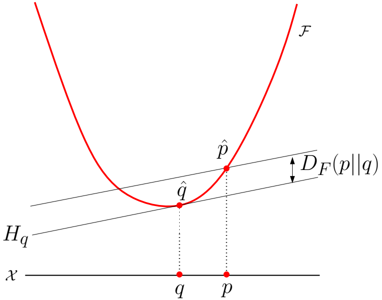
Quelle distance dans ces espaces ?

- La métrique de Fischer : fournit une structure riemanienne
- Divergences de Kullback-Leibler, Itakura-Saito, etc.

Divergences de Bregman

F une fonction strictement convexe et différentiable définie sur un ensemble convexe $\mathcal{X}$

$$D_F(\mathbf{p}, \mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p} - \mathbf{q}, \boldsymbol{\nabla}_F(\mathbf{q}) \rangle$$



D_F n'est pas une distance mais $D_F(\mathbf{p}, \mathbf{q}) \geq 0$ et $D_F(\mathbf{p}, \mathbf{q}) = 0$ ssi $\mathbf{p} = \mathbf{q}$

Examples de divergences de Bregman

- $F(x) = x^2$: Distance euclidienne au carré

$$D_F(\mathbf{p},\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p}-\mathbf{q}, \nabla_F(\mathbf{q})\rangle = \mathbf{p}^2 - \mathbf{q}^2 - \langle \mathbf{p}-\mathbf{q}, 2\mathbf{q}\rangle = \|\mathbf{p}-\mathbf{q}\|^2$$

- $F(p) = \sum p(x)\log_2 p(x)$ (entropie de Shannon)
  $D_F(p,q) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)}$ (divergence K-L)

- $F(p) = -\sum_x \log p(x)$ (entropie de Burg)
  $D_F(p,q) = \sum_x (\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - 1)$ (Itakura-Saito)

Examples de divergences de Bregman

- $F(x) = x^2$: Distance euclidienne au carré

$$D_F(\mathbf{p}, \mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p} - \mathbf{q}, \nabla_F(\mathbf{q}) \rangle = \mathbf{p}^2 - \mathbf{q}^2 - \langle \mathbf{p} - \mathbf{q}, 2\mathbf{q} \rangle = \|\mathbf{p} - \mathbf{q}\|^2$$

- $F(p) = \sum p(x) \log_2 p(x)$ (entropie de Shannon)
  $D_F(p, q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$ (divergence K-L)

- $F(p) = -\sum_x \log p(x)$ (entropie de Burg)
  $D_F(p, q) = \sum_x (\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - 1)$ (Itakura-Saito)

Examples de divergences de Bregman

- $F(x) = x^2$: Distance euclidienne au carré

$$D_F(\mathbf{p},\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p}-\mathbf{q}, \nabla_F(\mathbf{q})\rangle = \mathbf{p}^2 - \mathbf{q}^2 - \langle \mathbf{p}-\mathbf{q}, 2\mathbf{q}\rangle = \|\mathbf{p}-\mathbf{q}\|^2$$

- $F(p) = \sum p(x)\log_2 p(x)$ (entropie de Shannon)
  $D_F(p,q) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)}$ (divergence K-L)

- $F(p) = -\sum_x \log p(x)$ (entropie de Burg)
  $D_F(p,q) = \sum_x (\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - 1)$ (Itakura-Saito)

Bisecteurs

$$D_F(\mathbf{p},\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p}-\mathbf{q}, \boldsymbol{\nabla}_F(\mathbf{q})\rangle$$

Deux types de bisecteurs

$$\begin{aligned} H_{pq} : D_F(\mathbf{x},\mathbf{p}) &= D_F(\mathbf{x},\mathbf{q}) \quad \text{(hyperplan)} \\ H^*_{pq} : D_F(\mathbf{p},\mathbf{x}) &= D_F(\mathbf{q},\mathbf{x}) \quad \text{(hypersurface)} \end{aligned}$$

Diagrammes de Bregman

- Deux types de diagrammes de Bregman
- Par la dualité de Legendre : $D_F(\mathbf{x},\mathbf{y}) = D_{F^*}(\mathbf{y}',\mathbf{x}') \qquad (\mathbf{x}' = \nabla F(x))$

Bisecteurs

$$D_F(\mathbf{p},\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p}-\mathbf{q}, \nabla_F(\mathbf{q})\rangle$$

Deux types de bisecteurs

$$\begin{aligned} H_{pq} : D_F(\mathbf{x},\mathbf{p}) &= D_F(\mathbf{x},\mathbf{q}) \quad \text{(hyperplan)} \\ H^*_{pq} : D_F(\mathbf{p},\mathbf{x}) &= D_F(\mathbf{q},\mathbf{x}) \quad \text{(hypersurface)} \end{aligned}$$

Diagrammes de Bregman

- Deux types de diagrammes de Bregman
- Par la dualité de Legendre : $D_F(\mathbf{x},\mathbf{y}) = D_{F^*}(\mathbf{y}',\mathbf{x}') \qquad (\mathbf{x}' = \nabla F(x))$

Diagrammes de Bregman

[Boissonnat, Nielsen, Nock 2010]

Le diagramme de Bregman du 1er type de $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ est le diagramme de minimisation des n fonctions $D_F(\mathbf{x}, \mathbf{p}_i)$, $i = 1, \ldots, n$

On déduit de

$$\arg\min(D_F(\mathbf{x}, \mathbf{p}_i)) = \arg\max(h_i(\mathbf{x}) = \langle \mathbf{x} - \mathbf{p}_i, \mathbf{p}_i' \rangle - F(\mathbf{p}_i))$$

que le diagramme de Bregman du 1er type de $\mathcal{P}$ est affine

Le diagramme de Bregman du 2ième type of $\mathcal{P}$ est le diagramme de minimisation (courbe) des n fonctions $D_F(\mathbf{p}_i, \mathbf{x})$, $i = 1, \ldots, n$

Diagrammes de Bregman

[Boissonnat, Nielsen, Nock 2010]

Le diagramme de Bregman du 1er type de $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ est le diagramme de minimisation des n fonctions $D_F(\mathbf{x}, \mathbf{p}_i)$, $i = 1, \ldots, n$

On déduit de

$$\arg\min(D_F(\mathbf{x}, \mathbf{p}_i)) = \arg\max(h_i(\mathbf{x}) = \langle \mathbf{x} - \mathbf{p}_i, \mathbf{p}_i' \rangle - F(\mathbf{p}_i))$$

que le diagramme de Bregman du 1er type de $\mathcal{P}$ est affine

Le diagramme de Bregman du 2ième type of $\mathcal{P}$ est le diagramme de minimisation (courbe) des n fonctions $D_F(\mathbf{p}_i, \mathbf{x})$, $i = 1, \ldots, n$

Diagrammes de Bregman

[Boissonnat, Nielsen, Nock 2010]

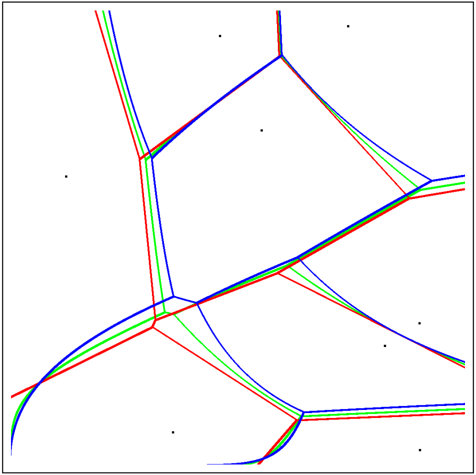
Le diagramme de Bregman du 1er type de $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ est le diagramme de minimisation des n fonctions $D_F(\mathbf{x}, \mathbf{p}_i)$, $i = 1, \ldots, n$

On déduit de

$$\arg\min(D_F(\mathbf{x}, \mathbf{p}_i)) = \arg\max(h_i(\mathbf{x}) = \langle \mathbf{x} - \mathbf{p}_i, \mathbf{p}_i' \rangle - F(\mathbf{p}_i))$$

que le diagramme de Bregman du 1er type de $\mathcal{P}$ est affine

Le diagramme de Bregman du 2ième type of $\mathcal{P}$ est le diagramme de minimisation (courbe) des n fonctions $D_F(\mathbf{p}_i, \mathbf{x})$, $i = 1, \ldots, n$



Diagrammes de Bregman et diagrammes de Laguerre

The diagramme de Bregman du 1er type d'un ensemble de n sites de $\mathcal{P}$ est identique au diagramme de Laguerre de n hypersphères euclidiennes centrées aux points $\mathbf{p}_i'$

$$
\begin{aligned}
& D_F(\mathbf{x}, \mathbf{p}_i) \leq D_F(\mathbf{x}, \mathbf{p}_j) \\
\Longleftrightarrow\quad & -F(\mathbf{p}_i) - \langle \mathbf{x} - \mathbf{p}_i, \mathbf{p}_i' \rangle) \leq -F(\mathbf{p}_j) - \langle \mathbf{x} - \mathbf{p}_j, \mathbf{p}_j' \rangle) \\
\Longleftrightarrow\quad & \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{p}_i' \rangle - 2F(\mathbf{p}_i) + 2\langle \mathbf{p}_i, \mathbf{p}_i' \rangle \leq \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{p}_j' \rangle - 2F(\mathbf{p}_j) + 2\langle \mathbf{p}_j, \mathbf{p}_j' \rangle \\
\Longleftrightarrow\quad & \langle \mathbf{x} - \mathbf{p}_i', \mathbf{x} - \mathbf{p}_i' \rangle - r_i^2 \leq \langle \mathbf{x} - \mathbf{p}_j', \mathbf{x} - \mathbf{p}_j' \rangle - r_j^2
\end{aligned}
$$

où $r_l^2 = \langle \mathbf{p}_l', \mathbf{p}_l' \rangle + 2(F(\mathbf{p}_l) - \langle \mathbf{p}_l, \mathbf{p}_l' \rangle)$

Diagrammes de Bregman et diagrammes de Laguerre

The diagramme de Bregman du 1er type d'un ensemble de n sites de $\mathcal{P}$ est identique au diagramme de Laguerre de n hypersphères euclidiennes centrées aux points $\mathbf{p}_i'$

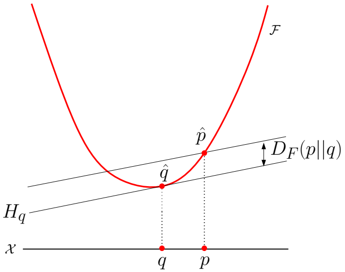
$$\begin{aligned}
& D_F(\mathbf{x},\mathbf{p}_i) \leq D_F(\mathbf{x},\mathbf{p}_j) \\
\iff\quad & -F(\mathbf{p}_i) - \langle \mathbf{x}-\mathbf{p}_i, \mathbf{p}_i' \rangle) \leq -F(\mathbf{p}_j) - \langle \mathbf{x}-\mathbf{p}_j, \mathbf{p}_j' \rangle) \\
\iff\quad & \langle \mathbf{x},\mathbf{x}\rangle - 2\langle \mathbf{x},\mathbf{p}_i'\rangle - 2F(\mathbf{p}_i) + 2\langle \mathbf{p}_i,\mathbf{p}_i'\rangle \leq \langle \mathbf{x},\mathbf{x}\rangle - 2\langle \mathbf{x},\mathbf{p}_j'\rangle - 2F(\mathbf{p}_j) + 2\langle \mathbf{p}_j,\mathbf{p}_j'\rangle \\
\iff\quad & \langle \mathbf{x}-\mathbf{p}_i', \mathbf{x}-\mathbf{p}_i'\rangle - r_i^2 \leq \langle \mathbf{x}-\mathbf{p}_j', \mathbf{x}-\mathbf{p}_j'\rangle - r_j^2
\end{aligned}$$

où $r_l^2 = \langle \mathbf{p}_l', \mathbf{p}_l' \rangle + 2(F(\mathbf{p}_l) - \langle \mathbf{p}_l, \mathbf{p}_l' \rangle)$

Sphères de Bregman

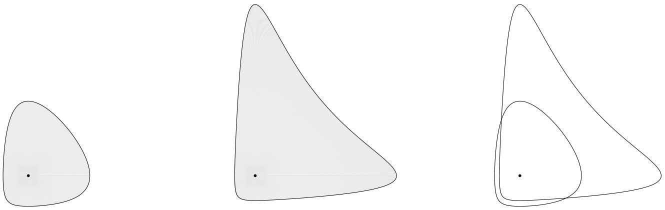
Définition : $\sigma(\mathbf{c}, r) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{x}, \mathbf{c}) = r\}$

Lemme Le relèvement $\hat{\sigma}$ sur $\mathcal{F}$ d'une sphère de Bregman σ is contenue dans un hyperplan H_σ



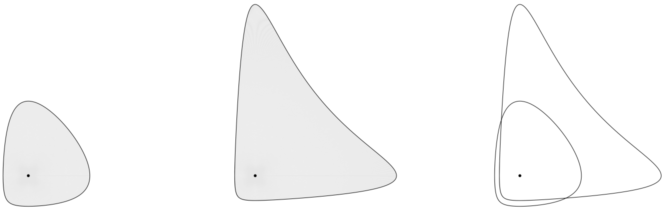
Inversement, l'intersection de tout hyperplan H avec $\mathcal{F}$ se projette verticalement sur une sphère de Bregman

Unions de boules de Bregman des types 1 et 2



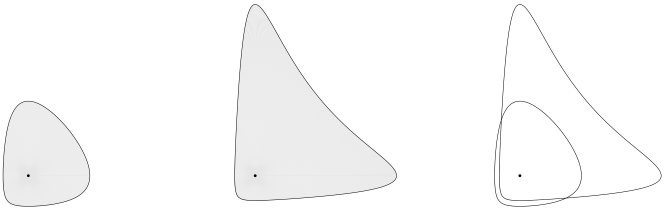
- La complexité combinatoire et algorithmique d'une union de boules de Bregman est la même que pour les boules euclidiennes
- Le nerf d'un ensemble fini de boules de Bregman du 1er type a le même type d'homotopie que leur union (théorème du nerf)
- Il va en va de même pour les unions de boules du 2ième type (via la transformation de Legendre qui est un homéomorphisme (si F est une fonction de Legendre)

Unions de boules de Bregman des types 1 et 2



- La complexité combinatoire et algorithmique d'une union de boules de Bregman est la même que pour les boules euclidiennes
- Le nerf d'un ensemble fini de boules de Bregman du 1er type a le même type d'homotopie que leur union (théorème du nerf)
- Il va en va de même pour les unions de boules du 2ième type (via la transformation de Legendre qui est un homéomorphisme (si F est une fonction de Legendre)

Unions de boules de Bregman des types 1 et 2



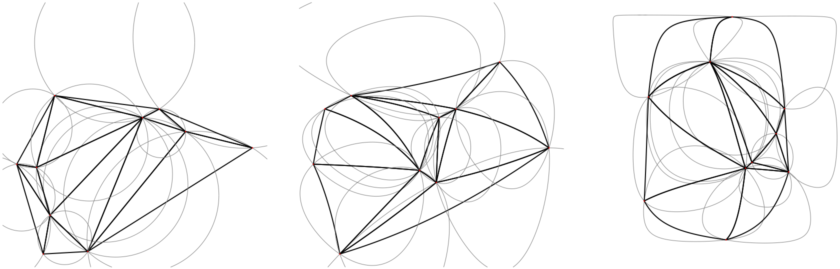
- La complexité combinatoire et algorithmique d'une union de boules de Bregman est la même que pour les boules euclidiennes
- Le nerf d'un ensemble fini de boules de Bregman du 1er type a le même type d'homotopie que leur union (théorème du nerf)
- Il va en va de même pour les unions de boules du 2ième type (via la transformation de Legendre qui est un homéomorphisme (si F est une fonction de Legendre)

Triangulations de Bregman

Définition : le nerf du diagramme de Bregman du 1er type

Primal space		Gradient space
1st type $BVD(\mathcal{P})$	=	Laguerre diagram of $(\mathcal{P}')$
		$\updownarrow *$
geodesic $BT(\mathcal{P})$	$\leftrightarrow$	regular triangulation of $(\mathcal{P}')$
$\updownarrow$		
$BT(\mathcal{P})$		

Examples



(a) Ordinary Delaunay (b) Exponential loss (c) Hellinger-like divergence

Propriétés des triangulations de Bregman

- La projection verticale de l'enveloppe convexe inférieure de $\hat{\mathcal{P}}$ est une réalisation de $BT(\mathcal{P})$ ($\hat{\mathcal{P}}$ est le relèvement de $\mathcal{P}$ sur le graphe $\mathcal{F}$ de F)
- Propriété caractéristique : La sphère de Bregman circonscrite à tout simplexe de $BT(\mathcal{P})$ est vide
- Optimalité : $BT(\mathcal{P}) = \min_{T \in \mathcal{T}(\mathcal{P})} \max_{\tau \in T} r(\tau)$
  ($r(\tau) =$ radius of the smallest Bregman ball containing τ)
  [Rajan]
- Filtration : on peut définir une filtration de Čech ou de Bregman (analogue à la filtration de Delaunay)

Conclusion

- Bien que les divergences de Bregman ne soient pas des métriques, on peut étendre beaucoup de techniques de l'espace euclidien
- Une fois encore, les diagrammes de Laguerre se révèlent utiles
- Applications
  - k-moyennes : l'algorithme de Lloyd s'applique [Banerjee et dol. 2005]
  - Plus petite boule englobante [Nielsen & Nock 2006]
  - Réduction de dimension [Carter 2009]
  - Analyse topologique des données ?