Convergence de la constante de Cheeger de graphes de voisinage.

Ery Arias-Castro [a] & Bruno Pelletier [b] & Pierre Pudlo [c]

[a] Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, USA.
eariasca@ucsd.edu

[b] Department of Mathematics
IRMAR — UMR CNRS 6625
Université Rennes II
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France
bruno.pelletier@univ-rennes2.fr

[c] I3M, UMR CNRS 5149
Université Montpellier II
Place Eugène Bataillon
34095 Montpellier, France
pierre.pudlo@univ-montp2.fr

Résumé

Nous nous intéressons dans ce travail aux ensembles minimisant la constante de Cheeger d'un sous-ensemble $\mathcal{M}$ de $\mathbb{R}^d$. Cette dernière minimise le rapport d'un périmètre à un volume parmi tous les sous-ensembles de $\mathcal{M}$. Étant donné un n-échantillon issu de la mesure uniforme sur $\mathcal{M}$, nous introduisons une version régularisée de la conductance du graphe de voisinage construit sur l'échantillon. Nous établissons alors la convergence de la conductance régularisée vers la constante de Cheeger de $\mathcal{M}$. En outre, nous montrons la convergence des suites de partitions optimales du graphe vers les ensembles de Cheeger de $\mathcal{M}$ pour la topologie de $L^1(\mathcal{M})$.

Mots-clés **— Statistique mathématique, Inégalités isopérimétriques, Graphe, Classification non-supervisée, U-statistiques, Processus empiriques.**

Abstract

We focus in this work on the sets minimizing the Cheeger constant of a subset $\mathcal{M}$ of $\mathbb{R}^d$. This latter minimizes the ratio of a perimeter to a volume among all subsets of $\mathcal{M}$. Given an n-sample drawn from the uniform measure on $\mathcal{M}$, we introduce a regularized version of the conductance of the neighborhood graph defined on the sample. Then, we establish the

convergence of the regularized convergence to the Cheeger constant of $\mathcal{M}$. In addition, we prove the convergence of the sequences of optimal graph partitions to the Cheeger sets of $\mathcal{M}$ for the topology of $L^1(\mathcal{M})$.

Keywords **— Mathematical Statistic, Isoperimetric inequalities, Graph, Clustering, U-statistics, Empirical processes.**

1 Introduction

Soit $\mathcal{M}$ un sous-ensemble compact de $\mathbb{R}^d$ de bord lisse $\partial\mathcal{M}$. La constante isopérimétrique de Cheeger de $\mathcal{M}$, notée $h(\mathcal{M})$, est définie par:

$$h(\mathcal{M}) = \inf_{A\subset\mathcal{M}} \frac{\mathrm{vol}_{d-1}(\partial A \cap \mathring{\mathcal{M}})}{\min\{\mathrm{vol}_d(A), \mathrm{vol}_d(A^c)\}}.$$

Cette constante mesure le degré d'homogénéité de $\mathcal{M}$, en ce sens qu'une valeur faible de $h(\mathcal{M})$ indique la présence d'un goulet d'étranglement dans $\mathcal{M}$. La constante de Cheeger est liée à la première valeur propre non nulle de l'opérateur de Laplace-Beltrami de $\mathcal{M}$ par l'inégalité suivante : $\lambda_1(\mathcal{M}) \geq h^2(\mathcal{M})/4$. Une majoration de $\lambda_1(\mathcal{M})$ par un polynôme en $h(\mathcal{M})$ a été obtenu par Buser (1982).

Une version discrète de la constante de Cheeger est également fréquemment utilisée en théorie des graphes. Étant donné un graphe $\mathcal{G} = (V, E)$, d'ensembles de sommets et d'arêtes respectivement V et E, la constante de Cheeger de $\mathcal{G}$, ou conductance $c(\mathcal{G})$, est définie par

$$c(\mathcal{G}) = \inf_{U\subset V} \frac{\displaystyle\sum_{u\in U}\sum_{v\in U^c} \mathbf{1}\{u\sim v\}}{\min\left\{\displaystyle\sum_{u,v\in U;u\neq v} \mathbf{1}\{u\sim v\}, \sum_{u,v\in U^c;u\neq v} \mathbf{1}\{u\sim v\}\right\}},$$

où $u \sim v$ s'il existe une arête de E entre les sommets u et v. La première valeur propre non nulle du Laplacien du graphe satisfait des inégalités similaires à celles du cas continu.

La conductance du graphe est une quantité utilisée dans l'étude du comportement d'une marche aléatoire simple sur un graphe, et qui permet d'obtenir des bornes sur les temps d'atteinte et de mixage de la marche. Elle est également utilisée comme heuristique de partitionnement en clustering pour le problème type suivant: partitionner un échantillon de donnée en deux sous-ensembles homogènes. Cette heuristique est à l'origine des algorithmes de partitionnement de graphes. Bien qu'étant un problème NP-hard, les algorithmes de clustering spectral permettent d'obtenir une approximation de la partition

optimale (voir par exemple Chung, 2007).

Étant donné un échantillon i.i.d. $X_1, \ldots, X_n$ issu de la mesure uniforme μ sur $\mathcal{M}$, nous nous intéressons dans ce travail à l'estimation de la constante de Cheeger de $\mathcal{M}$ et des partitions optimales, à partir de la conductance du graphe de voisinage dont les sommets sont les observations $X_1, \ldots, X_n$. Un tel graphe est construit en plaçant une arête entre deux observations X_i et X_j dès lors que $\mathrm{dist}(X_i, X_j) \leq r_n$, où r_n est une suite de nombres réels positifs. Dans ce but, nous considérons tout d'abord la collection $\mathcal{D}_n$ des sous-ensembles compacts D de $\mathbb{R}^d$ tels qu'une boule de rayon $r_n^{1-\delta}$, pour un certain $1/2 < \delta < 1$, roule librement dans D et dans D^c. Une telle condition prescrit les courbures de ∂D, ses courbures sectionnelles ne pouvant excéder $1/r_n^{1-\delta}$, mais est plus forte qu'une condition sur les courbures, le bord ∂D ne pouvant revenir trop près de lui-même. Nous formons ensuite la collection $\mathcal{A}_n$ des sous-ensembles de $\mathcal{M}$ obtenue par intersection de $\mathcal{D}_n$ avec $\mathcal{M}$, et satisfaisant une condition de régularité au bord de $\mathcal{M}$.

Nous définissons alors une version pénalisée $\hat{h}_n$ de la conductance du graphe par

$$\widehat{h}_n = \inf_{A \in \mathcal{B}_n} h_n(A), \quad \text{où } \mathcal{B}_n = \{A \in \mathcal{A}_n : \alpha_n \leq \mu(A) \leq \mu(\mathcal{M})/2\}, \tag{1}$$

$$h_n(A) = \frac{\displaystyle\sum_{i \neq j \leq n} \mathbf{1}_A(X_i)\mathbf{1}\left\{\mathrm{dist}(X_i, X_j) \leq r_n\right\}}{\displaystyle\sum_{i \neq j \leq n} \mathbf{1}_A(X_i)\mathbf{1}_{A^c}(X_j)\mathbf{1}\left\{\mathrm{dist}(X_i, X_j) \leq r_n\right\}},$$

et α_n est une suite de nombres réels positifs tendant vers 0 moins vite que r_n.

2 Résultats de convergence

Nous montrons tout d'abord, en utilisant des résultats de découplage de U-statistiques et de processus empiriques (de la Pena et Giné, 1999), la convergence uniforme de h_n sur la classe $\mathcal{A}_n$. Notons ω_d le volume de la boule unité de $\mathbb{R}^d$, $\pi_d(\eta)$ le volume de la calotte $\{x : \|x\| \leq 1, x_1 \geq \eta\}$ et $\gamma = \int_0^1 \pi_d(\eta)\mathrm{d}\eta$ la moyenne des volumes des calottes. Précisément, si $r_n \to 0$ et $nr_n^{d+2-\delta} \to \infty$, alors

$$\sup_{A \in \mathcal{B}_n} \left| \frac{\omega_d}{\gamma} \frac{1}{r_n} h_n(A) - h(A) \right| \to 0$$

avec probabilité 1 lorsque $n \to \infty$. Nous en déduisons alors la convergence de la conductance pénalisée (normalisée par $1/r_n$) vers la constante de Cheeger de $\mathcal{M}$, i.e., sous les mêmes conditions, nous avons:

$$\lim_{n \to \infty} \frac{\omega_d}{\gamma} \frac{1}{r_n} \widehat{h}_n \to h(\mathcal{M}) \quad \text{p.s.}$$

L'ensemble $\mathcal{M}$ étant compact, en utilisant la semicontinuité inférieure de l'application $A \mapsto h(A)$ pour la topologie induite par la métrique $L^1(\mathcal{M})$, ainsi qu'un argument de compacité (voir Henrot et Pierre, 2005), nous en déduisons les résultats suivants. Soit $\widehat{A}_n \in \mathcal{B}_n$ minimisant h_n, i.e., $h_n(\widehat{A}_n) = \widehat{h}_n$. Les événements suivant sont de probabilité 1 :

(i) La suite $\widehat{A}_n$ est séquentiellement compacte. En particulier, elle admet des valeurs d'adhérence, i.e., il existe un sous-ensemble A_∞ de $\mathcal{M}$ et une sous-suite A_{n_k} tels que les fonctions indicatrices $\mathbf{1}_{A_{n_k}}$ convergent vers $\mathbf{1}_{A_\infty}$ dans $L^1(\mathcal{M})$.

(ii) Si A_∞ est une valeur d'adhérence de $\widehat{A}_n$, alors $h(A_\infty) = h(\mathcal{M})$.

Ainsi, presque sûrement, toute suite $\widehat{A_n}$, construite en minimisant h_n pour tout n, admet une sous-suite convergente, et ses valeurs d'adhérence sont des sous-ensembles de Cheeger de $\mathcal{M}$. Par suite, un algorithme de partitionnement de graphe fondé sur la minimisation de la conductance converge vers la partition optimale du support de la loi des observations.

Bibliographie

[1] Buser, P. (1982). A note on the isoperimetric constant. *Annales Scientifiques de l'E.N.S.*, Tome 15 (2), p. 213-230.
[2] Chung, F. (2007). Random walks and local cuts in graphs, *Linear Algebra and its Applications*, **443**, 22-32.
[3] de la Pena, V.H. and Giné, E. (1999). *Decoupling. From Dependence to Independence.* Springer-Verlag, New-York.
[4] Henrot, A. and Pierre, M. (2005). *Variation et optimisation de formes. Une analyse géométrique.* Springer-Verlag, Berlin Heidelberg.