Fonctions ζ et topologie stochastique

Daniel Perez [1,2,3]

[1] École normale supérieure

[2] Université Paris-Saclay

[3] DataShape, INRIA

29 septembre 2021

Table des matières

Notation

Notons X un espace topologique connexe, localement connexe par arcs et compact et $f : X \to \mathbb{R}$ une fonction continue.

On peut associer à f un arbre T_f de la manière suivante

Théorème (Arbre de surniveau, P.)

La fonction $d_f : X \times X \to \mathbb{R}$ définie par

$$d_f(x,y) := f(x) + f(y) - 2 \sup_{\gamma : x \mapsto y} \inf_{t \in [0,1]} f \circ \gamma \tag{1.1}$$

est une pseudo-distance sur X et satisfait

$$d_f(x,y) = 0 \iff x, y \in \{f = r\} \text{ et même comp. connexe de } \{f \geq r\}. \tag{1.2}$$

De plus, $T_f := X/\{d_f = 0\}$ muni de la topologie quotient est un arbre réel (enraciné).

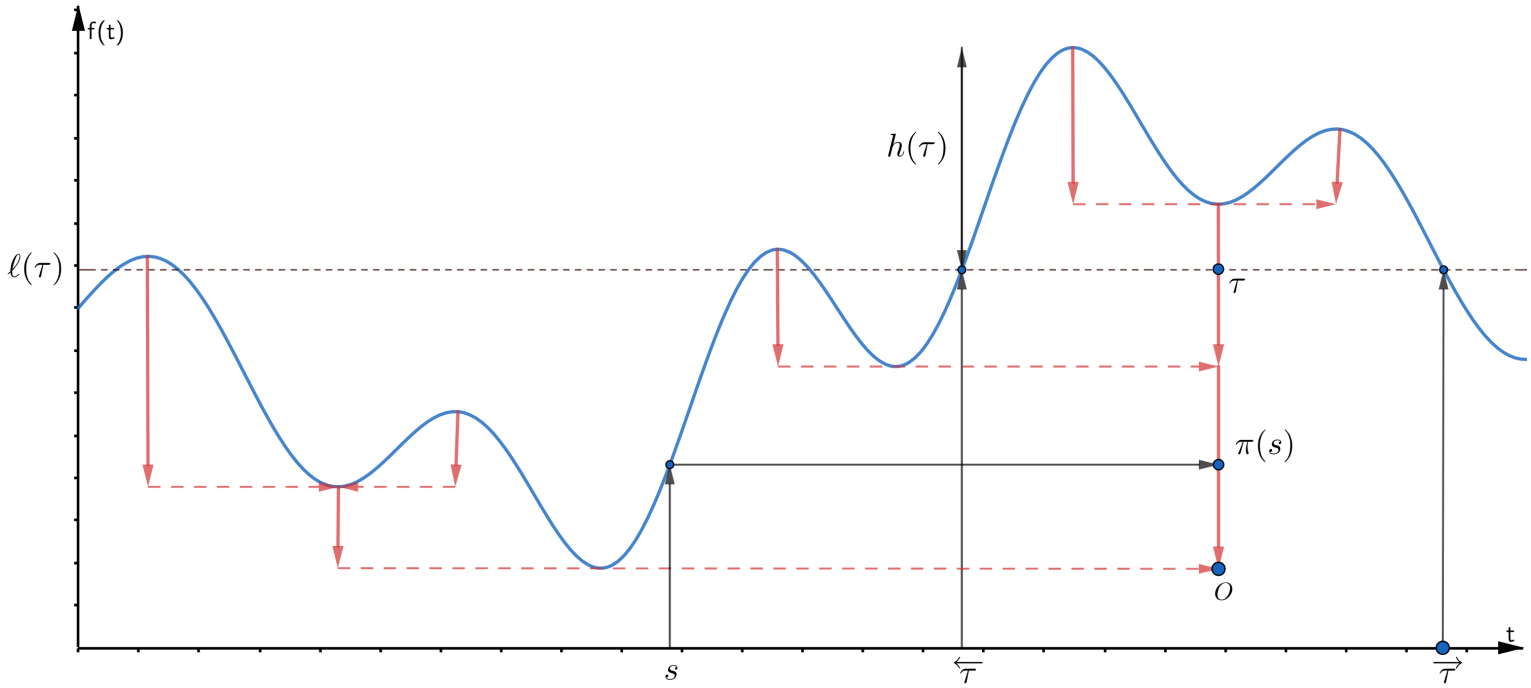


FIGURE – Arbre associé à une fonction sur $[0, 1]$ et quantités associées.

Notation

On notera $\pi_f : X \to T_f$ la projection canonique, et $h : T_f \to \mathbb{R}$ la fonction associant la distance entre la point τ et la feuille la plus haute au-dessus de τ.

Définition

Avec la même fonction h sur l'arbre on peut définir l'arbre ε-simplifié

$$T_f^\varepsilon := \{\tau \in T_f \mid h(\tau) \geq \varepsilon\} \tag{1.3}$$

Remarque

Par compacité et locale connexité de T_f, T_f^ε est toujours un arbre fini.

Définition

On définit N^ε comme étant le nombre de feuilles de l'arbre T_f^ε.

Construction du code-barres sur l'arbre

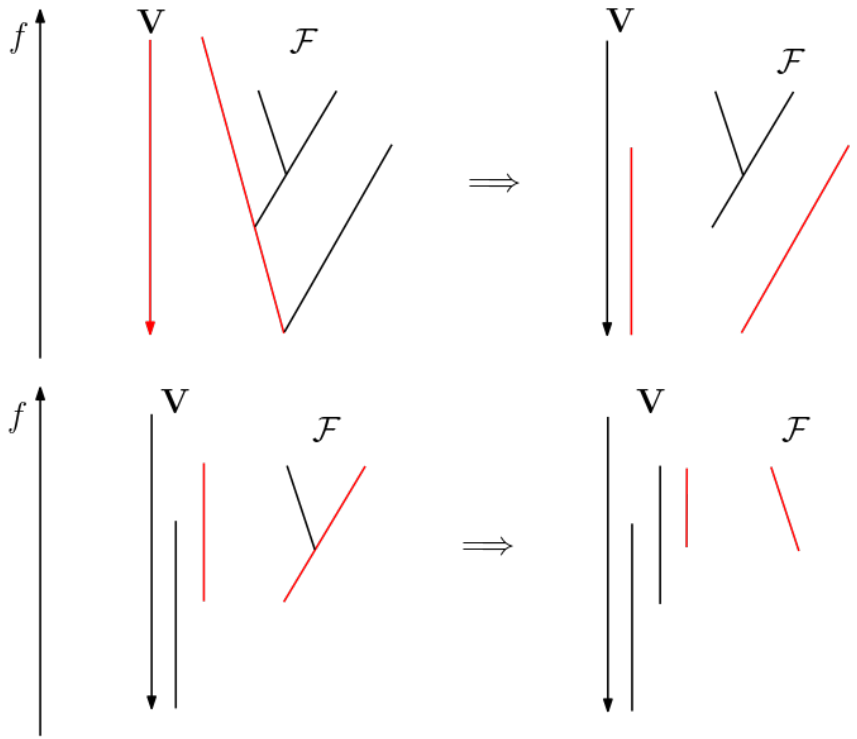


FIGURE – Construction du code-barres à partir de l'arbre.

Remarque

Si on fixe r, le code-barres quantifie le nombre de composantes connexes de $\{f \geq r\}$. En revanche, l'arbre contient plus d'informations que la donnée seule du code-barres.

Définition

Nous noterons N^ε le nombre de barres de $\mathcal{B}(f)$ ayant une longueur $\geq \varepsilon$.

Remarque

N^ε correspond au nombre de variations de taille au moins ε de f.

Définition

On définit la norme ℓ_p d'une fonction f par

$$\ell_p(f) := \left[\sum_{b \in \mathcal{B}(f)} \ell(b)^p \right]^{1/p}. \tag{2.4}$$

où $\mathcal{B}(f)$ est le code-barres de f.

Remarque

Il est commode de faire le saut conceptuel de considérer la fonction $\ell_p^p(f)$ pour $p \in \mathbb{C}$.

Proposition (P.)

Les fonctions N^ε et ℓ_p^p sont duales au sens où

$$\ell_p^p(f) = p \int_0^\infty \varepsilon^{p-1} N^\varepsilon \, d\varepsilon \tag{2.5}$$

$$N^\varepsilon = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} \varepsilon^{-p} \ell_p^p(f) \, \frac{dp}{p} \tag{2.6}$$

Démonstration.

On exprime

$$\ell(b)^p = p \int_0^{\ell(b)} \varepsilon^{p-1}\, d\varepsilon = p \int_0^{\infty} 1_{[0,\ell(b)]}(\varepsilon)\, \varepsilon^{p-1}\, d\varepsilon\,, \tag{2.7}$$

en applicant le théorème de Tonelli,

$$\ell_p^p(f) = p \int_0^{\infty} \underbrace{\left[\sum_{b \in \mathcal{B}(f)} 1_{[0,\ell(b)]}(\varepsilon) \right]}_{=: N^{\varepsilon}} \varepsilon^{p-1}\, d\varepsilon\,. \tag{2.8}$$

Par le théorème d'inversion de Mellin, on a la deuxième équation. □

Théorème (Picard, §3[8] et[6])

Pour une fonction continue $f : [0,1] \to \mathbb{R}$,

$$\mathcal{V}(f) = \mathcal{L}(f) = \limsup_{\varepsilon \to 0} \frac{\log N^{\varepsilon}}{\log(1/\varepsilon)} \vee 1 \tag{2.9}$$

où $a \vee b := \max\{a, b\}$ *et*

$$\mathcal{V}(f) := \inf\{p \mid \|f\|_{p-var} < \infty\} \quad \textit{et} \quad \mathcal{L}(f) := \inf\{p \mid \ell_p < \infty\}. \tag{2.10}$$

Remarque

A priori ℓ_p^p *est holomorphe sur le demi-plan* $\mathrm{Re}(p) > \mathcal{L}(f)$.

Maintenant, considérons f un processus stochastique (p.s. continu) sur X, *i.e.*

Définition

$f : \Omega \times X \to \mathbb{R}$ avec $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé.

Définition

La fonction ζ de f est définie par

$$\zeta_f(p) := \mathbb{E}\left[\ell_p^p(f)\right] \tag{3.11}$$

Remarque

Pour des processus "canoniques", ζ_f admet en fait un prolongement méromorphe à tout le plan complexe.

Remarque

Dorénavant $X = [0, t]$ et on autorisera f p.s. càdlàg.

Fait (Flajolet [4])

On a la correspondance suivante

$$\textit{Extensions méro. de } \zeta \textit{ (à gauche, à droite)}$$
$$\Longleftrightarrow$$
$$\textit{Développements (hyper)asymptotiques de } \mathbb{E}[N^{\varepsilon}] \,(\textit{lorsque } \varepsilon \to 0, \varepsilon \to \infty)\,.$$

Remarque

Le fait ci-haut s'applique de manière générale aux fonctions et leur transformées de Mellin.

Théorème (P.)

La fonction ζ du mouvement brownien B sur l'intervalle $[0,t]$ admet un prolongement méromorphe partout sur $\mathbb{C}$. De plus, elle est exactement égale à

$$\zeta_B(p) = \frac{4(2^p-3)}{\sqrt{\pi}} \left(\frac{t}{2}\right)^{\frac{p}{2}} \Gamma\left(\frac{p+1}{2}\right) \zeta(p-1) \tag{3.12}$$

pour tout p et elle admet un unique pôle simple en $p=2$ de résidu $[B]_t = t$.

Théorème (P.)

En fait, la fonction ζ de toute semimartingale S admet un pôle simple à $p=2$ de résidu $[S]_t$.

Proposition (P.)

Dans le cas du mouvement brownien, $\mathbb{E}[N^\varepsilon]$ admet les deux répresentations suivantes qui convergent bien pour grand et petit ε respectivement

$$\mathbb{E}[N^\varepsilon] = \frac{t}{2\varepsilon^2} + \frac{2}{3} + o(\varepsilon^n) \text{ lorsque } \varepsilon \to 0. \tag{3.13}$$

Démonstration.

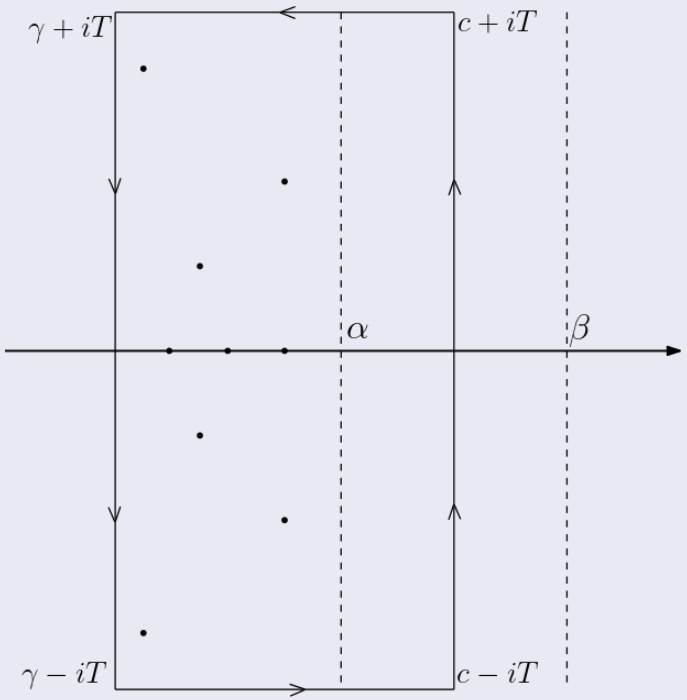
C'est la correspondance fondamentale. On applique l'inversion de Mellin à $\zeta_B(p)/p$.

$$\mathbb{E}[N^\varepsilon] = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} \varepsilon^{-p} \zeta_B(p) \, \frac{dp}{p} \tag{3.14}$$

On a deux pôles, un à $p = 2$ et un autre à $p = 0$. □

Suite.

On intègre le long du contour suivant



Par le théorème des résidus, on a le résultat. □

En fait on a mieux,

Proposition (P.)

Dans le cas du mouvement brownien, $\mathbb{E}[N^\varepsilon]$ admet les deux répresentations suivantes qui convergent bien pour grand et petit ε respectivement

$$\mathbb{E}[N^\varepsilon] = 4\sum_{k\geq 1}(2k-1)\,\mathrm{erfc}\left(\frac{(2k-1)\varepsilon}{\sqrt{2t}}\right) - k\ \mathrm{erfc}\left(\frac{2k\varepsilon}{\sqrt{2t}}\right) \tag{3.15}$$

$$= \frac{t}{2\varepsilon^2} + \frac{2}{3} + 2\sum_{k\geq 1}(2(-1)^k - 1)\frac{e^{-\frac{\pi^2 k^2 t}{2\varepsilon^2}}\, t}{\varepsilon^2}\left[1 + \frac{\varepsilon^2}{\pi^2 k^2 t}\right]. \tag{3.16}$$

Remarque

On peut le démontrer en prennant la transformée de Mellin inverse explicitement. On obtient la deuxième expression comme image inverse par la transformée de Mellin de ζ_B en utilisant l'équation fonctionnelle de la fonction ζ de Riemann.

Définition

Un processus stochastique f sur $\mathbb{R}_+$ est dit de Lévy si

1. *p.s. $f(0) = 0$;*
2. *Les accroissements de f sont indépendants, i.e. pour toute partition finie $(t_i)_i$ de $\mathbb{R}_+$, $(f(t_i) - f(t_{i-1}))_i$ sont mutuellement indépendants ;*
3. *Les accroissements de f sont stationnaires, i.e. pour tout $s < t \in \mathbb{R}_+$, $f(t-s) = f(t) - f(s)$ en distribution ;*
4. *f est continu en probabilité.*

Définition

Un processus de Lévy est dit α stable s'il existe α tel que pour tout t et tout c, $f(c^\alpha t) = cf(t)$ en distribution.

Remarque

En particulier, le mouvement brownien est un processus de Lévy 2-stable.

Théorème (P.)

Si f est un processus de Lévy α-stable, il existe une variable aléatoire U avec des moments finis telle que la fonction ζ du processus admet un prolongement méromorphe partout sur $\mathbb{C}$ avec un seul pôle simple en $p = \alpha$ de résidu $\mathbb{E}[U]\,\alpha t$.

Théorème (P.)

Pour un processus de Lévy α-stable sur $[0, t]$, on a le développement asymptotique suivant

$$\mathbb{E}[N^\varepsilon] = \frac{t}{\mathbb{E}[U]\,\varepsilon^\alpha} + \frac{\mathbb{E}\left[U^2\right]}{2\mathbb{E}[U]^2} + o(\varepsilon^{\alpha n}) \quad \textit{lorsque } \varepsilon \to 0\,, \tag{3.17}$$

pour tout $n \in \mathbb{N}$. De plus, p.s.

$$N^\varepsilon \sim \frac{t}{\mathbb{E}[U]\,\varepsilon^\alpha} \quad \textit{lorsque } \varepsilon \to 0\,. \tag{3.18}$$

Théorème (P., Théorème 3.1 [7])

Soit $\delta_N := \|X - X_N\|_{L^\infty} \to 0$ p.s. quand $N \to \infty$, alors pour tout $\varepsilon \geq 2\delta_N$

$$N_X^{\varepsilon+\delta_N} \leq N_{X_N}^{\varepsilon} \leq N_X^{\varepsilon-\delta_N} \tag{3.19}$$

De plus, si $\mathbb{E}\left[N_X^\varepsilon\right]$ est continue en ε, alors

$$N_{X_N}^\varepsilon \xrightarrow[N\to\infty]{L^1} N_X^\varepsilon \quad \text{et} \quad N_{X_N}^\varepsilon \xrightarrow[N\to\infty]{\mathbb{P}} N_X^\varepsilon \,,$$

à N fixé, ceci se traduit quantitativement par

$$\mathbb{E}\left[\left|N_X^\varepsilon - N_{\hat{X}}^\varepsilon\right|\right] \leq 2\omega_\varepsilon(\delta_N) \quad \text{et} \quad \mathbb{P}\left(\left|N_X^\varepsilon - N_{\hat{X}}^\varepsilon\right| \geq k\right) \leq \frac{2\omega_\varepsilon(\delta_N)}{k} \tag{3.20}$$

où ω_ε est le module de continuité de $\mathbb{E}\left[N_X^\varepsilon\right]$ sur l'intervalle $[\varepsilon - \delta_N, \varepsilon + \delta_N]$. Enfin, on a les inégalités suivantes

$$N_{X_N}^{\delta_n} \geq N_X^{2\delta_N} \quad \text{and} \quad N_X^{\delta_N} \geq N_{X_N}^{2\delta_N} \,.$$

Soit f un processus α-stable.

1. Échantillonner M chemins du processus stochastique f (par exemple à intervalles réguliers de taille $\frac{1}{N}$ pour un certain N) ;
2. Calculer le code-barres des chemins échantillonnés.
3. Pour *une certaine plage de valeurs suffisamment petites* ε, et pour une un $c > 1$, calculer

$$\hat{\alpha}_M := \log_c \left[\frac{\overline{N}_t^{\varepsilon/c} - \overline{N}_t^{2\varepsilon/c}}{\overline{N}_t^{\varepsilon} - \overline{N}_t^{2\varepsilon}} \right] . \tag{3.21}$$

   Ici, la notion de *une certaine plage de valeurs suffisamment petites* ε et la constante c dépendent toutes deux de N, avec la condition limite que lorsque $N \to \infty$, la borne inférieure de la plage de ε valides s'approche de 0.

En ignorant les contributions superpolynomiales l'argument à l'intérieur du log de l'estimateur est approx.

$$c^{\hat{\alpha}_M} \approx \frac{\frac{t}{\mathbb{E}[U](\varepsilon/c)^\alpha} - \frac{t}{\mathbb{E}[U](2\varepsilon/c)^\alpha}}{\frac{t}{\mathbb{E}[U]\varepsilon^\alpha} - \frac{t}{\mathbb{E}[U](2\varepsilon)^\alpha}} \approx c^\alpha . \tag{3.22}$$

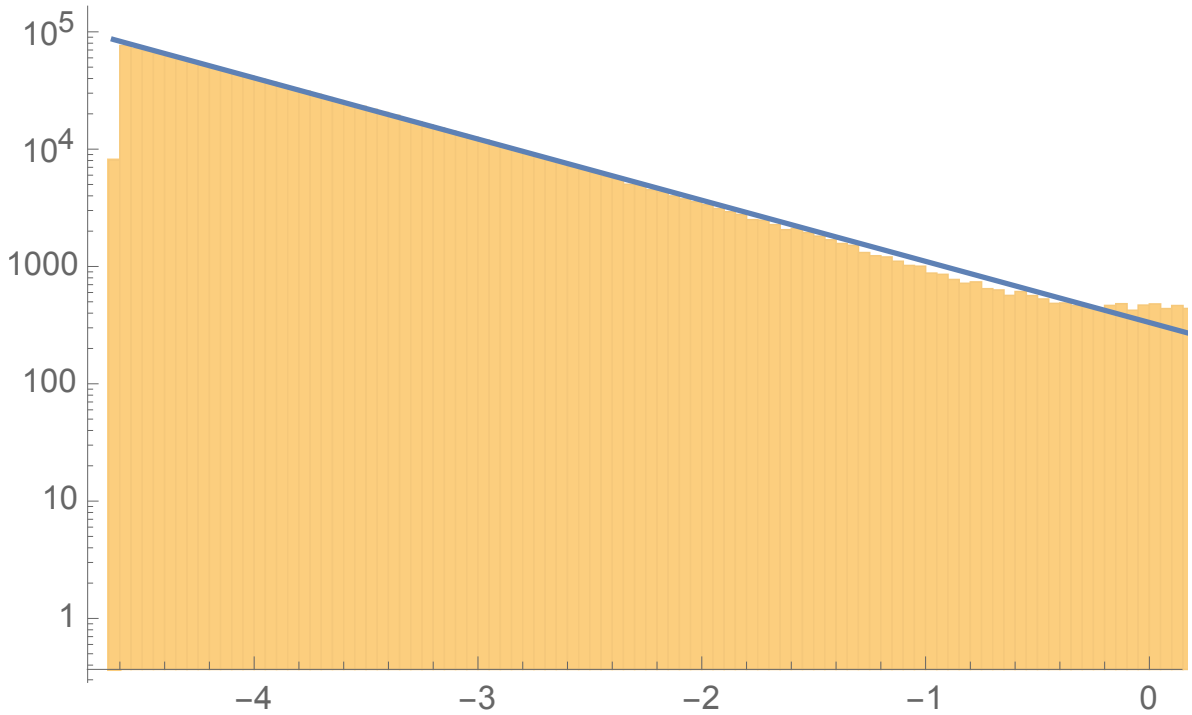


FIGURE – Test statistique pour un processus 1.2-stable. La pente de la droite est bien 1.2.

- Les prolongements analyitiques de ζ_f ne sont pas triviaux. Il serait intéressant d'étudier les séries de Dirichlet associées (déjà quelques résultats dans cette direction...) ;
- ℓ_p n'est pas stable par perturbations L^∞, mais sa variable duale N^ε l'est et semble être robuste statistiquement au vu de nos résultats pour les processus α-stables ;
- Ceci donne une explication qualitative sur le fait que ℓ_p soit mieux adapté dans le cadre de l'apprentissage statistique ;
- Il serait intéressant d'étudier les fonctions ζ de fonctions sur des espaces plus généraux de plus grande dimension.

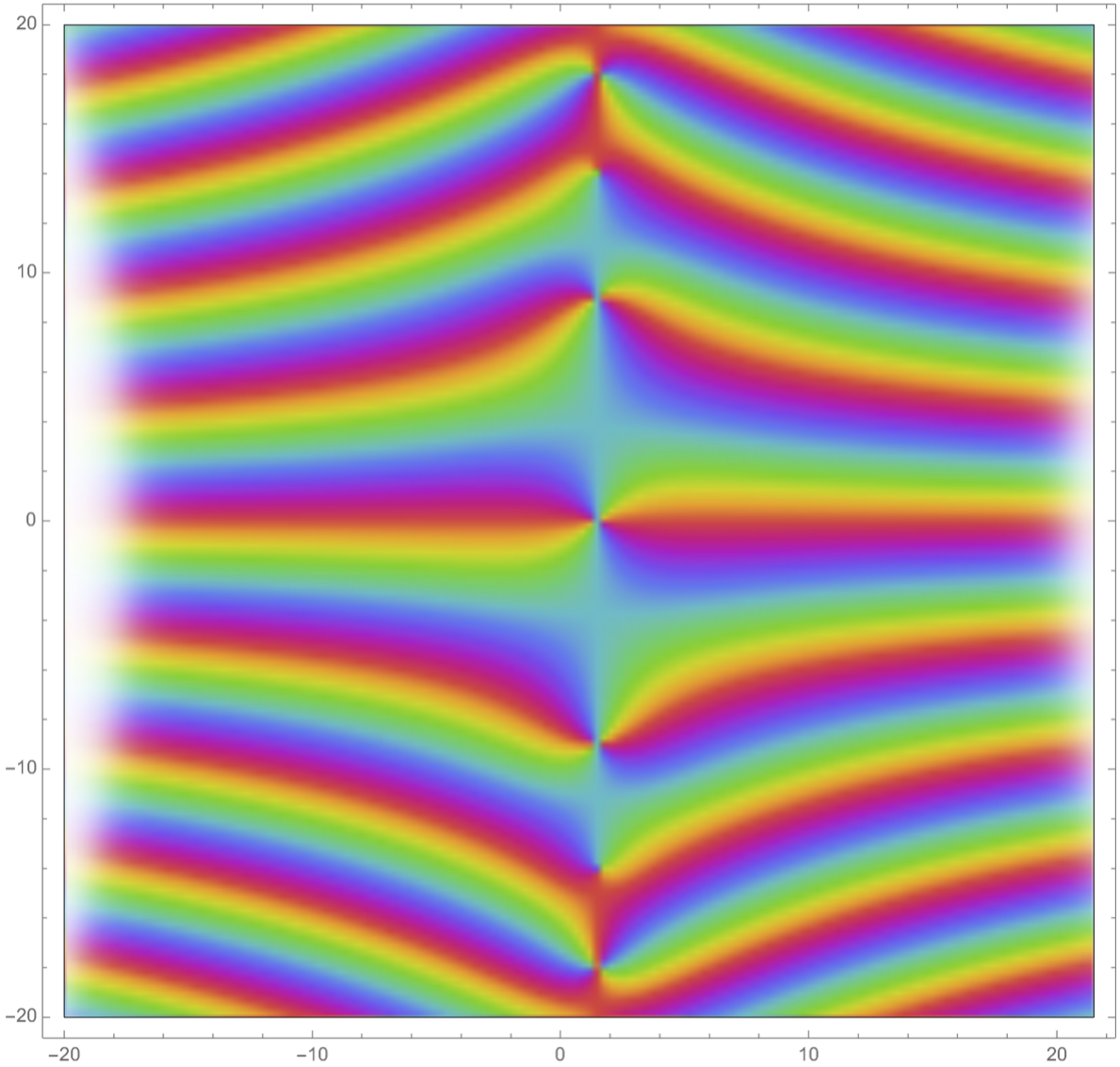


FIGURE – La fonction η_B.

Notation

Dans la suite, soit X une variété et $f : X \to \mathbb{R}$ une fonction continue.

Remarque

La fonction f induit une filtration de l'espace X par

$$X_r := \{f \geq r\} \tag{4.23}$$

Définition

Fixons un corps k. L'homologie persistante du couple (X, f) est un foncteur de la petite catégorie (partiellement ordonnée) $\mathbb{R}$ vers Vect_k

$$H_*(X, f) : r \mapsto H_*(X_r, k) \tag{4.24}$$

$$[s \to r] \mapsto [H_*(X_s, k) \to H_*(X_r, k)] \tag{4.25}$$

Théorème (Décomposition, [2, 5])

Pour les fonctions continues $f : X \to \mathbb{R}$, l'homologie persistante admet toujours une décomposition en des "modules d'intervalle"

$$H_*(X, f) := \bigoplus_{(a,b)} k[a, b[\tag{4.26}$$

où les modules d'intervalle sont des foncteurs $\mathbb{R} \to \mathrm{Vect}_k$ définis par :

$$k[a, b[: r \mapsto \begin{cases} k & si\ r \in [a, b[\\ 0 & sinon \end{cases} \tag{4.27}$$

$$[s \to r] \mapsto \begin{cases} \mathrm{id} & si\ s, r \in [a, b[\\ 0 & sinon \end{cases} \tag{4.28}$$

Remarque

Toute l'information de $H_*(X, f)$ est contenue dans les couples $(a, b) \in \mathbb{R}^2$.

Remarque

Dans la suite, nous ne considérerons l'homologie persistante qu'en degré 0. Dans ce cas, on s'intéresse juste aux composantes connexes.

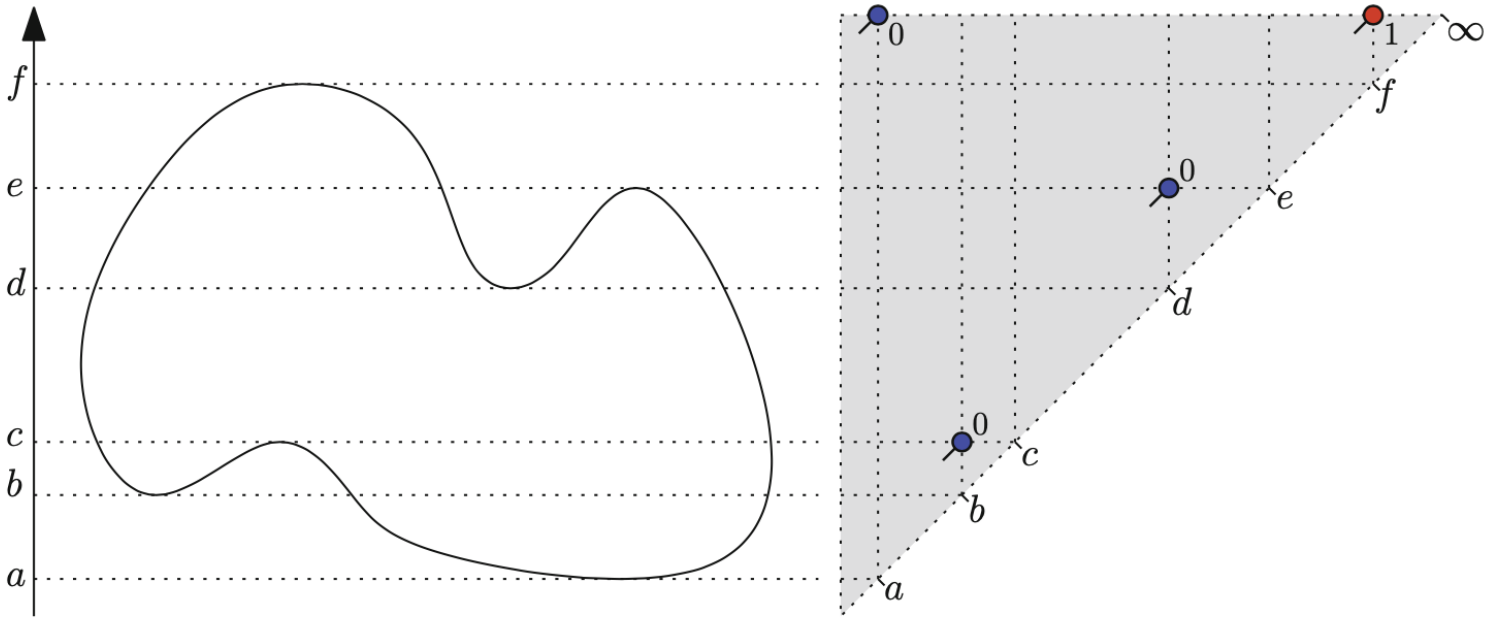


FIGURE – Diagramme de persistance associé à une filtration du cercle par ensembles de *sous-niveaux* [2].

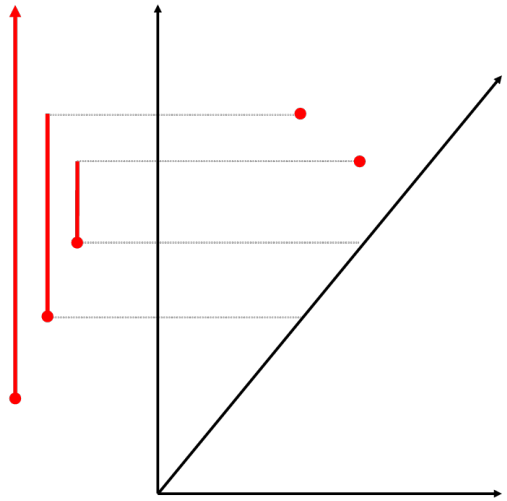


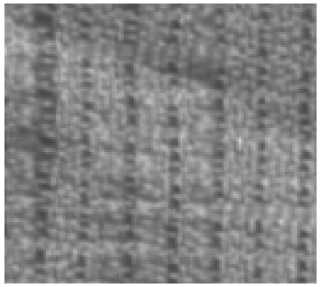
FIGURE – Diagrammes de persistance et code-barres.

Remarque

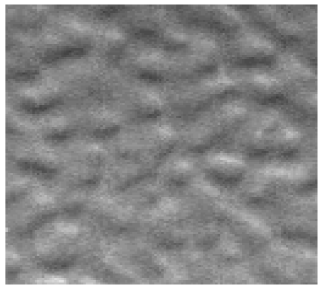
Il existe une distance naturelle sur l'espace des diagrammes appelée "distance bottleneck", stable par perturbations C^0 de f. Cette distance "oublie" les petites barres du code-barres.

Remarque

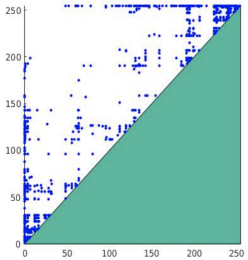
En pratique, les petites barres identifient les "phénomènes haute fréquence" comme la texture. Pour certaines applications, elles sont donc fondamentales.



Label = Canvas



Label = Carpet



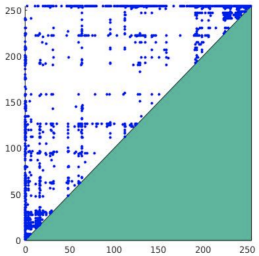


FIGURE – Diagrammes associés à différentes textures [1].

Définition

$$d_p(a,b) = \left[\inf_{\pi \in \Gamma(a,b)} \sum_{x \in a \cup \partial\Omega} d(x, \pi(x))^p \right]^{\frac{1}{p}} \tag{4.29}$$

où d est la norme q sur $\mathbb{R}^2$, $\Gamma(a,b)$ est l'ensemble des bijections entre $a \cup \partial\Omega$ et $b \cup \partial\Omega$.

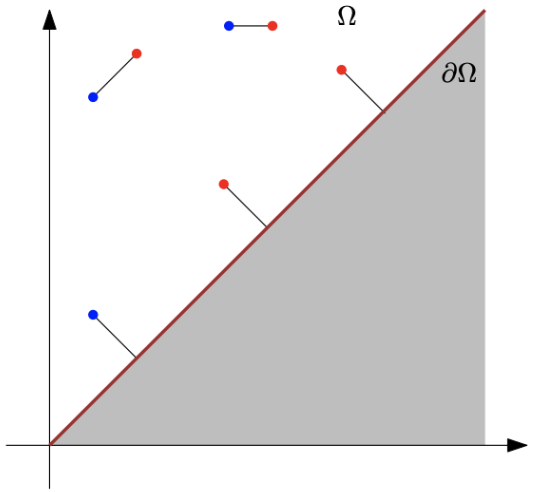


FIGURE – Matching partiel entre deux diagrammes a et b [3].

Remarque

ℓ_p est la distance entre le diagramme et la diagonale.

Théorème

Si $f = M + A$ est une semimartingale continue tel que pour tout $s \geq 1$

$$\mathbb{E}\left[[M]_t^s + \left(\int_0^t |dA|_s\right)^s\right] < \infty\,. \tag{4.30}$$

Alors dans L^s on a

$$N^\varepsilon \sim \frac{[f]_t}{2\varepsilon^2} \quad \textit{lorsque } \varepsilon \to 0\,. \tag{4.31}$$

En particulier ζ_f admet un pôle simple à $p = 2$ de résidu $[f]_t$.

Remarque

Si $f = B$, l'énoncé du théorème est vrai p.s.


M. Carriere, S. Oudot, and M. Ovsjanikov.
Sliced Wasserstein Kernel for Persistence Diagrams.
In *ICML 2017 - Thirty-fourth International Conference on Machine Learning*, pages 1–10, Sydney, Australia, Aug. 2017.


F. Chazal, V. de Silva, M. Glisse, and S. Oudot.
The Structure and Stability of Persistence Modules.
Springer International Publishing, 2016.


V. Divol and T. Lacombe.
Understanding the topology and the geometry of the persistence diagram space via optimal partial transport.
ArXiv, abs/1901.03048, 2019.


P. Flajolet, X. Gourdon, and P. Dumas.
Mellin transforms and asymptotics : Harmonic sums.
Theoretical Computer Science, 144(1) :3–58, 1995.


S. Y. Oudot.
Persistence Theory - From Quiver Representations to Data Analysis, volume 209 of *Mathematical surveys and monographs*.
American Mathematical Society, 2015.


D. Perez.
On C^0-persistent homology and trees.
https://arxiv.org/abs/2012.02634, Dec. 2020.


D. Perez.

On the persistent homology of almost surely C^0 stochastic processes.
https://arxiv.org/abs/2012.09459, Dec. 2020.

J. Picard.
A tree approach to p-variation and to integration.
The Annals of Probability, 36(6) :2235–2279, Nov 2008.