Introduction au Domaine de Recherche

—

Estimation de Métriques basées sur la Mesure

Jérôme Taupin

Mai 2025

Résumé

Quand un ensemble de données statistiques existe dans un espace de grande dimension, la distance euclidienne n'est pas toujours la plus adaptée pour évaluer la proximité réelle entre les objets initiaux représentés par deux points donnés. On s'intéresse ici à des déformations de la métrique euclidienne, appelées changements conformes, via une fonction positive qui dépend de la distribution sous-jacente aux données. Ce type de métriques a déjà été beaucoup étudiée ces dernières années et s'avère un outil puissant en science des données, par exemple pour traiter des problèmes d'apprentissage semi-supervisé ou de classification, là où la métrique euclidienne n'exploite pas efficacement la géométrie de la distribution sous-jacente aux données. La distance de Fermat est un exemple de métrique qui déforme l'espace en rapprochant les points dans les zones de forte densité de la mesure. Cependant, plusieurs problèmes se posent concernant son analyse théorique, et sa définition se restreint aux mesures possédant une densité. On introduit une variante de la distance de Fermat définie pour n'importe quelle mesure, qui consiste à remplacer la densité par la *distance-à-la-mesure* (DTM) introduite par (Chazal et al. 2011) dans le but de palier ces limitations. Il en découle une nouvelle métrique, appelée *Fermat-distance-à-la-mesure* (FDTM), pour laquelle un certain nombre de résultats de stabilité par rapport à la mesure et d'estimation ont déjà été montrés (Taupin et al. 2025). Ce document constitue une introduction à ces métriques ainsi qu'à divers outils de géométrie et de probabilités qui permettent d'analyser leur estimation.

1 Introduction

Je présente dans ce document un sujet sur lequel je vais travailler dans le cadre de ma thèse au Laboratoire de Mathématiques d'Orsay sous la direction de Frédéric Chazal. Ce sujet s'inscrit dans le domaine de l'analyse topologique des données, un domaine relativement récent qui s'intéresse à l'étude des propriétés topologiques et géométriques de données statistiques. L'utilisation de métriques adaptées à la distribution des données est un exemple d'outil typiquement utilisé pour inférer ce genre d'informations. De plus, cette démarche présente un intérêt particulier dans l'optique de réduction de dimension des données, un sujet très important en statistiques.

Changements conformes La géométrie conforme désigne l'étude de variétés munies d'une classe d'équivalence de métriques riemanniennes, où deux métriques sont équivalentes si elles sont égales modulo une déformation contrôlée par une fonction lisse appelée facteur conforme. Les résultats recherchés dans ce domaine sont souvent pour une variété donnée des questions d'existence d'un changement conforme particulier qui exhibe des propriétés remarquables, par exemple une courbure constante (Yamabe 1960). Ici, on emprunte le terme "changement conforme" (ou métrique conforme) pour désigner la déformation de la métrique riemannienne. Le point de vue est cependant assez différent : on fixe un changement conforme particulier et on cherche à obtenir des informations quantitatives sur la métrique associée et la géométrie de ses plus court chemins. Soit $\mathcal{M}$ une sous-variété différentielle connexe de $\mathbb{R}^D$.[1] Le changement conforme par une fonction strictement positive f est une distance sur $\mathcal{M}$ définie par

$$D_{\mathcal{M},f}(x,y) \stackrel{\text{def}}{=} \inf_{\gamma\in\Gamma_{\mathcal{M}}(x,y)} \int_0^1 f(\gamma(t))\|\dot\gamma(t)\|\mathrm{d}t \tag{1}$$

où $\Gamma_{\mathcal{M}}(x,y)$ est l'ensemble des chemins Lipschitz-continus $\gamma : [0,1] \longrightarrow \mathcal{M}$ reliant x à y.[2] On rappelle que pour un tel chemin γ, le théorème de Rademacher affirme que la dérivée $\dot\gamma$ est définie presque partout et la longueur euclidienne de γ est finie :

$$|\gamma| \stackrel{\text{def}}{=} \int_0^1 \|\dot\gamma\| < \infty\,.$$

Ce type de métrique s'inscrit dans le cadre des espaces de longueur tel que définis dans (Burago et al. 2001). En particulier, (Burago et al. 2001, Proposition 2.5.19) garantie l'existence quelque soient les extrémités x et y d'une *géodésique minimisante* — raccourci en géodésique dans le reste de ce document — c'est-à-dire d'un chemin γ réalisant l'infimum en Équation (1). Si f est la fonction constante égale à 1, $D_{\mathcal{M},1}$ coïncide avec la métrique riemannienne $D_{\mathcal{M}}$ induite par l'espace ambient $\mathbb{R}^D$. La Section 2 s'intéresse au problème d'estimation d'une métrique conforme à partir d'un nuage de points, introduisant notamment le concept géométrique de *reach* pour y parvenir. En pratique, on s'intéresse à des facteurs conformes f qui dépendent d'une mesure μ sur $\mathcal{M}$.

Distance de Fermat[3] Étant donnée une mesure de probabilité μ à densité ρ par rapport à la forme volume sur $\mathcal{M}$ — qui joue le rôle de mesure uniforme sur $\mathcal{M}$ — et un paramètre $\beta \geq 0$, la distance de Fermat *continue* est le changement conforme par $\rho^{-\beta}$:

$$F_{\mu,\beta}(x,y) \stackrel{\text{def}}{=} D_{\mathcal{M},\rho^{-\beta}}(x,y) = \inf_{\gamma} \int_\gamma \rho^{-\beta}\,. \tag{2}$$

D'autre part, pour un nuage de point $X \subset \mathbb{R}^D$ et un paramètre $\alpha > 1$, la distance Fermat *discrète* est définie comme la distance induite par le graphe pondéré de sommets X et de poids les distances euclidiennes à la

1. On pourrait se placer dans un cadre plus général que celui des variétés. Ici on s'assure simplement d'avoir un cadre commun pour les différents points abordés dans ce document.

2. On considère généralement l'ensemble plus large des chemins rectifiables, ce qui signifie essentiellement que la notion de longueur du chemin a du sens. Un chemin étant rectifiable si et seulement si il peut-être reparamétré en chemin Lipschitz–continu, ce choix n'a aucun impact ici.

3. Le terme *distance de Fermat* est inspiré du principe de Fermat en optique, qui affirme que le chemin emprunté par la lumière minimise l'intégrale de l'indice de réfraction. Ici, une puissance négative de la densité joue le rôle de cet indice de réfraction.

puissance α :

$$\hat{F}_{X,\alpha}(x,y) \stackrel{\text{def}}{=} \min_{x=x_1,x_2,\ldots,x_k=y} \sum_{i=1}^{k} \|x_i - x_{i+1}\|^\alpha \tag{3}$$

où l'infimum est pris sur l'ensemble des suites finies de points dans X d'extrémités x et y. Cette définition favorise les chemins composés de courtes arêtes, et donc les détours dans des zones de forte densité de points. Si X_n est un échantillon de n points tirés de manière i.i.d. à partir de la mesure μ, alors il est montré que sous des hypothèses appropriées la distance de Fermat discrète $\hat{F}_{X_n,\alpha}$ converge en un sens vers la distance de Fermat continue (Groisman et al. 2022) à mesure que le nombre de points augmente :

$$\lim_{n\to+\infty} n^\beta \hat{F}_{X_n,\alpha} = \lambda_{\alpha,d} F_{\mu,\beta} \qquad \text{p.s.} \tag{4}$$

où α et β sont reliés par la relation $\beta = \frac{\alpha-1}{d}$ et $\lambda_{\alpha,d}$ est une constante non connue. La capacité à estimer la distance de Fermat à partir d'une métrique très simple sur un échantillon de points est intéressante, cependant la constante $\lambda_{\alpha,d}$ qui apparaît dû à des phénomènes de percolation de premier passage constitue un obstacle à la bonne compréhension de cette convergence et de la variance du processus. En particulier, il n'existe aucune estimée de vitesse de convergence à moins d'imposer des hypothèses très strictes sur la régularité de la densité ρ (Fernández et al. 2023 ; Garcia Trillos et al. 2024). Par ailleurs, la définition de la distance de Fermat continue impose l'existence d'une densité f, ce qui restreint toute garantie de convergence de la distance de Fermat discrète aux mesures à densité. Ces limitations sont la motivation pour l'introduction de la FDTM.
Concernant les applications de la distance de Fermat,des exemples possibles sont l'application à l'estimation de Laplaciens de graphe pour de la segmentation d'images (Garcia Trillos et al. 2018) et l'estimation de diagrammes de persistence (Fernández et al. 2023). On s'attend à ce que la FDTM puisse être utilisée pour ces mêmes applications.

Distance-à-la-mesure La DTM d'une mesure μ sur $\mathbb{R}^D$ est une fonction s'interprétant comme la distance à une certaine fraction $m \in (0,1)$ de masse de μ. Elle se comporte ainsi similairement à l'inverse de la densité ρ de μ quand cette dernière existe, mais est définie en toute généralité et sur l'espace tout entier. Cette similitude est renforcée par le fait que la DTM converge (modulo une normalisation appropriée) vers $\rho^{-\beta}$ pour un certain β quand le paramètre m s'approche de 0 (Biau et al. 2011). Par ailleurs, la DTM dispose de plusieurs propriété de stabilité (Chazal et al. 2011) qui en font un bon candidat pour induire une métrique robuste aux perturbations de la mesure mais représentative de la géométrie des données. On définit donc la FDTM comme le changement conforme par la DTM.
La définition de ces objets est détaillée en Section 3. Plusieurs résultats ont déjà été obtenus sur la FDTM dans un cadre légèrement différent où le chemins peuvent quitter le support. Ces résultats sont disponible sous la forme d'une pré-publication (Taupin et al. 2025) et les points essentiels sont présentés en Section 3.2. Principalement, il est possible d'estimer la FDTM sous réserve d'hypothèses peu restrictives sur la mesure, et ce avec des garanties explicites et indépendantes de la dimension ambiante sur la vitesse de convergence. Enfin, on conclut en Section 3.3 en donnant une idée des résultats qu'on s'attend à obtenir sur la vitesse d'estimation de la FDTM à l'aide des différents résultats présentés dans ce document.

2 Estimation de la métrique riemannienne induite

Dans cette section, on fixe un domaine $M \subset \mathbb{R}^D$ et l'objectif est de réussir à estimer efficacement la distance $D_M(x, y)$ à partir d'un échantillon de points de M. Le raisonnement détaillé ici pourra ensuite être adapté au cas d'une métrique conforme tant que le facteur conforme est Lipschitz et minoré, comme ce sera le cas pour la DTM. Plusieurs résultats non publiés ont déjà été établis dans ce sens mais ne sont pas détaillés ici. Afin de contrôler la distortion des chemins sur le domaine, on introduit une hypothèse sur la régularité de M à l'aide de la notion de reach.

2.1 Ensembles de reach positif

On note $M^r = \{x \in \mathbb{R}^D : d(x, M) < r\}$ le r-épaississement de M où $d(x, M)$ désigne la distance euclidienne de x à l'ensemble M.

Définition 2.1 (Reach)**.** Soit $M \subset \mathbb{R}^D$. Le *reach* de M, noté τ_M, est le plus grand épaississement de M dont tous les points admettent une unique projection orthogonale sur M.

$$\tau_M \overset{\text{def}}{=} \sup \left\{ r \geq 0 : \forall x \in M^r, \exists! y \in M \mid d(x, M) = d(x, y) \right\}.$$

La projection orthogonale $\pi_M : M^{\tau_M} \to M$ est alors bien définie.

La notion de reach fut introduite dans l'article fondamental (Federer 1959) qui énonce un grand nombre de propriétés sur cet objet et l'utilise pour étudier les mesures de courbure. En effet, le reach permet notamment de généraliser la notion de convexité d'un ensemble et peut être relié à la notion de courbure. Par exemple :

- Une sphère de rayon r a un reach égal à r.
- Un ensemble est convexe si et seulement si son reach est infini.
- Une variété C^2 compacte a nécessairement un reach positif.
- Un ensemble possédant un angle concave est de reach nul, mais un ensemble de reach positif peut posséder un angle convexe, par exemple le carré (rempli).

Dans tout ce qui suit on suppose que M est de reach positif. Cette hypothèse est moins forte que la différentiabilité mais permet tout de même de montrer des résultats intéressants. Notamment, la projection sur un ensemble de reach positif est Lipschitzienne à l'intérieur du reach (Federer 1959).

Proposition 2.2. (Rataj et al. 2019, Lemme 4.7) Supposons $M \subset \mathbb{R}^D$ de reach positif et $r < \tau_M$. Alors la projection orthogonale π_M est Lipschitz sur M^r. Précisément, pour tous $x, y \in M^r$,

$$\|\pi_M(x) - \pi_M(y)\| \leq \frac{\tau_M}{\tau_M - r} \|x - y\|.$$

En particulier, Proposition 2.2 implique que π_M est continue sur M^{τ_M}. La constante de Lipschitz présentée est optimale : si $M = \tau S^1$, $x \in (\tau - r)S^1$ et $y = -x$, on a

$$\|\pi_M(x) - \pi_M(y)\| = \left\| \frac{\tau}{\tau - r} x - \frac{\tau}{\tau - r} y \right\| = \frac{\tau}{\tau - r} \|x - y\|$$

et on réalise donc le cas d'égalité dans Proposition 2.2. Par ailleurs, cette régularité permet de majorer la distance D_M entre deux points de M suffisamment proches.

Proposition 2.3. (Boissonnat et al. 2019, Lemme 3) Soient $x, y \in M$ tels que $\|x - y\| \leq \tau_M$. Alors

$$\|x - y\| \leq D_M(x, y) \leq 2\tau_M \arcsin\left(\frac{\|x - y\|}{\tau_M}\right).$$

Ce contrôle précis sur la distortion de la métrique permet de montrer que les géodésiques sont régulières. Dans le cas C^2 on a le résultat suivant.

Théorème 2.4. (Boissonnat et al. 2019, Lemme 4) Supposons M de classe C^2 et soit $\gamma : [0, |\gamma|] \to M$ une géodésique pour D_M paramétrée à vitesse unitaire, i.e., $\|\dot{\gamma}(t)\| = 1$ presque partout. Alors γ est de classe C^2 et pour tous $s, t \in [0, |\gamma|]$,

$$\|\dot{\gamma}(t) - \dot{\gamma}(s)\| \leq \frac{1}{\tau_M}|t - s|.$$

Théorème 2.4 montre qu'une géodésique pour la métrique riemannienne induite ne peut pas être plus courbée que la variété elle même. Cette régularité va permettre dans ce qui suit d'obtenir des bornes plus fines sur l'estimation de la trajectoire d'une géodésique.

2.2 Approximation de métrique conforme

Étant donné un nuage de points X approximant le domaine M, on se pose maintenant la question de l'approximation d'une géodésique à l'aide d'une ligne polygonale utilisant ces points. Le niveau d'approximation du nuage de point peut se mesurer via la distance de Hausdorff, i.e., la distance maximale entre un point d'un ensemble et son plus proche voisin dans l'autre ensemble. Si $X \subset M$,

$$d_{\mathrm{H}}(M, X) = \sup_{x \in M} \inf_{y \in X} \|x - y\|.$$

Rappelons que dans le cas de la métrique induite par l'espace ambient, la distance $D_M(x, y)$ coïncide avec la longueur euclidienne d'une géodésique entre x et y. Un raisonnement naturel pour approximer une géodésique consiste à la découper en segments de longueur environ égale à r et à projeter chaque étape sur X, puis à relier les projections obtenues. Le chemin γ_r ainsi obtenu reste constamment à distance au plus r de γ et sa longueur totale peut être bornée de la manière suivante de façon élémentaire :

$$\big||\gamma_r| - |\gamma|\big| \leq C \max\left(r, \frac{d_{\mathrm{H}}(M, X)}{r}\right) \tag{5}$$

où C est une constante indépendante de r et de X. Cette borne présente un compromis sur le choix de r : si r est trop grand, le découpage de γ est peu précis ; si r est trop petit, le découpage devient précis mais les déviations de γ_r autour de γ peuvent créer des détours sous-optimaux — voir Figure 1. Dans le cas général, ces fluctuations peuvent être accentuées par les potentielles irrégularités de la géodésique. Le choix optimal pour r dans Équation (5) est $r = d_{\mathrm{H}}(M, X)^{1/2}$ et entraîne une erreur d'approximation du même ordre. Sous des hypothèses de régularité sur γ comme celle donnée par Théorème 2.4, Équation (5) peut être améliorer en remplaçant $d_{\mathrm{H}}(M, X)/r$ par $d_{\mathrm{H}}(M, X)^2/r^2$.

Proposition 2.5. Soit $X \subset M$ un nuage de points et γ une géodésique pour D_M. Alors il existe une constante $C > 0$ telle que pour tout $r \in (0, |\gamma|)$ l'approximation polygonale γ_r de γ vérifie

$$\big||\gamma_r| - |\gamma|\big| \leq C \max\left(r, \frac{d_{\mathrm{H}}(M, X)^2}{r^2}\right) \tag{6}$$

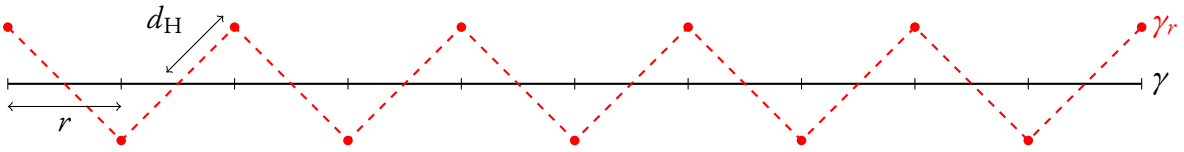

FIGURE 1 – En choisissant r de l'ordre de $d_{\mathrm{H}}(M, X)$, l'approximation polygonale peut osciller autour du chemin original et avoir une longueur significativement plus grande.

Proposition 2.5 est adaptée de (AARON et al. 2018, Théorème 1), dont les hypothèses sont vérifiées d'après Théorème 2.4 et par compacité de M. Le choix optimal de r devient $r = d_{\mathrm{H}}(M, X)^{2/3}$ et l'erreur est du même ordre. Avec une démarche légèrement différente, il est possible d'obtenir des résultats encore plus précis.

Proposition 2.6. (AAMARI et al. 2023, Proposition 5.3) Soit $M \subset \mathbb{R}^D$ de reach positif et $X \subset \mathbb{R}^D$ une partie telle que $d_{\mathrm{H}}(M, X) < \varepsilon \leq \frac{\tau_M}{2}$. Alors $M \subset X^\varepsilon$ et on a

$$\sup_{x \neq y \in M} \left| 1 - \frac{D_{X^\varepsilon}(x, y)}{D_M(x, y)} \right| \leq \frac{2\varepsilon}{\tau_M}.$$

Dans Proposition 2.6, l'erreur devient linéaire en $d_{\mathrm{H}}(M, X)$. À partir de cette inégalité, Aamari et al. obtiennent des bornes minimax-optimales[4] en $O\big((\log(n)/n)^{k/d}\big)$ sur la vitesse d'estimation de D_M dans le cas où M est une variété C^k ($k \geq 2$) de dimension d et de reach positif. Cet estimateur est cependant non réalisable en pratique puisque les chemins admissibles sont l'ensemble des chemins rectifiables à valeurs dans X^ε.

En pratique, on ne connaît pas la géodésique et on ne peut donc pas appliquer la construction précédente pour obtenir γ_r. On peut cependant calculer le plus court chemin de x à y sur le graphe X_n. Parmi les chemins possibles sont les $(\gamma_r)_{r>0}$ donc ce plus court chemin ne ne peut pas être beaucoup plus long que γ, et aucune trajectoire n'est meilleure que celle de γ donc le plus court chemin sur le graphe ne peut pas être beaucoup plus court que γ, à condition que les arêtes ne puissent réaliser des raccourcis trop important par rapport aux chemins permis par le domaine. En n'utilisant que des arêtes de longueur au plus un certain r approprié, on s'assure que $\|x_i - x_{i+1}\|$ soit proche de $D_M(x_i, x_{i+1})$ pour toutes les arêtes et donc que la distance sur le graphe approche $D_M(x, y)$. On définit donc l'estimateur suivant :

$$\hat{D}_{X,r}(x, y) \overset{\text{def}}{=} \min_{\substack{x = x_1, x_2, \ldots, x_k = y \\ \|x_i - x_{i+1}\| \leq r}} \sum_{i=0}^{k-1} \|x_i - x_{i+1}\|. \tag{7}$$

Le minimum est pris sur l'ensemble des familles de points de X allant de x à y et telles que deux points consécutifs sont séparés d'une distance au plus r. Pour prendre en compte un facteur conforme f, il est possible d'adapter Théorème 2.4 pour s'assurer de la régularité des géodésiques, et de bonnes hypothèses sur f permettent alors d'obtenir une approximation de $D_{M,f}$ à partir de γ_r avec une garantie similaire à Proposition 2.5 en utilisant les valeurs de f sur X.

2.3 Estimation du support par un tirage de points aléatoires

Dorénavant, on suppose que $X = X_n$ est le résultat d'un tirage i.i.d. de n points selon une mesure de probabilité μ dont le support est M. Section 2.2 appelle à contrôler $d_{\mathrm{H}}(M, X_n)$ et à choisir un r approprié

4. Par minimax-optimal, on entend que c'est la convergence la plus rapide possible dans le pire des cas.

pour en déduire la convergence de $\hat{D}_{X,r}(x,y)$ vers $D_M(x,y)$. Pour obtenir des garanties quantitatives sur $d_{\mathrm{H}}(M,X)$, on suppose que μ est suffisamment dense en tout point de M au sens de l'hypothèse suivante.

Hypothèse (A). μ est d-Ahlfors avec $d \geq 1$: Il existe $a, A > 0$ tels que pour tout $r \geq 0$,

$$ar^d \wedge 1 \leq \mu\big(\mathcal{B}(x,r)\big) \leq Ar^d\,. \tag{8}$$

Dans Hypothèse (A), d s'interprète comme une borne supérieure sur la dimension intrinsèque du support M de μ. On peut montrer que c'est effectivement le cas au sens de la *dimension de Minkowski*. En effet, considérons $\mathrm{cov}(M,r)$ le *nombre de recouvrement* de M, c'est-à-dire le nombre minimal de boules ouvertes de rayon r nécessaire pour recouvrir M. On a alors le résultat suivant :

Lemme 2.7. Soit μ une mesure d-Ahlfors de support M. Alors pour tout $0 < r \leq 2a^{-1/d}$,

$$\mathrm{cov}(M,r) \leq \frac{2^d}{a} r^{-d}\,. \tag{9}$$

Démonstration. Tout point $x \in M$ vérifie $\mu\big(\mathcal{B}(x,t)\big) \geq at^d$ pour tout $0 < t \leq a^{-1/d}$. Soit $\mathcal{A} = (x_1, x_2, \ldots, x_{|\mathcal{A}|})$ une famille maximale de points r-séparés de M, c'est-à-dire que les points sont deux-à-deux à distance au moins r et qu'aucun autre point de M ne peut être ajouté à $\mathcal{A}$ en conservant cette propriété. Alors $\mathcal{A}$ est un r-recouvrement de M, d'où $\mathrm{cov}(M,r) \leq |\mathcal{A}|$ et

$$1 \geq \mu\left(\bigsqcup_{x\in\mathcal{A}} \mathcal{B}\left(x, \frac{r}{2}\right)\right) = \sum_{x\in\mathcal{A}} \mu\left(\mathcal{B}\left(x,\frac{r}{2}\right)\right) \geq |\mathcal{A}| a \left(\frac{r}{2}\right)^d \geq \mathrm{cov}(M,r) a \left(\frac{r}{2}\right)^d,$$

ce qui conclut Équation (9). □

La dimension de Minkowski (supérieure) $\dim_{\mathrm{M}}$ est donnée par

$$\dim_{\mathrm{M}}(M) \stackrel{\mathrm{def}}{=} \limsup_{r\to 0} \frac{\log(\mathrm{cov}(M,r))}{\log(1/r)}\,.$$

Lemme 2.7 implique directement que sous Hypothèse (A), $\dim_{\mathrm{M}}(M) \leq d$. Si $M \subset \mathbb{R}^D$ est une sous-variété différentielle de dimension d, alors $\dim_{\mathrm{M}}(M) = d$ (Falconer 1990). Pour travailler dans un cadre plus général que celui des variétés différentielles, on peut utiliser la mesure de Hausdorff de dimension d, notée H^d, au lieu de la forme volume. Ces deux notions coïncident dans le cas des sous-variétés (Morgan 1988) à un facteur multiplicatif près. On peut alors montrer que si μ est à densité minorée par rapport à la forme volume sur une sous-variété M, alors Hypothèse (A) est vérifiée.

Lemme 2.8. Soit $M \subset \mathbb{R}^D$ une sous-variété de dimension d et μ une mesure à densité bornée dans $[a,A] \subset \mathbb{R}_+^*$ par rapport à la forme volume (normalisée) sur M. Alors μ est d-Ahlfors.

Démonstration (esquisse). Pour r suffisamment petit, $\mathcal{B}(x,r) \cap M$ est approximativement une boule de dimension d car M est une variété. Alors,

$$\mu(\mathcal{B}(x,r)) \geq a\frac{H^d\big(\mathcal{B}(x,r)\cap M\big)}{H^d(M)} \underset{r\to 0}{\sim} a\frac{H^d\big(\mathcal{B}_{\mathbb{R}^d}(0,r)\big)}{H^d(M)} = a\frac{H^d\big(\mathcal{B}_{\mathbb{R}^d}(0,1)\big)}{H^d(M)} r^d\,.$$

On en déduit que la minoration d'ordre r^d pour r petit, puis pour tout r par compacité de M. Le raisonnement est le même pour la majoration. □

Rappelons que X_n est un tirage i.i.d. de n points suivant μ. Si μ est d-Ahlfors, alors pour tout $x \in M$, $\mathcal{B}(x, r)$ doit contenir en moyenne au moins nar^d points de X_n. On peut ainsi contrôler la distance de x au nuage de points X_n — voir par exemple (Taupin et al. 2025, Lemme D.2). Pour tout $\varepsilon > 0$,

$$\mathbb{P}\big(d(x, X_n) > \varepsilon\big) \leq \exp\big(-na\varepsilon^d\big). \tag{10}$$

En considérant un r-recouvrement $\mathcal{A}$ de M avec r d'ordre $n^{-1/d}$, qui utilise un nombre de points d'ordre n d'après Lemme 2.7, on a

$$d_{\mathrm{H}}(M, X_n) \leq r + \max_{x \in \mathcal{A}} d(x, X_n)$$

et il en découle le résultat suivant, dont une preuve est donnée en (Taupin et al. 2025, Proposition D.4).

Proposition 2.9. Supposons que μ is d-Ahlfors de support M. Alors pour tout $n \geq 1$,

$$\mathbb{E}\big[d_{\mathrm{H}}(M, X_n)\big] \leq \frac{2}{d} \frac{\log(2e^{2d}n)}{(an)^{1/d}} \lesssim \frac{\log(n)}{n^{1/d}}. \tag{11}$$

Sous certaines conditions, il est théoriquement possible d'estimer une variété $M \subset \mathbb{R}^D$ de dimension d avec une vitesse minimax-optimale de $n^{-2/(d+2)}$ au sens de la distance de Hausdorff (Genovese et al. 2012). À ma connaissance, il n'existe pas d'estimateur réalisable en pratique qui atteigne cette vitesse, ou même une vitesse plus grande que $n^{-1/d}$.

3 Distance-à-la-mesure et FDTM

On s'intéresse maintenant au cas d'une fonction f particulière dépendant de la mesure μ. On a déjà évoqué le cas de la distance de Fermat en Section 1. Dans cette section on choisit plutôt pour f la DTM. Notons qu'ici le facteur conforme d_μ n'est pas connu et doit être lui-même estimé. À l'erreur géométrique étudiée précédemment s'ajoutera donc une nouvelle erreur relativement orthogonale d'estimation de la DTM, mais cette dernière est contrôlée assez efficacement.

3.1 Distance-à-la-mesure

Fixons deux paramètres $m \in (0, 1)$ et $p \geq 1$. La DTM de μ pour ces paramètres, introduite par (Chazal et al. 2011) est définie pour tout $x \in \mathbb{R}^D$ par

$$d_\mu(x) \stackrel{\text{def}}{=} \left(\frac{1}{m}\int_0^m \delta_{\mu,u}(x)^p \,\mathrm{d}u\right)^{1/p} \tag{12}$$

où

$$\delta_{\mu,u}(x) = \inf\big\{r > 0 \,:\, \mu\big(\mathcal{B}(x, r)\big) > u\big\} \tag{13}$$

est la pseudo-DTM. $\delta_{\mu,u}$ généralise la distance au support de μ — qui coïncide avec $\delta_{\mu,0}$ — et illustre à quel point x est loin d'une fraction de masse u de la mesure. La DTM, obtenue comme la moyenne p-Hölder de $\delta_{\mu,u}$ sur $u \in [0, m]$, se comporte intuitivement de manière similaire à la densité : $d_\mu(x)$ est grande quand x est dans une zone de faible densité ou est loin du support. (Biau et al. 2011) montre qu'on peut estimer

la densité de μ quand elle existe à l'aide de la DTM d'un échantillon aléatoire de μ. (Chazal et al. 2011) démontre plusieurs propriétés essentielles. Notamment, la DTM est 1-Lipschitz :

$$|d_\mu(x) - d_\mu(y)| \leq \|x - y\| \text{ pour tout } x, y \in \mathbb{R}^D \tag{14}$$

et stable par rapport à μ :

$$\|d_\mu - d_\nu\|_\infty \leq \frac{W_p(\mu, \nu)}{m^{1/p}} \tag{15}$$

où W_p est la distance de Wasserstein définie par

$$W_p(\mu, \nu) \overset{\text{def}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^D \times \mathbb{R}^D} \|x - y\|^p \pi(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}}$$

avec $\Pi(\mu, \nu)$ l'ensemble des mesures de probabilités sur $\mathbb{R}^D \times \mathbb{R}^D$ de marginales respectives μ et ν. La distance de Wasserstein est liée à la théorie du transport optimal. Elle représente le coût de déplacer la masse μ vers la masse ν en prenant en compte les distances contrairement à la distance en variation totale. Enfin, sous Hypothèse (A), il est possible d'estimer efficacement la DTM à partir d'un échantillon X_n de μ.

Proposition 3.1. (Chazal et al. 2016, Théorème 1) Supposons que μ est d-Ahlfors de support M et que $m \leq \frac{1}{2}$. Alors pour tout $n \geq \frac{1}{m}$,

$$\mathbb{E}\left[\left|d_\mu(x)^p - d_{\hat{\mu}_n}(x)^p\right|\right] \lesssim \frac{d}{\sqrt{n}} \tag{16}$$

où $\lesssim$ cache une constante dépendant de m et de μ.

À condition de connaître un minorant de d_μ — qui existe dès que μ ne possède aucun atome de masse au moins m et qui peut être explicité via Hypothèse (A) par exemple — Proposition 3.1 est également valide sans la puissance p avec une constante différente. De plus, la DTM empirique $d_{\hat{\mu}_n}(x)$ s'exprime comme une moyenne des distances aux mn plus proches voisins de x et est donc tout à fait calculable en pratique.

3.2 FDTM avec raccourcis

La FDTM est définie comme le changement conforme par la DTM[5] :

$$\mathcal{F}_{M,\mu}(x, y) \overset{\text{def}}{=} D_{M,d_\mu}(x, y) = \inf_{\gamma \in \Gamma_M(x,y)} \int_\gamma d_\mu \,. \tag{17}$$

Cette sous-section est essentiellement consacrée à faire un résumé des résultats de (Taupin et al. 2025) sur la FDTM. Rappelons que la DTM étant définie en tout point de $\mathbb{R}^D$, le domaine M peut être quelconque. Le cas qui nous intéresse dans le contexte d'estimation reste cependant celui où M est le support de μ. Dans (Taupin et al. 2025), on se place dans un cadre légèrement différent où les chemins admissibles

5. On pourrait également élever la DTM à une puissance $\beta \geq 1$, ce qui est fait dans (Taupin et al. 2025). On omet cette possibilité ici pour faciliter la lecture.

peuvent s'aventurer en dehors du domaine à condition de tracer des lignes droites. Notons $\Gamma'_M(x,y)$ l'ensemble des chemins de x à y sous cette contrainte plus faible et $\mathcal{F}'_{M,\mu}(x,y)$ la notion alternative de FDTM avec raccourcis qui y est associée. Ce choix permet d'obtenir des résultats quantitatifs en toute généralité, c'est-à-dire sans aucune hypothèse géométrique sur la régularité de M. On verra toutefois qu'il présente des limitations pratiques. De plus, les vitesses de convergence peuvent être améliorées sous des hypothèses de régularité assez légères sur le support. On reviendra donc en Section 3.3 à l'étude de la FDTM sans raccourcis donnée par Équation (17).
Le premier résultat important est le fait que la longueur des géodésiques pour cette métrique est bornée de manière indépendante de μ et M sans aucune hypothèse sur ces derniers. L'existence de ces géodésiques nécessite une légère adaptation du raisonnement de (Burago et al. 2001) évoqué dans le cas des métriques conformes en Section 1 mais découle des mêmes arguments fondamentaux.

Théorème 3.2. (Taupin et al. 2025, Corollaire 2.5) Soit $\gamma \in \Gamma'_M(x,y)$ une géodésique pour $\mathcal{F}'$. Alors

$$|\gamma| \lesssim \|x-y\| \tag{18}$$

où $\lesssim$ cache une constante explicite dépendant de m.

Ce contrôle uniforme sur la longueur des géodésiques est essentiel pour obtenir les résultats de stabilité et d'estimation qui suivent. Pour le montrer, on utilise de manière crucial la propriété suivante de la DTM, qui affirme que cette dernière ne peut prendre de faibles valeurs sur un ensemble trop large au sens du nombre de recouvrement.

Lemme 3.3. (Taupin et al. 2025, Lemme 3.1) Soit $\delta > 0$ et $L_{\mu,\delta} = \{x \in \mathbb{R}^D : d_\mu(x) < \delta\}$. Alors,

$$\mathrm{cov}\left(L_{\mu,\delta}, 4\delta\right) \leq \frac{2}{m}. \tag{19}$$

Démonstration. Pour tout $x \in L_{\mu,\delta}$, par définition

$$\delta^p > d_\mu(x)^p = \frac{1}{m}\int_0^m \delta_{\mu,u}(x)^p \mathrm{d}u \geq \frac{1}{m}\int_{m/2}^m \delta_{\mu,m/2}(x)^p \mathrm{d}u \geq \frac{1}{2}\delta_{\mu,m/2}(x)^p,$$

donc $\delta_{\mu,m/2}(x) < 2^{1/p}\delta \leq 2\delta$, ce qui implique par définition de la pseudo-DTM que $\mu\big(\mathcal{B}(x_i, 2\delta)\big) \geq \frac{m}{2}$. Considérons $(x_1, x_2, \ldots, x_N)$ une famille maximale de points 4δ-séparés de M, qui est également un 4δ-recouvrement de M. Comme μ est une mesure de probabilité il vient

$$1 \geq \mu\left(\bigsqcup_{i=1}^N \mathcal{B}(x_i, 2\delta)\right) = \sum_{i=1}^N \mu\big(\mathcal{B}(x_i, 2\delta)\big) \geq N\frac{m}{2} \geq \mathrm{cov}\left(L_{\mu,\delta}, 4\delta\right)\frac{m}{2}$$

d'où le résultat. □

Dans ce qui suit, on suppose pour que toutes les mesures et domaines considérés sont inclus dans une partie convexe et compacte K. Cette hypothèse n'est pas fondamentale et sert seulement à simplifier les énoncés.

Théorème 3.4. (Taupin et al. 2025, Théorème 2.7) Soient μ, ν des mesures à support dans K et $M \subset K$ un fermé. Alors

$$\left\|\mathcal{F}'_{M,\mu} - \mathcal{F}'_{M,\nu}\right\|_{\infty,K} \lesssim \mathrm{diam}(K)\, W_p(\mu,\nu) \tag{20}$$

où $\lesssim$ cache une constante explicite dépendant de m.

Théorème 3.4 est une conséquence directe de Théorème 3.2 et Équation (15). De même, Théorème 3.2 combiné au raiseonnement décrit en Section 2.2 permet d'obtenir la stabilité de $\mathcal{F}'$ par rapport au domaine.

Théorème 3.5. (Taupin et al. 2025, Théorème 2.8) Soient μ une mesure à support dans K et $M, M' \subset K$ deux fermés. Alors

$$\left\|\mathcal{F}'_{M,\mu} - \mathcal{F}'_{M',\mu}\right\|_{\infty,K} \lesssim \operatorname{diam}(K)^{\frac{3}{2}} \sqrt{d_{\mathrm{H}}(M, M')} \tag{21}$$

où $\lesssim$ cache une constante explicite dépendant de m.

Dans le cas pratique où M est le support de μ et où on tire un nuage de n points X_n, un estimateur naturel apparaît alors sous la forme de la FDTM $\mathcal{F}'_{X_n,\hat{\mu}_n}$ de la mesure empirique $\hat{\mu}_n = \frac{1}{n}\sum_{x\in X_n} \delta_x$, qui est de fait la métrique du graphe sur X_n pondéré par les poids

$$w_{x,y} = \int_{[x,y]} d_{\hat{\mu}_n} \tag{22}$$

À partir des résultats de stabilité, la convergence de $\mathcal{F}'_{X_n,\hat{\mu}_n}$ se déduit alors par

$$\begin{aligned}\left\|\mathcal{F}'_{M,\mu} - \mathcal{F}'_{X_n,\hat{\mu}_n}\right\|_{\infty,K} &\leq \left\|\mathcal{F}'_{M,\mu} - \mathcal{F}'_{M,\hat{\mu}_n}\right\|_{\infty,K} + \left\|\mathcal{F}'_{M,\hat{\mu}_n} - \mathcal{F}'_{X_n,\hat{\mu}_n}\right\|_{\infty,K} \\ &\lesssim \operatorname{diam}(K)\|d_\mu - d_{\hat{\mu}_n}\|_{\infty,K} + \operatorname{diam}(K)^{\frac{3}{2}}\sqrt{d_{\mathrm{H}}(M, M')}\,.\end{aligned}$$

Sous Hypothèse (A), Proposition 2.9 s'applique. On obtient finalement la vitesse de convergence suivante.

Théorème 3.6. (Taupin et al. 2025, Théorème 2.9) Supposons que μ est de support $M \subset K$ et est d-Ahlfors, $d_\mu \geq \sigma > 0$ et $m \leq \frac{1}{2}$. Alors pour tout $n \geq \frac{1}{m}$,

$$\mathbb{E}\left[\left\|\mathcal{F}'_{M,\mu} - \mathcal{F}'_{X_n,\hat{\mu}_n}\right\|_{\infty,K}\right] \lesssim \frac{\log(n)}{n^{\frac{1}{2d}}} \tag{23}$$

où $\lesssim$ cache une constante multiplicative dépendant de m, μ, $\operatorname{diam}(K)$ et D.

On montre également (Taupin et al. 2025, Théorème 2.10) que n'importe quel estimateur à une vitesse de convergence au mieux d'ordre $n^{-1/d}$ dans le pire des cas.

La FDTM avec raccourcis $\mathcal{F}'$ est donc un objet avec de très bonne propriétés théoriques. En revanche, elle présente plusieurs limitations pratiques. En effet, le calcul des poids donnés par Équation (22) nécessite de calculer des approximations d'intégrale de la DTM, qui est elle même relativement couteuse à évaluer. De plus, pour s'assurer d'évaluer tous les chemins possibles, il n'est pas possible de réduire le graphe à un graphe des k plus proches voisins (kNN) comme il se fait pour la distance de Fermat (Groisman et al. 2022). Certaines méthodes existent toutefois pour réduire la taille du graphe de manière appropriée, comme par exemples les graphes de Yao (Yao 1982), mais deviennent peu efficaces quand la dimension ambiante augmente. Pour ces raisons, on se ramène à l'étude de $\mathcal{F}_{M,\mu}(x, y)$ qui rentre dans le cadre de Section 2.

3.3 FDTM sans raccourcis

Cette dernière section est informelle et donne une idée de résultat d'estimation qu'on s'attend à pouvoir obtenir sur la FDTM sans raccourcis. En supposant qu'il soit possible d'adapter les résultats de Section 2 pour considérer le facteur conforme d_μ, on s'attend à pouvoir estimer $\mathcal{F}_{M,\mu}$ par un estimateur plus simple à calculer que dans le cas avec raccourcis. Si la dimension du support est d, Propositions 2.5 et 2.9 indiquent que le choix optimal pour r serait $r = d_{\mathrm{H}}(M, X_n)^{2/3} \lesssim n^{-2/3d}$ et on s'attend à ce que l'erreur d'estimation soit de cet ordre. Par ailleurs, le nombre de points à distance r d'un point du support est d'ordre $r^d n \lesssim n^{1/3}$ et le nombre total d'arêtes dans le graphe tronqué est d'ordre $n^{4/3}$. Ainsi, on s'attend à obtenir une vitesse de convergence de la forme

$$\left|\hat{\mathcal{F}}_{X_n,r}(x, y) - \mathcal{F}_{M,\mu}(x, y)\right| \lesssim n^{-2/3d}$$

pour une complexité temporelle en $O(n^{4/3})$. Comparé à la convergence de la distance de Fermat donnée par Équation (4), on aurait ainsi un résultat quantitatif bien plus fort, pour une complexité algorithmique raisonnable.

Références

Aamari, Eddie, Clément Berenfeld et Clément Levrard (juin 2023). « Optimal Reach Estimation and Metric Learning ». In : *The Annals of Statistics* 51.3, p. 1086-1108. ISSN : 0090-5364, 2168-8966.

Aaron, Catherine et Olivier Bodart (déc. 2018). « Convergence Rates for Estimators of Geodesic Distances and Fréchet Expectations ». In : *Journal of Applied Probability*.

Biau, Gérard et al. (1er jan. 2011). « A Weighted K-Nearest Neighbor Density Estimate for Geometric Inference ». In : *Electronic Journal of Statistics* 5. ISSN : 1935-7524.

Boissonnat, Jean-Daniel, André Lieutier et Mathijs Wintraecken (1er juin 2019). « The Reach, Metric Distortion, Geodesic Convexity and the Variation of Tangent Spaces ». In : *Journal of Applied and Computational Topology* 3.1, p. 29-58. ISSN : 2367-1734.

Burago, Dmitri, Yuri Burago et Sergei Ivanov (12 juin 2001). *A Course in Metric Geometry*. T. 33. Graduate Studies in Mathematics. Providence, Rhode Island : American Mathematical Society. ISBN : 978-1-4704-1794-9.

Chazal, Frédéric, David Cohen-Steiner et Quentin Mérigot (2011). « Geometric Inference for Measures Based on Distance Functions ». In : *Foundations of Computational Mathematics* 11.6, p. 733.

Chazal, Frédéric, Pascal Massart et Bertrand Michel (jan. 2016). « Rates of Convergence for Robust Geometric Inference ». In : *Electronic Journal of Statistics* 10.2, p. 2243-2286. ISSN : 1935-7524, 1935-7524.

Falconer, K. (sept. 1990). « Fractal Geometry : Mathematical Foundations and Applications. » In : *Biometrics*. T. 46. 3, p. 886. JSTOR : 2532125.

Federer, Herbert (1959). « Curvature Measures ». In : *Transactions of the American Mathematical Society* 93.3, p. 418-491. ISSN : 0002-9947, 1088-6850.

Fernández, Ximena et al. (1er jan. 2023). « Intrinsic Persistent Homology via Density-Based Metric Learning ». In : *Journal of Machine Learning Research* 24.1, 75 :3341-75 :3382. ISSN : 1532-4435.

Garcia Trillos, Nicolas et al. (30 jan. 2018). *Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs towards the Laplace–Beltrami Operator*. arXiv : 1801.10108.

Garcia Trillos, Nicolas et al. (1er jan. 2024). « Fermat Distances : Metric Approximation, Spectral Convergence, and Clustering Algorithms ». In : *Journal of Machine Learning Research* 25.1, 176 :8331-176 :8395. ISSN : 1532-4435.

Genovese, Christopher R et al. (2012). « Minimax Manifold Estimation ». In : *Journal of Machine Learning Research*.

Groisman, Pablo, Matthieu Jonckheere et Facundo Sapienza (fév. 2022). « Nonhomogeneous Euclidean First-Passage Percolation and Distance Learning ». In : *Bernoulli* 28.1, p. 255-276. ISSN : 1350-7265.

Morgan, Frank (1988). *Geometric Measure Theory : A Beginner's Guide*. San Diego, California : Academic Press. ISBN : 1-4832-7780-1.

Rataj, Jan et Martina Zähle (2019). « Sets with Positive Reach ». In : *Curvature Measures of Singular Sets*. Sous la dir. de Jan Rataj et Martina Zähle. Cham : Springer International Publishing, p. 55-86. ISBN : 978-3-030-18183-3.

Taupin, Jérôme et Frédéric Chazal (3 avr. 2025). *Fermat Distance-to-Measure : A Robust Fermat-like Metric*. arXiv : 2504.02381.

Yamabe, Hidehiko (1960). « On a Deformation of Riemannian Structures on Compact Manifolds ». In : *Osaka Mathematical Journal* 12.1, p. 21-37.

Yao, Andrew Chi-Chih (nov. 1982). « On Constructing Minimum Spanning Trees in K-Dimensional Spaces and Related Problems ». In : *SIAM Journal on Computing* 11.4, p. 721-736. ISSN : 0097-5397.