Régression par processus Gaussiens pour des entrées graphes en grande dimension

Raphaël CARPINTERO PEREZ [1,2] & Sébastien DA VEIGA [3] & Josselin GARNIER [2] & Brian STABER [1]

[1] *Safran Tech, Digital Sciences & Technologies Department, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France,*
{raphael.carpintero-perez,brian.staber}@safrangroup.com
[2] *Centre de Mathématiques Appliquées, Ecole Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France,*
{raphael.carpintero-perez,josselin.garnier}@polytechnique.edu
[3] *Univ Rennes, Ensai CNRS, CREST - UMR 9194, F-35000 Rennes, France,*
sebastien.da-veiga@ensai.fr

Résumé. Les algorithmes d'apprentissage statistique appliqués à des données sous formes de graphes ont suscité une grande attention dans des domaines tels que la biochimie, les systèmes de recommandation sociaux et, très récemment, l'apprentissage de simulations basées sur la physique. Les méthodes à noyau, et plus particulièrement la régression par processus Gaussiens, sont particulièrement appréciées car elles sont efficaces lorsque la taille de l'échantillon est faible et lorsqu'il est nécessaire de quantifier les incertitudes de prédiction. Dans cet exposé, nous présentons le noyau entre graphes Sliced Wasserstein Weisfeiler-Lehman (SWWL) qui considère des graphes avec des attributs continus attachés aux sommets. Nous combinons les itérations continues de Wesifeiler Lehman et du transport optimal entre distributions de probabilités empiriques avec la distance de sliced Wasserstein afin de définir une fonction noyau définie positive avec une faible complexité de calcul. Ces deux propriétés permettent de considérer des graphes avec un grand nombre de sommets, ce qui était auparavant une tâche délicate.

Mots-clés. Apprentissage statistique, processus Gaussiens, noyaux entre graphes, grande dimension, transport optimal, métamodèles pour la simulation numérique

Abstract. Machine learning algorithms applied to graph data have garnered significant attention in fields such as biochemistry, social recommendation systems, and very recently, learning physics-based simulations. Kernel methods, and more specifically Gaussian process regression, are particularly appreciated since they are powerful when the sample size is small, and when uncertainty quantification is needed. In this talk, we introduce the Sliced Wasserstein Weisfeiler-Lehman (SWWL) graph kernel which handles graphs with continuous node attributes. We combine continuous Wesifeiler Lehman iterations and an optimal transport between empirical probability distributions with the sliced Wasserstein distance in order to define a positive definite kernel function with low computational complexity. These two properties make it possible to consider graphs with a large number of nodes, which was previously a tricky task.

Keywords. Machine learning, Gaussian processes, graph kernels, high dimension, optimal transport, surrogate models for numerical simulation

1 Introduction

Ce travail porte sur la régression par processus Gaussiens indexés par des graphes de grande dimension et possédant des attributs continus.

Dans ce contexte, il est fondamental d'avoir accès à une fonction noyau défini positif entre graphes et dont la complexité est raisonnable afin de pouvoir traiter des graphes en grande dimension (de l'ordre de plusieurs dizaines de milliers de sommets).

Il existe de nombreuses approches qui permettent de définir des noyaux entre graphes, comme l'indiquent les multiples articles de synthèse sur les noyaux entre des graphes (Nikolentzos et al., 2021; Borgwardt et al., 2020). Les approches traditionnelles se concentrent principalement sur la structure d'adjacence des graphes, et peu prennent en compte les possibles *attributs continus* ou sont applicables avec des graphes *de grande taille* et *creux*.

Plus récemment des approches utilisant du transport optimal (Peyré et al., 2019) ont été proposées comme le noyau Wasserstein Weisfeiler Lehman de Togninalli et al. (2019) mais ont une complexité trop élevée et ne permettent pas d'assurer un noyau défini positif. Mais contrairement à la distance de Wasserstein, la distance de sliced Wasserstein (Bonneel et al., 2015) injectée dans des noyaux usuels donne bien des noyaux définis positifs comme l'a mis en avant Meunier et al. (2022) dans le cas de distributions empiriques.

L'approche que nous proposons (Carpintero Perez et al., 2024) est résumée dans la Figure 1. Elle repose sur deux ingrédients: les embeddings de Weisfeiler-Lehman (WL) et la distance de sliced Wasserstein.

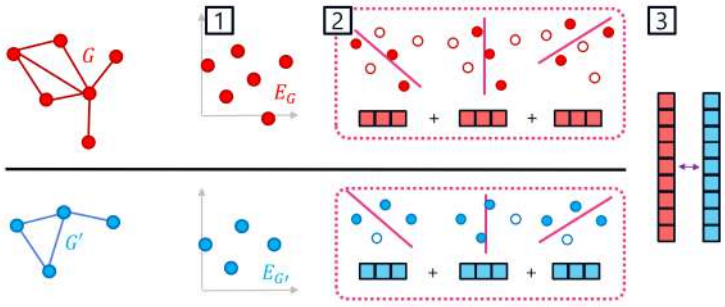

Figure 1: Noyau SWWL. Etape 1: embeddings des graphes. Etape 2: embeddings des quantiles projetés (EQP). Etape 3: distances Euclidiennes entre EQP.

2 Notations et présentation du problème

On considère la tâche d'apprentissage d'une fonction $f : \mathcal{G} \to \mathcal{Y}$ où $\mathcal{Y} = \{0, 1\}$ pour les tâches de classification, et $\mathcal{Y} = \mathbb{R}$ pour les tâches de régression. Ici, $\mathcal{G}$ désigne un ensemble de graphes non orientés et éventuellement pondérés ayant des attributs continus. Chaque graphe $G \in \mathcal{G}$ peut donc être représenté comme $G = (V, E, w, \mathbf{F})$ où V est l'ensemble des

sommets et E est un ensemble de paires de sommets, dont les éléments sont appelés arêtes. Les attributs à d dimensions des sommets sont regroupés dans la matrice (de taille $|V| \times d$) $\mathbf{F} = (\mathbf{F}_u)_{u \in V}$. Les poids des arêtes sont attribués par la fonction $w : E \to \mathbb{R}$. Le voisinage d'un sommet $u \in V$ est donné par $\mathcal{N}(u) = \{v \in V : \{u, v\} \in E\}$ et son degré est noté $\deg(u) = |\mathcal{N}(u)|$.

On suppose que l'on dispose d'un ensemble de données $\mathcal{D}$ constitué de N observations $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^N$, où les graphes G_i en entrée peuvent différer en termes de nombres de sommets et de matrices d'adjacence, c'est-à-dire qu'il est possible d'avoir $|V_i| \neq |V_j|$ et/ou $E_i \neq E_j$ pour certains $(i, j) \in \{1, \ldots, N\}^2$. Pour approcher la fonction f on utilise la régression par processus Gaussien. On suppose une loi a priori Gaussienne sur la fonction f (processus Gaussien). Après avoir conditionné selon les observations, la loi a posteriori est à nouveau Gaussienne, ce qui permet d'obtenir les prédictions (grâce à la moyenne) et les incertitudes de prédiction (grâce à la covariance). Ce processus Gaussien est entièrement déterminé par sa moyenne et sa covariance, correspondant à la fonction noyau. On souhaite donc construire une fonction noyau $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ définie positive. On renvoie le lecteur à Williams and Rasmussen (2006) pour plus de détails sur la régression par processus Gaussiens.

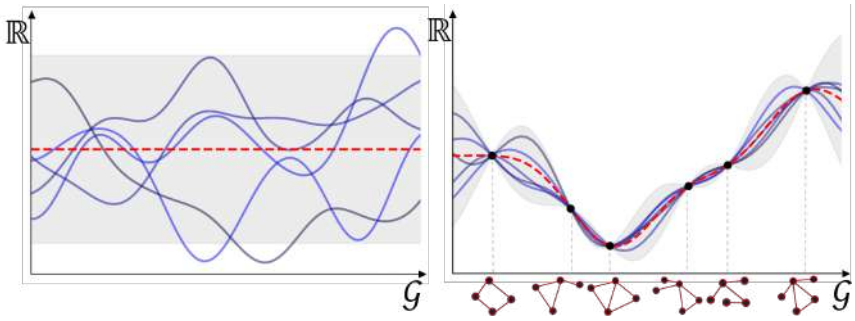

Figure 2: Illustration de la régression par processus Gaussiens pour des entrées de type graphes. Gauche: échantillons de la distribution a priori. Droite: échantillons de la distribution a posteriori après conditionnement sur les observations (les entrées sont des graphes ici).

3 Noyau Sliced Wasserstein Weisfeiler-Lehman

La méthodologie proposée repose sur la distance de sliced Wasserstein (Bonneel et al., 2015) qui est basée sur des projections aléatoires et l'expression analytique de la distance de Wasserstein 1D sous forme de quantiles. L'autre ingrédient correspond aux itérations de Weisfeiler-Lehman (WL) continues (Togninalli et al., 2019) qui permettent d'obtenir un embedding des sommets des graphes prenant en compte les attributs continus ainsi que la structure d'adjacence. Les différentes étapes sont détaillées après la définition du noyau.

Définition 1 (Noyau SWWL)**.** *Soit $P \geqslant 1$ le nombre de projections, $Q \geqslant 2$ le nombre de quantiles, et $H \geqslant 0$ le nombre d'itérations de WL continues. Le noyau SWWL (illustré sur*

la Figure 1) est défini pour $G, G' \in \mathcal{G}$ par

$$k_{\mathrm{SWWL}}(G, G') = \exp\left(-\gamma \widehat{SW}^2_{2,P,Q}(\mu_G, \mu_{G'})\right) , \tag{1}$$

où $\mu_G = |V|^{-1} \sum_{u \in V} \delta_{\mathbf{E}^G_u}$ est la mesure empirique associée à l'embedding de WL continu $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$ de G avec H iterations (voir la Définition 2) et $\gamma > 0$ est un paramètre de précision.

Le noyau SWWL possède plusieurs propriétés clés pour la régression par processus Gaussiens.

Propriété 1. *Il existe une feature map ϕ dans un espace de dimension PQ (voir l'Equation (9)) tel que le noyau SWWL peut s'écrire de la façon suivante:*

$$k_{\mathrm{SWWL}}(G, G') = \exp\left(-\gamma ||\phi(\mu_G) - \phi(\mu_{G'})||_2^2\right) . \tag{2}$$

L'équation (2) montre que la construction de la matrice de Gram $K = (k(G_i, G_j)_{ij}$ peut se décomposer en deux parties: (1) calcul des embeddings $\phi(\mu_G)$ et (2) assemblage de la matrice de pseudo-distances. On notera que la complexité temporelle de la dernière étape est indépendante des nombres de sommets dans les graphes.

Propriété 2. *Soient δ le degré moyen des sommets et n le nombre moyen de sommets. La complexité temporelle requise pour assembler la matrice de Gram $N \times N$ avec le noyau donné par l'Equation (1) est*

$$\mathcal{O}(NH\delta n + NPn(\log(n) + H) + N^2PQ) .$$

Propriété 3. *Le noyau SWWL est défini positif.*

On détaille désormais la construction de la feature map ϕ de l'Equation (2).

Embeddings de Weisfeiler-Lehman continus. Les embeddings de WL ont initialement été proposés pour des graphes ayant des sommets avec des labels discrets, mais ils ont été étendus aux attributs continus par Togninalli et al. (2019). De façon intuitive, les itérations de WL continues mettent à jour les attributs des sommets en agrégeant l'information des voisins. Après h itérations, les embeddings capturent l'information du h-voisinage des sommets.

Définition 2 (Embeddings de WL continus)**.** *Soit $G = (V, E, w, \mathbf{F})$ un graphe possédant les attributs continus $\mathbf{F} = (\mathbf{F}_u)_{u \in V}$, $\mathbf{F}_u \in \mathbb{R}^d$. Les itérations de Weisfeiler-Lehman continues sont définies récursivement pour $h \in \mathbb{N}$ par*

$$\mathbf{F}^{(h+1)}_u = \frac{1}{2}\left(\mathbf{F}^{(h)}_u + \frac{1}{\deg(u)} \sum_{v \in \mathcal{N}(u)} w(u, v)\mathbf{F}^{(h)}_v\right) , \tag{3}$$

avec $\mathbf{F}^{(0)}_u = \mathbf{F}_u$ pour $u \in V$. Etant donné un nombre d'itérations $H \geqslant 0$, l'embedding par sommets de WL continu du graphe G est la concaténation des itérations de WL continues aux étapes $0, 1, \ldots, H$, que l'on note $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$, $\mathbf{E}^G_u \in \mathbb{R}^{(H+1)d}$.

Transport optimal. Une fois que les embeddings de WL continus $\mathbf{E}^G$ ont été obtenus pour tous les graphes, les distributions empiriques associées peuvent être considérées et les distances de sliced Wasserstein peuvent être calculées entre toutes les paires de distributions empiriques pour construire un noyau.

La distance de sliced Wasserstein (Bonneel et al., 2015) moyenne sur la sphère unité les distances de Wasserstein unidimensionnelles entre les distributions projetées. Cela réduit dans un premier temps la complexité puisque la distance de Wasserstein 1D peut être calculée en $\mathcal{O}(n\log(n))$, et cela donne par ailleurs un noyau de substitution défini positif. La distance de sliced Wasserstein est en effet Hilbertienne (Meunier et al., 2022). On rappelle l'expression de la distance de Wasserstein unidimensionnelle avant de définir la distance de sliced Wasserstein.

Définition 3 (Distance de Wasserstein unidimensionnelle)**.** *Soit $r \geqslant 1$. La r-distance de Wasserstein pour des mesures sur $\mathbb{R}$ a l'expression suivante:*

$$W_r(\mu,\nu) = \left(\int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^r dt\right)^{\frac{1}{r}} \tag{4}$$

où $F_\mu(x) = \mu((-\infty, x]), x \in \mathbb{R}$ est la fonction de répartition et $F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geqslant t\}, t \in [0,1]$ est la fonction de répartition inverse (fonction quantile).

Définition 4 (Distance de sliced Wasserstein)**.** *Soient $s \geqslant 1$, $r \geqslant 1$. La r-distance de sliced Wasserstein est définie par*

$$SW_r(\mu,\nu) := \left(\int_{\mathbb{S}^{s-1}} W_r(\theta^*_\sharp\mu, \theta^*_\sharp\nu)^r d\sigma(\theta)\right)^{\frac{1}{r}}, \tag{5}$$

où $\mathbb{S}^{s-1}$ est la sphère unité $(s-1)$-dimensionnelle, σ est la distribution uniforme sur $\mathbb{S}^{s-1}$, $\theta^* : \mathbf{x} \in \mathbb{R}^s \mapsto \langle \mathbf{x}, \theta\rangle$ la fonction projection dans la direction $\theta \in \mathbb{S}^{s-1}$, $\theta^*_\sharp\mu$ la mesure image de μ par θ^*, et W_r est la r-distance de Wasserstein unidimensionnelle.

En pratique, une estimation de Monte-Carlo est réalisée en tirant P directions $\theta_1, \ldots, \theta_P$ uniformément dans $\mathbb{S}^{s-1}$.

Embedding des quantiles projetés. Le calcul de la distance de Wasserstein 1d entre les mesures empiriques consiste généralement à trier les points projetés, puis à additionner une puissance des distances euclidiennes entre les valeurs aux mêmes rangs. Ici, nous proposons plutôt d'utiliser $Q \ll n$ quantiles équidistants, qui ne dépendent pas des tailles n des distributions empiriques pouvant varier. Cette stratégie diffère des implémentations usuelles (Flamary et al., 2021), où une grille de quantiles est choisie pour chaque paire de distributions en entrée. Plus précisément, soit $0 = t_1 < \cdots < t_\ell < \cdots < t_Q = 1$ une suite de Q points équirépartis sur $[0,1]$, et soit $\theta \in \mathbb{S}^{s-1}$ une direction de projection. L'estimation à Q quantiles de la distance de Wasserstein 1d entre les mesures images $\mu_\theta := \theta^*_\sharp\mu$ et $\nu_\theta := \theta^*_\sharp\nu$ s'écrit

$$\tilde{W}_{r,Q}(\mu_\theta,\nu_\theta) = \left(\frac{1}{Q}\sum_{\ell=1}^{Q}\left|F_{\mu_\theta}^{-1}(t_\ell) - F_{\nu_\theta}^{-1}(t_\ell)\right|^r\right)^{\frac{1}{r}} \tag{6}$$

où $F_\mu(x) = \mu(]-\infty, x]), x \in \mathbb{R}$ est la fonction de répartition et $F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geqslant t\}, t \in [0,1]$ est la fonction de répartition inverse. La distance de sliced Wasserstein donnée par l'Equation (5) est finalement estimée de la façon suivante:

$$\widehat{SW}_{r,P,Q}(\mu,\nu) = \left(\frac{1}{P}\sum_{p=1}^{P} \tilde{W}_{r,Q}(\mu_{\theta_p}, \nu_{\theta_p})^r\right)^{\frac{1}{r}}, \tag{7}$$

où $\theta_1, \ldots, \theta_P$ désignent les P directions de projections tirées uniformément dans $\mathbb{S}^{s-1}$. En combinant les équations (6) et (7), cette estimation peut s'écrire comme

$$\widehat{SW}_{r,P,Q}(\mu,\nu) = ||\phi(\mu) - \phi(\nu)||_r \,, \tag{8}$$

où $||.||_r$ est la r-norme dans $\mathbb{R}^{PQ}$, et ϕ la feature map explicite

$$\phi_{p+P(q-1)}(\mu) = (PQ)^{-1/r} F^{-1}_{\mu_{\theta_p}}(t_q) \tag{9}$$

pour $p = 1, \ldots, P$ et $q = 1, \ldots, Q$. On appelle cette feature map *embedding des quantiles projetés* (EQP). Il fournit une représentation PQ-dimensionnelle de toute distribution de probabilité μ dans $\mathbb{R}^s$. Dans la définition 1, la distribution de probabilité d'intérêt est la distribution empirique associée aux embeddings de WL continus $\mathbf{E}^G = (\mathbf{E}^G_u)_{u \in V}$ des graphes $G \in \mathcal{G}$ avec H itérations, ce qui donne une dimension $s = H(d+1)$. On obtient de ce fait la propriété 1.

4 Expériences

Deux tâches ont été considérées pour tester le noyau SWWL.

Une première concerne la classification de petits graphes avec moins de 100 sommets correspondant à des molécules (Morris et al., 2020) afin de valider les performances du noyau en le comparant à des noyaux de l'état de l'art (en utilisant des séparateurs à vaste marge).

Une seconde concerne la régression par processus Gaussiens pour des graphes provenant de simulations numériques en dynamique et mécanique des fluides basées sur des maillages. La figure 3 montre trois graphes d'entrée provenant du jeu de données Tensile2d [1]. On démontre par ces expériences que le noyau SWWL peut aisément manipuler des graphes à plus de 10^5 sommets en quelques secondes ou minutes là ou d'autres approches ne passent pas à l'échelle.

Pour ces deux tâches, on obtient des scores de classification/régression équivalents aux méthodes de l'état de l'art comparées mais avec des temps de calculs très inférieurs. On étudie en particulier l'influence des hyperparamètres du modèle: nombre d'itérations, nombre de quantiles et nombre de projections sur les erreurs de prédiction.

[1] `https://plaid-lib.readthedocs.io/en/latest/source/data_challenges/tensile2d.html`

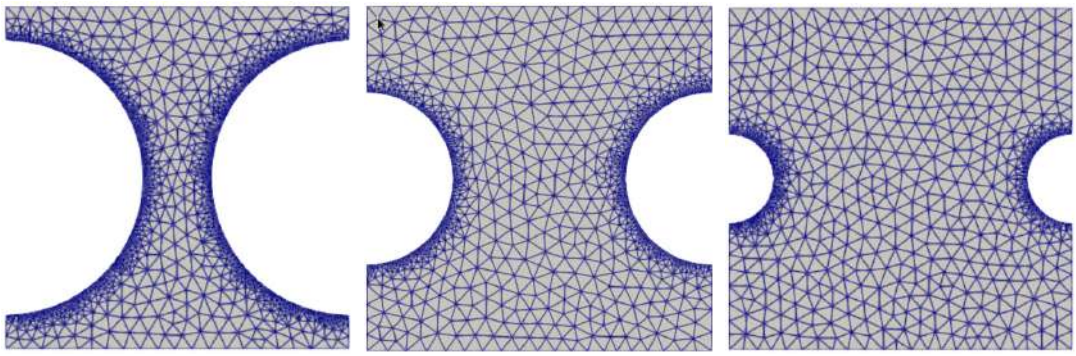

Figure 3: Trois graphes provenant de la version sous-échantillonnée du jeu de données Tensile2d. Les nombres de sommets et les connectivités changent entre les échantillons.

Remerciements

Ce travail de recherche est mené dans le cadre du projet SAMOURAI (Simulation Analytics and Meta-model-based solutions for Optimization, Uncertainty and Reliability AnalysIs) financé par l'Agence Nationale de la Recherche (ANR-20-CE46-0013).

References

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., Rieck, B., et al. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends® in Machine Learning*, 13(5-6):531–712.

Carpintero Perez, R., da Veiga, S., Garnier, J., and Staber, B. (2024). Gaussian process regression with Sliced Wasserstein Weisfeiler-Lehman graph kernels. `https://arxiv.org/abs/2402.03838`.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Meunier, D., Pontil, M., and Ciliberto, C. (2022). Distribution regression with sliced Wasserstein kernels. In *International Conference on Machine Learning*, pages 15501–15523. PMLR.

Morris, C., Kriege, N. M., Bause, F., Kersting, K., Mutzel, P., and Neumann, M. (2020). Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*.

Nikolentzos, G., Siglidis, G., and Vazirgiannis, M. (2021). Graph kernels: A survey. *Journal of Artificial Intelligence Research*, 72:943–1027.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Togninalli, M., Ghisu, E., Llinares-López, F., Rieck, B., and Borgwardt, K. (2019). Wasserstein Weisfeiler-Lehman graph kernels. *Advances in neural information processing systems*, 32.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT press.