Statistical learning algorithms for geometric and topological data analysis

Thomas Bonis

NNT : 2016SACLS459

Thèse de Doctorat
de
L'Université Paris-Saclay
préparée à
L'Université Paris-Sud
au sein de
Inria Saclay Ile-de-France

ÉCOLE DOCTORALE N°580
Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Mathématiques et Informatique

par

M. Thomas BONIS

Algorithmes d'apprentissage statistique pour l'analyse géométrique et topologique de données

Thèse présentée et soutenue à Palaiseau, le 1er décembre 2016

Après avis des rapporteurs :

M. Mikhail BELKIN	Rapporteur	Ohio State University
M. Giovanni PECCATI	Rapporteur	University of Luxembourg

Composition du jury :

M. Laurent DECREUSEFOND	Président du jury	Télécom Paristech
M. Giovanni PECCATI	Rapporteur	University of Luxembourg
M. Guillaume CHARPIAT	Examinateur	Inria Saclay Ile-de-France
M. Jérôme DEDECKER	Examinateur	Université Paris Descartes
M. Frédéric CHAZAL	Directeur de thèse	Inria Saclay Ile-de-France

Remerciements

Je souhaiterais tout d'abord remercier Frédéric Chazal pour son encadrement, et surtout sa patience envers ma prose, lors de ces années de thèse. Merci à Giovanni Peccati et Mikhail Belkin pour leur intért envers mes travaux et pour avoir pris le temps de rapporter cette thèse. Je remercie aussi Guillaume Charpiat, Jérôme Dedecker et Laurent Deucreusefond pour leur participation au jury de thèse. Je remercie également Steve Oudot, avec qui ce fut un plaisir de travailler. Je souhaite remercier Michel Ledoux et Yvik Swan pour les discussions enrichissantes que j'ai eu avec eux au sujet de mes travaux sur la méthode de Stein. Un grand merci aussi à Christine Biard pour m'avoir guidé au travers des méandres administratifs.

Ces trois années chez[1] Inria n'auraient sans doute pas été les même sans les autres membres de l'équipe et assimilés, notamment Etienne[2] mais aussi Ilaria[3], Clément, Mathieu, Dorian, Eddie, Claire, Mickal et tous les autres. Sur un autre registre, je voudrais remercier Terry Pratchett dont les livres, et leur traduction par Patrick Couton, auront réussi à rendre agréable mes nombreuses heures de trajets en RER et de qui je garderai un goût prononcé pour les notes de bas de page. Enfin, un petit mot pour Adélaïde qui m'a supporté[4] lors de la rédaction du présent manuscrit.

[1] Sic.

[2] Alias "Jules-de-chez-Smith-en-face".

[3] Victime d'une aversion inexpliquée envers les poireaux, vraisemblablement causée par un ancien traumatisme.

[4] Ceci n'est pas un anglicisme.

Contents

Chapter 1

Introduction en français

1.1 Motivation

Les algorithmes d'analyse de données sont utilisées dans de nombreux domaines tels que les moteurs de recherche, la publicité en ligne ou encore le sport. Récemment, on parle de plus en plus des potentielles applications de l'analyse de données dans la santé par exemple. Ces nouvelles applications ont des enjeux tels que l'on aura besoin d'une profonde compréhension des algorithmes utilisés et, autant que possible, de garanties théoriques permettant de donner des garanties sur les résultats obtenus par les algorithmes utilisés. D'un autre côté, les bases de données que l'on a à traiter sont de plus en plus grandes tandis que les données elle-même sont de plus en plus difficiles à traiter : on a donc besoin d'algorithmes plus complexes, souvent plus difficiles à analyser. Par exemple, les performances de nombreux algorithmes diminuent très vite lorsque la dimension des données augmente, un phénomène appelé malédiction de la dimension. Cependant, il est possible que des données étant à première vue de grande dimension vivent en réalité sur une structure de faible dimension : une sous-variété plongée dans un espace de grande dimension. Il est alors capital de pouvoir travailler dans cette structure de faible dimension plutôt que dans l'espace ambiant. Supposons que nos données sont des variables aléatoires indépendantes tirées selon une mesure μ dont le support est une variété plongée. N'ayant pas directement accès à la variété, on peut vouloir construire un maillage de cette variété pour l'approcher. Néanmoins le calcul d'un tel maillage risque bien souvent d'être trop long. En pratique, on utilisera plutôt un graphe de voisinage, obtenu en prenant comme sommets les points de données et en ajoutant une arête entre deux points s'ils sont suffisamment proches. Ces graphes portent le nom de graphes géométriques aléatoires. On peut

alors supposer que ce graphe contient la structure de la variété, la prochaine étape consiste alors à traiter l'information contenue dans ce graphe en utilisant des algorithmes d'analyse de graphe. En analyse de graphes, une des approches les plus efficaces consiste à utiliser les propriétés de marches aléatoires pour comprendre la structure générale du graphe étudié. Afin de comprendre le comportement ce type d'algorithme dans notre cas, il suffit donc d'étudier les propriétés des marches aléatoires sur des graphes géométriques aléatoires. Cependant, ces marches aléatoires sont difficiles à étudier directement. Néanmoins, il est possible de mieux les comprendre dans un cadre asymptotique : lorsque le nombre de points de données augmente, et sous certaines hypothèses concernant la façon dont on construit les graphe géométriques aléatoire, les marches aléatoires sur ces graphes convergent vers un processus de diffusion dont le générateur infinitésimal dépend à la fois de la variété et de la mesure μ. On peut donc s'attendre à ce que le comportement de marches aléatoires soit influencé à la fois par la structure de la variété plongée et par la structure de μ. Ainsi, même si le support de μ n'est pas une variété mais plus simplement l'espace ambiant, les approches utilisant des propriétés de marches aléatoires peuvent quand même être utilisées pour obtenir des algorithmes d'analyse de données non linéaires. Dans une première partie de cette thèse, nous allons utiliser la convergence des marches aléatoires sur des graphes géométriques aléatoires pour proposer un nouvel algorithme de partitionnement de données flou. Puis nous allons approfondir les résultats concernant la convergence des mesures invariantes de ces marches aléatoires, souvent utilisées dans des algorithmes d'analyse de graphe. Outre les mesures invariantes, les algorithmes d'analyse de graphes utilisent souvent deux autres propriétés de marches aléatoires : le Laplacien de graphe et les temps d'atteinte.

Laplacien de graphe Les Laplaciens de graphe sont des opérateurs agissant sur les fonctions définies sur les sommets d'un graphe. Ici, on se concentre sur un type particulier de Laplacien de graphe, appelé Laplacien de marche aléatoire, défini comme étant l'inverse du générateur d'une marche aléatoire sur un graphe. Les Laplaciens de graphe sont utilisés dans plusieurs types d'algorithmes comme le partitionnement de données [57], la réduction de dimension [7] ou encore la classification semi-supervisée [19]. Puisque des marches aléatoires sur des graphes géométriques aléatoires converge vers des processus de diffusion, on peut s'attendre à ce que leur générateur convergence vers le générateur infinitésimal du processus de diffusion, entranant la convergence des Laplaciens de graphe correspondants. Cette convergence peut être utilisée de plusieurs manières. Dans un premier

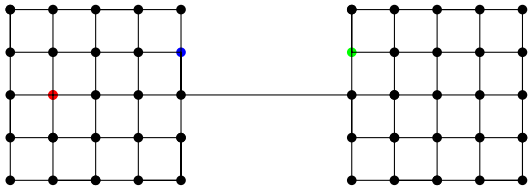


Figure 1.1: Deux composantes fortement connectées liée par une seule arête.

temps, on a cherché à montrer la convergence de l'action des Laplaciens sur certains ensembles de fonctions [9, 39, 45]. La convergence des marches aléatoires vers des processus de diffusion est en fait obtenue via une telle convergence [81]. D'autres travaux se sont ensuite concentrés sur la convergence du spectre des Laplaciens de graphe [8, 38].

Temps d'atteinte Il est aussi possible d'étudier la structure d'un graphe en utilisant le temps requis par une marche aléatoire pour passer d'un sommet x à un autre sommet x'. Cette quantité est appelée temps d'atteinte de x à x', notée $H(x, x')$. Les temps d'atteinte peuvent être utilisés pour définir une métrique sur le graphe, appelée distance de commutation, en prenant comme distance entre deux sommets x et x' la somme $H(x, x') + H(x', x)$. La distance de commutation est plus informative que des distances plus simples telles que la distance de plus court chemin car elle est influencée par la structure globale du graphe. Par exemple, considérons le graphe présenté dans la Figure 1.1, formé de deux composantes fortement connectées reliées par une seule arête. A priori, on souhaiterait qu'un sommet appartenant à la composante de gauche, tel que le sommet bleu, soit plus proche d'un autre sommet appartenant à la même composante comme le sommet rouge, que d'un sommet appartenant à la composante de droite, comme le sommet vert. Pour la distance de commutation, le sommet bleu est bien plus proche du sommet rouge que du vert, mais ce n'est pas le cas pour la distance de plus court chemin. Cette capacité à prendre en compte la structure complète du graphe explique le succès de cette distance dans de nombreuses applications comme le plongement de graphe [73, 68], le partitionnement de données [90] ou encore la classification semi-supervisée [91]. Cependant, lorsque l'on considère des graphes géométriques aléatoires, la convergence des marches aléatoires nous déconseille d'utiliser ce type de quantité. En effet, en dimension supérieure à un, le temps d'atteinte d'un processus de diffusion à un point fixé est infini. Il n'existe donc pas de quantité naturelle vers lesquelles les temps d'atteinte sur des graphes géométriques aléatoires risquent pourraient converger, le risque étant que les temps d'atteinte convergent vers des quantités peu informatives. Ce phénomène a été étudié

dans [85], où les auteurs prouvent que la distance de commutation entre deux sommets converge vers une quantité dépendant essentiellement des structures locales du graphe autour des deux sommets: la distance ne capture donc pas la structure du graphe.

Mesure invariante La mesure invariante est une autre propriété des marches aléatoires couramment utilisée en analyse de graphes. L'algorithme PageRank [65], qui classe des pages web via la mesure invariante d'une marche aléatoire sur le graphe ayant pour sommets les pages web et pour arêtes les hyperliens, est probablement l'exemple le plus connu d'un algorithme utilisant cette propriété. En ce qui concerne l'étude de la mesure invariante d'une marche aléatoire sur un graphe on peut distinguer deux cas suivant si le graphe est orienté ou non. Quand le graphe considéré est un graphe non orienté, la mesure invariante peut être facilement étudiée : elle est égale au degré des sommets du graphe. Cependant, quand le graphe est orienté, la mesure invariante de la marche aléatoire sur le graphe ne peut pas être calculée facilement. Ceci est problématique car de nombreux graphes intéressants sont orientés, comme le graphe de pages web décrit précédemment ou les graphes de plus proches voisins, souvent utilisés en analyse de données. Dans le cas de graphes géométriques aléatoires, on peut espérer que la mesure invariante des marches aléatoires convergent vers la mesure invariante du processus de diffusion limite, que le graphe soit orienté ou non. Si la mesure μ selon laquelle les données sont générées admet une densité f, la mesure invariante du processus de diffusion limite a une densité qui dépend de f. La mesures invariante d'une marche aléatoire sur un graphe géométrique aléatoire pourrait donc être utilisée pour calculer un estimateur de la densité f. Un tel estimateur pourrait être calculé en utilisant uniquement la structure du graphe, répondant ainsi à un problème posé dans [84]. Bien que cet estimateur de densité serait sans doute peu intéressant quand on a accès aux coordonnées des points de données, en quel cas il est possible d'utiliser des estimateurs de densité plus classiques, avoir accès à un tel estimateur nous garantirait qu'utiliser un graphe pour analyser les données n'entrane pas de perte majeure d'information. En effet, si on peut retrouver f à partir du graphe, alors ce graphe contient quasiment toute l'information initiale. Un résultat de convergence faible pour les mesures invariantes de marches aléatoires sur des graphes géométriques aléatoires a été obtenu dans [44].

1.2 Contributions

En général, la convergence des marches aléatoires sur des graphes géométriques aléatoires est utilisée pour obtenir des garanties théoriques pour des algorithmes utilisant certaines propriétés de ces marches. Dans cette thèse, nous commençons par suivre une démarche inverse et utilisons cette convergence pour proposer un nouvel algorithme de partitionnement de données flou. Le but du partitionnement de données est de répartir les échantillons de données en paquets en espérant que les échantillons appartenant à un paquet similaire aient des propriétés similaires. Cependant, cette approche n'apporte qu'une connaissance limitée de la structure des données car elle n'apporte a priori aucune information sur les relations entre les différents paquets : lesquels sont proches, à quel point ces paquets sont bien séparés etc. Afin d'obtenir une compréhension plus profonde des données, il faut donc identifier les interfaces entre les différents paquets. Considérons par exemple le cas de l'étude de l'espace des conformations d'une protéine. D'un côté, il est possible de détecter les conformations stables d'une protéine en utilisant du partitionnement de données [29]. D'un autre côté comprendre la structure des interfaces entre paquets permettrait d'identifier les étapes probables de transition entre deux conformations stables. On pourrait simplement définir ces interfaces comme l'ensemble de points dont le voisinage contient des points appartenant à plusieurs paquets différents. Néanmoins, les interfaces obtenues en utilisant cette définition peuvent être très instables. Le partitionnement de données flou semble un meilleur outil pour répondre à ce problème. Plutôt que de placer chaque échantillon de données dans un paquet, on cherche à attribuer à chaque échantillon des coefficients d'appartenance à chacun des paquets. Les interfaces correspondent alors simplement aux échantillons de données dont les coefficients d'appartenance sont bien répartis. Ces interfaces sont bien plus stables que les interfaces obtenues via la première définition proposée. Dans ce travail, on se concentre sur le partitionnement de données obtenu par une approche de recherche de modes. Dans cette approche, on suppose que les échantillons de données sont des variables aléatoires indépendantes tirées selon une mesure ayant une densité f et on identifie les paquets de données aux bassins d'attraction des maxima locaux de f : on utilise donc un flot de gradient pour partitionner les données. Ici, on propose d'obtenir un partitionnement de données flou en utilisant une version perturbée de ce flot de gradient prenant la forme d'un processus de diffusion dépendant de la densité f. Nous approchons ce flot de gradient en utilisant une marche aléatoire sur un graphe géométrique aléatoire calculé en utilisant un estimateur de f. En pratique on obtient une version floue de l'algorithme

de recherche de mode "Topological Mode Analysis Toolbox" (ToMATo) [22] et, en utilisant les résultats théoriques existant pour cet algorithme, nous obtenons des garanties pour notre algorithme flou. Nous obtenons aussi des résultats expérimentaux encourageant pour notre algorithme sur des jeux de données synthétiques et réelles. En pratique, notre algorithme n'a besoin que d'un graphe géométrique aléatoire et d'un estimateur de la densité f. Ainsi, on pourrait l'utiliser sur des données ayant directement une structure de graphe, supposé être un graphe géométrique aléatoire, à condition d'être capable de calculer un estimateur de densité. Comme vu précédemment, un tel estimateur peut être obtenu en utilisant la mesure invariante de la marche aléatoire sur le graphe. Cependant, en général, on ne peut qu'obtenir la convergence faible de cette mesure invariante vers une mesure à partir de laquelle on peut estimer f. Améliorer ce résultat en obtenant une vitesse de convergence de la mesure invariante constitue le second apport de cette thèse.

Notre approche se base sur la méthode de Stein, initialement développée pour obtenir des vitesses de convergence pour le théorème central limite [52]. Plus tard, Barbour [5] a montré que cette approche pouvait être généralisée pour comparer la mesure invariante d'une chane de Markov réversible à la mesure invariante d'un processus de diffusion. En fait, sous certaines hypothèses techniques, ces mesures sont proches si le générateur de la chane de Markov correctement renormalisé et le générateur infinitésimal d'un processus de diffusion le sont. Il y a deux problèmes avec cette approche : elle est difficile à utiliser dans un cadre multidimensionnel et elle requiert une chane de Markov réversible. En effet, dans notre cas, la chane de Markov que nous souhaitons utiliser est la marche aléatoire sur le graphe qui n'est réversible que si le graphe est non orienté, auquel cas la mesure invariante de cette marche aléatoire peut être directement calculée. En nous basant sur des résultats récents obtenus dans [52], nous proposons une nouvelle manière de borner la distance de Wasserstein entre deux mesures ν et μ où μ est une mesure réversible pour un opérateur différentiel $\mathcal{L}_\mu$, satisfaisant certaines conditions techniques, et ν est la mesure invariante d'une chane de Markov, réversible ou non. Plus précisément, on montre que si le générateur de la chane de Markov renormalisé et $\mathcal{L}_\nu$ sont proches alors la distance de Wasserstein d'ordre deux entre μ et ν est faible. En utilisant ce résultat, nous sommes capables de quantifier la convergence des mesures invariantes de marches aléatoires sur des graphes géométriques aléatoires et nous prenons comme exemple le cas des graphes des k plus proches voisins. Notre résultat peut aussi être utilisé pour étudier un algorithme de Monte-Carlo basique. Enfin, il est possible de raffiner notre résultat dans le cas où μ est la mesure gaussienne. Nous pouvons ainsi donner des vitesses

de convergence pour la distance de Wasserstein pour le théorème central limite.

La dernière partie de cette thèse n'est pas liée aux marches aléatoires sur des graphes mais utilise le concept d'homologie persistante utilisé par l'algorithme de partitionnement de données ToMATo pour faire de la reconnaissance de formes. De nombreux algorithmes ont été proposés pour traiter automatiquement des bases de données de formes 3D. Parmi ces algorithmes, certains des plus performants se basent sur l'approche dite "sac-de-mots". Dans cette approche, on commence par extraire un ensemble de descripteurs de chacune des formes de la base de données considérée. Ces descripteurs sont ensuite quantifiés par des vecteurs dont chaque coordonnée correspond à un "mot". L'information contenue dans chaque mot est ensuite extraite et compressée lors d'une étape dite de "pooling" : pour chaque forme, on considère la moyenne, "sum-pooling", ou le maximum, "max-pooling", de chaque mot sur la forme correspondante. On a donc caractérisé chaque forme par un vecteur et on peut ainsi classifier ces formes en utilisant l'ensemble des vecteurs obtenus dans des algorithmes de classification classiques. Dans l'idéal, toutes les étapes de ce processus devraient être robustes vis-à-vis des transformations que les formes peuvent subir : translations, rotations ou encore changement d'échelle, tout en étant suffisamment discriminantes. Maintenant, considérons les mots non pas comme un simple ensemble de valeurs mais comme des fonctions définies sur un graphe de voisinage, par exemple si on utilise un maillage pour calculer les descripteurs on peut utiliser le graphe fourni par le maillage et associer à chaque sommet la valeur du mot obtenu pour le descripteur correspondant. Le max-pooling consiste alors à construire un vecteur contenant les maxima globaux de chacune des fonctions mots. On serait alors tenté de rendre le max-pooling plus discriminant en considérant tous les maxima locaux de chaque fonction mot plutôt que de se restreindre aux maxima globaux. Cependant, une telle approche serait fortement instable : perturber les fonctions mots, même faiblement, peut créer de nombreux maxima locaux qui ne sont pas forcément caractéristiques de la forme 3D considérée. Pour pallier ce problème, nous proposons d'utiliser le concept d'homologie persistante 0-dimensionnelle, ou proéminence, afin de sélectionner les maxima locaux les plus stables par rapport aux perturbations des fonctions mots. Nous appelons la procédure obtenue "Topological Pooling", et nous montrons qu'elle est plus efficace que le schéma de max-pooling classique.

1.3 Plan de la thèse

Dans le Chapitre 3, nous introduisons le concept de marches aléatoires sur des graphes et les algorithmes d'analyse de graphes utilisant ces marches. Nous définissons ensuite les processus de diffusion et présentons leurs propriétés en lien avec ce travail. Nous introduisons ensuite les graphes géométriques aléatoires et nous présentons les résultats de convergence existant pour les marches aléatoires sur ces graphes.

Dans le Chapitre 4, nous présentons le cadre de la recherche de mode en partitionnement de données ainsi que l'algorithme ToMATo. Nous présentons ensuite notre algorithme de partitionnement de données flou utilisant une marche aléatoire sur un graphe géométrique aléatoire approprié. Nous obtenons ensuite des garanties théoriques pour notre algorithme et nous montrons son efficacité sur des jeux de données synthétiques et réelles.

Dans le Chapitre 5, nous utilisions une approche basée sur la méthode de Stein pour borner la distance de Wasserstein entre deux mesures et nous donnons plusieurs applications de nos résultats. Nous commençons par donner des vitesses de convergence pour le théorème central limite. Nous montrons ensuite comment nos résultats peuvent être utilisés pour quantifier la convergence de mesures invariantes de marches aléatoires sur des graphes géométriques aléatoires en prenant comme exemple le cas des graphes de k plus proches voisins. Enfin, nous étudions les performances d'un algorithme basique de type Monte-Carlo.

Dans le Chapitre 6, nous présentons l'approche "sac-de-mots" pour la reconnaissance de formes ainsi que les forces et faiblesses des méthodes de pooling usuelles. Nous présentons notre nouvelle méthode de pooling basée sur l'homologie persistante 0-dimensionnelle et étudions ses performances sur le jeu de données SHREC 2014.

Enfin, dans le Chapitre 7, nous présentons différentes pistes de nouvelles recherches pour mieux comprendre la convergence de marches aléatoires sur des graphes géométriques aléatoires.

Chapter 2

Introduction

2.1 Motivation

Data analysis algorithms are used in a large number of applications such as search engines, online advertisement, sports, etc. The range of these applications is to become even broader as new promising applications in other domains such as health care are arising. As the stakes involved in these new applications are higher, we need a strong understanding of the algorithms we are using and, if possible, theoretical guarantees ensuring the correctness of these algorithms. Another challenge posed by these new applications is the increasing complexity of the data to process, either due to the sheer amount of data available or to the nature of the data itself. The dimensionality of the data is one such source of complexity: the efficiency of learning algorithms drops quickly when the dimension of the data increases, a phenomenon called the curse of dimensionality. Fortunately, in some cases the data happens to be drawn from a lower dimensional structure, a manifold, embedded in a higher dimensional space. It is then crucial to take advantage of this structure. Suppose the data takes the form of a point cloud of independent and identically distributed random variables drawn from a measure μ with density f supported on a manifold. Since we do not have access to the manifold itself, a natural idea to use the manifold structure would be to build a mesh from the data point cloud and then perform data analysis using this mesh. As computing an actual mesh might be too computationally expensive, one can instead compute a graph using the data points as vertices and adding an edge between two vertices if they are sufficiently close. Such a graph is called a random geometric graph. At this point, we can expect the graph to capture the structure of the manifold, thus the next step is processing the information contained in the graph

through graph analysis algorithms. An efficient approach to graph analysis consists in using the properties of a random walk to gain insight of the full graph structure. Obtaining theoretical guarantees, such as consistency, for algorithms using random geometric graphs can then be achieved by understanding the properties of random walks on these graphs. Although these are complex objects, it turns out it is possible to gain some insight on their properties in an asymptotic setting. Indeed, when the number of data samples goes to infinity, and under some rather natural assumption on the way the graphs are built, random walks on random geometric graphs converge to a diffusion process whose generator depends on both the Laplacian of the manifold and the sampling density f. Since the limiting diffusion does not only depend on the manifold structure but also on f, it makes sense to use random walks on random geometric graphs to perform non-linear data analysis even when μ has a full-dimensional support. Moreover, the convergence of random walks on random geometric graphs to diffusion processes can also be exploited to study the consistency of random-walk based algorithms on such graphs. Our work will consist in using the convergence of random walks on random geometric graphs to propose a new data analysis algorithm and to provide a better understanding of the convergence of a property of random walks often used by data analysis algorithm: the invariant measure. Outside of the invariant measure, random walk-based data analysis usually rely on two properties: the graph Laplacian and hitting times. Let us discuss the consequences of the convergence of random walks to diffusion processes for these three properties.

Graph Laplacian Graph Laplacians are operators acting on functions of the vertices of a graph G. Here we focus on the random walk graph Laplacian $\mathcal{L}$ of G which is simply the opposite of the generator of a random walk on G. As the spectrum of $\mathcal{L}$ contains information regarding the connectivity of G, it is often used by Spectral clustering algorithms [57]. Graph Laplacians are also used in many other applications such as dimensionality reduction using Diffusion maps [7] or semi-supervised classification [19]. Since random walks on random geometric graphs converge to diffusion processes, we can expect the generators of the random walks to converge to the infinitesimal generators of the limiting diffusion processes. There are two main approaches to obtain such a convergence. The first approach is to show the convergence of $\mathcal{L}\phi$ for ϕ belonging to a set of functionals [9, 39, 45]. The convergence of random walks to diffusion processes is actually obtained using this type of convergence [81]. The second approach focuses on proving the convergence of the spectrum of the graph Laplacian to the spectrum of

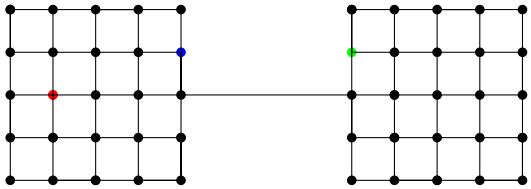


Figure 2.1: Two highly connected components linked by a single edge.

the infinitesimal generator of the limiting diffusion process [8, 38].

Hitting times It is also possible to study the structure of the graph by using the time required for a random walk to travel from a vertex x to another vertex x'. This quantity is called hitting time from x to x', and we denote it by $H(x, x')$. Hitting times can be used to define a metric on the graph called commute distance, where the commute distance between two vertices x, x' is simply the sum $H(x, x') + H(x', x)$. The commute distance is more informative than simpler distances such as the shortest path distance as it takes into account the full graph structure. For instance, let us consider the graph drawn in Figure 2.1 which is composed of two highly connected components linked by a single edge. Intuitively, we expect the blue vertex and the red vertex to be close, since they belong to the same highly connected component, while being far away from a vertex from the other component such as the green vertex. While this is true for the commute distance, it is not the case for the shortest path distance. Hence, the commute distance is used with great success in many different applications such as graph embedding [73, 68], clustering [90] or semi-supervised learning [91]. To study the consistency of hitting times, it would seem natural to compare hitting times of random walks on random geometric graphs to hitting times of the limiting diffusion processes. However, in dimension greater than one, the hitting time of a diffusion process to a given point is infinite, thus we cannot expect hitting times of random walks on random geometric graphs to converge to meaningful quantities. This ill-behaviour was described in [85], where the authors proved the hitting times between two vertices of a large random geometric graph depend mostly on local quantities alone, in which case hitting times do not capture the global structure of the graph.

Invariant measure Finally, it is possible to use the invariant measure of a random walk on a connected graph to process data. The most famous algorithm using this approach is probably the PageRank algorithm [65] which ranks web-pages using the invariant measure of a random walk on

the graph obtained by taking web pages as vertices and hyperlinks as edges. When studying the invariant measure of a random walk on a graph, there are two cases. When the graph undirected, the invariant measure can be directly computed: it is proportional to the degree function. On the other hand, it is hard to say anything about the invariant measure of a random walk on a directed graph. This is problematic since nearest neighbor graphs, which are quite popular in data analysis due to their sparsity, are directed. In our case, one may wonder whether the convergence of random walks to diffusion processes implies the convergence of the invariant measure of the random walks to the invariant measure of the diffusion processes. Let us note that the invariant measure of the limiting diffusion process depends can be used to recover the underlying density f. Hence, the invariant measure of a random walk on a random geometric graph can be used to obtain an estimator of f. Moreover, this estimator can be computed using the graph structure alone, which would solve an open problem stated in [84]. Indeed, while this density estimator would not be interesting whenever we have access to the coordinates of the data points, in which case it is more efficient to use a standard density estimator, it would confirm we are not loosing information by working with the random geometric graph rather than the original point cloud. Indeed, if we are able to recover f, and thus μ, from the sole graph structure, then it means it contains most of the original information. However, the current results regarding the convergence of invariant measure of random walks on random geometric graphs are not sufficient to prove the consistency of such an estimator since the only known result is a weak convergence result proved in [44].

2.2 Contributions

As we have seen, the convergence of random walks on random geometric graphs is usually used to prove the consistency of data analysis algorithms. In a first part of this work, we propose to use this convergence to design a new algorithm for soft clustering. The objective of clustering algorithms is to partition the data in clusters, hoping that data belonging to a similar cluster shares similar properties. Overall, clustering provides a fairly limited knowledge of the structure of the data: while the partition into clusters is well understood, the interplay between clusters -respective locations, proximity relations, interactions- remains unknown. Identifying interfaces between clusters is the first step toward a higher-level understanding of the data, and it already play a prominent role in some applications such as the study of the conformations space of a protein, where a fundamental

question beyond the detection of metastable states is to understand when and how the protein can switch from one metastable state to another [29]. Clustering can be used in this context, for instance by defining the border between two clusters as the set of data points whose neighborhood, in the ambient space or in some neighborhood graph, intersects the two clusters, however this kind of information is by nature unstable with respect to perturbations of the data. Soft clustering appears as the appropriate tool to deal with interfaces between clusters. Rather than assign each data point to a single cluster, it computes a degree of membership to each cluster for each data point. The promise is that points close to the interface between two clusters will have similar degrees of membership to these clusters and lower degrees of membership to the rest of the clusters. Thus, compared to hard clustering, soft clustering uses a fuzzier notion of cluster membership in order to gain stability on the locations of the clusters and their boundaries. Here, we are interested in the mode-seeking approach to clustering. In mode-seeking, clusters are defined continuously as the basins of attractions of the modes of the density. In other words, clusters are defined through the limit of a gradient flow of the density. We propose to derive a soft clustering approach to a mode-seeking algorithm called Topological Mode Analysis Toolbox algorithm [22] by turning this gradient flow into a stochastic differential equation whose solution is a diffusion process depending on the density f. In practice, we approximate the behaviour of the diffusion process by a random walk on a random geometric graph computed using a density estimator. We then prove the consistency of our algorithm, study the impact of the various parameters on the output of our algorithm and evaluate the performance of our algorithms on several datasets. An interesting aspect of our algorithm is that it does not require the actual coordinates of the data points but merely a random geometric graph structure and a density estimator. Hence, this algorithm could directly be used as a graph analysis algorithm for random geometric graphs provided we are able to compute a density estimator using the graph structure alone. As we have mentioned earlier, it is possible to compute a density estimator from a random geometric graph using a random walk on the graph. However, the only result regarding the convergence of this invariant measure is expressed in terms of weak convergence. The second contribution of this thesis will consist in strengthening this result by quantifying this convergence.

Our approach is based on Stein's method, which was originally developed to bound the distance to the Gaussian measure to provide rates for the Central Limit Theorem [52]. In [5], it was shown Stein's method can actually be used to bound the distance between the invariant measure of a reversible Markov chain to the invariant measure of a diffusion process

using their (infinitesimal) generators. This seems close to our objective since a random walk on a graph is a Markov chain, however there are two main issues with using this result in our setting. First, it is difficult to use in a multidimensional setting. Second, it requires a reversible Markov chain and thus would only be helpful to study random walks on undirected graphs which we already know how to deal with. By adapting a recent result from the Stein's method literature [52], we provide a way to bound the 2-Wasserstein distance between the invariant measure ν of a Markov chain with generator $\mathcal{L}_\nu$ and a measure μ which is the reversible measure of a differential operator $\mathcal{L}_\mu$. More precisely, we show that the distance between μ and ν can be bounded using a discrepancy between the two operators $\mathcal{L}_\nu$ and $\mathcal{L}_\mu$. We are thus able to quantify the convergence of invariant measures of random walks on random geometric graphs: as an illustration, we compute the rate of convergence in the case of a k-nearest neighbor graph. Whenever μ is the Gaussian measure, we provide a more refined bound which allows us to give rates of convergence for the Wasserstein distance for the Central Limit Theorem, either for the $p \geq 2$-Wasserstein distance in the unidimensional setting or for the 2-Wasserstein distance distance in the multidimensional case. Finally, we also show how our result can be used to study a simple Markov Chain Monte Carlo algorithm.

The last work of this thesis is not related to random walks on random geometric graphs. Instead, we tackle the problem of shape recognition by using the concept of 0-dimensional persistent homology used by the Topological Mode Analysis Toolbox clustering algorithm. In order to automatically process 3D shapes databases, there exist multiple retrieval and classification algorithms. The bag-of-words approach, originally developed for image analysis, is one such classification pipeline used to process 3D shapes. Traditionally, the bag-of-words method relies on extracting an unordered collection of descriptors from the shapes we consider. Each descriptor is then quantized by a vector whose coordinates are called words. The information available in the words values is then extracted and summarized through a step called pooling, producing a vector usable by standard learning algorithms. Traditional pooling schemes consist in taking either the average of the value -sum-pooling- or the maximum value -max-pooling- of each words across the shape. Ideally, each step of this framework should be robust to standard transformations which may affect 3D shapes -translations, rotations or changes of scale- while being sufficiently discriminative. Here, we propose to put more structure in the bag-of-words approaches: rather than considering an unordered collection of word values, we instead view these values as a set of functions on the mesh of the shape. From here, the pooling step should extract information from a function opening up new avenues to

refine traditional pooling schemes. For instance, the max-pooling scheme, which consists in taking the global maximum of a word function, could be refined by considering all the local maxima of a word function instead. But such an approach would not be stable since perturbing a function, even slightly, can create many uninteresting multiple local maxima. To fix this issue, we use the concept of 0-dimensional persistent homology, or prominence, to obtain a pooling procedure which uses all the local maxima of the word functions and weights them regarding to their stability which respect to perturbations. Thus, we are able to generalize max-pooling to design a more discriminative scheme while retaining most of its stability. We provide experimental results proving the efficiency of this new pooling scheme.

2.3 Organization of the thesis

In Chapter 3, we present random walks on random graphs along with algorithms using these random walks to perform graph analysis. We then give a brief presentation of diffusion processes. Finally, we present random geometric graphs and known results on the convergence of random walks on random geometric graphs.

In Chapter 4, we present the mode-seeking framework and the Topological Mode Analysis Toolbox algorithm. We then show how this algorithm can be turned into a soft clustering algorithm by using a random walk on a properly weighted random geometric graph. We prove the consistency of this new algorithm and show its effectiveness on both synthetic and real datasets.

In Chapter 5, we obtain new bounds on the Wasserstein distance between two measures using Stein's method and provide several applications of these bounds. We first derive convergence rates in the Central Limit Theorem which hold in the multivariate setting. We then show how to apply our result to bound the Wasserstein distance between the invariant measure of a random walk and the invariant measure of a diffusion process. Using this result, we are able to give rates for the convergence of the invariant measure of a random walk on a random geometric graph. As an illustration, we study the special case of the k-nearest neighbor graph. Finally, we show how our results can be used to study a simple Monte Carlo algorithm.

In Chapter 6, we present the bag-of-word approach for shape recognition and discuss the strengths and weaknesses of traditional pooling procedure. We then provide a new pooling procedure based on 0-dimensional persistent homology and assess its performance on the SHREC 2014 dataset.

Finally, in Chapter 7, we highlight possible avenues for new research to

better understand the convergence of algorithms relying on random walks on random geometric graphs.

Chapter 3

Random walks on random geometric graphs

In this chapter, we formally introduce the concept of random walks on graphs as Markov chains and we provide an overview of the main random walk-based algorithms used in graph analysis. Then, we introduce diffusion processes, their relevant properties for our works and we show how to approximate a diffusion process through a Markov chain. Finally, we describe random geometric graphs and we explore the consequences of the convergence of random walks on random geometric graphs to diffusion processes.

3.1 Markov chains for graph analysis

A (time-homogeneous) Markov chain is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ taking values in a measurable space $(E, \mathcal{F})$ such that the behavior of X_n depends on X_{n-1} only. More formally we say that $(X_n)_{n\in\mathbb{N}}$ is a Markov Chain if

$$\begin{aligned}\forall n \in \mathbb{N}, \forall F \in \mathcal{F}, \mathbb{P}(X_{n+1} \in F \mid X_0, \ldots, X_n) &= \mathbb{P}(X_{n+1} \in F \mid X_n) \\ &= K(X_n)(F),\end{aligned}$$

where K is a function from E to the space of probability measures on E called the transition kernel of the Markov chain. A Markov chain is completely characterized by its state space E, the measure of X_0 and K. Whenever E is discrete, the measures $K(.)$ are completely characterized by their values on singletons. Hence, we can define K as a function from E^2 to $[0, 1]$ such that

$$\forall x, x' \in E, K(x, x') = \mathbb{P}(X_1 = x' \mid X_0 = x).$$

In the remainder of this Section, we assume E to be discrete.

Let $\mathcal{G}$ be a weighted graph with vertices $\mathcal{X}$ and weight function $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. A random walk on G is a Markov chain with state space $\mathcal{X}$ and transition kernel K defined by

$$\forall x, x' \in \mathcal{X}, K(x, x') = \frac{w(x, x')}{\sum_{x \in \mathcal{X}} w(x, x)}.$$

We say a graph is undirected if

$$\forall x, x' \in \mathcal{X}, w(x, x') = w(x', x).$$

Finally, the degree of a vertex x is given by

$$d(x) = \sum_{x' \in \mathcal{X}} w(x', x),$$

and the volume of a subset $A \subset \mathcal{X}$ is

$$vol(A) = \sum_{x \in A} d(x).$$

3.1.1 Graph Laplacian

Consider a Markov chain with state space E and transition kernel K. We call generator of the Markov chain the operator $\mathcal{L}$ such that, for any function $\phi : E \to \mathbb{R}$ and any state $x \in E$,

$$\mathcal{L}\phi(x) = \mathbb{E}[\phi(X_1)|X_0 = x] - \phi(x) = \sum_{x' \in \mathcal{X}} (\phi(x') - \phi(x))K(x, x').$$

In the case of a random walk on a graph, the opposite of this generator is called the graph Laplacian $L := -\mathcal{L}$. Actually, there exist three possible definitions of graph Laplacians depending on the normalization used [57]; the graph Laplacian defined here is usually called random walk graph Laplacian. Let us study the spectrum of L.

Proposition 1 (Proposition 3 [57])**.** *If $\mathcal{G}$ is undirected, then L is positive semi-definite and admits $|\mathcal{X}|$ positive real eigenvalues.*

These eigenvalues and the corresponding eigenvectors contain a great deal of information regarding the structure of the graph itself. As an example, let us consider the problem of clustering which consists in gathering vertices of the graph in clusters such that the intra-cluster connectivity is

high while the inter-cluster connectivity is small. The most elementary way to cluster a graph is to associate clusters with connected components. While there are more direct ways to recover the connected components of an undirected graph, they can be computed from the spectrum of the graph Laplacian.

Proposition 2 (Proposition 4 [57]). *If $\mathcal{G}$ is undirected, then the multiplicity of the eigenvalue 0 of L equals the number of connected components $A_1, \ldots, A_k$ in the graph. Moreover, the eigenspace of 0 is spanned by the indicator functions of these components.*

Since an eingenvalue equal to zero indicates a connected component, we could expect a small eingenvalue to indicate a subset of $\mathcal{X}$ that is close to being disconnected from the rest of the graph. In order to quantify how close $A \subset \mathcal{X}$ is close to being disconnected from the graph, we can consider the cut value

$$cut(A) = \frac{1}{2} \sum_{x \in A, x' \notin A} w(x, x').$$

This cost corresponds to the total weight of the edges we would need to remove in order to disconnect A from the rest of the graph. If this quantity is small, it means A and $\mathcal{X} - A$ are close to being disconnected. In order to define two clusters on a connected graph, it is tempting to assign the first cluster to a set A minimizing the cut value and the second cluster to $\mathcal{X} - A$. However, quite often, the set A minimizing this cut value is simply a vertex of the graph so this quantity alone cannot be used to obtain a good clustering. Instead, one possibility is to minimize a regularized quantity such as Ncut [76]

$$Ncut(A) = cut(A)(Vol(A)^{-1} + Vol(\mathcal{X} - A)^{-1}).$$

In this case, if A (or $\mathcal{X} - A$) is reduced to a point, then its volume will be low and $Ncut(A)$ large. To divide a graph in two clusters, one would like to find the set minimizing the Ncut value and define it as a cluster while the second cluster would correspond to the remainder of the graph. However, this minimization problem is NP-hard [86]. Fortunately, a relaxation of this minimization problem is easier to deal with: its solution is simply the second eigenvector of the graph Laplacian [57]. Moreover, if the associated eigenvalue is small, so is the corresponding relaxed quantity. This analysis can be generalized when we want to cluster $\mathcal{X}$ in k different subclusters in which case it involves the first k eigenvalues and eigenvectors of the graph Laplacian.

Algorithms using the spectral properties of graph Laplacians to perform clustering belong to the family of Spectral clustering algorithms [57]. The Spectral clustering algorithm corresponding to our definition of the graph Laplacian was introduced in [76]: one computes the first k (k being the number of clusters) eigenvectors of L, each vertex of the graph is then embedded into $\mathbb{R}^d$ with coordinates equal to the values of the eigenvectors on the vertex, the clustering is finally obtained using a k-means algorithm [43].

3.1.2 Hitting time and commute distance

Let $x \in E$ and let $(X_n^x)_{n\geq 0}$ be a random walk on the graph such that $X_0^x = x$. For any state $x' \in E$, we define the hitting time from x to x' by

$$\tau_{x,x'} = \inf_{n\geq 0}(X_n^x = x'),$$

and we denote by $H_{x,F}$ the mean hitting time from x to x'

$$H_{x,F} = \mathbb{E}[\tau_{x,x'}].$$

Finally, we define the commute distance between any two states x and x' as

$$C_{x,x'} = H_{x,x'} + H_{x',x}.$$

For a random walk on a connected graph, the commute distance between two vertices is always finite. Moreover, on an undirected graph, the commute distance can be computed using the generalized inverse of the graph Laplacian [57].

3.1.3 Invariant measure

A probability measure π is said to be an invariant measure for the Markov chain if

$$\forall x \in E, \pi(x) = \sum_{x'\in E} K(x', x)\pi(x').$$

We say a Markov chain to be irreducible if

$$\forall x, x' \in E, \exists n > 0, \mathbb{P}(X_n = x'|X_0 = x) > 0.$$

Furthermore, we say a Markov chain to be positive recurrent if

$$\forall x \in E, H_{x,x} < \infty.$$

Irreducibility and positive recurrence ensures the existence of a unique invariant measure for the Markov chain.

Proposition 3 (Theorem 1.7.7 [61])**.** *An irreducible Markov chain admits a unique invariant measure π if and only if it is positive recurrent. Moreover,*

$$\forall x \in E, \pi(x) = \frac{1}{H_{x,x} \sum_{x' \in E} H_{x,x}^{-1}}.$$

Computing the invariant measure of a random walk on an undirected graph is simple: it is proportional to the degree function of the graph.

Proposition 4. *The random walk on a connected and undirected graph admits a unique invariant measure π with*

$$\forall x \in \mathcal{X}, \pi(x) = \frac{\sum_{x' \in \mathcal{X}} w(x, x')}{\sum_{x', x'' \in \mathcal{X}} w(x', x'')}.$$

On the other hand, the invariant measure of a random walk on a directed graph cannot be deduced from local quantities alone. For instance, let us take $n \in \mathbb{N}$ and consider the random walk of the graph with vertices $\mathcal{X} = \{0, \ldots, n\}$ and weight function

$$\forall i, j \in \mathcal{X}, w(i, j) = \begin{cases} 1 \text{ if } j = 0 \\ 1 \text{ if } j = i + 1 \\ 0 \text{ otherwise} \end{cases}.$$

The invariant measure of a random walk on this graph is defined on any vertex $i \in \{1, \ldots, n\}$ by

$$\pi(i) = \frac{2^{n-i}}{2^{n+1} - 1}.$$

As we can see, the values of this invariant measure do not depend on local quantities such as the degree of a vertex.

For directed graphs, computing π requires solving a system of linear equations. However, the computations can be prohibitive whenever $|E|$ is large. Instead, it is sometimes possible to approximate π. We say a Markov chain to be aperiodic if

$$\forall x \in E, \exists n' \in \mathbb{N}, \forall n \geq n', \mathbb{P}(X_n = x | X_0 = x) > 0.$$

An approximation of π can be obtained through the measures of X_n for large values of n as long as the Markov chain is aperiodic, irreducible and positive recurrent, in which case it is said to be ergodic.

Proposition 5 (Theorem 1.8.3 [61])**.** *If E is connected and $(X_n)_{n \geq 0}$ is ergodic then*

$$\forall x \in E, \lim_{n \to \infty} \mathbb{P}(X_n = x) = \pi(x).$$

There is a large variety of results providing convergence rates of X_n to π, for example if the generator of the Markov chain is diagonalisable with eigenvalues $\lambda_{|E|} \leq \ldots \lambda_2 \leq \lambda_1 = 0$, then when $|\lambda_2|$ is large, the convergence of the measure of X_n to π is fast [53]. This is not a surprise: if we consider a random walk on a graph, when λ_2 is close to zero then the first non-zero eigenvalue of the graph Laplacian is small as well. This means there exists $A \subset \mathcal{X}$ such that the connectivity between A and the rest of the graph is small. Hence, the random walk will take a long time entering or leaving A.

Invariant measures of random walks on graphs are commonly used for graph analysis. For example, let $\mathcal{X}$ be the set of all web-pages and, for two web-pages x, x', let $a(x, x') = 1$ if the page x contains a link to the page x'. For $\alpha > 0$, let

$$w_\alpha(x, x') = \frac{\alpha}{|\mathcal{X}|} + (1 - \alpha)\frac{a(x, x')}{\sum_{x \in \mathcal{X}} a(x, x)}.$$

The random walk on the random graph with vertices $\mathcal{X}$ and weight function w_α can be used as a simple model of a person clicking randomly on links or, sometimes, jumping to a totally unrelated page. This random walk is irreducible and positive recurrent, it thus admits an invariant measure π. By Proposition 5, the value of this invariant measure on a vertex corresponds to the probability for the random surfer to be visiting the corresponding web-page after an infinite amount of time, it will thus tend to be higher for web-pages having many incoming hyperlinks, which can be assumed to be more relevant than other pages. Hence, the value of the invariant measure can be used to rank vertices of the graph and thus web-pages. This algorithm is a simpler version of the PageRank algorithm [65] which is used along many other ranking algorithms in the Google search engine. As the total number of web-pages is very large, this invariant measure is computed using bots actually performing the random walk.

3.2 From Markov chains to diffusion processes

In order to study the behaviour of algorithms based on random walks on graphs, we want to exploit the convergence of these random walks to continuous processes called diffusion processes. In this section, we introduce these processes and give a brief overview of their relevant properties for our work. A more detailed presentation of these processes can be found in [3, 34, 37, 79]. We start by introducing a few notations.

3.2.1 Notations for multivariate analysis

Let $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$, we denote by $x^{\otimes k} \in (\mathbb{R}^d)^{\otimes k}$ the tensor of order k of x:

$$\forall j_1, \dots, j_k \in \{1, \dots, d\}, (x^{\otimes k})_{j_1,\dots,j_k} = \prod_{i=1}^{k} x_{j_i}.$$

For any $x, y \in (\mathbb{R}^d)^{\otimes k}$, the Hilbert-Schmidt scalar product between x and y is

$$< x, y >= \sum_{i \in \{1,\dots,d\}^k} x_i y_i,$$

and, by extension,

$$\|x\|^2 =< x, x > .$$

Let $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$, we have

$$\begin{aligned}
\|x^{\otimes k}\|^2 &= \sum_{i \in \{1,\dots,d\}^k} (\prod_{j=1}^{k} x_{i_j})^2 \\
&= \sum_{i \in \{1,\dots,d\}^k} \prod_{j=1}^{k} x_{i_j}^2 \\
&= (\sum_{l=1}^{k} x_l^2)^k \\
&= \|x\|^{2k}.
\end{aligned}$$

Hence,

$$\|x^{\otimes k}\| = \|x\|^k.$$

For any smooth function ϕ and $x \in \mathbb{R}^d$, let $\nabla^k \phi \in (\mathbb{R}^d)^{\otimes k}$ such that

$$\forall j_1, \dots, j_k \in \{1, \dots, d\}, (\nabla^k \phi(x))_{j_1,\dots,j_k} = \frac{\partial^k \phi}{\partial x_{j_1} \dots \partial x_{j_k}}(x).$$

Finally, the Laplacian of a function ϕ is

$$\Delta \phi =< \nabla^2 \phi, I_d >= \sum_{i=1}^{d} \frac{\partial^2 \phi}{\partial^2 x_i}.$$

3.2.2 Diffusion processes

Let E be an open domain of $\mathbb{R}^d$. Let $(K_t)_{t \geq 0}$ be a family of kernels from E to the space of probability measures on E such that

- for any $x \in E$, and any Borel set $B \subset E$ such that $x \in B$, $K_0(x)(B) = 1$ ($K_0(x)$ is the Dirac measure centered on x) ;
- for any Borel set $B \subset E$ and any $t \geq 0$, $K_t(.)(B)$ is a measurable function;
- for any Borel set B and any $s, t \geq 0$, $K_{t+s}(x)(B) = \int_E K_t(y)(B)K_s(x)(dy)$.

This family is said to be a transition function and plays a similar role as the transition kernel for Markov chains. A family of random variables $(X_t)_{t\geq 0}$ taking values in E is a Markov process if, for every compactly supported continuous function ϕ,

$$\forall s_1 \leq \cdots \leq s_n \leq t, \mathbb{E}[\phi(X_t)|X_{s_1}, \ldots, X_{s_n}] = \int_E \phi(y)K_{t-s_n}(X_{s_n})(dy).$$

Remark that such a definition implies that $(X_t)_{t\geq 0}$ satisfies the Markov property:

$$\forall s_1 \leq \cdots \leq s_n \leq t, \mathbb{E}[\phi(X_t)|X_{s_1}, \ldots, X_{s_n}] = \mathbb{E}[\phi(X_t)|X_{s_n}].$$

Let us consider a simple case of a Markov process $(W_t)_{t\geq 0}$ for which

- $W_0 = 0$;
- $t \to W_t$ is almost surely continuous;
- For any $s \leq t$, $W_t - W_s$ is a d-dimensional centered normal random variable with variance $(t-s)I_d$;
- $W_t - W_s$ is independent of W_s.

W_t is called the d-dimensional Brownian motion. By Donsker's Theorem, we know that if $Z_1, \ldots, Z_n$ are i.i.d. random variables with mean 0 and variance 1 then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i 1_{\frac{i}{n} \leq t}$ converges to $(W_t)_{t\in[0,1]}$. In other words, we are able to approximate the trajectories of $(W_t)_{t\geq 0}$ by trajectories of a Markov Chain $(X_k)_{k\in\mathbb{N}}$ with state space $\mathbb{R}$ such that

$$X_{k+1} = X_k + \frac{1}{\sqrt{n}} Z'_k,$$

where the Z'_k are i.i.d. random variables with the same measure as Z_1. While this is a good start, we cannot expect random walks on random geometric graphs to always converge to a Brownian motion. On the other hand, the class of Markov processes is too large to study convergence of Markov chains.

Let C_0 be the space of real-valued, continuous and compactly supported functions on E. For any $\phi \in C_0$, any $x \in E$ and any $t > 0$, let

$$P_t\phi(x) = \mathbb{E}[\phi(X_t) \mid X_0 = x] = \int_E \phi(y) K_t(x)(dy).$$

We say $(P_t)_{t\geq 0}$ is the semigroup associated to $(X_t)_{t\geq 0}$ because

$$\begin{aligned} P_{t+s}\phi(x) &= \int_E \phi(y) K_{t+s}(x)(dy) \\ &= \int_{E\times E} \phi(z) K_s(x)(dz) K_t(z)(dy) \\ &= P_t(P_s\phi(x)). \end{aligned}$$

A family of bounded operators $(P_t)_{t\geq 0}$ acting on E is a Feller semigroup if

- $P_0 = Id$;
- $\forall t \geq 0, P_t 1 = 1$;
- $\forall s, t \geq 0 P_{t+s} = P_t P_s$;
- $\forall \phi \in C_0, t \geq 0, \|P_t\phi\|_\infty \leq \|\phi\|_\infty$;
- $\forall \phi \in C_0, t \geq 0, \phi \geq 0 \implies P_t\phi \geq 0$;
- $\forall \phi \in C_0, \lim_{t\to 0} \|(P_t - Id)f\|_\infty \to 0$.

A Markov process associated to a Feller semigroup is called a Feller process. Furthermore, given a Feller semigroup, it is always possible to define a Feller process associated to this semigroup.

In particular, a Feller semigroup is a strongly continuous positive contraction semigroup. These semigroups are well-studied objects [36] and can be characterized by an operator called infinitesimal generator. The infinitesimal generator $\mathcal{L}$ of a Feller semigroup $(P_t)_{t\geq 0}$ is an operator acting on

$$D(\mathcal{L}) = \{\phi \in C_0 \mid \lim_{t\geq 0} \frac{(P_t - Id)\phi}{t} \ exists\},$$

and such that, for any $\phi \in D(\mathcal{L})$,

$$\lim_{t\geq 0} \frac{(P_t - Id)\phi}{t} = \mathcal{L}\phi.$$

$D(\mathcal{L})$ is dense in C_0 and $\mathcal{L}$ is a linear operator. Moreover, by the Hille-Yosida Theorem, a linear operator satisfying some technical conditions is

also the generator of a Feller semigroup. Hence, there is a one to one correspondence between Feller semigroups and a set of linear operators. A diffusion process $(X_t)_{t\geq 0}$ is a Markov process associated to a Feller semigroup with infinitesimal generator $\mathcal{L}$ such that $\forall \phi \in D(\mathcal{L})$,

$$\mathcal{L}\phi = b.\nabla\phi + < a, \nabla^2\phi >,$$

where $b : E \to \mathbb{R}^d$ and $a = \frac{\sigma^T\sigma}{2}$ with $\sigma : E \to \mathbb{R}^d \times \mathbb{R}^d$ such that σ is symmetric and positive definite on E. For any diffusion process $(X_t)_{t\geq 0}$, there exists $(\tilde{X}_t)_{t\geq 0}$ such that the trajectories of $(\tilde{X}_t)_{t\geq 0}$ are almost surely continuous and $\forall t \geq 0, \mathbb{P}(X_t = \tilde{X}_t) = 1$. Hence, without loss of generality, we can assume the trajectories of diffusion processes to be almost surely continuous.

In order to give more intuition about diffusion processes, let us note that a diffusion process $(X_t)_{t\geq 0}$ is actually the solution to the following stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{3.1}$$

where W_t is a d-dimensional Brownian motion. The function b is called a drift and correspond to a deterministic trend for the trajectories of X_t whereas σ, called diffusion coefficient, controls the noise disrupting the trajectories. If σ were equal to zero then Equation 3.1 would be a purely deterministic ordinary differential equation.

3.2.3 Convergence of Markov chains to diffusion processes

In order to tackle the problem of convergence of Markov chains, we must first specify a suitable notion of convergence. Let μ and $(\nu_n)_{n\in\mathbb{N}}$ be measures on a space E. We say $(\nu_n)_{n\in\mathbb{N}}$ converges weakly to μ if, for any continuous and bounded function ϕ on E,

$$\int_E \phi d\nu_n \to \int_E \phi d\mu.$$

Weak convergence can actually be characterized in several ways.

Theorem 6 (Portmanteau Theorem)**.** *A family of probability measure $(\nu_n)_{n\in\mathbb{N}}$ converges weakly to another probability measure μ if and only if*

- *For any uniformly continuous function ϕ, $\int_E \phi d\nu_n \to \int_E \phi d\mu$;*
- *For any closed set $F \subset E$, $\limsup \nu_n(F) \leq \mu(F)$;*

- *For any open set* $F \subset E$, $\liminf \nu_n(F) \geq \mu(F)$;
- *For any Borel set* $F \subset E$ *such that* $\mu(\partial F) = 0$, $\nu_n(F) = \mu(F)$.

By extension, we say a family of random variables $(X_n)_{n\in\mathbb{N}}$ converges weakly to Y if the measures of $(X_n)_{n\geq 0}$ converges weakly to the measure of Y.

As we have seen, diffusion processes are continuous processes, so one may be tempted to obtain a convergence in the space of continuous trajectories of E. However, as we have seen with the example of Donsker's Theorem, an approximation built using a Markov chain is in general not continuous. Instead, given $T > 0$, we are going to work in the space of trajectories $[0, T] \to \mathbb{R}^d$ that are right-continuous and have left limits. This space is called the Skorokhod space and is equipped with the following metric

$$d(f, g) = \inf_{\epsilon}\{\exists \lambda \in \Lambda, \|\lambda\| \leq \epsilon, \sup_{t\leq T} |f(t) - g(\lambda(t))| \leq \epsilon\},$$

where Λ denotes the space of strictly increasing automorphisms of the unit segment $[0, 1]$, and where $\|\lambda\|$ is

$$\|\lambda\| = \sup_{s\neq t} \left|\log\left(\frac{\lambda(t) - \lambda(s)}{t - s}\right)\right|.$$

Consider a family of Markov chains $(X_n^s)_{n\in\mathbb{N},s\geq 0}$ defined on discrete state spaces $E_s \subset E$ with initial states $X_0^s = x_s$ and transition kernels K^s. For $x \in E_s$, $\gamma > 0$ and $s > 0$, let

- $b^s(x) = \frac{1}{s}\int_{x'\in E_s}(x' - x)K^s(x, dx')$;
- $a^s(x) = \frac{1}{2s}\int_{x'\in E_s}(x' - x)(x' - x)^T K^s(x, dx')$;
- $\Delta_s^\gamma = \frac{1}{s}K^s(x, \mathcal{B}(x, \gamma)^c)$,

where $\mathcal{B}(x, \gamma)^c$ is the complementary of the ball of radius γ centered in x.

Theorem 7 (Theorem 7.1 [34])**.** *Suppose* $(X_t)_{t\geq 0}$ *is a diffusion process on* $E \subset \mathbb{R}^d$ *with* $X_0 = x_0$ *and generator* $\mathcal{L} = b.\nabla + < a, \nabla^2 >$ *such that* b *and* a *are continuous on* E. *If,*

(i) $\lim_{s\to 0} \sup_{x\in E_s} \|b^s - b\|_\infty = 0$;

(ii) $\lim_{s\to 0} \sup_{x\in E_s} \|a^s - a\|_\infty = 0$;

(iii) $\forall \gamma > 0, \lim_{s\to 0} \sup_{x\in E_s} \Delta_s^\gamma = 0$;

(iv) $\lim_{s\to 0} \|x_s - x_0\|_\infty = 0$,

then, for any $T > 0$*, the continuous time process* $(X^s_{s\lfloor t/s\rfloor})_{t\geq 0}$ *converges weakly in* $D([0,T],\mathbb{R}^d)$ *to* $(X_t)_{t\geq 0}$*. Furthermore, the convergence is uniform with respect to* $x_0 \in U$*.*

This result tells us that, if the expected direction of a step of the Markov chain is close to b, its second moment is close to a and the length of the jump is small, then the trajectory of the Markov chain is close to the trajectory of the diffusion process. It is possible to relate these assumptions to a convergence of the generator of the Markov chain.

Proposition 8. *Suppose there exists* $R > 0$ *such that*

$$\forall s > 0, \forall x \in E_s, \int_{x'\in E_s, \|x'-x\|\geq R} K(x, dx') = 0.$$

If assumptions $(i) - (iii)$ *of Theorem 7 hold, then*

$$\lim_{s\to 0} \sup_{\phi, \|\nabla\phi\|, \|\nabla^2\phi\|, \|\nabla^3\phi\|\leq 1} \sup_{x\in E_s} |(\frac{1}{s}\mathcal{L}_s - \mathcal{L})\phi(x)| \to 0.$$

Proof. In order to simplify the notations, we prove the result in dimension one. Let ϕ be a smooth function with bounded first three derivatives. By Taylor's expansion we have, for any $\gamma > 0$,

$$\begin{aligned}\frac{1}{s}\mathcal{L}_s\phi(x) &= \frac{1}{s}\int_{E_s} (\phi(x') - \phi(x))K_s(x, dx') \\ &= \frac{1}{s}\left[\int_{E_s} \phi'(x)(x'-x) + \phi''(x)\frac{(x'-x)^2}{2} + \left(\int_x^{x'} \phi^{(3)}(y)\frac{(y-x)^2}{2}dy\right)\right] K_s(x)(dx').\end{aligned}$$

Therefore,

$$\begin{aligned}\left(\frac{1}{s}\mathcal{L}_s - \mathcal{L}\right)\phi(x) = (b^s - b)(x)\phi'(x) + (a^s - a)(x)\phi''(x) \\ + \int_{E_s}\int_x^{x'} \phi^{(3)}(y)\frac{(y-x)^2}{2}dyK_s(x)(dx').\end{aligned}$$

Since ϕ has bounded derivatives, for any $\gamma > 0$,

$$\left|\left(\frac{1}{s}\mathcal{L}_s - \mathcal{L}\right)\phi\right| \leq |b^s - b| + |a^s - a| + \left(\frac{\gamma^3}{6} + \frac{4R^3}{3}\Delta^\gamma_s\right).$$

Therefore, if assumptions $(i) - (iii)$ are verified, then

$$\lim_{s\to 0}\sup_{x\in\mathcal{X}}\left|\left(\frac{1}{s}\mathcal{L}_s - \mathcal{L}\right)\phi(x)\right| = 0.$$

□

3.2.4 Invariant measure

We say a measure μ to be an invariant measure for a diffusion process associated to a semigroup $(P_t)_{t\geq 0}$ with infinitesimal generator $\mathcal{L}$ if, for any function ϕ in C_0 and any $t > 0$,

$$\int_E P_t \phi d\mu = \int_E \phi \mu.$$

Equivalently, for any $\phi \in D(\mathcal{L})$,

$$\int_E \mathcal{L}\phi d\mu = 0.$$

If this measure is finite, that is

$$\int_E d\mu < \infty,$$

then we assume it is a probability measure. An invariant measure is said to be reversible, if for any two functions $\phi, \phi' \in C_0$,

$$\int_E \phi P_t \phi' d\mu = \int_E \phi' P_t \phi d\mu,$$

or for two compactly supported smooth functions ϕ, ϕ',

$$\int_E \phi \mathcal{L} \phi' d\mu = \int_E \phi' \mathcal{L} \phi d\mu.$$

According to Section 1.11.3 [3], an absolutely continuous measure $\mu = h d\lambda$ is reversible for P_t if and only if

$$b_i = \sum_{j=1}^{n} \frac{\partial a_{i,j}}{\partial x_i} + a_{i,j} \frac{\partial \log h}{\partial x_i}.$$

Furthermore, since

$$\frac{1}{2}(\mathcal{L}(\phi\phi') - \phi\mathcal{L}\phi' - \phi'\mathcal{L}\phi) =< \nabla\phi, a\nabla\phi' >,$$

if μ is reversible then it satisfies the following integration by parts formula

$$-\int_E \phi \mathcal{L} \phi' d\mu = \int_E \frac{1}{2}(\mathcal{L}(\phi\phi') - \phi\mathcal{L}\phi' - \phi'\mathcal{L}\phi) = \int_E < \nabla\phi, a\nabla\phi' > d\mu.$$

When μ is reversible, it is after interesting to consider $(P_t)_{t\geq 0}$ and $\mathcal{L}$ as operators acting on the Hilbert space $L_2(\mu)$ in which case they are bounded and symmetric operators. This is going to be especially useful when studying the spectral properties of $\mathcal{L}$.

3.2.5 Spectrum of the infinitesimal generator

As we have seen, many graph analysis algorithms use the spectrum of the graph Laplacian, and thus of the generator of the random walk on the graph. The convergence of the outputs of these algorithms can thus be obtained through the convergence of the spectra of generators of random walks on graphs. Should these random walks converge to diffusion processes, one may expect the spectra of the corresponding generators to converge to the spectra of infinitesimal generators of the limiting diffusion processes, should it exists.

We say a bounded operator from a Banach space X to another Banach space Y is compact if the image of the unit ball of X is relatively compact in Y. Let $\mathcal{L}$ be the infinitesimal generator of a semigroup $(P_t)_{t\geq 0}$ acting on a Hilbert space $\mathcal{H}$.

Theorem 9 (Theorem A.6.4 [3]). *If $-\mathcal{L}$ is positive and there exists $t > 0$ such that P_t is a compact operator, then $\mathcal{L}$ has a discrete spectrum with eigenvalues $\cdots \leq \lambda_n \leq \cdots \leq \lambda_1 = 0$. Moreover, $\lambda_n \to_{n\to\infty} -\infty$.*

Let us note that, although we are only going to deal with diffusion processes in $\mathbb{R}^d$, this results holds even if $\mathcal{L}$ and $(P_t)_{t\geq 0}$ are defined as operators acting on functions defined on manifolds. In particular, if the manifold is compact, $(P_t)_{t\geq 0}$ is compact as well. Hence, results dealing with convergence of graph Laplacians for random geometric graphs are usually obtained in a compact manifold setting.

3.2.6 Hitting times

Given $F \subset E$, let

$$\tau_F = \inf_{t\geq 0}(X_t \in F).$$

For the simplest case of diffusion process, the Brownian motion $(W_t)_{t\geq 0}$, we have the following result.

Proposition 10 (Corollary 2.26 [60]). *If $d > 1$, then for any $T > 0$ and $x \in \mathbb{R}^d$, $\mathbb{P}(\exists t \in (0, T], W_t = x) = 0$.*

Thus, in dimension greater than one, a Brownian motion does not hit any fixed point in finite time. Hence, we can expect this is going to be the case for general diffusion processes. This is a major difference with respect to random walks on graphs which always hit any given vertex of the graph in finite time.

3.3 Random geometric graphs

Let M be a p-dimensional manifold embedded in $\mathbb{R}^d$ and let ν be a measure with support M and strictly positive smooth density f. Let $X_1, \dots, X_n$ be i.i.d. random variables drawn from ν and let $\mathcal{X}_n = \{X_1, \dots, X_n\} \in \mathbb{R}^d$. Consider a function $W_{\mathcal{X}_n} : \mathbb{R}^d \to \mathbb{R}^+$, a radius function $r_{\mathcal{X}_n} : \mathbb{R}^d \to \mathbb{R}^+$ and a bandwidth $h_n \in \mathbb{R}^+$. Let $K : \mathbb{R}^+ \to \mathbb{R}^+$ be a decreasing function, called kernel, such that

- $K > 0$;
- $\int_{\mathbb{R}^p} K(\|x\|)dx = 1$;
- $\int_{\mathbb{R}^p} \|x\|^2 K(\|x\|)dx < \infty$.

We call *random geometric graph* the graph $\mathcal{G}$ with vertices $\mathcal{X}_n$ and weight function

$$\forall x, x' \in \mathcal{X}, w(x, x') = W_{\mathcal{X}_n}(x') \frac{1}{h_n^d} K\left(\frac{\|x' - x\|}{h_n r_{\mathcal{X}_n}(x)}\right).$$

Random geometric graphs encompass many graphs used in data analysis such as the ϵ-graph, obtained by taking $r_{\mathcal{X}_n} = W_{\mathcal{X}_n} = 1$ and $h_n = \epsilon$. Taking $k \in \mathbb{N}$, $K = 1_{[0,1]}, W_{\mathcal{X}_n} = 1$ and

$$h_n r_{\mathcal{X}_n}(x) = \inf(r \in \mathbb{R}^+ | \sum_{y \in \mathcal{X}_n} 1_{\|y-x\| \leq r} \geq k),$$

we obtain a graph such that each point is linked to its k-nearest neighbors. Such a graph is called k-nearest neighbor graph and is a widely used type of graph as it is very sparse, however since this graph is directed, it is harder to analyze than ϵ-graphs.

3.3.1 Convergence of random walks on random geometric graphs

For simplicity, we will assume in this Section that points are not drawn from a manifold embedded in $\mathbb{R}^d$ but rather from the flat torus $T = (\mathbb{R}/\mathbb{Z})^d$ and we suppose random geometric graphs are built using the Riemaniann metric of T.

As h_n goes to zero and n goes to infinity, the length of the jumps of the random walks get smaller and smaller. Hence, if the first and second moments of the jumps also converge, the random walk itself should converge

to a diffusion process. Moreover, since the structure of the graph is going to be influenced by ν, it is reasonable to expect the limiting diffusion process to depend on ν as well. This dependency is made explicit by the following result.

Theorem 11 (Theorem 3 [81])**.** *Suppose $h_n \to 0$ and $\frac{nh_n^{d+2}}{\log n} \to 0$. Moreover, assume there exists deterministic smooth quantities $\tilde{W}, \tilde{r}$ such that*

$$\sup_{x\in T} \sup_{\epsilon\in\mathbb{R}^d, \|\epsilon\|\leq h(n)} |W_n(x+\epsilon) - \tilde{W}(x) - \nabla\tilde{W}.\epsilon| = o(h(n)^2);$$
$$\sup_{x\in T} \sup_{\epsilon\in\mathbb{R}^d, \|\epsilon\|\leq h(n)} |r_n(x+\epsilon) - \tilde{r}(x) - \nabla\tilde{r}.\epsilon| = o(h(n)^2).$$

Then, for any $T > 0$, the random walk on the random geometric graph converges weakly in $D([0,T])$ to a diffusion process with generator

$$\mathcal{L}\phi = \tilde{r}^2(\nabla \log(f\tilde{W}).\nabla + \frac{1}{2}\Delta).$$

The approximation timestep is $s = \frac{C}{h^{d+2}}$, where C is a constant depending on d and K.

For ϵ-graphs, the corresponding generator is simply $\mathcal{L} = \nabla \log f.\nabla + \frac{1}{2}\Delta$, so trajectories of $(X_t)_{t\geq 0}$ are attracted by high density regions. This result is readily obtained using Taylor's expansion along with concentration inequalities and Theorem 7. As an example, we prove this result for the case of the k-nearest neighbor graph via the following Lemma, which will prove useful in Chapter 5.

Lemma 12. *Let $k > 0$, and pose*

$$s = \left(\frac{k}{n}\right)^{2/d} \frac{\int_{\|x\|\leq 1} x_1^2 dx}{\left(\int_{\|x\|\leq 1} 1dx\right)^{1+2/d}};$$
$$\tilde{r} = f^{-1/d}.$$

For any $x, x' \in \mathcal{X}$, we pose

$$K^n(x,x') = \frac{w(x,x')}{\sum_{x''} w(x,x'')},$$

where w is the w function corresponding to a k-nearest neighbor graph. There exists a constant C such that, with probability $1 - \frac{C}{n}$,

(i) $\sup_{x\in\mathcal{X}_n} \|\frac{1}{s}\sum_{x'\in\mathcal{X}_n}(x'-x)K^n(x,x')-f(x)^{-2/d}\nabla f(x)\| \leq C(\frac{\sqrt{\log n}n^{1/d}}{k^{1/2+1/d}}+\left(\frac{k}{n}\right)^{2/d})$;

(ii) $\sup_{x\in\mathcal{X}_n} \|\frac{1}{s}\sum_{x'\in\mathcal{X}_n}(x'-x)^T(x'-x)K^n(x,x')-f(x)^{-2/d}I_d\| \leq C(\sqrt{\frac{logn}{k}}+\left(\frac{k}{n}\right)^{2/d})$;

(iii) $\sup_{x\in\mathcal{X}_n} \|\frac{1}{s}\sum_{x'\in\mathcal{X}_n}(x'-x)^{\otimes 3}K^n(x,x')\| \leq C(\frac{\sqrt{\log n}k^{1/d}}{n^{1/d}k^{1/2}}+\left(\frac{k}{n}\right)^{2/d})$;

(iv) $\exists n_0, \forall n > n_0, \forall x \in \mathcal{X}_n \sum_{x'\in\mathcal{X}_n, \|x-x'\|\geq C(\frac{k}{n})^{1/d}} K^n(x,x') = 0$.

Proof. Let $x \in T$. In the remainder of this proof, C denotes a generic constant depending only on d and f. For any $r > 0$, we denote by $\mathcal{B}(x,r)$ the ball centered in x with radius r. Let $P_r = \int_{\mathcal{B}(x,r)} \mu(dx)$ and, for $k > 0$, we pose $V_k = \int_{\mathcal{B}(0,1)} x_1^k dx$. Let N_r be the number of points in $\mathcal{B}(x,r)$. For any $0 < \epsilon < 1$, Chernoff's bound yields

$$P(|N_r - nP_r| \geq n\epsilon P_r) \leq 2e^{-\frac{\epsilon^2 nP_r}{3}}. \tag{3.2}$$

Take $r_M = \left(\frac{2k}{nV_0 \min_{y\in T} f(y)}\right)^{1/d}$, we have $P_{r_M} \geq \frac{2k}{n} + C\left(\frac{k}{n}\right)^{1+2/d}$. Hence, for $\frac{k}{n}$ sufficiently small, with probability greater than $1-\frac{1}{n^2}$, $N_{r_M} \geq k$. Thus the k-th nearest neighbor of x is at most at distance r_M of x. Applying a union-bound, this is true for any $x \in \mathcal{X}_n = \{X_1, \ldots, X_n\}$ with probability $1-\frac{1}{n}$ so we proved (iv).

Let us now prove (i), we have

$$\begin{aligned}
&\mathbb{E}[(X_i-x)1_{X_i\in\mathcal{B}(x,r)}] \\
&\quad= \int_{\mathcal{B}(x,r)} (y-x)\mu(dy) \\
&\quad= \int_{\mathcal{B}(x,r)} (y-x)f(y)dy \\
&\quad= \int_{\mathcal{B}(x,r)} (y-x)f(x) + (y-x)^{\otimes 2}\nabla f(x) + \frac{(y-x)^{\otimes 3}\nabla^2 f(x)}{2} + Cr^4 dy \\
&\quad= V_2 r^{d+2}\nabla f(x) + Cr^{d+4}.
\end{aligned}$$

Therefore, letting $b_1 = \sum_{X_i\in\mathcal{B}(x,r)} X_i - x$ and applying Bernstein's inequality,

$$P\left(\left|b_1 - nV_2 r^{d+2}\nabla f(x)\right| \geq C\left(r\sqrt{nP_r \log n} + nr^{d+4}\right)\right) \leq \frac{2}{n^2}.$$

Taking $r = \left(\frac{k}{nV_0 f(x)}\right)^{1/d}$, we have $|P_r - \frac{k}{n}| \leq C\left(\frac{k}{n}\right)^{1+2/d}$. Hence, by Equation 3.2, $|N_r - k| \leq C(\sqrt{k \log n} + \frac{k^{1+2/d}}{n^{2/d}})$ holds with probability $1 - \frac{1}{n^2}$. Thus, for $b_2 = \sum_{X_i \in \mathcal{B}(x,\tilde{r})} X_i - x$, we have

$$|b_1 - b_2| \leq C r_M \left(\sqrt{k \log n} + \frac{k^{1+2/d}}{n^{2/d}}\right).$$

Putting everything together, we have, with probability $1 - \frac{C}{n^2}$

$$\begin{aligned}
&\|\frac{1}{ks} \sum_{X_i \in \mathcal{B}(x,\tilde{r})} (X_i - x) - f^{-2/d} \nabla \log f\| \\
&\qquad = \|\frac{b_2}{ks} - f^{-2/d} \nabla \log f\| \\
&\qquad \leq \left\|\frac{b_1}{ks} - f^{-2/d} \nabla \log f\right\| + C\left(\frac{\sqrt{\log n} n^{2/d}}{k^{1/2+2/d}} + \frac{1}{k}\right) \\
&\qquad \leq C\left(\frac{\sqrt{\log n} n^{1/d}}{k^{1/2+1/d}} + \left(\frac{k}{n}\right)^{2/d}\right).
\end{aligned}$$

Using a union bound, we obtain the uniform convergence of $\frac{1}{s}\sum_{x' \in \mathcal{X}_n}(x' - x)K^n(x, x')$ over $\mathcal{X}_n$ with probability $1 - \frac{C}{n^2}$. Bounds (ii) and (iii) can be obtained in the same way. □

3.3.2 Hitting times

By Proposition 10, we know that, in dimension greater than one, the Brownian motion does not hit fixed points in finite time and we cannot expect diffusion processes to do otherwise. Hence, while random walks on random geometric graphs converge to diffusion processes, there are no continuous quantities we can expect hitting times and commute distances to converge to. Therefore, the limiting behaviour of these quantities might not be of interest. In [85], it is proved the mean hitting time between two vertices x and x' of a random geometric graph converges to a quantity depending on the local structure of the graph around x' alone. Hence, hitting times fail to describe the general structure of large random geometric graphs. Different ways to correct this issue for the commute distance have been proposed, see for example the amplified commute distance [85].

3.3.3 Invariant measure

Suppose a family of random walks converge to a diffusion process admitting a unique invariant probability measure μ. Can we expect the invariant

measures of the random walks to converge to μ?

Let us start with an ϵ-graph. Random walks on such random geometric graphs converge to diffusion processes with infinitesimal generator $\mathcal{L} = \nabla \log f.\nabla + \frac{\Delta}{2}$. This generator has an invariant probability measure μ with a density proportional to f^2. By Proposition 4, the invariant measure of a connected and undirected graph is proportional to its degree function. An ϵ-graph being undirected, the invariant measure π of a random walk on such a graph is

$$\forall x \in \mathcal{X}, \pi(x) = \frac{d(x)}{Vol(\mathcal{X})} = \frac{\sum_{i=1}^n K(\frac{\|X_i - x\|}{h_n})}{\sum_{i,j=1}^n K(\frac{\|X_i - X_j\|}{h_n})}.$$

Actually, π is proportional to the value of a standard kernel density estimator.

Proposition 13. *Suppose the density f is supported on the flat torus T and has bounded second derivative. Then there exists $C > 0$ such that, with probability $1 - ne^{\frac{n\epsilon^2}{2h_n^d}}$,*

$$\sup_{x \in \mathcal{X}} \left\| \sum_{i=1}^n \frac{1}{h_n^d} K\left(\frac{\|X_i - x\|}{h_n}\right) - f(x) \right\| \leq \epsilon + Ch_n^2.$$

Proof. Let $x \in T$ and let X be a random variable drawn from ν. Since T is compact and f is smooth, its second derivative is bounded. Thus, Taylor's expansion yields

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{h_n^d} K\left(\frac{\|X_i - x\|}{h_n}\right)\right] \\
&= \int_T \frac{1}{h_n^d} K\left(\frac{\|x' - x\|}{h_n}\right) f(x')dx' \\
&= \int_T \frac{1}{h_n^d} K\left(\frac{\|x' - x\|}{h_n}\right) (f(x) + (x' - x).\nabla f(x) + O(\|x' - x\|^2))dx'.
\end{aligned}$$

By our assumptions on the kernel K, we thus have

$$\mathbb{E}\left[K\left(\frac{\|X_i - x\|}{h_n}\right)\right] = f(x) + O(h_n^2).$$

Therefore, by Hoeffding's inequality, we have, with probability $1 - e^{\frac{n\epsilon^2}{2h_n^d}}$,

$$\left\| \sum_{i=1}^n K\left(\frac{\|X_i - x\|}{h_n}\right) - f(x) \right\| \leq \epsilon + O(h_n^2).$$

We then conclude the proof using the union bound inequality on all $x \in \mathcal{X}_n$. □

Hence, as long as $h_n \to 0$ and $\frac{n}{\log n} h_n^d \to \infty$, there exists $C > 0$ such that $\lim_{n\to\infty} \sup_{x\in\mathcal{X}_n} |\pi(x) - C\frac{f(x)}{n}| = 0$. Hence, for any open set A of E,

$$\pi(A) \approx C \sum_{x\in\mathcal{X}_n\cap A} \frac{f(x)}{n} \approx C \int_A f(x)d\mu \approx C \int_A f^2(x)dx.$$

Therefore, π converges to μ. While this gives an idea regarding why algorithms relying on invariant measures of random walks on such graphs actually work well in practice, it also guarantees these random geometric graphs do retain most of the information contained in the input data since it is possible to recover f, and thus ν, from the sole graph structure.

One may now wonder whether a similar result holds for directed random geometric graphs such as the k-nearest neighbor graph. By Theorem 11, the random walk on a k-nearest neighbor graph converges to a diffusion process with infinitesimal generator $\mathcal{L} = f^{-2/d}\left(\nabla \log f.\nabla + \frac{\Delta}{2}\right)$. This diffusion process admits an invariant measure μ with density proportional to $f^{1+2/d}$. In [44], the authors prove the weak convergence of the invariant measures of random walks on random geometric graphs. Applying this result to k-nearest neighbor graphs, we obtain that π converges weakly to μ as long as $\frac{k^{d+2}}{(n\log(n))^2} \to \infty$. Moreover, it is conjectured that, similarly to the undirected case, the convergence holds as long as $\frac{k}{\log n} \to \infty$ and is pointwise.

Chapter 4

Soft-clustering

In this Chapter, we propose a new soft clustering algorithm based on the mode seeking algorithm ToMATo [22] which relies on hitting times of random walks on random geometric graphs. As we have seen in Section 3.1.2, we cannot expect consistency of hitting times to fixed points. We solve this issue by using hitting times to a set rather than hitting times to a point, allowing us to prove the consistency of our algorithm. We finally provide some experimental results for our algorithm on both synthetic and real data. This work was done in collaboration with Steve Oudot. We thank Cecilia Clementi for providing the Alanine dipeptide dataset.

4.1 Mode-seeking

Let us assume the data points $\mathcal{X}_n = \{X_1, \dots, X_n\}$ are i.i.d. random variables drawn from a measure ν with smooth density f. We assume f is a Morse function, i.e. f is a smooth function with non-degenerate critical points such that all the critical values are distinct, with a finite number of critical points. For $x \in \mathbb{R}^d$, we define the gradient flow induced by f with initial state x

$$\begin{cases} dy^x(t) = \nabla f(u(t)) \\ y^x(0) = x \end{cases} . \tag{4.1}$$

Let $v_1, \dots, v_m$ be the set of local maxima of f, the sets $(M_i)_{i\in\{1,\dots,m\}}$ defined by

$$\forall i \in \{1, \dots, m\}, M_i = \{x \in \mathbb{R}^d \mid \lim_{t\to\infty} y^x = v_i\},$$

are called modes of f. These modes define a partition of $\mathbb{R}^d \backslash U$, where U is a set of measure zero containing points whose gradient flow lead to

saddle points. The mode-seeking approach consists in associating clusters to modes of f.

Of course, since we do not have access to the density f itself, we do not have access to its modes either. A simple way to turn this framework into a clustering algorithm would consist in computing a density estimator $\hat{f}$ and define clusters as the modes of $\hat{f}$. However, modes of $\hat{f}$ do not necessarily correspond to modes of f. Mode seeking algorithms such as [24, 22, 27, 28, 31, 48] aim at recovering the modes of f from $\hat{f}$ to perform clustering. From a statistical point of view, the analysis of mode inference was developed recently in [23, 2, 25, 26]. Let us present one of these mode seeking algorithms: the Topological Mode Analysis Tool (ToMATo) algorithm [22].

4.1.1 ToMATo

The ToMATo algorithm relies on the concept of 0-dimensional persistent homology, which we present below, in order to estimate the modes of f.

4.1.2 0-dimensional persistent homology of the superlevel sets of a function

0-dimensional persistent homology was introduced in the context of Topological Data Analysis under the name size theory [82] and was later generalized by the persistent homology theory (see [35, 93] for more details). Here we are interested in the 0-dimensional persistent homology of the superlevel sets of a function, also called prominence.

Let G be an unweighted graph with vertices V and edges E, and suppose there exists an ordering on V. Consider a function $\phi : V \to \mathbb{R}$. For any $\alpha \in \mathbb{R}$, we define the superlevel sets of ϕ by

$$V_\alpha = \{x \in V \mid \phi(x) \geq \alpha\}.$$

For any $x \in G$, we denote by V_α^x the connected component containing x in the subgraph G_α with vertices V_α and edges E_α such that

$$\forall x, x' \in V_\alpha, (x, x') \in E_\alpha \iff (x, x') \in E.$$

Let $b_\phi(x)$ be the highest α such that $x \in V_\alpha$ (i.e. $\phi(x)$). For $\alpha < b_\phi(x)$, we use the ordering on V to define

$$D_{\alpha,x} := \max\{x' \in V_\alpha^x \mid \phi(x') \geq \phi(x)\}.$$

We define $d_\phi(x)$ as the highest α such that $D_{\alpha,x} \neq x$ and we pose $D_{\phi,x} = D_{d_\phi(x),x}$. The prominence $p_\phi(x)$ of x with respect to ϕ is then defined as $b_\phi(x) - d_\phi(x)$. If a vertex x is not a local maximum of ϕ on the graph, i.e.

$$\exists x' \in V, \phi(x) > \phi(x),$$

then it has a prominence equal to zero. Replacing G by $\mathbb{R}^d$, it is possible to define a similar notion of prominence in the continuous domain.

The prominence information is usually encoded as a collection of points in the plane with coordinate $(b_\phi(x), d_\phi(x))$ called persistence diagram. These diagrams are endowed with a natural metric called the bottleneck distance involving the notion of partial matching. A partial matching M between two diagrams Δ_1 and Δ_2 is a subset of $\Delta_1 \times \Delta_2$ such that each point of Δ_1 and Δ_2 appears at most once in M. The bottleneck cost $C(M)$ of a partial matching M between two diagrams Δ_1 and Δ_2 is the infimum of $\delta \geq 0$ such that

- For any $(p_1, p_2) \in M$, $||p_1 - p_2||_\infty \leq \delta$;
- For any other point (b, d) of Δ_1 or Δ_2, $b - d \leq 2\delta$.

The bottleneck distance between two diagrams D_1 and D_2, is then defined as

$$d_B(\Delta_1, \Delta_2) = \inf_\delta \{\delta \mid \exists M, C(M) \leq \delta\}$$

Intuitively, the bottleneck distance can be seen as the cost of a minimum perfect matching between persistence diagrams with possibility to match points to the diagonal $y = x$. A remarkable property of persistence diagrams is their stability, proved in [30] and [21]. Let ϕ and ϕ' be two functions defined on the vertices of the same graph and $\Delta_\phi, \Delta'_\phi$ be the persistence diagrams of the superlevel-sets of ϕ and ϕ', applying these stability results, we obtain the following

$$d_B(\Delta_\phi, \Delta'_\phi) \leq ||\phi - \phi'||_\infty = \sup_{x \in V} |\phi(x) - \phi'(x)| \tag{4.2}$$

Computation As 0-dimensional persistence encodes the evolution of the connectivity of the superlevel-sets of a function, it can be computed using a simple variant of a Union-find algorithm. In practice, we use Algorithm 1 described in [22], with parameter τ set to infinity. This algorithm has close to linear complexity in the number of vertices of the meshes; more precisely it has complexity $O(|V| \log(|V|) + |V|\alpha(|V|))$ where α is the inverse of the Ackermann function.

4.1.3 The Algorithm

The algorithm starts by computing an unweighted graph with vertices $\mathcal{X}_n$, while it is not mandatory, the graph is usually either an ϵ-graph with kernel $K = 1_{[0,1]}$ or a k-nearest neighbors graph. We then compute the values of a density estimator $\hat{f}$ on $\mathcal{X}_n$. In order to pinpoint local maxima of $\hat{f}$ which correspond to local maxima of f, we use the prominence of the vertices of the graph with respect to the function $\hat{f}$. In this case, the higher the prominence of a local maximum, the more likely it is to correspond to an actual maximum of the underlying density f and thus to indicate a mode of f. Given a prominence threshold $\tau > 0$, we pose

$$\Delta_\tau(x) = \begin{cases} x \text{ if } p_{\hat{f}}(x) > \tau \text{ or } x = D_{\hat{f},x} \\ \Delta_\tau(D_{\hat{f},x}) \text{ otherwise.} \end{cases}$$

If $v_1, \ldots, v_K$ denote the vertices of G with prominences higher than τ, the algorithm outputs K clusters $C_1, \ldots, C_K$ where

$$\forall i \in \{1, \ldots, K\}, C_i = \{x \in \mathcal{X}_n \mid \Delta_\tau(x) = v_i\}.$$

Moreover, if $\hat{f}(v_i) \leq \tau$, then the points belonging to C_i are considered as outliers. Upon a proper choice of τ with respect to n and as long as the density estimator converges, this clustering procedure can be shown to be consistent [22].

4.2 Soft clustering

A natural way to turn the mode seeking approach into a soft-clustering algorithm is to add some randomness in Equation 4.1. Here, we use the following stochastic differential equation

$$dY_t = \frac{1}{\beta}\nabla(\log f)dt + dW_t,$$

where W_t is a d-dimensional Brownian motion and β is a strictly positive temperature parameter controlling the amount of noise introduced in the gradient flow. If it exists, the solution to this stochastic differential equation is a diffusion process with generator

$$\mathcal{L} = \frac{1}{\beta}\nabla(\log f).\nabla + \frac{1}{2}\Delta.$$

Fortunately, by Theorem 11, this diffusion process can be approximated by a random walk on the (properly weighted) random geometric graph used

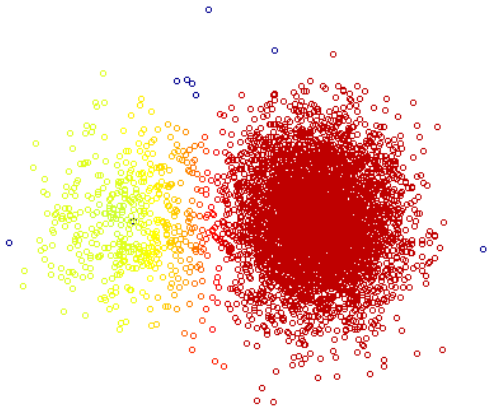


Figure 4.1: Soft-clustering output for an unbalanced mixture of Gaussian measures. Red colors corresponds to the right cluster, blue to the left one, hence green points have similar membership to both clusters.

by ToMATo. From here, one possibility would be to consider the first local maximum of the density encountered by the random walk. When $\beta = 1$, such an approach can be compared to the soft clustering procedure proposed in [24]. In this work, the authors use a statistical method to select relevant local maxima of $\hat{f}$ in $\mathbb{R}^d$ denoted $m_1, \dots, m_K$. Then, they build an ϵ-graph with vertices $\mathcal{X}_n \cup \{m_1, \dots, m_K\}$ and consider the first local maximum m_i hit by the random walk on the graph. However, as emphasized in Section 3.3.2, we cannot expect consistency regarding the amount of time required to hit a given point in the random geometric graph. Thus, if we apply this method to a dataset consisting in two unbalanced Gaussian measures in Figure 4.1, the obtained soft memberships are definitely wrong. In order to circumvent this issue, we assign a zone of high density to each cluster, called cluster core and computed using ToMATo, and we look at the first cluster core encountered by the random walk.

4.3 Our Algorithm

The input of the algorithm is a finite set of points $\mathcal{X}_n = \{X_1, \cdots, X_n\}$, a radius $\epsilon > 0$ –or an integer k if one wants to use a nearest neighbors graph– , a density estimator $\hat{f}$, a prominence threshold $\tau > 0$ and a temperature parameter $\beta > 0$.

Our algorithm then proceeds as follows. It first computes an ϵ-graph (or a k-nearest neighbors graph) G from $\mathcal{X}_n$ then runs the ToMATo algorithm using G. The output of ToMATo is a set of K clusters $C_1, \cdots, C_K$ each corresponding to a local maximum of $\hat{f}$ on the graph denoted by $x_1, \dots x_K$. We define the i-th cluster core as the connected component containing X_i within the subgraph of G spanned by those vertices x' such that $\hat{f}(x') > \hat{f}(x_i) - \tau/2$, in other words, the i-the cluster core is equal to $F^{x_i}_{\hat{f}(x_i)-\tau/2}$. We then compute a random geometric graph $\tilde{G}$ with the same parameters used to compute G except for

$$W_{\mathcal{X}_n}(x) = \hat{f}(x)^{1/\beta - 1}. \tag{4.3}$$

Finally, we compute the soft-membership values $\mu_1, \dots, \mu_K$ by solving the linear system $A^T\mu = \mu$, where the matrix A is defined by:

$$A_{kl} = \begin{cases} \delta_{kl} & \text{if } \exists i, X_k \in \mathcal{C}_{i,n} \\ K_{\tilde{G}}(X_k, X_l) & \text{otherwise,} \end{cases}$$

where $K_{\tilde{G}}$ is the transition kernel of the random walk on the graph $\tilde{G}$. The output of the algorithm is the set of those soft-membership values.

4.3.1 Parameters selection

Density estimator, window size, kernel and prominence threshold

These four parameters are tied to the classical mode-seeking framework. The density estimator can be linked to the window size in practice, for instance by using a density kernel estimator, as is done e.g. in Mean-Shift [27] and its successors. This not only reduces the number of parameters to tune in practice, but it also gives a way to select the window size ϵ –or k– using standard parameter selection techniques for density estimation, which is done for example in [24]. Finally, the prominence threshold τ can be selected by running ToMATo with prominence threshold equal to 0 in order to obtain the distribution of prominences of the vertices within the neighborhood graph G. An adequate value of τ can then be inferred by looking for a gap in the distribution. This procedure is detailed in [22].

Temperature parameter

As for temperature or fuzziness parameters in other soft clustering algorithms, it is not clear how β should be selected. Outputs corresponding

to large values of β will tend to have smooth interfaces between clusters while small values of β will encourage quick transitions from one cluster to another. β can also be interpreted as a trade-off between the respective influences of the Euclidean metric and of the density: when β is small, the output of our algorithm is mostly guided by the density and therefore close to the output of the hard mode seeking algorithm; by contrast, when β is large, the definition of the cluster cores is the only influence of the density on the output of the algorithm. In practice, one may get insights into the choice of β by looking at the evolution of a certain measure of fuzziness for the clustering output across a range of values of β. We elaborate on this in Section 4.4.

4.3.2 Convergence guarantees

In this Section we provide theoretical guarantees for our soft-clustering scheme by exploiting the convergence of random walks on random geometric graphs to diffusion processes.

As usual in mode-seeking, we assume our input data points $\mathcal{X}_n = \{X_1, ..., X_n\}$ to be i.i.d. random variables drawn from some unknown probability density f over $\mathbb{R}^d$. We also assume that f and $\nabla \log f$ are smooth and Lipschitz continuous over $\mathbb{R}^d$. This condition is sufficient to ensure the existence of a diffusion process $(X_t)_{t\geq 0}$, with $X_0 = x \in \mathbb{R}^d$, associated to the infinitesimal generator

$$\mathcal{L} = \frac{1}{\beta}\nabla \log f.\nabla + \frac{1}{2}\Delta.$$

Let us assume the unweighted graph G used by our algorithm is an ϵ-graph. For $x \in \mathcal{X}_n$, let $M^{x,\epsilon}$ be the random walk on the weighted graph $\tilde{G}$ whose initial state is the closest neighbors of x in $\mathcal{X}_n$ (break ties arbitrarily). The weighting we use to create $\tilde{G}$ trajectories was designed such that $M^{x,\epsilon}$ converges weakly to $(X_t)_{t\geq 0}$. From here, convergence of the cluster cores to continuous sets is sufficient to prove the consistency of the algorithm. Formally, letting $v_1, \ldots, v_K$ be the local maxima of f with prominence higher than τ, we can define the limit cluster cores as

$$\forall i \in \{1, \ldots, K\}, \tilde{\mathcal{C}}_i = F^{v_i}_{\tau/2},$$

and we define $\tilde{\mu}_i(x)$ as the probability for the diffusion process with generator $\mathcal{L}$ started at x to hit $\tilde{\mathcal{C}}_i$ before any other $\tilde{\mathcal{C}}_j$.

In order to obtain a consistency result for our algorithm, we require these continuous cluster cores to satisfy some assumptions. For any $C \subset \mathbb{R}^d$ and

any $\delta > 0$, let us pose

$$C^\delta = \{x \in \mathbb{R}^d \mid \exists y \in C, \|x - y\| \leq \delta\}.$$

Assumption 14. *The boundary of the $\tilde{\mathcal{C}}_i$ are smooth. Moreover, for any $\delta > 0$ there exist $\delta' > 0$ such that*

$$F^{v_i}_{\tau/2+\delta} \subset (F^{v_i}_{\tau/2})^{\delta'} = \tilde{\mathcal{C}}_i^{\delta'},$$

and

$$F^{v_i}_{\tau/2} = \tilde{\mathcal{C}}_i \subset (F^{v_i}_{\tau/2-\delta})^{\delta'}.$$

Theorem 15. *Let $\beta > 0, \tau > 0$ and suppose the cluster cores verify Assumption 14. Let $\epsilon : \mathbb{N} \to \mathbb{R}^+$ be a decreasing window size such that $\lim_{n\to\infty} \epsilon(n) = 0$ while $\lim_{n\to\infty} \frac{\epsilon(n)^{d+2} n}{\log n} = \infty$. Suppose the density estimator $\hat{f}_n$ satisfies, for any compact set $U \subset \mathbb{R}^d$ and any $\eta > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\sup_{x\in U} \sup_{h\in\mathbb{R}^d, \|h\|\leq\epsilon(n)} |\hat{f}_n^{1/\beta-1}(x+h) - (f^{1/\beta-1}(x) + \nabla f^{1/\beta-1}.h)| \geq \epsilon(n)^2\eta) = 0.$$

Almost surely, there exists n_0 such that, for $n \geq n_0$, the number of clusters output by ToMATo is equal to K. Let $(\mu_{n,i})_{n\geq n_0, i\in\{1,\dots,K\}}$ be the output of our algorithm. There exists a family of permutation $(\pi_n)_{n\geq n_0}$ such that, for any compact set $U \subset \mathbb{R}^d$, any $\eta > 0$ and any $i \in \{1, \dots, K\}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in U} |\tilde{\mu}_i(x) - \mu_{n,\pi_n(i)}(x)| \geq \eta\right) = 0.$$

When the input graph is a k-nearest neighbors graph, the trajectories of the random walk converge to a diffusion process $(X_t)'_{t\geq 0}$ with generator

$$\mathcal{L}' = f^{-2/d}(\frac{1}{\beta}\nabla \log f.\nabla + \frac{1}{2}\Delta).$$

Since trajectories of $(X'_t)_{t\geq 0}$ correspond, up to a time-change, to trajectories of $(X_t)_{t\geq 0}$, a similar result can be obtained.

4.4 Experiments

We first illustrate the effect of the temperature parameter β on the clustering output using synthetic data. We then apply our method on three UCI repository datasets and on simulated protein conformations data. In all our experiments we use a k-nearest neighbors graph along with a distance to measure density estimator [11] computed using the k-nearest neighbors.

4.4.1 Synthetic data

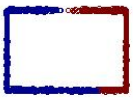
The first dataset is presented in Figure 4.2a and is composed of two high-density clusters connected by two links. The bottom link is sampled from a uniform density while the top link is sampled from a density that has a gap in-between the two clusters. Standard mode seeking algorithms will have a hard time clustering the bottom link as the separation between the two modes of f is extremely smooth. Thus, ToMATo missclusters most of the bottom link (see Figure 4.2b). We display the results of our algorithm for three values of $\beta : \beta = 0.2$ in Figure 4.2c, $\beta = 1$ in Figure 4.2d and $\beta = 2$ in Figure 4.2e. As we can see from the output of the algorithm, for small values of β, the temperature parameter is not large enough to compensate for the influence of the noise in the density estimation: the result obtained is really close to hard clustering. Large values of β do not give enough weight to the density function which leads to a smooth transition between the two clusters on the top link despite the density gap. Intermediate values of β seem to give more satisfying results. In order to gain intuition regarding which value of β one should use, it is possible to look at the evolution of a fuzziness value for the clustering. For example, one can consider a notion of clustering entropy:

$$H = \sum_i \sum_j \mu_j(X_i) \log(\mu_j(X_i)), \tag{4.4}$$

which gets lower when the fuzziness of the clustering increases. As we can see in Figure 4.2f, the evolution of H with respect to β presents three distinct plateaus corresponding to the three behaviours highlighted earlier.

The second dataset we consider is composed of two interleaved spirals—see Figure 4.3. An interesting property of this dataset is that the head of each spiral is close –in Euclidean distance– to the tail of the other spiral. Thus, the two clusters are well-separated by a density gap but are close in terms of Euclidean metric. We use our algorithm with two different values of β: $\beta = 1$ and $\beta = 0.3$, we also run a spectral fuzzy C-means algorithm on a subsampling of this dataset. The first thing we want to emphasize is that the output result of spectral C-means and our algorithm using $\beta = 1$ are similar, this is to be expected as both algorithms rely on the properties of the same diffusion operator. On the other hand, giving more weight to the structure of the density by setting $\beta \simeq 0.3$, we correctly recover the two clusters.

(a) The data.

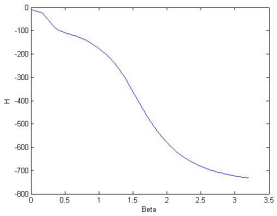


(b) Output of ToMATo.

(c) Output for $\beta = 0.2$.

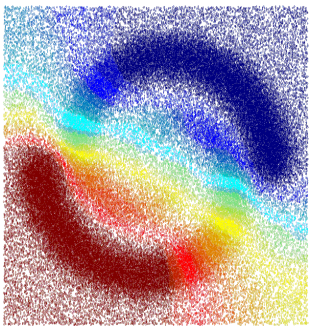
(d) Output for $\beta = 1$.

(e) Output for $\beta = 2$.

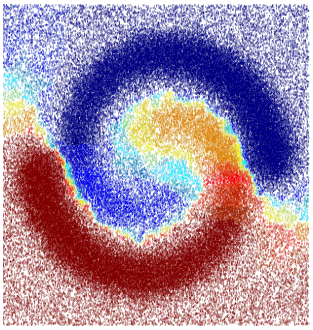


(f) Evolution of H with respect to β.

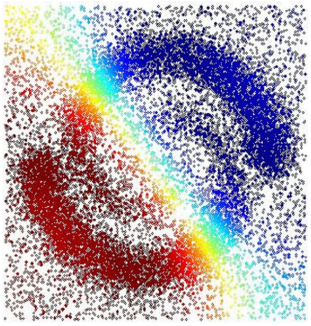
Figure 4.2: Output of our algorithm on a simple dataset composed of two overlapping clusters. For soft clustering, green corresponds to an equal membership to both clusters.



(a) Output for $\beta = 1$.



(b) Output for $\beta = 0.28$.



(c) soft Spectral Clustering.

Figure 4.3: Experiments on a cluttered spirals dataset.

4.4.2 UCI datasets

In order to perform a quantitative evaluation of our soft clustering scheme, we evaluate it in a classification scenario on a few datasets from the UCI repository: the Pendigits dataset (10,000 points and 10 classes), the Waveset dataset (5,000 points and 3 classes) and the Landsite Satellite dataset (6,435 points and 7 classes). We preprocess each dataset by renormalizing the various coordinates so they have unit variance. Then, for each dataset, we run our algorithm with various values of the parameter β between 0.1 and 5, but a single value of k. We select the prominence threshold τ using a prominence gap. Since there are two possible prominence gaps for the Pendigits and the Landsite Satellite datasets, we run the experiments twice using both thresholds and indicate the corresponding number of clusters K in our results. As a baseline, we use the fuzzy C-means algorithm with fuzziness parameters between 1.2 and 5 using the same number of clusters as detected by the ToMATo algorithm. We also consider the soft clustering algorithm proposed by [24], for which the cluster cores are reduced to a single point and $\beta = 1$. Let $X_1, \dots, X_n$ denote our sample points and $Y_1, \dots, Y_n$ their respective labels taking values in $\{1, \dots, K'\}$. We propose an automatic selection of β by computing the values of the clustering entropy H for multiple values of β and by selecting

$$\beta = \arg\max \frac{dH}{d\beta},$$

in other words we take β maximizing the slope of H. In order to evaluate hard clustering algorithms, it is common to use the purity measure defined by

$$P = \max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} 1_{\mu_j(X_i)=1} 1_{Y_i=\pi(j)},$$

where π is a map from the set of clusters $\{1, \dots, K\}$ to the set of labels $\{1, \dots, K'\}$. As this measure is not adapted to soft clustering, we consider instead a quantity we call ϵ-*entropic purity* and defined by

$$HP_\epsilon = \max_{\pi} \frac{1}{n} \sum_{i} \log \left(\epsilon + \sum_{j,\pi(j)=Y_i} \tilde{\mu}_j(X_i) \right),$$

for some $\epsilon > 0$. The ϵ parameter is used to prevent the quantity from exploding due to possible outliers. This quantity can be viewed as an approximation of $\mathbb{E}[\log(\epsilon + \sum_{j,\pi(j)=Y} \mu_j(X))]$, for X a random variable drawn from a

Algorithm/Data(K)	Wave(3)	Pend(9)	Pend(13)	Sat(5)	Sat(9)
Ours, optimal β	-0.57	-0.40	-0.37	-0.42	-0.27
Ours, automatic β	-0.59	-0.40	-0.38	-0.42	-0.28
fuzzy C-means	-1.1	-0.84	-0.58	-0.64	-0.35
[24] algorithm	-0.73	-0.47	-0.45	-0.44	-0.32

Table 4.1: Entropic purity obtained by soft clustering algorithms on UCI datasets.

measure with density f. For any random variables $X \in \mathbb{R}^d, Y \in \{1, \ldots, K\}$ and any $\epsilon > 0$, we have

$$\underset{f \in \mathbb{R}^d \to \mathbb{R}^K, \|f\|_1 = 1}{\arg\max} \mathbb{E}[\log(\epsilon + f(X))] = (1-\epsilon)^{-1}(\mathbb{P}(Y = j \mid X))_{1 \leq j \leq K} - \epsilon.$$

Thus, for small values of ϵ, a soft clustering minimizing the ϵ-entropic purity recovers the conditional probabilities of the labels of the data points with respect to their coordinates. Hence, this extension of the traditional purity can be useful to evaluate soft clustering

We provide the best 0.1-entropic purity obtained by each algorithm on all datasets in Table 4.1.

Alanine-dipeptide conformations. We now turn to the problem of clustering protein conformations. We consider the case of the alanine-dipeptide molecule. Our dataset is composed of $1,420,738$ protein conformations, each one represented as a 30-dimensional vector. The metric used on this type of data is the root-mean-squared deviation (RMSD). The goal of soft clustering in this case is twofold: first, to find the right number of clusters corresponding to metastable states of the molecule; second, to find the conformations lying at the border between different clusters, as these represent the transition phases between metastable states. It is well-known that conformations of alanine-dipeptide only have two relevant degrees of freedom, so it is possible to project the data down to two dimensions (called a Ramachadran plot) to have a comfortable view of the clustering output. In order to highlight interfaces between clusters, we only display the second highest membership function. As we can see in Figure 4.4 there are 5 clusters and 6 to 7 interfaces.

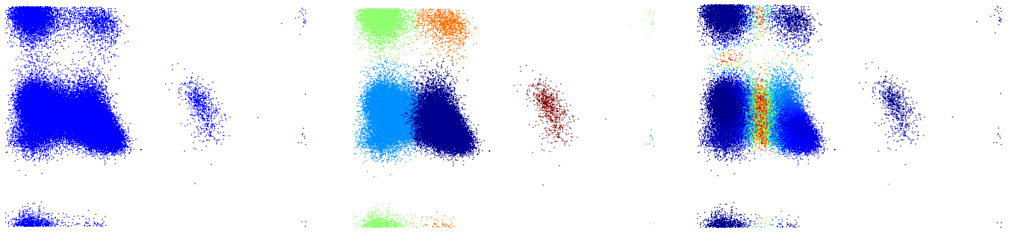


Figure 4.4: From left to right: (a) the dataset projected on the Ramachadran plot, (b) ToMATo output, (c) second highest membership obtained with our algorithm for $\beta = 0.2$

4.5 Proofs

4.5.1 Weak-Convergence

Let $x \in \mathbb{R}^d$, we denote by $(Y_t^x)_{t\geq 0}$ the diffusion process with infinitesimal generator

$$\mathcal{L} = \frac{1}{\beta}\nabla \log f.\nabla + \frac{1}{2}\Delta$$

and initial state $Y_0^x = x$. We start by proving the following result.

Proposition 16. *Let $\epsilon : \mathbb{N} \to \mathbb{R}^+$ be a decreasing function such that $\lim_{n\to\infty} \epsilon(n) = 0$ and $\lim_{n\to\infty} \frac{\epsilon(n)^{d+2}n}{\log n} = \infty$. Suppose our estimator $\hat{f}_n$ satisfies, for any compact set $C \subset \mathbb{R}^d$ and any $\eta > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\sup_{x\in U} \sup_{h\in\mathbb{R}^d, \|h\|\leq\epsilon(n)} |\hat{f}_n^{1/\beta-1}(x+h) - (f^{1/\beta-1}(x) + \nabla f^{1/\beta-1}.h)| \geq \epsilon(n)^2\eta) = 0.$$

Then, for any $T, \eta > 0$, for any compact set $U \subset \mathbb{R}^d$, and for any Borel set B of $D([0,T], \mathbb{R}^d)$ such that $\mathbb{P}(Y^y \in \partial B) = 0$ for all $y \in U$, there exists $s(n) > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}(\sup_{x\in U} |\mathbb{P}(M^{x,n}_{s(n)\lfloor t/s(n)\rfloor} \in B) - \mathbb{P}(Y_t^x \in B)| \geq \eta) = 0.$$

The proof relies on Theorem 11 along with a proper control of boundary effects.

Let $\mathcal{X}_n = (X_1, ..., X_n)$ and let $F_\alpha = \{x \in \mathbb{R}^d \mid f(x) \geq \alpha\}$ be the α superlevel-set of f. Throughout the course of the proof, the notation $M^{x,\epsilon}$ stands for the continuous time process $M^{x,\epsilon}_{s\lfloor t/s\rfloor}$. Let T and η be strictly positive reals.

For $\alpha > 0$, F_α is closed as f is continuous. Moreover, since f is Lipschitz continuous, $\lim_{\|x\|_2\to\infty} f(x) = 0$. Hence, F_α is also bounded and thus

compact. Following the proof of Theorem 11, there exists $s(n) > 0$ such that

(i) $\lim_{n\to\infty} \mathbb{P}(\sup_{y\in\mathcal{X}_n\cap F_\alpha} |b^s - \frac{\nabla f}{\beta f}| \leq \nu) = 0$,

(ii) $\lim_{n\to\infty} \mathbb{P}(\sup_{y\in\mathcal{X}_n\cap F_\alpha} |a^s - \frac{I_d}{2}| \leq \nu) = 0$,

(iii) $\sup_{y\in\mathcal{X}_n} \Delta_s^\epsilon = 0$,

(iv) $\lim_{n\to\infty} \mathbb{P}(\|M_0^{x,\epsilon} - x\| \leq \nu) = 0$.

Thus, the assumptions (i)-(iv) of Theorem 7 are verified on F_α which implies that $M^{x,n}$ approximates correctly Y^x as long as it does not leave F_α. More precisely, for $\alpha > 0$, let $B_\alpha = \{w \in D([0,T],\mathbb{R}^d) \mid \forall t, w(t) \in F_\alpha\}$. Since $\nabla \log f$ is Lipschitz continuous, there exists $\alpha > 0$ such that, for any $x \in U$, $\mathbb{P}(Y^x \in B_\alpha) \geq 1 - \eta/4$ (i.e. $(Y_t^x)_{t\geq 0}$ does not explode in finite time [34]). Since f is continuous, B_α is an open set. Therefore, there exists $n_0 > 0$ such that, for any $n > n_0$, using Theorem 7 along with Theorem 6,

$$\sup_{x\in U} \mathbb{P}(M^{x,n} \in B_\alpha) \geq \sup_{x\in U} \mathbb{P}(Y^x \in B_\alpha) - \eta/4 \geq 1 - \eta/2.$$

Therefore, for any Borel set B,

$$\sup_{x\in U} |\mathbb{P}(M^{x,n} \in B) - \mathbb{P}(M^{x,n} \in B \cap B_\alpha)| \leq \eta/2.$$

Thus, we only need to approximate trajectories that do not leave F_α to obtain a good approximation of $\mathbb{P}(M^{x,n} \in B)$. Applying Theorem 7 and Theorem 6 on these trajectories, we obtain

$$\sup_{x\in U} |\mathbb{P}(M^{x,n} \in B) - \mathbb{P}(Y^x \in B)| \leq \eta.$$

Every step of the proof hold with high probability as n tends to infinity, so the proof of Proposition 16 is complete.

4.5.2 Proof of Theorem 15

Since f is $\mathcal{C}^1$-continuous, the $\tilde{\mathcal{C}}_i$ are compact sets of $\mathbb{R}^d$ and are disjoint sets. Let β and τ be strictly positive real numbers, $U \subset \mathbb{R}^d$ be a compact set and $i \in \{1,\ldots,K\}$. Let $\mathcal{B}(x,r)$ denotes the ball of radius $r > 0$ centered on $x \in \mathbb{R}^d$. For $\delta > 0$, we pose $\tilde{\mathcal{C}}_i^\delta = \cup_{x\in\tilde{\mathcal{C}}_i} \mathcal{B}(x,\delta)$ and $\tilde{\mathcal{C}}_i^{-\delta} = \tilde{\mathcal{C}}_i \setminus \cup_{x\notin\tilde{\mathcal{C}}_i} \mathcal{B}(x,\delta)$. Let η be a strictly positive real and, for $x \in \mathbb{R}^d$,

- $\mu^+_{i,\delta}(x)$ be the probability that Y^x hits $\tilde{\mathcal{C}}^{\delta}_i$ before any other $\tilde{\mathcal{C}}^{-\delta}_j$;
- $\mu^-_{i,\delta}(x)$ be the probability that Y^x hits $\tilde{\mathcal{C}}^{-\delta}_i$ before any other $\tilde{\mathcal{C}}^{\delta}_j$.

Let us show that, for any i, a trajectory entering $\tilde{\mathcal{C}}^{\delta}_i$ has a high probability to enter $\tilde{\mathcal{C}}_i$ if δ is small enough. Since the $\tilde{\mathcal{C}}_i$ are closed and disjoint there exists $\delta_0 > 0$ such that the $\tilde{\mathcal{C}}^{\delta_0}_i$ are disjoints. Moreover, since the $\tilde{\mathcal{C}}_i$ have smooth boundaries, there exists $\delta^+_i > 0$ such that if $d(x, \tilde{\mathcal{C}}_i) \leq \delta^+_i$ then, the probability for Y^x to hit $\tilde{\mathcal{C}}_i$ before exiting $\tilde{\mathcal{C}}^{\delta_0}_i$ is at least $1-\eta/8$. Similarly, if a trajectory of Y^x enters $\tilde{\mathcal{C}}_i$, then it enters $\tilde{\mathcal{C}}^{-\delta}_i$ with high probability. More precisely there exists δ^-_i such that if a trajectory of Y^x hits $\tilde{\mathcal{C}}_i$, then it hits $\tilde{\mathcal{C}}^{-\delta}_i$ with probability at least $1 - \eta/8$.

Let $\delta = \min(\delta^+_j, \delta^-_j)$, since Y^x satisfies the strong Markov property, we have

- $\mu^+_{i,\delta}(x) - \mu_i(x) \leq \eta/4$,
- $\mu_i(x) - \mu^-_{i,\delta}(x) \leq \eta/4$.

The next step is to show that the approximation of $\mu^+_{i,\delta}$ provided by the Markov chain is correct. For $T > 0$, let

$$B = \{w \in D([0,\infty], \mathbb{R}^d) \mid \exists \tau \text{ such that } w(\tau) \in \tilde{\mathcal{C}}^{\delta}_i \\ \text{and } \forall t < \tau, w(t) \in \mathbb{R}^d \setminus \cup_j \tilde{\mathcal{C}}^{-\delta}_j\},$$

$$B_T = \{w \in D([0,T], \mathbb{R}^d) \mid \exists \tau \text{ such that } w(\tau) \in \tilde{\mathcal{C}}^{\delta}_i \\ \text{and } \forall t < \tau, w(t) \in \mathbb{R}^d \setminus \cup_j \tilde{\mathcal{C}}^{-\delta}_j\}.$$

We define the stopping time

$$\tau(Y) = \inf_t Y \in \tilde{\mathcal{C}}^{\delta}_i \cup_{j\in\{1,\dots,K\}, j\neq i} \tilde{\mathcal{C}}^{-\delta}_j.$$

Since $\mathcal{C}_i \subset \mathbb{R}^d$ and $\mathbb{R}^d$ has a single connected component, we have that $\mathbb{P}(\tau(Y^x_t) < \infty) = 1$, in particular that means that there exists T_0 such that, for any $T \geq T_0$, $\mathbb{P}(\tau(Y^x) \leq T) \geq 1 - \eta/6$. Using Proposition 16, we have that, with high probability with respect to $\mathcal{X}_n$,

$$\mathbb{P}(\tau(M^{x,n}) \leq T) \geq \mathbb{P}(\tau(Y^x) \leq T) - \eta/6 \geq 1 - \frac{1}{3}\eta.$$

Hence, we have

$$\mathbb{P}(M^{x,n} \in B \setminus B_T) + \mathbb{P}(Y^x \in B \setminus B_T) \\ \leq \mathbb{P}(\tau(M^{x,n}) > T) + \mathbb{P}(\tau(Y^x) > T) \leq \eta/2.$$

Since $\mathbb{P}(Y^x \in \partial B_T) = \mathbb{P}(Y^x \in \partial B) = 0$, we can apply Proposition 16 on the set B_T, and obtain

$$\sup_{x \in U} |\mathbb{P}(M^{x,n} \in B_T) - \mathbb{P}(Y^x \in B_T)| \leq \eta/4.$$

Combined with our previous result, we obtain:

$$\sup_{x \in U} |\mathbb{P}(M^{x,n} \in B) - \mathbb{P}(Y^x \in B)| \leq 3\eta/4.$$

Let $\mathcal{C}_1, \ldots, \mathcal{C}_{K(n)}$ be the cluster cores used by the algorithm and computed with the density estimator $\hat{f}$. These cluster cores are approximations of the sets $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_K$ obtained using the same computation with the true density f. By our assumptions on the convergence of $\hat{f}$, Theorems 9.2 and 11.1 [22] guarantee $\lim_{n\to\infty} K(n) = K$ almost surely. Hence, we can assume that n is sufficiently large for $K(n)$ to be equal to K. By our assumptions on the cluster cores and the convergence of $\hat{f}$, Theorem 10.1 [22] guarantees there exists π such that, almost surely,

$$\forall \delta > 0, \forall i \in \{1, \ldots, K\}, \lim_{n\to\infty} \mathbb{P}(\tilde{\mathcal{C}}_i^{-\delta} \subset \mathcal{C}_{\pi(i)} \subset \tilde{\mathcal{C}}_i^{\delta}) = 1. \tag{4.5}$$

Without loss of generality, we can assume $\pi(i) = i$. Hence, $\mathbb{P}(M^{x,n} \in B) \geq \lim_{n\to\infty} \mu_{n,i}(x)$ and

$$\lim_{n\to\infty} \mu_{n,i}(x) - \mu_i(x) \leq \lim_{n\to\infty} \mu_{n,i}(x) - \mu_{i,\delta}^{+}(x) + \eta/4 \leq \eta.$$

Similarly,

$$\mu_i(x) - \lim_{n\to\infty} \mu_{n,i}(x) \leq \eta,$$

concluding the proof.

Chapter 5

Stein's method for diffusion approximation

At the end of Section 3.3.3, we have seen it is possible to obtain the weak convergence of the invariant measure of random walks on random geometric graphs to the invariant measure of the limiting diffusion process. In this Section, we improve this result by quantifying this convergence in terms of Wasserstein distance of order 2 (Proposition 28). In order to obtain this bound, we adapt the approach of [52] which relies on Stein's method and derive new ways to bound Wasserstein distances between measures (Theorems 18, 21 and 23). We then deal with possible applications of our results. We first obtain convergence rates for the Central Limit Theorem (Theorem 25). Then, we bound the distance between the invariant measures of a random walk and of a diffusion process. This bound can be used to solve our initial problem and to quantify the convergence of the invariant measure of a random walk on a random geometric graph. To illustrate this result, we tackle the case of the k-nearest neighbor graph (Proposition 28). Finally, we show how our bound can also be used to study the complexity of a simple Monte Carlo algorithm (Proposition 29).

5.1 An introduction to Stein's method

Stein's method corresponds to a family of approaches used to bound distances between measures. It was introduced by Charles Stein in 1972 [77] to bound the distance to the Gaussian measure γ. His approach relied on the following idea: since the Gaussian measure is the only measure such

that, for any test function (i.e. compactly supported smooth function) ϕ,

$$\int (x\phi(x) - \nabla\phi)\,\gamma(dx) = 0, \tag{5.1}$$

we can expect that if a measure ν satisfies

$$\int (x\phi(x) - \nabla\phi)\,\nu(dx) \approx 0,$$

then ν should be close to γ.

Barbour [5] generalized this idea by replacing the Gaussian measure by a measure μ assumed to be the invariant measure of a diffusion process with infinitesimal generator $\mathcal{L}_\mu = b.\nabla + < a, \nabla^2 >$. Indeed, for any suitable function ϕ,

$$\int \mathcal{L}_\mu \phi \mu(dx) = 0.$$

In the case of the Gaussian measure the corresponding generator is $\mathcal{L}_\gamma = -x.\nabla + \Delta$, hence this equation generalizes Stein's identity by replacing ϕ by $\nabla\phi$. Thus, if a measure ν satisfies

$$\int \mathcal{L}_\mu \phi \nu(dx) \approx 0,$$

for ϕ belonging to some set of functional, we can expect ν to be close to μ. In order to show that such a condition holds there are two possibilities. The first one is to "solve the Stein's Equation" by computing $\mathcal{L}_\mu\phi$. But this approach is not always feasible, in particular in the multi-dimensional setting. Instead, let us assume that there exists $\mathcal{L}_\nu$ such that

$$\int \mathcal{L}_\nu \phi \nu(dx) = 0,$$

in which case we say ν is invariant under $\mathcal{L}_\nu$. Then, for any compactly supported smooth function ϕ,

$$\int \mathcal{L}_\mu \phi \nu(dx) = \int (\mathcal{L}_\mu - \mathcal{L}_\nu)\phi\nu(dx).$$

Thus if $\mathcal{L}_\nu$ is close to $\mathcal{L}_\mu$, then $\int \mathcal{L}_\mu\phi\nu(dx)$ should be close to zero and ν should be close to μ.

In our case, deriving such an operator is rather straightforward: if ν is the invariant measure of a Markov chain with generator $\mathcal{L}_\nu$ then ν is invariant under $\mathcal{L}_\nu$. Using this type of operators in Stein's method can be

related to approach of [72] in which the author uses the a pair (X, X') of random variables drawn from ν as such a pair of variable corresponds to using a Markov chain with transition kernel defined by

$$\forall x \in E, \forall F \subset \mathcal{F}, K(x, F) = \mathbb{E}[X' \in F | X = x].$$

Let us note there are many other ways to obtain an operator $\mathcal{L}_\nu$ under which ν is invariant such as the original method of exchangeable pairs [78], biasing techniques [4, 41] or the Stein kernel [52]. For a more complete presentation of Stein's method and results obtained using this framework, we invite the reader to consult the following survey [20].

5.2 Distances and divergences between measures

So far, we have only been interested into weak convergence of measures. Unfortunately, since this convergence is not associated to a distance, it cannot be quantified. Let us browse through potential candidates distances we could use to quantify the convergence between measures.

Let μ and ν be two measures defined on a domain E of $\mathbb{R}^d$ and let $\mathcal{B}$ denotes the class of Borel sets of $\mathbb{R}^d$. A classical distance between μ and ν is the total variation distance:

$$TV(\mu, \nu) = \max_B |\mu(B) - \nu(B)| = \sup_{\|\phi\|_\infty \leq 1} |\int_E \phi d\mu - \int_{\mathbb{R}^d} \phi d\nu|.$$

If $d\nu = h d\mu$, one can also use the relative entropy of ν with respect to μ, also called Kullback-Lieber divergence,

$$H(\nu|\mu) = \int_E h \log h d\mu.$$

Whenever h is smooth, one can also use the Fisher information of ν with respect to μ

$$I(\nu|\mu) = \int_E \frac{\|\nabla h\|^2}{h} d\mu.$$

Unfortunately, these quantities are not appropriate for our task. Indeed, if the total variation distance between a discrete measure and an absolutely continuous measure is always one (the maximum is obtained by taking B to be the support of the discrete measure). Similarly, the relative entropy and Fisher information of a discrete measure with respect to a continuous

one is not defined. Instead, we are going to focus on another family of distances called Wasserstein distances. Let μ and ν be two measures on E, a measure π on $E \times E$ is a transport plan between ν and μ if $\pi(., E) = \mu(.)$ and $\pi(E, .) = \nu(.)$. For any $p > 0$, we say a measure μ has finite p-th moment if $\int_E \|x\|^p \mu(dx) < \infty$. Let $p \geq 1$ and suppose μ and ν have finite p-th moment, the p-Wasserstein distance between μ and ν on E is

$$W_p^p(\mu, \nu) = \inf_{\pi} \left(\int_{E \times E} \|x - y\|^p \pi(dx, dy) \right)^{1/p}$$

where the infimum is taken over all transport plans between μ and ν. These distances quantify the weak convergence and are adapted to compare discrete and continuous measures.

Theorem 17 (Theorem 6.8 [83])**.** *Let $p \geq 1$ and let μ and $(\nu_n)_{n \in \mathbb{N}}$ be measures on $\mathbb{R}^d$ with finite p-th moment. The following statements are equivalent*

- $\lim_{n \to \infty} W_p(\nu_n, \mu) = 0$*;*
- $(\nu_n)_{n \geq 0}$ *converges weakly to μ and*

$$\forall 1 \leq q \leq p, \lim_{n \to \infty} \int_{\mathbb{R}^d} \|x\|^q \nu_n(dx) = \int_{\mathbb{R}^d} \|x\|^q \mu(dx).$$

Recently, Ledoux, Nourdin and Peccati [52] have provided a way to bound the 2-Wasserstein distance between two measures μ, invariant under an operator

$$\mathcal{L}_\mu = b.\nabla + < a, \nabla^2 >,$$

and ν, invariant under an operator of the form

$$\mathcal{L}_\nu = b.\nabla + < \tau_\nu, \nabla^2 >,$$

in which case τ_ν is called a Stein kernel.

Hence, if we manage to adapt their result and define $\mathcal{L}_\nu$ as the generator of a Markov chain, our objective would be achieved. As mentioned in the end of Section 5.1, ν is the invariant measure of a Markov chain with transition kernel K if and only if there exists a pair (X, X') of random variables with measure ν such that, for any $x \in \mathbb{R}^d$, the measure of $\mathbb{E}[X' \mid X = x]$ is equal to $K(x)$. The generator of the corresponding Markov chain is then given by

$$\mathcal{L}_\nu \phi(x) = \mathbb{E}[\phi(X') - \phi(X) \mid X = x].$$

Since it leads to clearer notations, we will use pairs of random variables rather than Markov chains in the remainder of this Chapter.

5.3 The approach

Let us first introduce some notations and recall notations introduced in Section 3.2.1. Let $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$, we denote by $x^{\otimes k} \in (\mathbb{R}^d)^{\otimes k}$ the tensor of order k of x,

$$\forall j_1, \ldots, j_k \in \{1, \ldots, d\}, (x^{\otimes k})_{j_1,\ldots,j_k} = \prod_{i=1}^{k} x_{j_i}.$$

For any $x, y \in (\mathbb{R}^d)^{\otimes k}$ and any symmetric positive-definite $d \times d$ matrix A, let

$$< x, y >_A = \sum_{l,j \in \{1,\ldots,d\}^k} x_l y_j \prod_{i=1}^{k} A_{j_i,l_i},$$

and, by extension,

$$\|x\|_A^2 = < x, x >_A .$$

For any smooth function ϕ and $x \in \mathbb{R}^d$, let $\nabla^k \phi \in (\mathbb{R}^d)^{\otimes k}$ where

$$\forall j_1, \ldots, j_k \in \{1, \ldots, d\}, (\nabla^k \phi(x))_{j_1,\ldots,j_k} = \frac{\partial^k \phi}{\partial x_{j_1} \ldots \partial x_{j_k}}(x).$$

Let E be a convex domain of $\mathbb{R}^d$ and ν and μ be two measures with support E. Suppose μ is a reversible measure for the diffusion process with generator $\mathcal{L}_\mu = b.\nabla + < a, \nabla^2 >$ where b and a are smooth on E and a is symmetric positive-definite on all of E.

Let $(P_t)_{t \geq 0}$ be the Markov semigroup with infinitesimal generator $\mathcal{L}_\mu$. For any measure $d\eta = h d\mu$, let $d\eta_t = P_t h d\mu$. We first assume that $d\nu = h d\mu$ and $I_\mu(\nu_t) < \infty$ for any $t > 0$.

Since μ is the invariant measure of $\mathcal{L}_\mu$, under reasonable assumptions, ν_t converges to μ as t goes to infinity. We can thus control the distance between μ and ν by controlling the distance between between ν_t and ν. The latter can be achieved via the following inequality (see [83]),

$$\frac{d^+}{dt} W_2(\nu, \nu_t) \leq I_\mu(\nu_t)^{1/2}, \tag{5.2}$$

along with a bound on $I_\mu(\nu_t)$. We have

$$I_\mu(\nu) = \int_E \frac{\|\nabla h\|_a^2}{h} d\mu = \int_E < \nabla h, \nabla \log h >_a \, d\mu.$$

If we write $v_t = \log(P_t h)$,

$$I_\mu(\nu_t) = \int_E < \nabla P_t h, \nabla v_t >_a \, d\mu.$$

Since μ is reversible, it satisfies the following integration by parts formula: for any smooth compactly supported functions f and g,

$$\int_E < \nabla f, \nabla g >_a d\mu = -\int_E f\mathcal{L}_\mu g d\mu.$$

Since $I_\mu(\nu)$ is finite and $P_t h\mathcal{L}_\mu v_t$ can be shown to be integrable by the results of the following sections, we can apply this integration by parts formula to obtain

$$I_\mu(\nu_t) = \int_E < \nabla P_t h, \nabla v_t >_a d\mu = -\int_E P_t h\mathcal{L}_\mu v_t d\mu.$$

Using the symmetry of μ with respect to P_t and the commutativity of P_t and $\mathcal{L}_\mu$,

$$I_\mu(\nu_t) = -\int_E hP_t\mathcal{L}_\mu v_t d\mu = -\int_E \mathcal{L}_\mu P_t v_t d\nu.$$

Now, suppose there exists an operator $\mathcal{L}_\nu$ such that,

$$\int_E \mathcal{L}_\nu P_t v_t d\nu = 0,$$

then

$$I_\mu(\nu_t) = \int_E (\mathcal{L}_\nu - \mathcal{L}_\mu) P_t v_t d\nu.$$

In [52], $\mathcal{L}_\nu$ is given by the Stein kernel but it can be defined in many other ways. For example, as mentioned at the end of Section 5.1, if (X, X') is a couple of random variables drawn from ν then, taking

$$\mathcal{L}_\nu \phi(x) = \frac{1}{s}\mathbb{E}\left[\phi(X') - \phi(X) | X = x\right],$$

we have

$$\int_{\mathbb{R}^d} \mathcal{L}_\nu P_t v_t d\nu = 0.$$

Now, suppose $P_t v_t$ is real analytic on E, then for any $s > 0$,

$$\mathcal{L}_\nu P_t v_t(x) = \frac{1}{s}\mathbb{E}\left[\sum_{k=1}^{\infty} < \frac{(X' - X)^{\otimes k}}{k!}, \nabla^k P_t v_t > | X = x\right],$$

In which case,

$$\begin{aligned} I_\mu(\nu_t) =& \mathbb{E}\left[< \mathbb{E}[X' - X \mid X] - b(X), \nabla P_t v_t(X) >\right] \\ &+ \mathbb{E}\left[< \mathbb{E}[\frac{(X' - X)^{\otimes 2}}{2} \mid X] - a(X), \nabla^2 P_t v_t >\right] \\ &+ \mathbb{E}\left[\sum_{k=3}^{\infty} < \mathbb{E}[\frac{(X' - X)^{\otimes k}}{k!} \mid X], \nabla^k P_t v_t(X) >\right]. \end{aligned} \tag{5.3}$$

The last step of the approach consists in using the regularizing properties of the semigroup P_t in order to bound the last equation by a quantity involving $P_t\|\nabla v_t\|_a^2$. Then, since $\mathbb{E}[P_t\|\nabla v_t\|_a^2(X)]^{1/2} = I_\mu(\nu_t)^{1/2}$ and $I_\mu(\nu_t)$ is finite, we obtain a bound on $I_\mu(\nu_t)^{1/2}$ and conclude. Let us note that, since a is positive-definite on all of E, the bounds we derive on $\|\nabla^k P_t v_t\|_a$ imply $P_t v_t$ is real analytic on all of E [47].

In order to deal with discrete measures, let us note that, for $\epsilon > 0$, ν_ϵ is well-defined. Thus, if it has a finite Fisher information with respect to μ, we can apply the previous approach to any ν_ϵ and let ϵ go to 0 to obtain a bound on $W_2(\nu, \mu)$ even though ν is discrete.

Our goal in the remainder of this section will thus consist in providing bounds for Equation 5.3. We start with the Gaussian case where such a bound can be directly obtained using the integral representation of $(P_t)_{t\geq 0}$ and integrations by parts. We then deal with more general measures μ and derive a bound using Gamma calculus.

5.3.1 Gaussian case

Let $d\mu = d\gamma = (2\pi)^{-d/2} e^{-\frac{|x|^2}{2}} dx$ be the Gaussian measure in $\mathbb{R}^d$. γ is the invariant measure of $\mathcal{L}_\gamma = -x.\nabla + \Delta$ and the associated semigroup $(P_t)_{t\geq 0}$ is the Ornstein-Uhlenbeck semigroup. Let ϕ be a smooth function with compact support on $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, $P_t\phi$ admits the following representation

$$P_t\phi(x) = \int_{\mathbb{R}^d} \phi(xe^{-t} + \sqrt{1 - e^{-2t}}y) d\gamma(y).$$

Using an integration by part, we obtain

$$\begin{aligned}\nabla P_t\phi(x) &= e^{-t} \int_{\mathbb{R}^d} \nabla\phi(xe^{-t} + \sqrt{1 - e^{-2t}}y) d\gamma(y) \\ &= \frac{e^{-t}}{\sqrt{1 - e^{-2t}}} \int_{\mathbb{R}^d} y\phi(xe^{-t} + \sqrt{1 - e^{-2t}}y) d\gamma(y).\end{aligned}$$

For any $k > 0$ and any $i \in \{1, \ldots, d\}^k$, let H_i be the multivariate Hermite polynomial of index i,

$$H_i = (-1)^k e^{\frac{|x|^2}{2}} \frac{\partial^k}{\partial x_{i_1} \ldots \partial x_{i_k}} e^{-\frac{|x|^2}{2}}.$$

Multiple integrations by parts yield

$$(\nabla^k P_t\phi(x))_i = \frac{e^{-kt}}{(1 - e^{-2t})^{k/2}} \int_{\mathbb{R}^d} H_i(y)\phi(xe^{-t} + \sqrt{1 - e^{-2t}}y) d\gamma(y).$$

Hermite polynomials form an orthogonal basis of $L_2(\gamma)$ with norm

$$\forall i \in \{1,\dots,d\}^k, \|H_i\|_\gamma^2 = \int_{\mathbb{R}^d} H_i^2(y)d\gamma(y) = \prod_{j=1}^{d}\left(\sum_{l=1}^{k}\delta_{i_l,j}\right)!.$$

Hence,

$$\begin{aligned}
&\sum_{k=1}^{\infty}\sum_{i\in\{1,\dots,d\}^{k-1}} \frac{e^{2kt}(1-e^{-2t})^{k-1}}{\|H_i\|_\gamma^2}(\nabla^k P_t\phi)_i^2(x)\\
&\quad=\sum_{k=1}^{\infty}\sum_{i\in\{1,\dots,d\}^{k-1}} \frac{e^{2kt}(1-e^{-2t})^{k-1}}{\|H_i\|_\gamma^2}\left(\int_{\mathbb{R}^d}(\nabla^k\phi)_i(xe^{-t}+\sqrt{1-e^{-2t}}y)d\gamma(y)\right)^2\\
&\quad=\sum_{k=1}^{\infty}\sum_{i\in\{1,\dots,d\}^{k-1}} \left(\int_{\mathbb{R}^d}\frac{H_i}{\|H_i\|_\gamma}\nabla\phi(xe^{-t}+\sqrt{1-e^{-2t}}y)d\gamma(y)\right)^2\\
&\quad=\int_{\mathbb{R}^d}\|\nabla\phi(xe^{-t}+\sqrt{1-e^{-2t}}y)\|^2 d\gamma(y)\\
&\quad=P_t\|\nabla\phi\|^2(x).
\end{aligned}$$

Let us pose

$$\begin{aligned}
S(t) =&e^{-2t}\mathbb{E}\left[\left\|\mathbb{E}[\frac{X'-X}{s}\mid X]+X\right\|^2\right]\\
&+\frac{e^{-4t}}{1-e^{-2t}}\mathbb{E}\left[\left\|\mathbb{E}[\frac{(X'-X)^{\otimes 2}}{2s}\mid X]-I_d\right\|^2\right]\\
&+\sum_{k=3}^{\infty}\sum_{i\in\{1,\dots,d\}^{k-1}}\frac{e^{-2kt}\|H_i\|_\gamma^2}{(sk!)^2(1-e^{-2t})^{k-1}}\mathbb{E}\left[\|\mathbb{E}[(X'-X)_i^{\otimes k}\mid X]\|^2\right].
\end{aligned}$$

Applying Cauchy-Schwarz's inequality to Equation 5.3 yields

$$I_\gamma(\nu_t)\le S(t)^{1/2}\mathbb{E}[P_t\|\nabla v_t(X)\|^2]^{1/2} = S(t)^{1/2}I_\gamma(\nu_t)^{1/2}. \tag{5.4}$$

We have thus bounded $I_\gamma(\nu_t)$. Now, according to Equation 5.2, integrating our bound on $I_\gamma(\nu_t)$ for $t\in\mathbb{R}^+$ would yield a bound on $W_2(\nu,\gamma)$. However, we encounter integrability issues for the higher order terms for small values of t. To circumvent this issue, we use a family of couplings $(X, X'_t)_{t\ge 0}$ interpolating between $X'_0 = X$ and $X'_\infty = X'$, ensuring the problematic terms are 0 at $t = 0$. Finally, for any discrete measure ν, ν_t has finite Fisher information as long as the second moment of ν is finite (see Remark 2.1 [62]). We are now ready to state the first result of this work.

Theorem 18. *Let ν be a measure on $\mathbb{R}^d$ with finite second moment and let X and $(X'_t)_{t\geq 0}$ be random variables drawn from ν. For any $s > 0$,*

$$W_2(\nu, \gamma) \leq \int_0^\infty \sqrt{S(t)}dt,$$

with

$$\begin{aligned} S(t) =& e^{-2t}\mathbb{E}\left[\left\|\mathbb{E}[\frac{X'_t - X}{s} \mid X] + X\right\|^2\right] \\ &+ \frac{e^{-4t}}{1-e^{-2t}}\mathbb{E}\left[\left\|\mathbb{E}[\frac{(X'_t - X)^{\otimes 2}}{2s} \mid X] - I_d\right\|^2\right] \\ &+ \sum_{k=3}^{\infty}\sum_{i\in\{1,\dots,d\}^{k-1}} \frac{e^{-2kt}\|H_i\|_\gamma^2}{(sk!)^2(1-e^{-2t})^{k-1}}\mathbb{E}\left[\|\mathbb{E}[(X'_t - X)_i^{\otimes k} \mid X]\|^2\right]. \end{aligned}$$

5.3.2 General case

In general, $(P_t)_{t\geq 0}$ does not admit a closed form formula so we cannot rely on a direct approach. Let us first apply Cauchy-Schwarz's inequality to Equation 5.3 in order to obtain

$$\begin{aligned} I_\mu(\nu_t) \leq& \mathbb{E}\left[\|\mathbb{E}[X' - X \mid X] - b(X)\|^2_{a^{-1}(X)}\right]^{1/2}\mathbb{E}[\|\nabla P_t v_t(X)\|^2_{a(X)}]^{1/2} \\ &+ \mathbb{E}\left[\|\mathbb{E}[\frac{(X'-X)^{\otimes 2}}{2} \mid X] - a(X)\|^2_{a^{-1}(x)}\right]^{1/2}\mathbb{E}[\|\nabla^2 P_t v_t(X)\|^2_{a(X)}]^{1/2} \\ &+ \sum_{k=3}^{\infty}\mathbb{E}\left[\|\mathbb{E}[\frac{(X'-X)^{\otimes k}}{k!} \mid X]\|^2_{a^{-1}(x)}\right]^{1/2}\mathbb{E}[\|\nabla^k P_t v_t(X)\|^2_{a(X)}]^{1/2}. \end{aligned} \tag{5.5}$$

Our objective is to bound $\|\nabla^k P_t v_t\|_a^2$ by a quantity involving $P_t\|\nabla v_t\|_a^2$ using the framework of Γ-calculus described in [3]. This approach relies on the iterated gradients Γ_i, defined recursively for any smooth functions f, g by

$$\begin{aligned} &\Gamma_0(f,g) = fg; \\ &\Gamma_{i+1}(f,g) = \frac{1}{2}\left[\mathcal{L}_\mu(\Gamma_i(f,g)) - \Gamma_i(\mathcal{L}_\mu f, g) - \Gamma_i(f, \mathcal{L}_\mu g)\right]. \end{aligned}$$

The triple (E, μ, Γ_1) is called a Markov triple, a structure extensively studied in [3]. In particular if there exists $\rho \in \mathbb{R}$ such that

$$\Gamma_2 \geq \rho\Gamma_1,$$

the Markov triple is said to satisfy a *curvature-dimension inequality*, or $CD(\rho, \infty)$ condition, under which $(P_t)_{t\geq 0}$ has many interesting properties. For instance, it is known that, under a $CD(\rho, \infty)$ condition, $(P_t)_{t\geq 0}$ satisfies the following gradient bound (see e.g. Theorem 3.2.3 [3])

$$\|\nabla P_t f\|_a^2 \leq e^{-2\rho t} P_t(\|\nabla f\|_a^2),$$

Remark 19. Under a $CD(\rho, \infty)$ condition, if $\nu = hd\mu$, then, according to Theorem 5.5.2 [3],

$$I_\mu(\nu_t) \leq \frac{2\rho}{1 - e^{-2\rho t}} \left(P_t(h \log h) - P_t h \log(P_t h)\right).$$

Thus, if ν_ϵ has finite entropy with respect to μ for any $\epsilon > 0$, then $I_\mu(\nu_t)$ is finite for any $t > 0$.

In the proof of Theorem 4.1 [52], the authors show that, under a $CD(\rho, \infty)$ condition for $\rho > 0$ and assuming there exists $\kappa, \sigma > 0$ such that $\Gamma_3 \geq \kappa\Gamma_2$ and $\Gamma_2 \geq \sigma\|\nabla^2 f\|_a$,

$$\|\nabla^2 P_t f\|_a^2 \leq \frac{\kappa}{\sigma(e^{\kappa t} - 1)} P_t \|\nabla f\|_a^2. \tag{5.6}$$

We could use a similar approach and suppose that for any $k > 1$ there exists some κ_k and σ_k such that $\Gamma_{k+1} \geq \kappa_k \Gamma_k$ and $\Gamma_k \geq \sigma\|\nabla^k f\|_a$ in order to bound $\|\nabla^k P_t f\|_a$. However, such assumptions would be quite restrictive in practice. Instead, we derive bounds relying on a simple $CD(\rho, \infty)$ condition.

Proposition 20. *Suppose that $\mathcal{L}$ satisfies a $CD(\rho, \infty)$ condition for $\rho \in \mathbb{R}$. Then, for any $k \in \mathbb{N}^\star$, $t > 0$ and any smooth compactly supported function ϕ,*

$$\|\nabla^k P_t \phi\|_a \leq f_k(t) \sqrt{P_t \|\nabla \phi\|_a^2},$$

where

$$f_k(t) = \begin{cases} e^{-\rho t \max(1, k/2)} \left(\frac{2\rho d}{e^{2\rho t/(k-1)} - 1}\right)^{(k-1)/2} & \text{if } \rho \neq 0 \\ t^{(1-k)/2} \text{ if } \rho = 0. \end{cases}$$

Unfortunately, our bound is not dimension-independent as one could expect from Equation 5.6. We believe this dependency to be an artifact of the proof. Nevertheless, by injecting these bounds in Equation 5.5, we obtain a bound on $I_\mu(\nu_t)^{1/2}$ leading to the following result.

Theorem 21. *Let ν be a measure on $\mathbb{R}^d$. Assume the entropy of ν_ϵ with respect to μ is finite for any $\epsilon > 0$ and let X and $(X'_t)_{t\geq 0}$ be random variables*

drawn from ν. If $\mathcal{L}_\mu$ satisfies a $CD(\rho,\infty)$ condition for $\rho \in \mathbb{R}$, then, for any $s > 0$, $T > 0$,

$$W_2(\nu,\nu_T) \leq \int_0^T e^{-\rho t}\mathbb{E}\left[\left\|\mathbb{E}\left[\frac{X_t'-X}{s} \mid X\right] - b(X)\right\|_{a^{-1}}^2\right]^{1/2} dt$$
$$+ \int_0^T f_2(t)\mathbb{E}\left[\left\|\mathbb{E}\left[\frac{(X_t'-X)^{\otimes 2}}{2s} \mid X\right] - a(X)\right\|_{a^{-1}}^2\right]^{1/2} dt$$
$$+ \sum_{k=3}^{\infty}\int_0^T \frac{f_k(t)}{sk!}\mathbb{E}[\left\|\mathbb{E}[(X_t'-X)^{\otimes k} \mid X]\right\|_{a^{-1}}^2]^{1/2}dt,$$

where the functions $(f_k)_{k\geq 1}$ are defined in Proposition 20.

If $\rho > 0$, we can set T to infinity to bound $W_2(\nu,\mu)$. On the other hand, if $\rho \leq 0$, it is still possible to bound $W_2(\nu,\mu)$ as long as ν_t converges exponentially fast to ν.

Lemma 22. *Suppose there exists κ such that for any measure η and any $t > 0$, we have*

$$W_2(\eta_t,\mu) \leq e^{-\kappa t}W_2(\eta,\nu).$$

Then, for any $T > 0$,

$$W_2(\nu,\mu) \leq \frac{W_2(\nu,\nu_T)}{1-e^{-\kappa T}}.$$

Proof. Indeed, we have

$$\begin{aligned} W_2(\nu,\mu) &\leq W_2(\nu,\nu_t) + W_2(\nu_t,\mu) \\ &\leq W_2(\nu,\nu_t) + e^{-\kappa t}W_2(\nu,\mu). \end{aligned}$$

□

Such an exponential convergence to μ can be verified under weaker conditions than a $CD(\rho,\infty)$ inequality for $\rho > 0$. For example, if a is the identity matrix and b is the gradient of some potential V then this assumption is satisfied whenever V is strongly convex outside a bounded set C with bounded first and second order derivatives on C [42], which is equivalent to satisfying a $CD(\rho_1,\infty)$ condition for some $\rho_1 \in \mathbb{R}$ and having $\Gamma_2 \geq \rho_2\Gamma_1$ with $\rho_2 > 0$ outside of C. An extension of this result for more general a and for manifolds is proposed in [87].

5.4 Gaussian measure in dimension one

The Stein kernel can also be used to bound the Wasserstein distance of order $p \geq 1$ between a measure ν and the Gaussian measure γ under a stronger definition. Let X be a random variable drawn from ν, we say that τ_ν is a strong Stein kernel for ν if

$$\mathbb{E}\left[-X\phi(X) + \tau_\nu(X)\nabla\phi(X)\right] = 0$$

for every compactly supported smooth function ϕ.

In dimension one, if τ_ν is a Stein kernel for ν, then it satisfies the previous condition, hence we can expect our coupling approach to be able to replace the Stein kernel. Let $\mu = \gamma_1$ be the one-dimensional Gaussian measure. For $k \in \mathbb{N}$, we denote by H_k the k-th Hermite polynomial,

$$H_k = (-1)^k e^{\frac{|x|^2}{2}} \frac{d^k e^{-\frac{|x|^2}{2}}}{dx^k}.$$

First, a modification of the proof of Lemma 2 from [64] yields the general estimate

$$\frac{d^+}{dt} W_p(\nu, \nu_t) \leq \left(\int_{\mathbb{R}} |v_t'|^p d\nu_t\right)^{1/p}. \tag{5.7}$$

Let us provide a version of v_t'. Let (X, X') be random variables drawn from ν and Z be a Gaussian random variable. For $t > 0$, let $F_t = e^{-t}X + \sqrt{1 - e^{-2t}}Z$ and consider the function ρ_t defined for any $x \in \mathbb{R}$ as follows

$$\rho_t(x) = \mathbb{E}\left[e^{-t}\left(\frac{X' - X}{s} + X\right) + \frac{e^{-2t}}{\sqrt{1 - e^{-2t}}}\left(\frac{(X' - X)^2}{2s} - 1\right) H_1(Z) | F_t = x\right] + \mathbb{E}\left[\sum_{k=3}^{\infty} \frac{e^{-kt}}{s\sqrt{1 - e^{-2t}}^{k-1}} \frac{(X' - X)^k}{k!} H_{k-1}(Z) | F_t = x\right].$$

For any compactly supported smooth function $\phi : \mathbb{R} \to \mathbb{R}$, we obtain, after successive integrations by parts with respect to Z,

$$\mathbb{E}[\rho_t(F_t)\phi(F_t)] = \mathbb{E}\left[e^{-t}X\phi(F_t) - e^{-2t}\phi'(F_t)\right] + \mathbb{E}\left[\sum_{k=1}^{\infty} e^{-kt} \frac{(X' - X)^k}{sk!} \phi^{(k-1)}(F_t)\right].$$

Let Φ be a primitive function of ϕ, by the results of Section 5.3.1, $\mathbb{E}[\Phi(F_t) \mid X = x] = P_t\Phi(x)$ is real analytic. Hence, since X' and X have the same

measure we have

$$\mathbb{E}\left[\sum_{k=1}^{\infty} e^{-kt}\frac{(X'-X)^k}{sk!}\phi^{(k-1)}(F_t)\right] = \frac{1}{s}\mathbb{E}[\Phi(e^{-t}X' + \sqrt{1-e^{-2t}}Z) - \Phi(e^{-t}X + \sqrt{1-e^{-2t}}Z)] = 0.$$

Therefore,

$$\begin{aligned}\mathbb{E}[(\rho_t(F_t) - F_t)\phi(F_t)] &= \mathbb{E}[(-F_t + e^{-t}X)\phi(F_t) - e^{-2t}\phi'(F_t)] \\ &= \mathbb{E}[-(1-e^{-2t})\phi'(F_t) - e^{-2t}\phi'(F_t)] \\ &= -\mathbb{E}[\phi'(F_t)].\end{aligned}$$

Therefore, ρ_t satisfies the characterization of v_t' presented in Equation 2.28 [62]: it is thus a version of v_t'. We are thus able to bound

$$\left(\int_{\mathbb{R}} |v_t'|^p d\nu_t\right)^{1/p} = \mathbb{E}[|\rho_t(F_t)|^p]^{1/p}$$

using the $L_p(\gamma_1)$-norm of the Hermite polynomials $\|H_k\|_{p,\gamma_1}^p = \int_{\mathbb{R}} |H_k|^p d\gamma_1$. Injecting this bound in Equation 5.7, we are able to bound $W_p(\nu, \gamma_1)$.

Theorem 23. *Let ν be a measure on $\mathbb{R}$ and let X and $(X_t)_{t\geq 0}$ be random variables drawn from ν. We have, for any $p \geq 1$, $s > 0$,*

$$\begin{aligned}W_p(\nu, \gamma_1) \leq &\int_0^\infty e^{-t}\mathbb{E}\left[\left|\mathbb{E}\left[\frac{X_t' - X}{s} \mid X\right] + X\right|^p\right]^{1/p} dt \\ &+ \int_0^\infty \frac{e^{-2t}\|H_1\|_{p,\gamma_1}}{\sqrt{1-e^{-2t}}}\mathbb{E}\left[\left|\mathbb{E}\left[\frac{(X_t'-X)^2}{2s} \mid X\right] - 1\right|^p\right]^{1/p} dt \\ &+ \sum_{k=3}^{\infty}\int_0^\infty \frac{e^{-kt}\|H_{k-1}\|_{p,\gamma_1}}{s\sqrt{1-e^{-2t}}^{k-1}k!}\mathbb{E}\left[\left|\mathbb{E}[(X_t'-X)^k \mid X]\right|^p\right]^{1/p} dt.\end{aligned}$$

Remark 24. [50] gives the asymptotic of the p-norm of Hermite polynomials with respect to the Gaussian measure, more precisely there exist constants $C(p)$ such that

$$\|H_k\|_p \leq \begin{cases} C(p)\sqrt{k!}k^{-1/4}(1 + O(k^{-1})) \text{ if } 0 < p < 2 \\ C(p)\sqrt{k!}(p-1)^{k/2}(1 + O(k^{-1})) \text{ if } p > 2 \end{cases}$$

5.5 Applications

5.5.1 Central Limit Theorem

Let $X_1, \dots, X_n$ be i.i.d. random variablestaking values in $\mathbb{R}^d$ such that $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^{\otimes 2}] = I_d$. Let ν_n be the measure of $S_n = n^{-1/2} \sum_{i=1}^n X_i$. According to the Central Limit Theorem, ν_n should converge to the Gaussian measure γ, our objective in this section is to provide a bound of $W_p(\nu_n, \gamma)$ for some $p \geq 2$. Let $X_1', \dots X_n'$ be independent copies of $X_1, \dots, X_n$ and let I be a uniform random variable on $\{1, \dots, n\}$. For any $t > 0$, we pose

$$S_{n,t}' = S_n + n^{-1/2}(X_I' - X_I) 1_{\|X_I'\|, \|X_I\| \leq \sqrt{n(e^{2t}-1)}}.$$

By construction, for any $t > 0$, $S_{n,t}'$ is drawn from ν_n. Using this coupling and applying either Theorem 18 or Theorem 23 with timestep $s = \frac{1}{n}$, see Section 5.6.2 for the detailed computations, we obtain the following result.

Theorem 25. *Let $X_1, \dots, X_n$ be i.i.d. random variables in $\mathbb{R}^d$ with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^{\otimes 2}] = I_d$. There exists a universal constant $C > 0$, such that if $\mathbb{E}[\|X_1\|^{2+m}] < \infty$ for some $m \in [0, 2]$, then,*

$$W_2(\nu_n, \gamma) \leq C \begin{cases} n^{-m/4} \mathbb{E}[\|X_1\|^{2+m}]^{1/2} + o(n^{-m/4}) \text{ if } m < 2 \\ n^{-1/2} \|\mathbb{E}[X_1^{\otimes 2} \|X_1\|^2]\| \text{ if } m = 2 \end{cases}.$$

Let $p \geq 2$. If $d = 1$, there exists a universal constant C_p such that if $\mathbb{E}[|X_1|^{p+q}] < \infty$ for some $q \in [0, p]$, then taking $m \leq \min(4, p+q) - 2$,

$$W_p(\nu_n, \gamma) \leq C_p \left(n^{-m/4} (\mathbb{E}[|X_1|^{2+m}]^{1/2} + o(1_{m<2})) + n^{-1/2+(2-q)/2p} \mathbb{E}[|X_1|^{p+q}]^{1/p} \right).$$

Remark 26. Taking $d = 1, p \geq 2$ in the previous result, we have, using Hölder's inequality,

$$\begin{aligned} \mathbb{E}[|X_1|^4]^{1/2} &= \mathbb{E}[X_1^{2-4/p} X_1^{2+4}]^{1/2} \\ &= \mathbb{E}[(X_1^2)^{1-2/p} (|X_1|^{p+2})^{p/2}]^{1/2} \\ &\leq \mathbb{E}[X_1^2]^{1-2/p} \mathbb{E}[|X_1|^{p+2}]^{1/p} \\ &\leq \mathbb{E}[|X_1|^{p+2}]^{1/p}. \end{aligned}$$

Therefore, as long as $\mathbb{E}[|X_1|^{p+2}] < \infty$, Theorem 25 gives

$$W_p(\nu_n, \gamma) \leq C_p n^{-1/2} \mathbb{E}[|X_1|^{p+2}].$$

The one-dimensional result completes a result obtained by [70] who considered the case $1 \leq p \leq 2, m = 2$ and generalizes a result obtained by [74] treating the case $p > 2, m = 0$. [12] also recovered the case $p = 2, m = 2$ using an entropic approach and recently proved the case $m = 2$ for any $p > 2$ [13]. To our knowledge, the multidimensional result is new although the entropic approach from [12] might be generalized to the multidimensional setting at the expense of stronger assumptions on the moments of the variables.

5.5.2 Diffusion approximation

Let μ be the invariant measure of the diffusion process with infinitesimal generator $\mathcal{L}_\mu = b.\nabla + < a, \nabla^2 >$. Consider a discretization of this diffusion process by a Markov chain M with transition kernel K and invariant measure π and let s be the timestep of this discretization. Let X be a random variable drawn from π and let ξ be a random jump from X. Then for any $t > 0, T > 0$,

$$X_t = X + 1_{t \geq T} \xi$$

and X follow the same law. Applying Theorem 21 using $(X_t)_{t \geq 0}$ yields the following result.

Corollary 27. *Under the assumptions Theorem 21, we have, for any $T_1 > T_2 > 0$,*

$$\begin{aligned} W_2(\pi, \pi_{T_1}) \leq & \int_0^{T_2} e^{-\rho t} \mathbb{E}[\|b(X)\|_{a^{-1}}^2]^{1/2} + f_2(t) dt \\ & + \int_{T_2}^{T_1} e^{-\rho t} \mathbb{E}\left[\left\| \mathbb{E}\left[\frac{\xi}{s} \mid X \right] - b(x) \right\|_{a^{-1}}^2 \right]^{1/2} dt \\ & + \int_{T_2}^{T_1} f_2(t) \mathbb{E}[\left\| \mathbb{E}\left[\frac{\xi^{\otimes 2}}{2s} \mid X \right] - a(x) \right\|_{a^{-1}}^2 \Bigg]^{1/2} dt \\ & + \int_{T_2}^{T_1} \sum_{k=3}^{\infty} f_k(t) \mathbb{E}\left[\left\| \mathbb{E}[\frac{\xi^{\otimes k}}{k! s} \mid X] \right\|_{a^{-1}}^2 \right]^{1/2} dt, \end{aligned}$$

where the functions $(f_k)_{k \geq 1}$ are defined in Proposition 20.

Remark the quantities involved in this Corollary seems natural as they are rather similar to the quantities appearing the Theorem 7 dealing with the convergence of Markov chains to diffusion processes.

Density approximation on k-nearest neighbor graphs

Let $X_1, \dots, X_n$ be i.i.d. random variables on the flat torus $T = (\mathbb{R}/\mathbb{Z})^d$ drawn from a measure μ with smooth density f. Let π be the invariant measure of a random walk on a k-nearest neighbor graph built on $\mathcal{X}_n = X_1, \dots, X_n$. As we have seen in Section 3.3, the random walk on the k-nearest neighbor graph converges to a diffusion process with generator

$$\mathcal{L} = f^{-2/d}(\nabla \log f.\nabla + \frac{1}{2}\Delta),$$

which admits a reversible measure with density proportional to $f^{1+2/d}$. The corresponding approximation timestep is

$$s = \left(\frac{k}{n}\right)^{2/d} \frac{\int_{\|x\|\leq 1} x_1^2 dx}{\left(\int_{\|x\|\leq 1} 1 dx\right)^{1+2/d}}.$$

While T is not a domain of $\mathbb{R}^d$, the arguments used in Theorem 21 still hold, let us check its assumptions. As T is compact and f is smooth and strictly positive, $f^{-2/d}\nabla \log f$ and $f^{-2/d}$ are smooth, hence, a $CD(\rho, \infty)$ condition is verified for some $\rho \in \mathbb{R}$. Moreover, for any $\epsilon > 0$, π^ϵ is a measure with strictly positive smooth density and thus finite Fisher information with respect to $\tilde{\mu}$. Finally, the assumption of Lemma 22 is verified thanks to Corollary 2.2 [87].

Using Lemma 12 and Corollary 27 along with Lemma 22, we obtain the following result.

Proposition 28. *There exists $C > 0$ such that, with probability $1 - \frac{C}{n}$,*

$$W_2(\pi, \tilde{\mu}) \leq C\left(\frac{\sqrt{logn}n^{1/d}}{k^{1/2+1/d}} + \left(\frac{k}{n}\right)^{1/d}\right).$$

Analysis of lower order schemes for the Langevin Monte Carlo algorithm

Quite often in Bayesian statistics, one is interested in sampling points from a probability measure $d\mu = e^{-u}dx$ on $\mathbb{R}^d$. Many Monte-Carlo algorithms have been proposed and analyzed to solve this task. We want to show how our result can be used to study the convergence rate of a simple Monte-Carlo algorithm.

The measure μ is a reversible measure for the diffusion process with infinitesimal generator

$$\mathcal{L}_\mu = -\nabla u.\nabla + \Delta.$$

Since, under some assumptions on μ, the measure of Y_t converges to μ as t goes to infinity, one may want to sample points from μ by approximating Y_t. Using the Euler-Maruyama approximation with timestep s, we discretize Y_t using a Markov chain M with $M^0 = 0$ and transitions given by

$$M^{n+1} = M^n - s\nabla u(M^n) + \sqrt{2s}\mathcal{N}_n,$$

where $\mathcal{N}_1, \ldots, \mathcal{N}_n$ is a sequence of independent normal random variables with mean 0 and covariance matrix I_d. If the timestep is small enough, the invariant measure π of the Markov chain, should be close to μ. Hence, for n large enough, the measure of M^n should be close to its invariant measure and thus be close to μ. Approximate sampling for μ using this approach is known as the Langevin Monte-Carlo (LMC) algorithm [71].

One may then wonder how large n should be to achieve a given accuracy. Answering this question is linked to the choice of s as this parameter must satisfy some trade-off: large values lead to a poor approximation of μ by π, but the smaller s is, the larger the number of iterations required for the measure of M^n to be close to π. Recently, [32] proved that whenever μ is a strictly log-concave measure (i.e. satisfying a $CD(\rho, \infty)$ condition for $\rho > 0$), the LMC algorithm can reach an ϵ accuracy in total variation distance in $O(\epsilon^{-2}(d^3 + d\log(1/\epsilon))$ steps. For the Wasserstein distance, this complexity was later improved to $O(\epsilon^{-1}\sqrt{d}\log(1/\epsilon))$ by [33]. A second order discretization, called the Ozaki discretization, was also considered in [32]. Under this scheme, the number of iterations required to achieve an ϵ accuracy in total variation distance is smaller than $O(\epsilon^{-1}\dim(d + \log(1/\epsilon))^{3/2})$. Here, we propose to do the opposite by considering an example of a smaller order scheme with non-normal increments. Let $(B_n)_{n\geq 0}$ be independent multivariate Rademacher random variables, and consider the following scheme

$$M^{n+1} = M^n - s\nabla u(M^n) + \sqrt{2s}B_n. \tag{5.8}$$

Let μ be a log-concave measure, i.e. $d\mu = e^{-u}d\lambda$ and there exists $\rho > 0$ such that

$$\forall x \in \mathbb{R}^d, < \nabla u(x) - \nabla u(y), (x - y) > \geq \rho\|x - y\|_2^2.$$

Taking $\Gamma_1(f, g) =< \nabla f, \nabla g >$, this is equivalent to saying the Markov Triple $(\mathbb{R}^d, \mu, \Gamma_1)$ satisfies a $CD(\rho, \infty)$ condition for $\rho > 0$. Moreover, as shown in Subsection 5.6.3, π has finite second moment which implies, by Theorem 5.1 [1], that π^ϵ has finite entropy with respect to μ for $\epsilon > 0$. Together with Remark 19 this implies π^ϵ has finite Fisher information with respect to μ for any $\epsilon > 0$. Let X be a random variable drawn from π and

ξ be an increment from state X. By computations of Subsection 5.6.3, if μ is log-concave and ∇u is Lipschitz continuous, then there exists $C > 0$ such that

- $\mathbb{E}[\frac{\xi}{s} - \nabla u(X) \mid X] = 0$;
- $\mathbb{E}[\|\mathbb{E}[\frac{\xi^{\otimes 2}}{2s} - I_d \mid X]\|^2]^{1/2} \leq C(sd)^{1/2}$;
- $\mathbb{E}[\|\mathbb{E}[\frac{\xi^{\otimes 3}}{s} \mid X]\|^2]^{1/2} \leq Csd$;
- $\forall k > 3, \mathbb{E}[\|\mathbb{E}[\frac{\xi^{\otimes k}}{s} \mid X]\|^2]^{1/2} \leq C^k s^{k/2-1} d^{(k-1)/2}$.

Applying Corollary 27 with $T = sd^2$ allows us to bound $W_2(\pi, \mu)$.

Proposition 29. *Suppose μ is log-concave. Then, if $\|\nabla u\| \leq L$,*

$$W_2(\pi, \mu) \leq O(d^2 s^{1/2}).$$

Let us note that, using the coarse Ricci curvature framework introduced in [63], it is possible to show that M^n converges exponentially fast to π. Hence, using our result, it is possible to show that $O(\epsilon^{-2} d^4 \log(1/\epsilon))$ iterations are required to achieve an ϵ accuracy in Wasserstein distance between the measure sampled by the LMC algorithm and μ. We believe our result to be suboptimal due to the dependency on the dimension of the function f_k defined in Proposition 20, we conjecture the correct complexity to be $O(\epsilon^{-2} d^2 \log(1/\epsilon))$.

5.6 Proofs

5.6.1 Proof of Proposition 20

By Theorem 3.2.4 [3], under a $CD(\rho, \infty)$, we have for any compactly supported smooth function ϕ,

$$\|\nabla P_t \phi\|_a \leq e^{-\rho t} P_t \|\nabla \phi\|_a. \tag{5.9}$$

In order to prove the Proposition, we need to find an equivalent to the integration by parts used in the Gaussian case.

Lemma 30. *Suppose $\mathcal{L}$ satisfies a $CD(\rho, \infty)$ condition, then for all compactly supported smooth function ϕ, and any $t > 0$,*

$$\|\nabla P_t \phi\|_a^2 \leq \frac{2\rho}{e^{2\rho t} - 1} P_t |\phi|^2.$$

Proof. Let $t > 0$, for any $0 \leq s \leq t$ let

$$\Lambda(s) = P_s(\Gamma_0(P_{t-s}\phi)),$$

the first two derivatives of this function are

$$\Lambda'(s) = 2P_s(\Gamma_1(P_{t-s}\phi));$$
$$\Lambda''(s) = 4P_s(\Gamma_2(P_{t-s}\phi)).$$

By our assumption, $\Lambda''(s) \geq 2\rho\Lambda'(s)$. Hence, by Gronwall's Lemma, $\Lambda'(s) \geq e^{2\rho s}\Lambda'(0)$. Now, we have

$$\begin{aligned}
\Gamma_1(P_t\phi) &= \frac{2\rho}{e^{2\rho t}-1}\int_0^t e^{2\rho s}\Gamma_1(P_t\phi)ds \\
&= \frac{2\rho}{e^{2\rho t}-1}\int_0^t e^{2\rho s}\Lambda'(0)ds \\
&\leq \frac{2\rho}{e^{2\rho t}-1}\int_0^t \Lambda'(s)ds \\
&\leq \frac{2\rho}{e^{2\rho t}-1}(P_t(\Gamma_0(\phi)) - \Gamma_0(P_t\phi)) \\
&\leq \frac{2\rho P_t(\Gamma_0(\phi))}{e^{2\rho t}-1}.
\end{aligned}$$

□

Let $(e_1, \ldots, e_d)$ be an orthonormal basis of $\mathbb{R}^d$ with respect to the a-scalar product $< ., . >_a$.

Lemma 31. *For any smooth function ϕ and any $k > 0$, we have*

$$\|\nabla^k\phi\|_a = \sup_{\alpha\in\mathbb{R}^d, \|\alpha\|=1} \sum_{i=1}^d \alpha_i \|\nabla^{k-1} < \nabla\phi, e_i >_a \|.$$

Proof. By the duality of the a-norm, we have

$$\begin{aligned}
\|\nabla^k\phi\|_a &= \sup_{h\in(\mathbb{R}^d)^{\otimes k}, \|h\|_a=1} < \nabla^k\phi, h >_a \\
&= \sup_{\alpha\in\mathbb{R}^d, \|\alpha\|=1} \sum_{i=1}^d \sup_{h\in(\mathbb{R}^d)^{\otimes k-1}, \|h\|_a=1} < \nabla^k\phi, \alpha_i e_i \otimes h >_a \\
&= \sup_{\alpha\in\mathbb{R}^d, \|\alpha\|=1} \sum_{i=1}^d \sup_{h\in(\mathbb{R}^d)^{\otimes k-1}, \|h\|_a=1} \alpha_i < \nabla^{k-1} < \nabla\phi, e_i >_a, h >_a \\
&= \sup_{\alpha\in\mathbb{R}^d, \|\alpha\|=1} \sum_{i=1}^d \alpha_i \|\nabla^{k-1} < \nabla\phi, e_i >_a \|_a.
\end{aligned}$$

□

Let us prove Proposition 20 by induction. Take $x \in \mathbb{R}^d$ and let ϕ be a compactly supported smooth function. The inequality holds for $k = 1$ by Equation 5.9. Now, suppose it is true for some $k \in \mathbb{N}$. By Lemma 31 we only need to bound

$$\|\nabla^{k+1} < \nabla P_t\phi, e_i >_a \|_a = \lim_{\epsilon \to 0} \|\nabla^k \left(P_t\phi(x + \epsilon a e_i) - P_t\phi(x)\right) \|_a$$

for any e_i. Let $\epsilon > 0$ and let $(X_t)_{t\geq 0}$ and $(\tilde{X}_t)_{t\geq 0}$ be two diffusion processes with infinitesimal generator $\mathcal{L}_\mu$ started respectively at x and $x + \epsilon a e_1$. Let $t \geq 0$ and let π_ϵ be a coupling between X_t and $\tilde{X}_t$. We have

$$P_t\phi(x + \epsilon a e_1) - P_t\phi(x) = P_t\left(\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right),$$

Applying the induction hypothesis,

$$\left\|\nabla^k P_t\left(\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right)\right\|_a^2 \leq e^{-\rho t \max(2,k)\frac{k-1}{k}} \left(\frac{2\rho d}{e^{2\rho dt/k} - 1}\right)^{k-1} P_{t\frac{k-1}{k}} \left\|\nabla P_{t/k}\left(\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right)\right\|_a^2,$$

and, applying Lemma 30,

$$\left\|\nabla^k P_t\left(\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right)\right\|_a^2 \leq e^{-\rho t(k-1)} d^{k-1} \left(\frac{2\rho}{e^{2\rho dt/k} - 1}\right)^{k} P_t \left|\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right|^2.$$

By Theorem 2.2 [49], Equation 5.9 implies that we can take π_ϵ such that, for any $y \in \mathbb{R}^d$, $\sup_{y\in\mathbb{R}^d} \|\pi_\epsilon(y) - id\|_{a^{-1}} \leq \epsilon e^{-\rho t} + o(\epsilon)$, therefore

$$\lim_{\epsilon\to 0}\left|\frac{\phi \circ \pi_\epsilon - \phi}{\epsilon}\right| = \lim_{\epsilon\to 0}\left|\frac{< \nabla\phi, a^{-1}(\pi_\epsilon - id) >_a + o(\|\pi_\epsilon - id\|)}{\epsilon}\right| \leq e^{-\rho t}\|\nabla\phi\|_a.$$

Since a similar result holds for any e_i, we have, using Lemma 31,

$$\|\nabla^{k+1} P_t\phi\|_a^2 \leq e^{-\rho t(k+1)} d^{k-1} \left(\frac{2\rho}{e^{2\rho dt/k} - 1}\right)^{k} \sup_{\alpha\in\mathbb{R}^d, \|\alpha\|=1} \left(\sum_{i=1}^d \alpha_i \sqrt{P_t\|\nabla\phi\|^2}\right)^2.$$

Finally, since the supremum is obtained for $\alpha_1 = \cdots = \alpha_d = \frac{1}{\sqrt{d}}$, the proof is complete.

5.6.2 Proof of Theorem 25

Let k be a positive integer. For any $x \in (\mathbb{R}^d)^{\otimes k}$, we pose

$$\|x\|_p^p = \sum_{i \in \{1,\dots,d\}^k} |x_i|^p.$$

Let Z be a random variable in $(\mathbb{R}^d)^{\otimes k}$. For any $l \in \{1,\dots,d\}^k$, we pose

$$(Z)_l = Z_{l_1,\dots,l_k}.$$

Before starting the proof of Theorem 25, we first need to derive a multidimensional version of the Rosenthal inequality.

Lemma 32. *Let $k > 0, p \geq 2$ and suppose $Z_1,\dots,Z_n$ are independent random variables taking values in $(\mathbb{R}^d)^{\otimes k}$, then*

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|_p^p\right]^{1/p} &\leq C_p\left(n\|\mathbb{E}[Z]\|_p + n^{1/2}\mathbb{E}[\|Z\|_p^2]^{1/2} + n^{1/p}\mathbb{E}[\|Z\|_p^p]^{1/p}\right)\\ &\leq C_p\left(n\|\mathbb{E}[Z]\| + n^{1/2}\mathbb{E}[\|Z\|^2]^{1/2} + n^{1/p}\mathbb{E}[\|Z\|^p]^{1/p}\right).\end{aligned}$$

Proof. By definition, we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|_p^p\right] = \sum_{l \in \{1,\dots,d\}^k} \mathbb{E}\left[\left|\sum_{i=1}^n (Z_i)_l\right|^p\right].$$

We pose $Z = Z_1$. For any $l \in \{1,\dots,d\}^k$, we know by Rosenthal's inequality (see [15]) that there exists $C_p > 0$ such that

$$\mathbb{E}\left[\left|\sum_{i=1}^n (Z_i)_l\right|^p\right] \leq C_p\left(n^p|\mathbb{E}[(Z)_l]|^p + n^{p/2}\mathbb{E}[(Z)_l^2]^{p/2} + n\mathbb{E}[|(Z)_l|^p]\right).$$

Hence, denoting by Z^2 the random variables taking values in $(\mathbb{R}^d)^{\otimes k}$ such that $(Z^2)_{l_1,\dots,l_k} = (Z)^2_{l_1,\dots,l_k}$,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|_p^p\right] \leq C_p\left(n^p\|\mathbb{E}[Z]\|_p^p + n^{p/2}\|\mathbb{E}[Z^2]\|_{p/2}^{p/2} + n\mathbb{E}[\|Z\|_p^p]\right).$$

Using Jensen's inequality, we have

$$\|\mathbb{E}[Z^2]\|_{p/2} \leq \mathbb{E}[\|Z^2\|_{p/2}] \leq \mathbb{E}[\|Z\|_p^2].$$

Finally, we conclude the proof by remarking that $\|.\|_p \leq \|.\|$. □

We are now ready to start the proof of Theorem 25. Let us pose $X = X_1$, $X' = X_1'$, and $\alpha(t) = e^{2t} - 1$. In the remainder of this proof, we are going to show there exist $C > 0, C_p > 0$ such that

- $\int_0^\infty e^{-t}\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n\|_p^p]^{1/p}dt$,
- $\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\mathbb{E}\left[\left\|\mathbb{E}\left[n\frac{(S'_{n,t}-S_n)^{\otimes 2}}{2} \mid S_n\right] - I_d\right\|_p^p\right]^{1/p} dt$,
- and $\sum_{k=3}^\infty \frac{\|H_k\|_{p,\gamma}}{k!}\int_0^\infty \frac{e^{-kt}}{\sqrt{1-e^{-2t^{k-1}}}}\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n)^{\otimes k} \mid S_n]\|_p^p]^{1/p}dt$,

are bounded by

$$C_p\Bigg(n^{-1/2+(2-q)/2p}\mathbb{E}[\|X\|^{p+q}]^{1/p} + n^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2}$$
$$+\begin{cases}\max(n^{-m/2}, n^{-1/2})\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|\left|\log\left(\frac{n\sqrt{d}}{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}\right)\right| & \text{if } n \le \frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}}e^{C/(1-m)} \\ \frac{1}{1-m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^2 & \text{else if } m < 1 \\ \frac{1}{1-m}\sqrt{d}^{1-1/m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^{2/m} & \text{otherwise} \\ - & \end{cases}\Bigg).$$

Theorem 25 is then obtained using these bounds in either Theorem 18 or Theorem 23 and remarking that

$$\begin{aligned}\mathbb{E}[\|X\|^{2+m}] &= \mathbb{E}\left[\sum_{i\in\{1,\dots,d\}}(X)_i^2\left(\sum_{j\in\{1,\dots,d\}}(X)_j^2\right)^{m/2}\right] \\ &= \sum_{i\in\{1,\dots,d\}}\mathbb{E}\left[(X)_i^2\left(\sum_{j\in\{1,\dots,d\}}(X)_j^2\right)^{m/2}\right] \\ &\le d^{1/2}\left(\sum_{i\in\{1,\dots,d\}}\mathbb{E}\left[(X)_i^2\left(\sum_{j\in\{1,\dots,d\}}(X)_j^2\right)^{m/2}\right]^2\right)^{1/2} \\ &\le d^{1/2}\left(\sum_{i,k\in\{1,\dots,d\}}\mathbb{E}\left[(X)_i(X)_k\left(\sum_{j\in\{1,\dots,d\}}(X)_j^2\right)^{m/2}\right]^2\right)^{1/2},\end{aligned}$$

leads to

$$\mathbb{E}[\|X\|^{2+m}] \le d^{1/2}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|. \tag{5.10}$$

Remark 33. For $d > 1$ and $p > 2$, these bounds are not optimal. For instance, whenever p is an even integer, one can replace all 2-norms $\|.\|$ by p-norms $\|.\|_p$ by taking

$$S'_{n,t} = S_n + n^{-1/2}(X'_I - X_I)1_{\|X'_I\|_p, \|X_I\|_p \leq \sqrt{n(e^{2t}-1)}}.$$

In the remainder of this proof, C_p denotes a generic constant depending only on p and C a generic universal constant. For any $t \geq 0$, we have, by definition of $S'_{n,t}$,

$$S'_{n,t} - S_n = \frac{1}{\sqrt{n}}(X'_I - X_I)1_{\|X_I\|, \|X'_I\| \leq \sqrt{n\alpha(t)}}.$$

Moreover, since I is independent from S_n, then, for any integer $k > 0$,

$$\begin{aligned}\mathbb{E}[n(S'_{n,t} - S_n)^{\otimes k} \mid S_n] &= \mathbb{E}\left[n^{1-k/2}\mathbb{E}_I[(X'_I - X_I)1_{\|X_I\|, \|X'_I\| \leq \sqrt{n\alpha(t)}}] \mid S_n\right] \\ &= n^{-k/2}\mathbb{E}\left[\sum_{i=1}^{n}(X'_i - X_i)^{\otimes k}1_{\|X_i\|, \|X'_i\| \leq \sqrt{n\alpha(t)}} \mid S_n\right].\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n &= \mathbb{E}[n(S'_{n,t} - S_n) + S_n \mid S_n] \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\left[\sum_{i=1}^{n}(X'_i - X_i)1_{\|X_i\|, \|X'_i\| \leq \sqrt{n\alpha(t)}} + X_i \mid S_n\right].\end{aligned}$$

Let us pose

$$Z = \mathbb{E}[(X' - X)1_{\|X\|, \|X'\| \leq \sqrt{n\alpha(t)}} + X].$$

Since X' is independent from S_n, $\mathbb{E}[X' \mid S_n] = \mathbb{E}[X'] = 0$. Hence,

$$\mathbb{E}[X'1_{\|X\|, \|X'\| \leq \sqrt{n\alpha(t)}} \mid S_n] = -\mathbb{E}[X'1_{\max \|X\|, \|X'\| \geq \sqrt{n\alpha(t)}} \mid S_n].$$

Therefore

$$\mathbb{E}[Z \mid S_n] = \mathbb{E}[(X - X')1_{\max \|X\|, \|X'\| \geq \sqrt{n\alpha(t)}} \mid S_n],$$

and

$$\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n\|_p^p]^{1/p} = n^{-1/2}\mathbb{E}\left[\left\|\mathbb{E}\left[\sum_{i=1}^{n}(X_i - X'_i)1_{\max \|X_i\|, \|X'_i\| \geq \sqrt{n\alpha(t)}} \mid S_n\right]\right\|_p^p\right]^{1/p}.$$

Applying Jensen's inequality to get rid of the conditional expectation,

$$\mathbb{E}[\|\mathbb{E}[n(S'_{n,t}-S_n) \mid S_n]+S_n\|_p^p]^{1/p} \leq n^{-1/2}\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - X'_i)1_{\max\|X_i\|,\|X'_i\|\geq\sqrt{n\alpha(t)}}\right\|_p^p\right]^{1/p}.$$

Let us pose

$$Y = (X - X')1_{\max\|X\|,\|X'\|\geq\sqrt{n\alpha(t)}}.$$

Since the $(X_i)_{1\leq i\leq n}, (X'_i)_{1\leq i\leq n}$ are i.i.d. random variables, so are the $((X_i - X'_i)1_{\max\|X_i\|,\|X'_i\|\geq\sqrt{n\alpha(t)}})_{1\leq i\leq n}$. Hence, we can apply Lemma 32 to obtain

$$\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n\|_p^p]^{1/p} \leq C_p(n^{1/2}\|\mathbb{E}[Y]\| + \mathbb{E}[\|Y\|^2]^{1/2} + n^{1/p-1/2}\mathbb{E}[\|Y\|^p]^{1/p}).$$

Since X and X' follow the same law, $\mathbb{E}[Y] = 0$. On the other hand, we have

$$\begin{aligned}\mathbb{E}[\|Y\|^p]^{1/p} &= \mathbb{E}[\|X - X'\|^p 1_{\max\|X\|,\|X'\|\geq\sqrt{n\alpha(t)}}]^{1/p}\\ &\leq 2\mathbb{E}[\|X\|^p 1_{\max\|X\|,\|X'\|\geq\sqrt{n\alpha(t)}}]^{1/p}\\ &\leq 2\mathbb{E}[\|X\|^p 1_{\|X\|\geq\sqrt{n\alpha(t)}}]^{1/p}\\ &\leq 2(n\alpha(t))^{-q/2p}\mathbb{E}[\|X\|^{p+q}]^{1/p},\end{aligned}$$

and, similarly,

$$\mathbb{E}[\|Y\|^2]^{1/2} \leq 2(n\alpha(t))^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2}.$$

Overall, we obtained

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n\|_p^p]^{1/p} \leq\\ &C_p\left(n^{-m/4}\alpha(t)^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2} + n^{-1/2+(2-q)/2p}\alpha(t)^{-q/2p}\mathbb{E}[\|X\|^{p+q}]^{1/p}\right).\end{aligned}$$

Finally, using the bound $\alpha(t) \geq 2t$ in the following integral,

$$\begin{aligned}\int_0^\infty e^{-t}\mathbb{E}[\|\mathbb{E}[n(S'_{n,t} - S_n) \mid S_n] + S_n\|_p^p]^{1/p}dt \leq\\ C_p\left(n^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2} + n^{-1/2+(2-q)/2p}\mathbb{E}[\|X\|^{p+q}]^{1/p}\right).\end{aligned}$$

Let us now tackle the second order term. We have

$$\begin{aligned}\mathbb{E}\left[n\frac{(S'_{n,t} - S_n)^{\otimes 2}}{2} \mid S_n\right] - I_d &= \mathbb{E}\left[\frac{(X'_I - X_I)^{\otimes 2}}{2}1_{\|X_I\|,\|X'_I\|\leq\sqrt{n\alpha(t)}} \mid S_n\right] - I_d\\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \frac{(X'_i - X_i)^{\otimes 2}}{2}1_{\|X_i\|,\|X'_i\|\leq\sqrt{n\alpha(t)}} - I_d \mid S_n\right].\end{aligned}$$

Again, taking

$$Y = \frac{(X'-X)^{\otimes 2}}{2} 1_{\|X\|,\|X'\| \leq \sqrt{n\alpha(t)}} - I_d$$

and using a combination of Jensen's inequality and Lemma 32, we obtain

$$\mathbb{E}\left[\left\| \mathbb{E}\left[n \frac{(S'_{n,t} - S_n)^{\otimes 2}}{2} \mid S_n \right] - I_d \right\|_p^p \right]^{1/p} \leq$$
$$C_p \left(\| \mathbb{E}[Y] \| + n^{-1/2} \mathbb{E}\left[\|Y\|^2 \right]^{1/2} + n^{1/p-1} \mathbb{E}\left[\|Y\|^p \right]^{1/p} \right).$$

First, since $\mathbb{E}[X^{\otimes 2}] = \mathbb{E}[X'^{\otimes 2}] = I_d$,

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{(X'-X)^{\otimes 2}}{2} 1_{\max \|X\|,\|X'\| \geq \sqrt{n\alpha(t)}} \right].$$

For two $x, y \in \left(\mathbb{R}^d\right)^{\otimes k}$, we denote by $< x, y >$ the corresponding Hilbert-Schmidt scalar product between x and y. Letting Z and Z' be two random variables such that X, X', Z, Z' are i.i.d., we have

$$\begin{aligned}
\| \mathbb{E}[Y] \| &= \sqrt{\left\langle \mathbb{E}\left[(X'-X)^{\otimes 2} 1_{\max \|X\|,\|X'\| \geq \sqrt{n\alpha(t)}} \right], \mathbb{E}\left[(Z'-Z)^{\otimes 2} 1_{\max \|Z\|,\|Z'\| \geq \sqrt{n\alpha(t)}} \right] \right\rangle} \\
&= \sqrt{\mathbb{E}\left[\langle (X'-X)^{\otimes 2}, (Z'-Z)^{\otimes 2} \rangle 1_{\max \|X\|,\|X'\| \geq \sqrt{n\alpha(t)}} 1_{\max \|Z\|,\|Z'\| \geq \sqrt{n\alpha(t)}} \right]} \\
&= \sqrt{\mathbb{E}\left[\langle (X'-X), (Z'-Z) \rangle^2 1_{\max \|X\|,\|X'\| \geq \sqrt{n\alpha(t)}} 1_{\max \|Z\|,\|Z'\| \geq \sqrt{n\alpha(t)}} \right]} \\
&\leq C \sqrt{\mathbb{E}\left[< X, Z >^2 1_{\max \|X\|,\|X'\| \geq \sqrt{n\alpha(t)}} 1_{\max \|Z\|,\|Z'\| \geq \sqrt{n\alpha(t)}} \right]} \\
&\leq C(n\alpha(t))^{-m/2} \sqrt{\mathbb{E}\left[< X, Z >^2 \max(\|X\|, \|X'\|)^m \max(\|Z\|, \|Z'\|)^m \right]} \\
&\leq C(n\alpha(t))^{-m/2} \sqrt{\mathbb{E}\left[< X, Z >^2 (\|X\|^m + \|X'\|^m + \|Z\|^m + \|Z'\|^m) \right]} \\
&\leq C(n\alpha(t))^{-m/2} \|\mathbb{E}[X^{\otimes 2}(\|X\|^m + \|X'\|^m)]\| \\
&\leq C(n\alpha(t))^{-m/2} (\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\| + \mathbb{E}[X^{\otimes 2}\|X'\|^m]\|)
\end{aligned}$$

Since X and X' are independent,

$$\begin{aligned}
\|\mathbb{E}[X^{\otimes 2}\|X'\|^m]\| &= \sqrt{d}\mathbb{E}[\|X\|^m] \\
&\leq d^{-1/2}\mathbb{E}[\|X\|^2]\mathbb{E}[\|X\|^m] \\
&\leq d^{-1/2}\mathbb{E}[\|X\|^{2+m}]^{2/(2+m)}\mathbb{E}[\|X\|^{2+m}]^{m/(2+m)} \\
&\leq d^{-1/2}\mathbb{E}[\|X\|^{2+m}],
\end{aligned}$$

and, since we know by Equation 5.10 that $\mathbb{E}[\|X\|^{2+m}] \leq \sqrt{d}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|$, we obtain

$$\|\mathbb{E}[Y]\| \leq C(n\alpha(t))^{-m/2}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|.$$

Let us remark that this causes integration issues when $m \geq 1$, as

$$\frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\alpha(t)^{-m/2} = \frac{1}{\alpha(t)^{(m+1)/2}} \sim_{t\to 0} \frac{1}{(2t)^{(m+1)/2}}.$$

In order to deal with this problem, we remark that, replacing m by 0 in the previous bound, we can also obtain

$$\|\mathbb{E}[Y]\| \leq C\|\mathbb{E}[X^{\otimes 2}]\| \leq C\sqrt{d}.$$

Then, taking some $0 < t_0 < 1$,

$$\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\|\mathbb{E}[Y]\|dt \leq$$
$$\int_0^{t_0} \frac{C\sqrt{d}e^{-2t}}{\sqrt{1-e^{-2t}}}dt + Cn^{-m/2}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\| \int_{t_0}^\infty \alpha(t)^{-(m+1)/2}.$$

Let $m \neq 1$. Since $\alpha(t) \geq 2t$ and $\alpha(t) \geq e^t$ for $t \geq 1$, we can decompose the previous bound further:

$$\int_{t_0}^\infty \alpha(t)^{-(m+1)/2} \leq \int_{t_0}^1 (2t)^{-(m+1)/2}dt + \int_1^\infty e^{-t}dt \leq C\left(1 + \frac{1-t_0^{(1-m)/2}}{1-m}\right).$$

At this point, there are three possibilities:

- If $t_0^{(1-m)/2} \leq 1/2$, then $\int_{t_0}^\infty \alpha(t)^{-(m+1)/2} \leq \frac{C}{1-m}$.
- If $t_0^{(1-m)/2} \geq 3/2$, then $\int_{t_0}^\infty \alpha(t)^{-(m+1)/2} \leq \frac{Ct_0^{(1-m)/2}}{1-m}$.
- If not, then $|\log(1/t_0)(m-1)| \leq C$ and

$$\begin{aligned}|1 - t_0^{(1-m)/2}| &= |1 - e^{\frac{1-m}{2}\log(t_0)}| \\ &\leq C|(m-1)\log(1/t_0)|.\end{aligned}$$

Thus

$$\frac{1-t_0^{(1-m)/2}}{1-m} \leq \log(1/t_0).$$

Therefore, taking $t_0 = \frac{1}{n}\left(\frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}}\right)^{2/m}$, there exists $C > 0$ such that

$$\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\|\mathbb{E}\left[Y\right]\|dt$$
$$\leq C\begin{cases}\max(n^{-m/2}, n^{-1/2})\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|\left|\log\left(\frac{n\sqrt{d}}{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}\right)\right| & \text{if } n \leq \frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}}e^{C/(1-m)} \\ \frac{1}{1-m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^2 & \text{else if } m < 1 \\ \frac{1}{1-m}\sqrt{d}^{1-1/m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^{2/m} & \text{otherwise}\end{cases}.$$

Let us now deal with the moments of Y. We have

$$\begin{aligned}\mathbb{E}[\|Y\|^p] &\leq 2^p(\mathbb{E}[\|\frac{(X'-X)^{\otimes 2}}{2}\|^p] + \|I_d\|^p) \\ &\leq 2^p(\mathbb{E}[\|X\|^p + d) \\ &\leq 2^p(\mathbb{E}[\|X\|^{2p}1_{\|X\|\leq\sqrt{n\alpha(t)}}] + \mathbb{E}[\|X\|^2]) \\ &\leq 2^p(1 + (n\alpha(t))^{(p-q)/2})\mathbb{E}[\|X\|^{p+q}].\end{aligned}$$

and

$$\mathbb{E}[\|Y\|^2] \leq 4(1 + (n\alpha(t))^{m-2})\mathbb{E}[\|X\|^{2+m}].$$

Putting everything together,

$$\int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\mathbb{E}\left[\left\|\mathbb{E}\left[n\frac{(S'_{n,t}-S_n)^{\otimes 2}}{2} \mid S_n\right] - I_d\right\|_p^p\right]^{1/p} dt \leq$$
$$C_p\Bigg(n^{-1/2+(2-q)/2p}\mathbb{E}[\|X\|^{p+q}] + n^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2}$$
$$+\begin{cases}\max(n^{-m/2}, n^{-1/2})\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|\left|\log\left(\frac{n\sqrt{d}}{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}\right)\right| & \text{if } n \leq \frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}}e^{C/(1-m)} \\ \frac{1}{1-m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^2 & \text{else if } m < 1 \\ \frac{1}{1-m}\sqrt{d}^{1-1/m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^{2/m} & \text{otherwise}\end{cases}\Bigg).$$

We are now left with dealing with the higher order terms. For $k > 2$, let

$$Y = \mathbb{E}[(X'-X)^{\otimes k}1_{\|X\|,\|X'\|\leq\sqrt{n\alpha(t)}} \mid S_n].$$

Then, by a combination of Jensen's inequality and Lemma 32,

$$\mathbb{E}[\|\mathbb{E}[n(S'_{n,t}-S_n)^{\otimes k} \mid S_n]\|_p^p]^{1/p} \leq$$
$$n^{1-k/2}\|\mathbb{E}[Y]\| + n^{1/2-k/2}\mathbb{E}[\|Y\|^2]^{1/2} + n^{1/p-k/2}\mathbb{E}[\|Y\|^p]^{1/p}.$$

First, we have

$$\begin{aligned}\mathbb{E}[\|Y\|^p] &\leq \mathbb{E}[\|X'-X\|^{kp}1_{\|X\|,\|X'\|\leq\sqrt{n\alpha(t)}}] \\ &\leq 2^{kp}\mathbb{E}[\|X\|^{kp}1_{\|X\|\leq\sqrt{n\alpha(t)}}] \\ &\leq 2^{kp}(n\alpha(t))^{((k-1)p-q)/2}\mathbb{E}[\|X\|^{p+q}],\end{aligned}$$

and

$$\mathbb{E}[\|Y\|^2] \leq 4^k(n\alpha(t))^{k-1-m/2}\mathbb{E}[\|X\|^{2+m}].$$

Then, since X' and X are i.i.d., $\mathbb{E}[Y]=0$ for odd values of k. Let us now consider an even integer $k>2$. Denoting by Z and Z' two random variables such that X, X', Z, Z' are i.i.d. Following the computations performed to bound the second order term, we obtain

$$\begin{aligned}\|\mathbb{E}[Y]\| &= \mathbb{E}\left[< X'-X, Z'-Z >^k 1_{\|X\|,\|X'\|\leq\sqrt{n\alpha(t)}}1_{\|Z\|,\|Z'\|\leq\sqrt{n\alpha(t)}}\right]^{1/2} \\ &\leq 2^k\mathbb{E}\left[< X, Z >^k 1_{\|X\|,\|Z\|\leq\sqrt{n\alpha(t)}}\right]^{1/2} \\ &\leq 2^k\mathbb{E}\left[< X, Z >^2 \|X\|^{k-2}\|Z\|^{k-2}1_{\|X\|,\|Z\|\leq\sqrt{n\alpha(t)}}\right]^{1/2} \\ &\leq 2^k(n\alpha(t))^{(k-m-2)/2}\mathbb{E}\left[< X'-X, Z'-Z >^2 \|X\|^m\|Z\|^m\right]^{1/2} \\ &\leq 2^k(n\alpha(t))^{(k-m-2)/2}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|,\end{aligned}$$

and, similarly,

$$\|\mathbb{E}[Y]\| \leq 2^k(n\alpha(t))^{(k-2)/2}\|\mathbb{E}[X^{\otimes 2}]\| \leq 2^k(n\alpha(t))^{k/2-1}\sqrt{d}.$$

Then, using the same integration procedure we used to bound the second order term, we obtain

$$\begin{aligned}&\int_0^\infty \frac{e^{-kt}}{\sqrt{1-e^{-2t^{k-1}}}}\mathbb{E}[\|\mathbb{E}[n(S'_{n,t}-S_n)^{\otimes k} \mid S_n]\|_p^p]^{1/p}dt \\ &\qquad \leq C_p 2^k\Bigg(n^{-1/2+(2-q)/2p}\mathbb{E}[\|X\|^{p+q}] + n^{-m/4}\mathbb{E}[\|X\|^{2+m}]^{1/2} \\ &+\begin{cases}\max(n^{-m/2},n^{-1/2})\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|\left|\log\left(\frac{n\sqrt{d}}{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}\right)\right| & \text{if } n\leq\frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}}e^{C/(1-m)} \\ \frac{1}{1-m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^2 & \text{else if } m<1 \\ \frac{1}{1-m}\sqrt{d}^{1-1/m}\left(n^{-m/4}\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2}\right)^{2/m} & \text{otherwise}\end{cases}\Bigg).\end{aligned}$$

Then, since by Remark 24, $\sum_{k=0}^{\infty} \frac{2^k \|H_k\|_{p,\gamma}}{k!} < \infty$,

$$\sum_{k=0}^{\infty} \frac{\|H_k\|_{p,\gamma}}{k!} \int_0^{\infty} \frac{e^{-kt}}{\sqrt{1-e^{-2t}}^{k-1}} \mathbb{E}[\|\mathbb{E}[n(S'_{n,t}-S_n)^{\otimes k} \mid S_n]\|_p^p]^{1/p} dt$$

$$\leq C_p \Bigg(n^{-1/2+(2-q)/2p} \mathbb{E}[\|X\|^{p+q}] + n^{-m/4} \mathbb{E}[\|X\|^{2+m}]^{1/2}$$

$$+ \begin{cases} \max(n^{-m/2}, n^{-1/2}) \|\mathbb{E}[X^{\otimes 2}\|X\|^m]\| \left| \log\left(\frac{n\sqrt{d}}{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|} \right) \right| & \text{if } n \leq \frac{\|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|}{\sqrt{d}} e^{C/(1-m)} \\ \frac{1}{1-m} \left(n^{-m/4} \|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2} \right)^2 & \text{else if } m < 1 \\ \frac{1}{1-m} \sqrt{d}^{1-1/m} \left(n^{-m/4} \|\mathbb{E}[X^{\otimes 2}\|X\|^m]\|^{1/2} \right)^{2/m} & \text{otherwise} \end{cases} \Bigg),$$

which is the last bound required to conclude the proof.

5.6.3 Proof of Proposition 29

By construction,

- $\mathbb{E}[\frac{\xi}{s} - \nabla u(X) \mid X] = 0$;
- $\mathbb{E}[\|\mathbb{E}[\frac{\xi^{\otimes 2}}{2s} - I_d \mid X]\|^2]^{1/2} = \frac{s}{2} \mathbb{E}[\|\nabla u(X)\|^4]^{1/2}$;

Since $\nabla u(0)$ is assumed to be 0 and Lipschitz continuous, we have $\|\nabla u(X)\| \leq L_1 \|X\|$. Let us bound $\mathbb{E}[\|X\|^2]$. Since X and $X + \xi$ have the same law,

$$\begin{aligned} \mathbb{E}[\|X\|^2] &= \mathbb{E}[\|X+\xi\|^2] \\ &= \mathbb{E}[\|X + s\nabla u(X)\|^2] + 2ds \\ &\leq \mathbb{E}[\|X\|^2 + 2sX.\nabla u(X) + s^2 \|\nabla(u)(X)\|^2] + 2ds. \end{aligned}$$

And, since μ is log-concave,

$$\mathbb{E}[\|X\|^2] \leq (1 - 2s\rho + L_1^2 s^2) \mathbb{E}[\|X\|^2] + 2ds,$$

therefore

$$\mathbb{E}[\|X\|^2] \leq \frac{2ds}{2s\rho - L_1^2 s^2} \leq \frac{d}{\rho} + O(s).$$

Now, since μ is strongly log-concave and by construction of our increments, $\|X\| < \frac{\sqrt{2}}{\rho\sqrt{sd}}$. Therefore there exists $C > 0$ such that $\mathbb{E}[\|\nabla u(X)\|^4]^{1/2} \leq \left(\frac{d}{s}\right)^{1/2}$. Let B be a multivariate Rademacher random variable, for $k \geq 3$,

$$\mathbb{E}[|\mathbb{E}[(\xi)^{\otimes k}] \mid X|^2]^{1/2} = \sum_{j=0}^{k} \binom{k}{j} 2^{j/2} s^{k-j/2} \mathbb{E}[\|\nabla u(X)\|^{2(k-j)}]^{1/2} \|\mathbb{E}[B^{\otimes j}]\|$$

For $j > 0$, we have

$$\|\mathbb{E}[B^{\otimes j}]\| = d^{(j-1)/2},$$

and, for $j < \frac{k}{2}$, taking

$$\mathbb{E}[\|X\|^{2(k-2j)}]^{1/2} \leq s^{k-2j-1}\mathbb{E}[\|X\|^2]^{1/2} \leq \frac{s^{k-2j-1}d^{1/2}}{\rho} + O(s),$$

completes the proof.

Chapter 6

Topological Pooling

In this Chapter, we incorporate the concept of 0-dimensional persistence presented in Section 4.1.2 in the bag-of-words framework to perform shape recognition. This work was done in collaboration with Maks Ovsjanikov, Steve Oudot and Frédéric Chazal and was the subject of a publication in the 6th International Workshop on Computational Topology in Image Context [14].

6.1 The bag-of-words pipeline

The bag-of-words approach consists of three main steps: feature extraction, coding and pooling. We assume that the input to the pipeline is a set of M 3D-shapes $(G_i)_{i\in[1,M]}$ represented as triangle meshes with vertices $(V_i)_{i\in[1,M]}$.

Feature extraction aims at deriving a meaningful representation of the shape: the feature function denoted as $\mathcal{F}_i : V_i \to \mathbb{R}^N$. These functions are usually obtained by computing local descriptors (such as HKS [80], SIHKS [17], WKS [6], Shape-net features [59], etc.) on each vertex of the mesh.

The purpose of *coding* is to decompose the values of the functions $\mathcal{F}_i$ by projecting them on a set of points $W = (w_k)_{k\in[|1,K|]} \in \mathbb{R}^N$ called a *codebook*. This allows to replace each feature function by a family of functions $(C_i : V_i \to \mathbb{R}^K)_{i\in[1,M]}$, called the *word functions*. In other words, for a coding procedure $Coding$ and codebook W, the C_i are defined through

$$\forall x \in V_i, C_i(x) = Coding(\mathcal{F}_i(x), W).$$

There exist various coding methods, such as Vector Quantization [75], Sparse Coding [89], Locally Constrained Linear Coding [88], Fisher Kernel [66] or Supervector [92]. The codebook is usually computed using K-means but supervised codebook learning methods [88], [16] generally achieve bet-

ter accuracy. In the Sparse Coding approach, the one we use in this paper, W and C are computed on the training set following

$$\min_{(C_i)_{i\in[1,M]},W} \sum_{i=1}^{M} \sum_{x\in V_i} \left(\|\mathcal{F}_i(x) - WC_i(x)\|_2^2 + \lambda \|C_i(x)\|_1 \right),$$

with constraint $\|w_i\| \leq 1$ and regularization parameter λ. During the testing phase, the optimization is only performed on C with the codebook already computed.

The *pooling* step aims at summarizing properties of the family $(C_i)_{i\in[1,M]}$ and representing them through vectors $(\mathcal{P}_i)_{i\in[1,M]}$ which can then be used in standard learning algorithms such as SVM (Support Vector Machine). Usually, the pooling method depends on the coding scheme used. For Vector Quantization, one traditionally uses the mean of the word functions, an approach called sum-pooling

$$\mathcal{P}_i = \left(\frac{\sum_{x\in V_i}(C_i(x))_1}{|V_i|}, ..., \frac{\sum_{x\in V_i}(C_i(x))_K}{|V_i|}\right).$$

The other traditional pooling scheme, called max-pooling, was introduced along the Sparse Coding scheme by Yang *et al.* in [89]. With this pooling technique, word functions are summarized by their maxima

$$\mathcal{P}_i = (\max_{x\in V_i}(C_i(x))_1, ..., \max_{x\in V_i}(C_i(x))_K).$$

Several works have highlighted the improvement obtained using max-pooling rather than sum-pooling, both in terms of accuracy and, since it can be used with a linear kernel in learning algorithms, of computational scalability [89, 16]. The strength of max pooling is due in part to its remarkable robustness properties as it is invariant with respect to many transformations a shaper can undergo such that translations, rotations or changes of scale. On the other hand, the information captured by this pooling procedure is rather restricted.

One of the main assumptions made in designing and studying the bag-of-words approach is that the values of the word functions are i.i.d. random variables. Refinements of the max-pooling scheme have been proposed under this assumption: for instance [55] proposed to consider the k highest values for each words. However, the independence assumption of the word functions is unrealistic: the values of the word functions on close vertices of the mesh of a 3D-shape tend to be similar, as illustrated in Figure 6.1. Thus, in this example, the generalization proposed by [55] ends up capturing the same feature multiple times and providing multiple redundant

values. On the other hand, an important enhancement of the pooling procedure called Spatial pyramid matching [51] was proposed in the image processing literature. The idea behind this improvement was to perform pooling separately on different parts, quadrants, stripes, etc. of an image thus gaining some information regarding the spatial distribution of the word functions. This approach has drastically improved the performance of the bag-of-words procedures on multiple datasets, contradicting the identically distributed assumption. Let us note there exist adaptations of this technique to 3D shape [56, 54], but they are not as efficient as the original Spatial pyramid matching for images.

As we have seen, instead of considering word functions as an unordered collection of independent random values, it seems more reasonable to consider them as random functions defined on the vertices of a graph. We thus propose to use the 0-dimensional persistent homology to capture information regarding the global structure of the word functions which is not available for the traditional max-pooling approach. In this work, we aim at being able to use classification algorithms such as SVM or logistic regression requiring a Hilbert space structure, which the space of persistence diagrams lacks. One approach to tackle this issue is to make use of the "kernel trick" by using a positive-definite kernel in order to map the persistence diagrams into a Hilbert space. As recently shown by Reininghaus et al. [69], one cannot rely on natural distances such as the Wasserstein distance to build traditional distance-based kernels. This led the authors to propose a new non-linear. But using a non-linear kernel increases the computational complexity of classification procedures leading to scalability issues. Another approach to directly embed persistence diagrams into a Hilbert Space was proposed in [18] but this embedding is highly memory-consuming as it maps a single diagram into a set of functions. It is thus not appropriate for dealing with large datasets.

In this work, we propose to perform pooling by computing the persistence diagrams of each word function. We then map these persistence diagrams into $\mathbb{R}^d$ for some reasonable value of d –< 20– by considering the peaks with highest prominence. Since we provide a direct mapping of persistence diagrams into $\mathbb{R}^d$, we can use it for the pooling stage for the bag-of-words procedure and achieve good performance with respect to the classification phase. We call this pooling approach Topological Pooling. Since it relies on persistence diagrams, this method is stable with respect to most transformations the shape can undergo: translations, rotations, etc., as long as the descriptors used in input are also invariant to these transformations. Moreover, we show that this pooling approach is robust to perturbations of the descriptors. Finally we demonstrate the validity of our

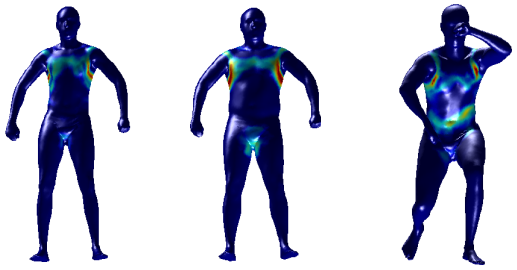


Figure 6.1: Example of a word function obtained on two different people in the same pose (left and middle) and on the same person in two different poses (middle and right).

approach compared to both sum-pooling and max-pooling by performing pose recognition on the SHREC 2014 dataset.

6.2 Using persistence diagrams for pooling

Given a persistence diagram Δ, let us recall that the prominence p of a point $(b, d) \in \Delta$ is defined by $p = b - d$. Given a function f on a graph G, we define the infinite-dimensional Topological Pooling vector of f with $i-th$ coordinate given by

$$TopoPool(f)_i = p_i(\Delta_f),$$

where $p_i(\Delta_f)$ is the i-th highest prominence of the points of Δ_f if there is at least i points in Δ_f and 0 otherwise. The stability of persistence diagrams given in Equation 4.2 implies the stability of our our pooling scheme.

Proposition 34. *Let G be a graph and f and g two real-valued functions on the vertices V of a graph G. Then, for any integer $i > 0$,*

$$|TopoPool(f)_i - TopoPool(g)_i| \leq 2 \sup_{x \in V} |f(x) - g(x)|$$

Of course, in practice we cannot use an infinite-dimensional vector so we simply consider a truncation of this vector keeping the first n coordinates, we denote such a truncated pooling vector by "TopoPool-n". Using the notations of Section 6.1, given some $n > 0$, the pooling vectors $(\mathcal{P}_i)_{1 \leq i \leq M}$ are vectors of dimension nK defined by

$$\mathcal{P}_i = (\text{TopoPool-n}((C_i)_1), ..., \text{TopoPool-n}((C_i)_K)).$$

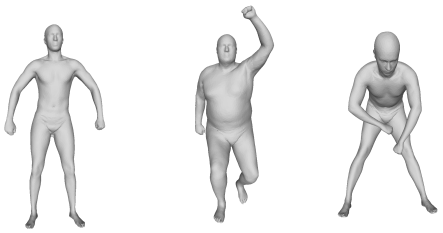


Figure 6.2: The real SHREC 2014 dataset

6.3 Experiments

In this section we evaluate the sum-pooling, the max-pooling and our topological pooling approaches on the SHREC 2014 dataset "Shape Retrieval of Non-Rigid 3D Human Models" [67], which we modify by applying a random rotation to each shape. The dataset is composed of 400 meshes corresponding to 40 subjects taking 10 different poses and we wish to classify each of these meshes with respect to the pose taken by the subject. We consider both SIHKS features [17] and curvature-based features corresponding to the unary features from [46] and composed of 64. We use Sparse Coding [89] for the coding step and the computation are performed using the SPAMS toolbox [58]. Finally, the classification is performed using a Support Vector Machine. We use 3 shapes per class for the training set, 2 for the validation set and 5 for the testing set. We compare the traditional sum-pooling with our TopoPool-n with different values for n –remark that $n = 1$ is equivalent to max-pooling– and under different codebook sizes. As a baseline, we also display the results obtained using a rigid Iterated Closest Point (ICP) [10] along with a 1-nearest neighbour classification, which aims at iteratively minimizing the distance between two point clouds through rigid deformations. In our case it corresponds to finding the correct rotation to align the shapes as two shapes in a similar pose are close, however the approach can fail if it gets stuck in a local minimum and is not able to recover the correct rotation. We run the experiment a hundred times, selecting the training and testing sets at random. We display the mean accuracy over the multiple runs in Table 6.1.

Overall, and especially for the SIHKS features, the Topological Pooling scheme outperforms both max-pooling and to the sum-pooling. In the case of curvature features, Topological Pooling and sum-pooling gives similar accuracy results for large codebooks but in the case of smaller codebooks, Topological pooling gives much better results. It is interesting to notice that the gap between the different pooling scheme decreases as the size of the codebook increases. Indeed, the smaller the codebook is, the richer each

Pooling / Codebook size	40	60	80	100	120	140	160	180	200
SIHKS features									
Sum-Pooling	0.53	0.56	0.60	0.60	0.58	0.62	0.61	0.60	0.60
TopoPool-1	0.46	0.55	0.53	0.54	0.58	0.59	0.63	0.64	0.64
TopoPool-5	0.69	0.71	0.69	0.70	0.73	0.70	0.74	0.73	0.72
TopoPool-10	0.70	0.71	0.71	0.69	0.72	0.71	0.73	0.74	0.72
TopoPool-15	0.72	0.73	0.71	0.70	0.74	0.71	0.74	**0.75**	0.71
TopoPool-20	0.72	0.73	0.70	0.72	0.73	0.72	0.73	0.75	0.73
Curvature features									
Sum-Pooling	0.80	0.80	0.84	0.85	0.88	0.88	0.87	0.88	0.89
TopoPool-1	0.39	0.56	0.56	0.57	0.64	0.69	0.69	0.73	0.76
TopoPool-5	0.63	0.79	0.80	0.80	0.82	0.85	0.86	0.87	0.86
TopoPool-10	0.74	0.85	0.85	0.86	0.86	0.87	0.89	0.89	0.88
TopoPool-15	0.78	0.85	0.87	0.87	0.88	0.89	0.89	**0.90**	0.90
TopoPool-20	0.79	0.88	0.88	0.88	0.88	0.89	0.90	0.90	0.89
ICP	0.55								

Table 6.1: Mean accuracy obtained on the SHREC 2014 dataset.

word function are in terms of topology –and thus the richer the corresponding persistence diagrams are–.

Chapter 7

Perspectives

In this thesis, we have mainly dealt with using and studying the convergence of random walks on random geometric graphs to diffusion processes. In this context, we first designed a new soft clustering algorithm for the mode seeking framework using such random walks and we provided a way to bound the 2-Wasserstein distance between the invariant measure of random walks on random geometric graphs and the invariant measure of the limiting diffusion process. Yet, there are still many questions left unanswered, either linked to the algorithm we have proposed or regarding the converge of random walks itself.

Soft-clustering algorithm. There are two main issues regarding our soft clustering algorithm. The first one is related to the choice of the temperature parameter β. While we have shown that the evolution of the fuzziness of the clustering through the clustering entropy can give some intuition regarding interesting values of β, it is not clear whether this approach would be efficient for more complicated datasets. Finding a way to automatically select β would prove extremely useful. The second shortcoming of our algorithm is linked to the construction of the cluster cores: our definition is rather ad hoc and it would be interesting to find a more canonical one.

Pointwise convergence of the invariant measure of random walks on random geometric graphs. In this thesis, we have obtained bounds in terms of Wasserstein distance for the convergence of the invariant measures of random walks on random geometric graphs to the invariant measures of diffusion processes. However, this result is still far from the pointwise convergence that would be required to prove we can compute a density estimator from the invariant measures of random walks on random geometric graphs. As highlighted by [44], weak convergence or 2-Wasserstein

distance can be turned into pointwise convergence as long as the invariant measure is sufficiently regular. This sounds reasonable as the structure of random geometric graphs itself is rather regular. Yet, it is unclear how such a result could be obtained.

Stochastic homogenization. Among the multiple open questions regarding the convergence of random walks on random geometric graphs, finding the minimal assumptions required for this convergence to hold is of first importance. As we have seen multiple times in this thesis, for the convergence to hold, the graph must be built in a certain way. Theorem 11 requires the window size $h(n)$ to satisfy $h_n \to 0$ and $\frac{nh_n^{d+2}}{\log n} \to 0$. But, if we consider the convergence of the invariant measures on an ϵ-graph, with $\epsilon = h_n$, it is sufficient that $h_n \to 0$ and $\frac{nh_n^d}{\log n} \to 0$ for the invariant measure to converge. Similarly, while most results regarding the convergence of the spectrum of graph Laplacians also requires the stronger assumption on the window size, it has been shown in [38] the weaker assumption was sufficient for this convergence to hold. Hence, one may wonder whether our results as well as other standard results regarding the convergence of random walks on random geometric graphs hold under the weaker assumption on the window size. In order to answer this question, it may be interesting to look into other approaches used to study random walks in random environments in general. For instance, results of stochastic homogenization [40] are of particular interest.

Stein's method and spectral convergence. It would also be interesting to see if Stein's method could be used to provide other results for the convergence of random walks to diffusion processes. For instance, let us consider the unidimensional Ornstein-Uhlenbeck semigroup $(P_t)_{t\geq 0}$ and let $\mathcal{L}_\gamma$ be its infinitesimal generator. Let $\mathcal{L}_\nu$ be the generator of a Markov chain with invariant measure ν and let ψ be the eigenvector corresponding to the first non-zero eigenvalue of $\mathcal{L}_\nu$. The approach Chapter 5 relies on the fact that, for any test function ϕ, $P_t\phi$ converges, as t goes to infinity, to $\int \phi d\gamma$. Now, let us recall that the first non-zero eigenvalue of $\mathcal{L}_\gamma$ is 1 and the corresponding eigenfunction is the identity. For any ϕ such that $\int \phi d\gamma = 0$ and any $x \in \mathbb{R}$, $e^t P_t\phi(x)$ converges, as t goes to infinity, to $x(\int \phi y d\mu(y))$. One may wonder whether it is possible to use this property to derive an interpolation scheme between ψ and x and obtain a bound on the distance between these two functions. More generally, we could wonder whether an approach similar to Stein's method could be used to prove the convergence of the spectra of a family of Markov chains to the spectrum of

a diffusion process. For random walks on random geometric graphs, such a result may be used to obtain new bounds in the spectral convergence of the graph Laplacian.

Bibliography

[1] Luigi Ambrosio, Giuseppe Savar, and Lorenzo Zambotti. Existence and stability for fokkerplanck equations with log-concave reference measure. *Probability Theory and Related Fields*, 145(3-4):517–564, 2009.

[2] M. Azizyan, Y.-C. Chen, A. Singh, and L. Wasserman. Risk Bounds For Mode Clustering. *ArXiv e-prints*, May 2015.

[3] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion operators*. Grundlehren der mathematischen Wissenschaften, Vol. 348. Springer, Jan 2014.

[4] Pierre Baldi and Yosef Rinott. On normal approximations of distributions in terms of dependency graphs. *Ann. Probab.*, 17(4):1646–1650, 10 1989.

[5] A. D. Barbour. Stein's method for diffusion approximations. *Probability Theory and Related Fields*, 84(3):297–322, 1990.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[7] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[8] Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In *In NIPS*, 2006.

[9] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. 74(8):1289–1308, 2008.

[10] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, February 1992.

[11] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodriguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011. http://imstat.org/ejs/.

[12] Sergey G. Bobkov. Entropic approach to e. rios central limit theorem for w2 transport distance. *Statistics and Probability Letters*, 83(7):1644–1648, 2013.

[13] Sergey G. Bobkov. Berry-esseen bounds and edgeworth expansions in the central limit theorem for transport distances. *preprint*, 2016.

[14] Thomas Bonis, Maks Ovsjanikov, Steve Oudot, and Frédéric Chazal. Persistence-based pooling for shape pose recognition. In *Computational Topology in Image Context - 6th International Workshop, CTIC 2016, Marseille, France, June 15-17, 2016, Proceedings*, pages 19–29, 2016.

[15] Stphane Boucheron, Gbor Lugosi, Pascal Massart, and Michel Ledoux. *Concentration inequalities : a nonasymptotic theory of independence*. Oxford university press, Oxford, 2013.

[16] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *In Proc. CVPR*, 2010.

[17] Michael M. Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *In Proc. CVPR*, 2010.

[18] Peter Bubenik. Statistical topology using persistence landscapes. *JMLR*, 16:77–102, 2015.

[19] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[20] S. Chatterjee. A short survey of Stein's method. *ArXiv e-prints*, April 2014.

[21] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. 2012.

[22] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6):41, 2013.

[23] Y.-C. Chen, C. R. Genovese, R. J. Tibshirani, and L. Wasserman. Nonparametric Modal Regression. *ArXiv e-prints, to appears in Annals of Statistics*, December 2014.

[24] Y.-C. Chen, C. R. Genovese, and L. Wasserman. A Comprehensive Approach to Mode Clustering. *ArXiv e-prints, to appears in Electronic Journal of Statistics*, June 2014.

[25] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density Level Sets: Asymptotics, Inference, and Visualization. *ArXiv e-prints*, April 2015.

[26] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Statistical Inference using the Morse-Smale Complex. *ArXiv e-prints*, June 2015.

[27] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, August 1995.

[28] Minsu Cho and Kyoung Mu Lee. Authority-shift clustering: Hierarchical clustering by authority seeking on graphs. In *CVPR*, pages 3193–3200. IEEE Computer Society, 2010.

[29] John D. Chodera, William C. Swope, Jed W. Pitera, and Ken A. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation*, 5(4):1214–1226, 2006.

[30] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proc. 21st ACM Sympos. Comput. Geom.*, pages 263–271, 2005.

[31] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[32] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. submitted 1412.7392, arXiv, December 2014.

[33] A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. *ArXiv e-prints*, May 2016.

[34] Richard Durrett. *Stochastic calculus : a practical introduction.* Probability and stochastics series. CRC Press, 1996.

[35] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction.* American Mathematical Society, 2010.

[36] R. Engel, K. nad Nagel. *A short course on operator semigroups*. Universitext. Springer-Verlag, 2006.

[37] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes : characterization and convergence*. Wiley series in probability and mathematical statistics. J. Wiley & Sons, New York, Chichester, 1986.

[38] N. García Trillos and D. Slepčev. A variational approach to the consistency of spectral clustering. *ArXiv e-prints*, August 2015.

[39] Evarist Gin and Vladimir Koltchinskii. *Empirical graph Laplacian approximation of LaplaceBeltrami operators: Large sample results*, volume Number 51 of *Lecture Notes–Monograph Series*, pages 238–259. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2006.

[40] Antoine Gloria. *Qualitative and quantitative results in stochastic homogenization*. Accreditation to supervise research, Université des Sciences et Technologie de Lille - Lille I, February 2012.

[41] Larry Goldstein and Gesine Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.*, 7(4):935–952, 11 1997.

[42] Arnaud Guillin, Ivan Gentil, and Franois Bolley. Convergence to equilibrium in wasserstein distance for fokker-planck equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012.

[43] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.

[44] Tatsunori B. Hashimoto, Yi Sun, and Tommi S. Jaakkola. Metric recovery from directed unweighted graphs. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

[45] Matthias Hein, Jean Y. Audibert, and Ulrike von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(Jun):1325–1368, 2007.

[46] Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. Learning 3D Mesh Segmentation and Labeling. *ACM Transactions on Graphics*, 29, 2010.

[47] Hikosaburo Komatsu. A characterization of real analytic functions. *Proc. Japan Acad.*, 36(3):90–93, 1960.

[48] W.L.G. Koontz, P.M. Narendra, and K. Fukunaga. A graph-theoretic approach to nonparametric cluster analysis. *IEEE Transactions on Computers*, 25(9):936–944, 1976.

[49] Kazumasa Kuwada. Duality on gradient estimates and wasserstein controls. *Journal of Functional Analysis*, 258(11):3758 – 3774, 2010.

[50] Lars Larsson-Cohn. L p-norms of hermite polynomials and an extremal problem on wiener chaos. *Arkiv fr Matematik*, 40(1):133–144, 2002.

[51] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, 2006.

[52] Michel Ledoux, Ivan Nourdin, and Giovanni Peccati. Stein's method, logarithmic sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015.

[53] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. Providence, R.I. American Mathematical Society, 2009. With a chapter on coupling from the past by James G. Propp and David B. Wilson.

[54] Chunyuan Li and A. Ben Hamza. Intrinsic spatial pyramid matching for deformable 3d shape retrieval. *IJMIR*, 2:261–271, 2013.

[55] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2486–2493, Washington, DC, USA, 2011. IEEE Computer Society.

[56] Roberto Javier Lpez-Sastre, A. Garca-Fuertes, Carolina Redondo-Cabrera, Francisco Javier Acevedo-Rodrguez, and Saturnino Maldonado-Bascn. Evaluating 3d spatial pyramids for classifying 3d shapes. *Computers and Graphics*, 37:473–483, 2013.

[57] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[58] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.

[59] Jonathan Masci, Davide Boscaini, Michael M. Bronstein, and Pierre Vandergheynst. Shapenet: Convolutional neural networks on non-euclidean manifolds. 2015.

[60] Peter Morters and Yuval Peres. *Brownian motion. Vol. 30.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[61] James R. Norris. *Markov chains.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998.

[62] Ivan Nourdin, Giovanni Peccati, and Yvik Swan. Entropy and the fourth moment phenomenon. *Journal of Functional Analysis*, 266(5):3170 – 3207, 2014.

[63] Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810 – 864, 2009.

[64] F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361 – 400, 2000.

[65] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[66] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[67] D. et al Pickup. Shrec'14 track: Shape retrieval of non-rigid 3d human models. EG 3DOR'14, 2014.

[68] Huaijun Qiu and Edwin R. Hancock. *Graph Embedding Using Commute Time*, pages 441–449. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[69] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A Stable Multi-Scale Kernel for Topological Machine Learning. In *CVPR*, 2015.

[70] Emmanuel Rio. Upper bounds for minimal distances in the central limit theorem. *Ann. Inst. H. Poincar Probab. Statist.*, 45(3):802–817, 08 2009.

[71] Gareth O. Roberts and Richard L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

[72] Adrian Röllin. A note on the exchangeability condition in steins method. *Statistics and Probability Letters*, 78(13):1800 – 1806, 2008.

[73] Marco Saerens, Franois Fouss, Luh Yen, and Pierre Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Lecture Notes in Artificial Intelligence*, pages 371–383. Springer-Verlag, 2004.

[74] A. I. Sakhanenko. Estimates in the invariance principle. *Proc. Inst. Math. Novosibirsk*, 5:2744, 1985.

[75] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[76] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[77] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602, Berkeley, Calif., 1972. University of California Press.

[78] Charles Stein and Institute of Mathematical Statistics. *Approximate computation of expectations*. Lecture notes-monograph series. Hayward, Calif. Institute of Mathematical Statistics, 1986.

[79] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Die Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, Heidelberg, New York, 1979.

[80] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, pages 1383–1392, 2009.

[81] Daniel Ting, Ling Huang, and Michael I. Jordan. An analysis of the convergence of graph laplacians. In *ICML*, 2010.

[82] Alessandro Verri, Claudio Uras, Patrizio Frosini, and Massimo Ferri. On the use of size functions for shape analysis. *Biological Cybernetics*, 70:99–107, 1993.

[83] Cédric Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[84] Ulrike von Luxburg and Morteza Alamgir. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, pages 225–233, 2013.

[85] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15:1751–1798, 2014.

[86] Dorothea Wagner and Frank Wagner. Between min cut and graph bisection. In *Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science*, MFCS '93, pages 744–750, London, UK, UK, 1993. Springer-Verlag.

[87] F.-Y. Wang. Exponential Contraction in Wasserstein Distances for Diffusion Semigroups with Negative Curvature. *ArXiv e-prints*, March 2016.

[88] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *in IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010.

[89] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *in IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009.

[90] L. Yen, L. Vanvyve, D. Wouters, F. Fouss, F. Verleysen, and M. Saerens. Clustering using a random-walk based distance measure. In *Proceedings of ESANN'2005*, 2005.

[91] Dengyong Zhou and Bernhard Schölkopf. *Learning from Labeled and Unlabeled Data Using Random Walks*, pages 237–244. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[92] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.

[93] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom*, 33:249–274, 2005.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Algorithmes d'apprentissage statistique pour l'analyse géométrique et topologique de données

Mots clés : Graphes géométriques aléatoires, marche aléatoires, partitionnement de données flou, méthode de Stein, homologie persistante, sac-de-mots

Résumé : Dans cette thèse, on s'intéresse à des algorithmes d'analyse de données utilisant des marches aléatoires sur des graphes de voisinage, ou graphes géométriques aléatoires, construits à partir des données. On sait que les marches aléatoires sur ces graphes sont des approximations d'objets continus appelés processus de diffusion.

Dans un premier temps, nous utilisons ce résultat pour proposer un nouvel algorithme de partitionnement de données flou de type recherche de modes. Dans cet algorithme, on définit les paquets en utilisant les propriétés d'un certain processus de diffusion que l'on approche par une marche aléatoire sur un graphe de voisinage. Après avoir prouvé la convergence de notre algorithme, nous étudions ses performances empiriques sur plusieurs jeux de données.

Nous nous intéressons ensuite à la convergence des mesures stationnaires des marches aléatoires sur des graphes géométriques aléatoires vers la mesure stationnaire du processus de diffusion limite. En utilisant une approche basée sur la méthode de Stein, nous arrivons à quantifier cette convergence. Notre résultat s'applique en fait dans un cadre plus général que les marches aléatoires sur les graphes de voisinage et nous l'utilisons pour prouver d'autres résultats : par exemple, nous arrivons à obtenir des vitesses de convergence pour le théorème central limite.

Dans la dernière partie de cette thèse, nous utilisons un concept de topologie algébrique appelé homologie persistante afin d'améliorer l'étape de "pooling" dans l'approche "sac-de-mots" pour la reconnaissance de formes 3D.

Title: Statistical learning algorithms for geometric and topological data analysis

Keywords: Random geometric graphs, random walks, soft clustering, Stein's method, persistent homology, bag-of-words

Abstract: In this thesis, we study data analysis algorithms using random walks on neighborhood graphs, or random geometric graphs. It is known random walks on such graphs approximate continuous objects called diffusion processes.

In the first part of this thesis, we use this approximation result to propose a new soft clustering algorithm based on the mode seeking framework. For our algorithm, we want to define clusters using the properties of a diffusion process. Since we do not have access to this continuous process, our algorithm uses a random walk on a random geometric graph instead. After proving the consistency of our algorithm, we evaluate its efficiency on both real and synthetic data.

We then deal tackle the issue of the convergence of invariant measures of random walks on random geometric graphs. As these random walks converge to a diffusion process, we can expect their invariant measures to converge to the invariant measure of this diffusion process. Using an approach based on Stein's method, we manage to obtain quantitfy this convergence. Moreover, the method we use is more general and can be used to obtain other results such as convergence rates for the Central Limit Theorem.

In the last part of this thesis, we use the concept of persistent homology, a concept of algebraic topology, to improve the pooling step of the bag-of-words approach for 3D shapes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France