Density-Based Diffusion for Soft Clustering

Thomas Bonis, Steve Oudot

Density-Based Diffusion for Soft Clustering

Thomas Bonis *
Inria Saclay
Geometrica Team
thomas.bonis@inria.fr

Steve Oudot
Inria Saclay
Geometrica Team
steve.oudot@inria.fr

Abstract

In this paper we advocate the use of diffusion processes guided by density to perform soft clustering tasks. Our approach interpolates between classical mode seeking and spectral clustering, being parametrized by a temperature parameter $\beta > 0$ controlling the amount of random motion added to the gradient ascent. In practice we simulate the diffusion process in the continuous domain by random walks in neighborhood graphs built on the input data. We prove the convergence of this scheme under mild sampling conditions, and we derive guarantees for the clustering obtained in terms of the cluster membership distributions. Our theoretical results are cooroborated by preliminary experiments on manufactured data and on real data.

1 Introduction

1.1 Context and motivation

The analysis of large and possibly high-dimensional datasets is becoming ubiquitous in the sciences. The long-term objective is to gain insight into the structure of measurement or simulation data, for a better understanding of the underlying physical phenomena at work. Clustering is one of the simplest ways of gaining such insight, by finding a suitable decomposition of the data into clusters such that data points within a same cluster share common (and, if possible, exclusive) properties.

Among the variety of existing approaches, mode seeking and spectral clustering are most relevant to our work. The first approach assumes the data points to be drawn from some unknown probability distribution and defines the clusters as the basins of attraction of the maxima of the density, requiring a preliminary density estimation phase [2, 3, 4, 6, 9]. The second approach computes diffusion distances between data points within some neighborhood graph, then applies the classical *K-means* algorithm in the new ambient space [12].

In the end, (hard) clustering methods such as above provide a fairly limited knowledge of the structure of the data. While the partition into clusters is well understood, their interplay (respective locations, proximity relations, interactions) remains yet to be unveiled. Identifying interfaces between clusters is the first step towards a higher-level understanding of the data, and it already plays a prominent role in some applications such as the study of the conformations space of a protein, where a fundamental question beyond the detection

*Partially funded by the DGA.

of metastable states is to understand when and how the protein can switch from one state to another [5]. Hard 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 location of the clusters and of their boundaries.

Spectral clustering can be turned into a soft clustering approach by using fuzzy C-means rather than the regular K-means on the reembedded data [11], or by considering soft normalized cuts instead of hard normalized cuts [8]. Meanwhile, mode seeking can be extended to do soft clustering in various ways, for instance by pertubing the density estimator and performing the clustering scheme multiple times, to detect the areas of instability and estimate the probability of each data point to belong to each cluster [15].

Although spectral clustering and mode seeking look quite different at first glance, they do share a common link. Spectral clustering is known to be related to random walks on graphs [13], and theoretical studies have highlighted the convergence of the Laplace operator of neighborhood graphs to continuous differential operators L of the form: $Lu = \frac{1}{2}\Delta u + \nabla \log f \cdot \nabla u$, where f is the density associated to the probability distribution from which the data points are drawn [14, 16]. The presence of the gradient $\nabla \log f$ actually relates to the mode seeking paradigm, as mode seeking studies the trajectories of the flow induced by the gradient vector field $x \mapsto \nabla \log f(x)$. Thus, the two types of approaches (spectral clustering and mode seeking) are connected, and a natural interpolation between them is given by operators of the form $Lu = \frac{\beta}{2}\Delta u + \nabla \log f.\nabla u$, where β is a strictly positive parameter. This is what this work is about.

1.2 Our approach

Assuming the input data points are drawn from a probability density f satisfying certain regularity conditions, for any temperature parameter $\beta > 0$ we can approximate the diffusion process solution to the following stochastic differential equation:

$$\begin{array}{rcl} dY_t^x & = & \nabla \log f(Y_t)dt + \sqrt{\beta} I_d dW_t \\ Y_0^x & = & x \text{ such that } f(x) > 0 \end{array} \quad (1)$$

by a random walk on a neighborhood graph computed on the data (Theorem 2). Parameter β allows to balance between mode seeking ($\beta \to 0$) and pure diffusion ($\beta \to \infty$). In particular, selecting $\beta = 1$ corresponds to following the traditional isotropic random walk, which is known to have strong bonds with spectral clustering [13].

Our soft clustering scheme proceeds then as follows. Suppose we are given as input a collection $C_1, ..., C_k$ of subsets of $\mathbb{R}^d$ corresponding to data we can reliably assign to a single cluster—these subsets are called *cluster cores*. We use the diffusion process solution to (1) to extend the clustering on the whole dataset in the following way: set the degree of membership of a data point x to the i-th cluster as the probability for the diffusion process starting at x to hit C_i before any other C_j. In particular, every data point that already belongs to a cluster core C_i at the beginning of the process gets a membership of 1 to the i-th cluster. The output is the collection of membership distributions over the dataset for clusters $i = 1$ to k.

Under suitable sampling conditions on the input, and modulo the use of reliable estimators $\hat{C}_i$ of the cluster cores C_i, we show that the cluster membership distributions, defined from the solution of (1), can be approximated using random walks on neighborhood graphs defined on the dataset (Corollary 3).

2 Background

A diffusion process on an open subset Ω of $\mathbb{R}^d$ is the solution of a Stochastic Differential Equation (SDE) of the following form:

$$\begin{aligned} dY_t &= b(Y_t)dt + \sigma(Y_t)dW_t \\ Y_0 &= P_0, \end{aligned} \tag{2}$$

where P_0 is the initial probability distribution on Ω, W_t is a d-dimensional brownian motion, $b : \Omega \to \mathbb{R}^d$ is a drift term guiding the trajectories, and $\sigma : \Omega \to \mathbb{R}^d \times \mathbb{R}^d$ is the diffusion coefficient controlling the amount of noise added to the trajectories. The solutions Y of such an equation are random variables taking values in the space of trajectories $C([0, \zeta], \mathbb{R}^d)$, where ζ is called the *explosion time* and corresponds to the (random) time at which a trajectory reaches the boundary of Ω. A SDE is said to be *well-posed* when there exists a unique solution and this solution has an infinite explosion time with probability 1.

Differently, the space in which convergence of Markov chains occurs is the Skorokhod space $D([0, T], \mathbb{R}^d)$, composed of the trajectories $[0, T] \to \mathbb{R}^d$ that are right-continuous and have left limits for some fixed $T > 0$. It is equipped with the following metric:

$$d(f, g) = \inf\{\epsilon \mid \exists \lambda \in \Lambda, \|\lambda\| \leq \epsilon, \sup_t |f(t) - g(\lambda(t))| \leq \epsilon\},$$

where Λ denotes the space of strictly increasing automorphisms of the unit segment $[0, 1]$, and where $\|\lambda\| = \sup_{s \neq t} \left|\log\left(\frac{\lambda(t)-\lambda(s)}{t-s}\right)\right|$. Standard results show that Markov chains converge *weakly* in $D([0, T], \mathbb{R}^d)$ to diffusion processes truncated at time T. Recall that a stochastic process M^s converges weakly in $D([0, T], \mathbb{R}^d)$ to Y as s tends to 0 if

$$\lim_{s \to 0} \mathbb{E}[\phi(M^s)] \to \mathbb{E}[\phi(Y)] \tag{3}$$

for any continuous and bounded function $\phi : D([0, T], \mathbb{R}^d) \to \mathbb{R}$. The convergence result that we will be using in the article is the following one. Consider a family of Markov chains $(M^{x_0,s})$ defined on discrete state spaces S_s with transition kernels K^s and initial states $M_0^{x_0,s} \in S_s$. For $x \in S_s$ and $\gamma > 0$, let

$$a^s(x) = \frac{1}{s} \sum_{y \in S_s} K^s(x, y)(y - x)(y - x)^T;$$

$$b^s(x) = \frac{1}{s} \sum_{y \in S_s} K^s(x, y)(y - x);$$

$$\Delta_s^\gamma = \frac{1}{s} K^s(x, \mathcal{B}(x, \gamma)^c),$$

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

Theorem 1 (Modified from Theorem 7.1 of [7])**.** *Let U be a compact subset of Ω. Assume the stochastic differential equation (2) is well-posed for any Dirac measure $P_0 = \delta_{x_0}$ with $x_0 \in U$. Assume also that the maps b and σ in (2) are continuous, and let $a = \sigma\sigma^T$. Assume further that for any $\gamma > 0$,*

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

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

(iii) $\lim_{s \to 0} \sup_{x \in S_s} \Delta_s^\gamma = 0$;

(iv) $\lim_{s \to 0} \sup_{x_0 \in U} \|M_0^{x_0,s} - x_0\|_\infty = 0$.

Then, for any $T > 0$, the continuous time process $M_{s\lfloor t/s \rfloor}^{x_0,s}$ converges weakly in $D([0, T], \mathbb{R}^d)$ to the solution of (2) with initial condition $P_0 = \delta_{x_0}$ as s tends to 0. Furthermore, the convergence is uniform with respect to $x_0 \in U$.

3 Random walks and diffusion processes

Let f be a probability density on $\mathbb{R}^d$, and let $\Omega = \{x \mid f(x) > 0\}$. We assume once and for all that f satisfies the following technical conditions:

- f is C^1-continuous over $\mathbb{R}^d$,
- $\lim_{\|x\|_2 \to \infty} f(x) = 0$,
- $\forall \alpha_0 > 0, \exists \alpha < \alpha_0, \forall x \in \Omega, f(x) = \alpha \implies \nabla f(x) \neq 0$,
- the SDE (1) over the domain Ω is well-posed.

Proving the well-posedness of (1) is beyond the scope of the paper, and reasonable sufficient conditions on f can be found e.g. in [1, 10].

Let $\mathcal{X}_n = X_1, ..., X_n$ be i.i.d random variables drawn from f, and let $\hat{f}_n$ be a density estimator. We will study the relationship between random walks on graphs built on $\mathcal{X}_n$ using $\hat{f}_n$ and the solution of (1), for a fixed temperature parameter $\beta > 0$.

Let $M^{x,h,n}$ be the Markov Chain whose initial state is the closest neighbour of x in $\mathcal{X}_n$ (break ties arbitrarily), and whose transition kernel K^h is

$$K^h(X_i, X_j) = \begin{cases} (1 + (\beta - 1)\frac{\hat{f}_n(X_i)}{\hat{f}_n(X_j)})Z_i & \text{if } \|X_i - X_j\|^2 \leq h \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where Z_i is the appropriate renormalization factor:

$$Z_i = \left(\sum_{j=1}^{n} 1_{\|X_i - X_j\|^2 \leq h} (1 + (\beta - 1)\frac{\hat{f}_n(X_i)}{\hat{f}_n(X_j)}) \right)^{-1}.$$

Under some conditions on the estimator $\hat{f}_n$, this graph-based random walk approximates the diffusion process in the continuous domain in the following sense: there exists s depending on h such that, as n tends to infinity, with high probability, $M^{x,h,n}_{s\lfloor t/s \rfloor}$ converges weakly to the solution of (1) in the Skorokhod space of trajectories. Formally, we prove the following theorem:

Theorem 2. *Suppose our estimator $\hat{f}_n$ satisfies:*

$$\lim_{n \to \infty} \mathbb{P}(\|f - \hat{f}_n\|_\infty \geq \epsilon) = 0.$$

Then, for any $T, \epsilon > 0$, for any compact set $U \subset \Omega$, and for any Borel set B of $D([0,T], \mathbb{R}^d)$ such that $\mathbb{P}(Y^x \in \partial B) = 0$ for all $x \in U$, there exists $h_0 > 0$ such that for all $h \leq h_0$,

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

where $s = \frac{\Gamma(1+\frac{d}{2})h^2}{\beta \Gamma(1+\frac{d+1}{2})2}$.

Note that we fix h before letting n go to infinity in the theorem. Indeed, letting h depend on n and decrease too quickly as n goes to infinity may prevent the convergence of the sequence of Markov chains. This phenomenon appears in the course of the proof of the theorem. This aspect is of utmost importance if one wants to extend the technique using k-nn graphs instead of neighborhood graphs. In this case indeed, the standard k-nn assumptions ($\frac{k_n}{n} \to 0$ and $n \to \infty$) are insufficient to ensure convergence.

Let us emphasize also that the result can be adapted to hold for random walks defined on weighted neighborhood graphs, i.e. random walks that jump from X_i to X_j with a probability proportional to $(1 + (\beta - 1)\frac{f(X_i)}{f(X_j)})V_h(\|X_j - X_i\|_2)$, where $V_h : \mathbb{R}^+ \to \mathbb{R}^+$ is a decreasing weighting scheme. The adaptation requires extra conditions on V_h so as to satisfy the conditions of Theorem 1.

4 Application to soft clustering

The results of Section 3 make it possible to derive a soft variant of mode seeking. In the continuous setting, our input is a collection of cluster cores $C_1, ..., C_k$ corresponding to subsets of the data that are known to belong surely to a single cluster. We assume these cores to be well-separated (i.e. $C_i \cap C_j = \emptyset$ for all $i \neq j$) compact sets of $\mathbb{R}^d$. Furthermore, we assume their boundary satisfies the following cone condition: for any $i \in \{1, \cdots, k\}$ and any $z \in \partial C_i$ there exists a truncated cone $O_{z,h,\alpha}$ belonging to C_i, where z is the base of the cone, h is its length and α is its angle.

For any point x of $\mathbb{R}^d$ and any cluster core C_i, we define the degree of membership of x to the i-th cluster to be the probability $\mu_i(x)$ for the solution Y^x of SDE (1) initialized at x to hit C_i before any other cluster core—in the event that it hits any cluster cores at all. Formally, we let $\tau_x = \inf_t Y_t^x \in \cup_i C_i$ be the stopping time of the solution, and we define

$$\mu_i(x) = \mathbb{P}(\tau_x < \infty \text{ and } Y_{\tau_x}^x \in C_i).$$

We also let $\mu_0(x) = \mathbb{P}(\tau_x = \infty)$ be the probability that the solution does not hit any cluster cores at all. The output of the soft-clustering is the collection $(\mu_i(x))_{0 \leq i \leq k}$ for every point $x \in \Omega$. This is the continuous version of the algorithm.

In the discrete setting, the cluster cores are estimated using the input data points. For this we suppose we have access to estimators $\hat{C}_{i,n}$ of the C_i that satisfy the following property:

$$\forall \delta > 0, \lim_{n \to \infty} \mathbb{P}(C_i^{-\delta} \subset \hat{C}_{i,n} \subset C_i^{\delta}) = 1, \tag{5}$$

where $C_i^{\delta} = \cup_{x \in C_i} \mathcal{B}(x, \delta)$ and $C_i^{\delta} = C^i \setminus \cup_{x \notin C_i} \mathcal{B}(x, \delta)$. We then compute approximations $\hat{\mu}_{i,h,n}$ of μ_i using trajectories of the Markov Chains $M^{x,h,n}$ and the $\hat{C}_{i,n}$ instead of the solution of (1) and the C_i. Under suitable assumptions on the estimators $\hat{f}_n$, the results of Section 3 imply that the $\hat{\mu}_{i,h,n}(x)$ are good approximations of the $\mu_i(x)$:

Corollary 3. *Let $\beta > 0$ and $i \in [|1, k|]$. Suppose that $\hat{f}_n$ satisfies*

$$\lim_{n \to \infty} \mathbb{P}(\|f - \hat{f}_n\|_\infty \geq \epsilon) = 0.$$

and that $\hat{C}_{i,n}$ satisfies (5). Then, for any compact set $U \subset \Omega$, for any $\epsilon > 0$, there exists h_0 such that if $h \leq h_0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{x \in U} |\hat{\mu}_{i,h,n}(x) - \mu_i(x)| \geq \epsilon\right) = 0.$$

The choice of parameter β depends on how much the clusters in the ground truth overlap. In cases where they are well separated, small values of β should be preferred. On the contrary, large values of β should be used when there are large overlapping areas. Another way to interpret β is as a trade-off between the respective influence of the metric and of the density in the diffusion process. When β is small, the output of our algorithm is mostly guided by the density and therefore close to the output of mode seeking algorithms. By contrast, when β is large, our algorithm becomes oblivious to the density and and thus has a behaviour close to that of K-means algorithms. We will elaborate on these points in Section 5.

Practical considerations

Computation of the $\hat{\mu}_{i,h,n}$. In practice, the value of $\hat{\mu}_{i,h,n}$ for $i > 0$ is computed by solving the linear system $A^T\mu = \mu$ with $\mu_k = 1$ if X_k belongs to $\hat{C}_{i,n}$ and $\mu_k = 0$ if the intersection of the connected component of X_k in the neighborhood graph and $\hat{C}_{i,n}$ is zero. A is defined in the following way:

$$A_{kl} = \begin{cases} K^h(X_k, X_l) & \text{if } X_k \text{ belongs to none of the cluster cores;} \\ \delta_{kl} & \text{if } X_k \in \hat{C}_{i,n}; \\ 0 & \text{otherwise,} \end{cases}$$

where K^h is defined as in (4). $\hat{\mu}_0(x)$ is then defined as $1 - \sum_{i=1}^{k} \hat{\mu}_i(x)$.

When one has to deal with large amounts of data for which solving the linear system directly is too expensive, it is possible to use an iterative scheme. Set $\hat{\mu}_i^0 = 1_{\hat{\mu}_i \in \hat{C}_{i,n}}$, then define the sequence $\hat{\mu}_i^n$ by $\hat{\mu}_i^{n+1} = A^T\hat{\mu}_i^n$. This sequence converges to the solution of $A^T\mu = \mu$.

Density Estimation. The definition of our transition kernel K^h requires to be able to:

- estimate distances in the ambient space,
- estimate the value of the density f at the data points.

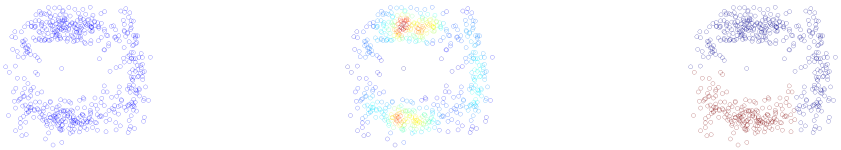
While both operations can be performed using standard tools in Euclidean space $\mathbb{R}^d$, their implementation becomes tricky when the data lie in more general ambient spaces. In such situations, it is often the case that the input is given in the form of a proximity or similarity graph, (a subset of) which we can use as our neighborhood graph. It is then desireable to derive a density estimator from the sole graph structure, which our framework permits.

In the case where the density f is non-zero on a single connected component U, the solution Y_t^x of (2) has a unique stationary distribution ρ provided that x belongs to U. In our case, ρ turns out to be equal to $f^{2/\beta}$. Therefore, it is natural to use the square root of the stationnary distribution (if it exists) of the istotropic random walk (case $\beta = 1$) on the graph as a proxy for f. Using the stationary distribution of a graph in mode seeking was already proposed in [4], but without providing the relationship between random walks and diffusion processes. When dealing with multiple connected components, the stationary distribution of $M^{x,h,n}$ is not uniquely defined, so we simply take the uniform distribution on our data to initialize the Markov chain before computing its limit distribution iteratively.

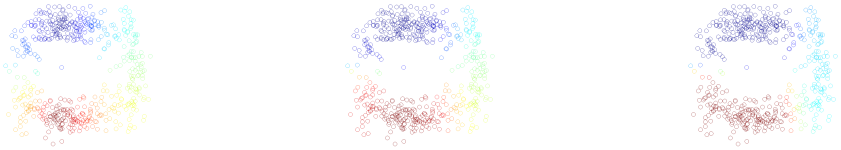
Selection of the cluster cores. In practice we use the hard mode seeking algorithm ToMATo [2] to select the cluster cores. ToMATo considers the local maxima of the density estimator in G, and for each one of them it computes a measure of *prominence*. It then selects the subset $\{v_1, ..., v_k\}$ of the local maxima with prominence higher than a user-defined threshold $\kappa > 0$, and it outputs k clusters where the i-th cluster is obtained by merging the basin of attraction of v_i with the basins of attraction of some of the discarded local maxima. Given an extra parameter ν such that $\kappa > \nu > 0$, consider the subgraph of G spanned by those vertices x such that $\{x \in \mathbb{R}^d \mid \hat{f}_n(x) > \hat{f}_n(v_i) - \nu\}$. We define the i-th cluster core $\hat{C}_{i,n}$ as the connected component of this subgraph that contains v_i. It can be readily checked from the analysis of ToMATo that $\hat{C}_{i,n}$ is indeed included in the i-th cluster.

5 Experiments

Below we illustrate the effect of the temperature parameter β on the clustering output on syntehtic data, before considering the application of our method on protein conformations data.



(a) The data. (b) Our density estimator. (c) Output of ToMATo.



(d) Our output for $\beta = 10$. (e) Our output for $\beta = 1$. (f) Output for $\beta = 0.4$.

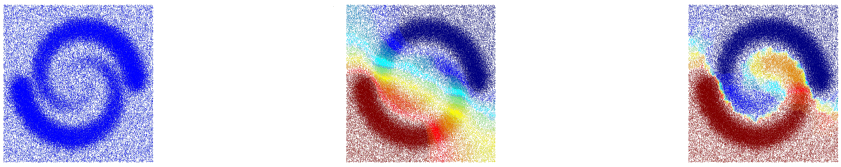
Figure 1: Output of our algorithm on a simple dataset composed of two overlapping clusters. For soft-clustering green corresponds to an equal membership to both clusters.

Synthetic data. The first dataset is presented in Figure 1a. The ground truth tells us that the output should be composed of two clusters. Moreover, the underlying density function being symmetric with respect to the horizontal line $y = 0$, it is desireable that the ouutput be also symmetric. Note that there is no symmetry of the density with respect to the vertical line $x = 0$: indeed, there is a large density gap between the clusters on the left-hand side, while there is a large overlapping area between them on the right-hand side. The soft clustering output is expected to show a small area of transition between the two clusters on the left-hand side, and on the contrary a large transition area on the right-hand side.

To compute the density, estimator we used the stationary density of a random walk on the graph, as proposed in Section 4. The output of the computation is displayed in Figure 1b. Not surprisingly, the density possesses two high local maxima corresponding to the two clusters. However, the overlap of between both clusters creates a small density peak on the right-hand side. Standard mode seeking algorithms can be mislead by this peak: for instance, the ToMATo algorithm [2] merges the basins of attraction of the small peak on the right into one of the two clusters, resulting in a non-symetrical clustering output, see Figure 1c. We display the results of our algorithm for three values of β: 10 in Figure 1d, 1 in Figure 1e and 0.4 in Figure 1f. As we can see from the output of the algorithm for the small value of β (0.4), the amount of noise injected in our trajectory is not large enough to counter the influence of the third density peak, so the result obtained is really close to hard clustering. Much larger values of β (10) do not give enough weight to the density function, so a highly smoothed transition between the two clusters on the left part. Intermediate values of β (1) seem to give more satisfying results.

The second dataset we consider is composed of two interleaved spirals, presented in Figure 2a. An interesting property of this dataset is that the head of each spiral is close to the tail of the other spiral. Thus, the two clusters are well-separated by a density gap but not by the Euclidean metric. We use our algorithm with two different values of β: 1 (Figure 2b) and 0.3 (Figure 2c). We can see that for $\beta = 1$, the density gap between the two spirals is not strong enough to counter the proximity of the two clusters in the Euclidean metric. On the other hand, for $\beta = 0.3$ we recover the two different clusters as we give more weight to the density.

Protein conformations data. We now turn to the problem of clustering protein conformations, as we briefly introduced it in Subsection 1.1. We consider the case of the alanine-dipeptide molecule. Our dataset (courtesy of C. Clementi and W. Zheng) is composed of $100,000$ protein conformations, each of which



(a) The data. (b) Our output for $\beta = 1$. (c) Our output for $\beta = 0.3$.

Figure 2: Output of our algorithm on a simple dataset composed of two interleaved spirals.

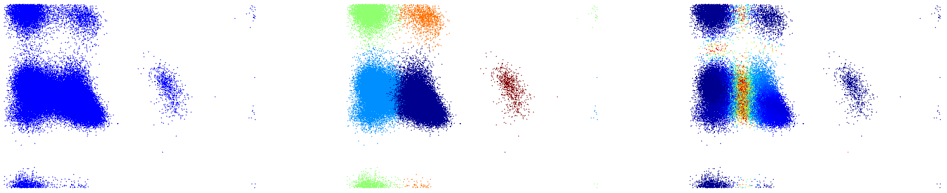


Figure 3: From left to right: (a) the dataset projected on the Ramachadran plot, (b) ToMATo output, (c) second highest membership obtained with our algorithm for $\beta = 0.2$

represented as a 30-dimensional vector. The metric we use for this type of data is the root-mean-squared deviation (RMSD). The purpose of soft-clustering in this case is twofold: on the one hand, we want to find the right number of clusters corresponding to metastable states of the molecule; on the other hand, we want to find the conformations lying at the border between different clusters, as these can help understand the mechanisms of transitions between metasable states. It is well-known that the conformations of alanine-dipeptide have only two relevant degrees of freedom, it is thus possible to project the data to have a good visualization of the clustering output with this representation (which is also called a Ramachadran plot), see Figure 3. To evaluate the output of our algorithm, we only display the second highest membership obtained as it indicates interfaces between two clusters. As we can see there are 5 clusters and 4 main interfaces. The important observation is that, although the borders between hard clusters are located roughly in the expected areas, the exact delimitation of each cluster is arbitrary and irregular. The soft clustering procedure has a regularizing effect, and it localizes the entire expected interface regions.

6 Conclusion

We have motivated the use of diffusion processes guided by density in the context of soft clustering, both from a theoretical and from a practical point of view. Our approach allows to interpolate between spectral clustering and mode seeking, moreover it is related to K-means clustering when the temperature β is infinite. Our theoretical results are stated and proven in Euclidean space $\mathbb{R}^d$, however they should be extendable to Riemannian manifolds provided the manifold curvature is taken into account in the analysis. Meanwhile, the use of graph diffusion for density estimation has given promising practical results and thus calls for further theoretical investigation.

Acknowledgements. The authors wish to thank Cecilia Clementi and her student Wenwei Zheng for providing the alanine-dipeptide conformation data used in Figure 3.

References

[1] Sergio Albeverio, Yuri Kondratiev, and Michael Röckner. Strong feller properties for distorted brownian motion and applications to finite particle systems with singular interactions. In *Finite and infinite dimensional analysis in honor of Leonard Gross (New Orleans, LA, 2001)*, volume 317 of *Contemp. Math.*, pages 15–35. Amer. Math. Soc., Providence, RI, 2003.

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

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

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

[5] 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.

[6] 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.

[7] Richard Durrett. *Stochastic calculus : a practical introduction*. Probability and stochastics series. CRC Press, Boca Raton, 1996. Edition revue de Brownian motion and martingales in analysis cop. 1984.

[8] Rong Jin, Chris Ding, and Feng Kang. A probabilistic approach for optimizing spectral clustering. In *In Advances in Neural Information Processing Systems 18*, 2005.

[9] 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.

[10] N.V. Krylov and M. Röckner. Strong solutions of stochastic equations with singular time dependent drift. *Probab. Theory Relat. Fields*, 131(2):154–196, 2005.

[11] Rocco Langone, Raghvendra Mall, and Johan A. K. Suykens. Soft kernel spectral clustering. In *IJCNN*, pages 1–8. IEEE, 2013.

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

[13] Marina Maila and Jianbo Shi. A random walks view of spectral segmentation. In *AI and STATISTICS (AISTATS) 2001*, 2001.

[14] Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *in Advances in Neural Information Processing Systems 18*, pages 955–962. MIT Press, 2005.

[15] P. Skraba, M. Ovsjanikov, F. Chazal, and L. Guibas. Persistence-based segmentation of deformable shapes. In *CVPR Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, page to appear, June 2010.

[16] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, April 2008.

A Proof of Theorem 2

We give an overview of the proof in Section A.1, then we deal with the technical details in Section A.2. But first, we provide complementary background material that will be used in the proof.

Weak convergence as defined in (3) can be characterized in different ways via the Portmanteau theorem. In particular, a stochastic process M^s converges weakly to a diffusion process Y in $D([0,T],\mathbb{R}^d)$ as s tends to 0 if and only if

$$\lim_{s\to 0} \mathbb{P}(M^s \in B) = \mathbb{P}(Y \in B) \tag{6}$$

for any Borel set B such that $\mathbb{P}(Y \in \partial B) = 0$. This equivalence allows to rewrite Theorem 1 as follows:

Corollary 4. *Let U be a compact subset of Ω. Assume the stochastic differential equation (2) is well-posed for any Dirac measure $P_0 = \delta_{x_0}$ with $x_0 \in U$. Assume also that the maps b and σ in (2) are continuous, and let $a = \sigma\sigma^T$. Let $Y_t^{x_0}$ be the solution of the SDE (2) with initial condition $P_0 = \delta_{x_0}$. Let also B be a Borel set in $D([0,T],\mathbb{R}^d)$ for some $T > 0$ such that $\mathbb{P}(Y^{x_0} \in \partial B) = 0$ for all $x_0 \in U$, and let $\epsilon > 0$. Then, there exist parameters ν and γ such that*

$$\sup_{x_0 \in U} |\mathbb{P}(M^{x_0,s}_{s\lfloor t/s \rfloor} \in B) - \mathbb{P}(Y_t^{x_0} \in B)| \leq \epsilon$$

whenever the following conditions are met:

(i) $\sup_{x\in S_s} \|a^s - a\|_\infty \leq \nu$;

(ii) $\sup_{x\in S_s} \|b^s - b\|_\infty \leq \nu$;

(iii) $\sup_{x\in S_s} \Delta_s^\gamma \leq \nu$;

(iv) $\sup_{x_0\in U} \|M_0^{x_0,s} - x_0\|_\infty \leq \nu$.

A.1 Proof overview

We denote by $\mathcal{X}_n = (X_1, ..., X_n)$ the i.i.d sampling which is also the state space of M_s. 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 $M^{x,h,n}$ stands for the continuous time process $M^{x,h,n}_{s\lfloor t/s \rfloor}$. Let T and ϵ be strictly positive reals, $s = \frac{\Gamma(1+\frac{d}{2})h^2}{\beta\Gamma(1+\frac{d+1}{2})2}$. For $\alpha > 0$, let $B_\alpha = \{w \in D([0,T],\mathbb{R}^d) \mid \forall t, w(t) \in F^\alpha\}$. Using Lemma 5, there exists α such that, for any $x \in U$, $\mathbb{P}(Y^x \in B_\alpha) \geq 1 - \epsilon/4$.

To obtain a good approximation of the trajectories in F^α using $M^{x,h,n}$, we only need to check assumptions (i)-(iv) of Corollary 4 on F^α for the diffusion process solution of 2.. Let us show that these assumptions are verified.

F^α is closed as f is continuous and it is also bounded as $\lim_{\|x\|_2\to\infty} f(x) = \infty$, it is thus compact. Since f is continuous, there exists $r > 0$ such that f is strictly positive on the r-offset of F^α: $F^{\alpha,r} = \cup_{x\in F^\alpha} \mathcal{B}(x,r)$. Since $F^{\alpha,r}$ is also compact, for $\kappa > 0$, there exists h_1 such that if $h < h_1$ then for all $x, y \in F^{\alpha,r}$ such that $d(x,y) \leq h_\epsilon$ we have $1 + (\beta - 1)\frac{f(x)}{f(y)} \geq \kappa$ Hence using our hypothesis on $\hat{f}_n$, we can apply Lemma 7 on F^α to obtain that if $h < h_1$, in high probability we have that $1 + (\beta - 1)\frac{f(x)}{f(y)} \geq 0$ for all $x, y \in F^{\alpha,r}$ such that $d(x,y) \leq h_\epsilon$.

Let K^s be the transition kernel of $M^{x,h,n}$, If h is smaller than h_1 then, we have:

$$|a^s(X_i) - \beta I_d| = |\frac{1}{s}\sum_j K^s(X_i, X_j)(X_j - X_i)(X_j - X_i)^T - \beta I_d|,$$

$$|b^s(X_i) - \frac{\nabla f(X_i)}{f(X_i)}| = |\frac{1}{s}\sum_j K^s(X_i, X_j)(X_j - X_i) - \frac{\nabla f(X_i)}{f(X_i)}|,$$

Combined with our assumption on $\hat{f}_n$, we can use Lemma 8 along with a Hoeffding inequality and a union bound on the points of the compact set F^α and obtain that as long as h is smaller than $h_0 \leq h_1$,

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

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

(iii) $\sup_{y\in\mathcal{X}_n\cap F^\alpha} \Delta_s^h = 0,$

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

Thus, the assumptions (i)-(iv) of Corollary 4 are verified on F^α for the diffusion process solution of (2).

Since f is continuous, B_α is an open set, therefore using Portmanteau's Theorem we can derive, in a similar way that we derived Corollary 4, that for $h \leq h_2$

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

in high probability. Therefore, for any Borel set B,

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

Thus, we only need to approximate trajectories that do not leave F^α to obtain a good approximation of $\mathbb{P}(M^{x,h,n} \in B)$. So we can apply Corollary 4 on these trajectories with an accuracy of $\epsilon/2$ to obtain, in high probability,

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

Every step of the proof hold with a probability that decreases to 0 as n tends to infinity, thus the proof of Theorem 2 is complete.

A.2 Technical Lemmas

Lemma 5. *Let U be a compact set of $\mathbb{R}^d$ such that f is bounded away from zero on U. Let $B_\alpha = \{w \in D([0,T],\mathbb{R}^d) \mid \forall t, w(t) \in F^\alpha\}$. For any $T > 0$ and $\epsilon > 0$, there exists $\alpha > 0$ such that*

$$\sup_{x\in U} \mathbb{P}(Y_t^x \in B_\alpha) \geq 1 - \epsilon$$

Proof. This is just another way to say that the diffusion process does not explode in finite time. □

Lemma 6. *For all $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$, we have that:*

$$(i) \int_{\|\lambda\|_2\leq R} <\lambda, u> \lambda d\lambda = \frac{\pi^{d/2} R^{d+2}}{2\Gamma(1 + \frac{d+1}{2})} u$$

Similarly, for all (u, v) *such that* $\|u\|_2 = \|v_2\| = 1$*,*

$$(ii) \int_{\|\lambda\|_2 \leq R} < \lambda, u >< \lambda, v > d\lambda = \frac{\pi^{d/2} R^{d+2}}{2\Gamma(1 + \frac{d+1}{2})} < u, v >$$

Proof. If $d = 1$, the result is trivial. If $d > 2$, Consider an orthonormal basis $(u, e_1, ..., e_{d-1})$ of $\mathbb{R}^d$. The i_{th} coordinate of $\int_{\|\lambda\|_2 \leq R} < \lambda, u > \lambda d\lambda$ in this basis is given by:

$$I_i = \int_{\sum_i x_i^2 = R^2} x_1 x_i dx_1 ... dx_d$$

For symmetry reasons, for $i > 1$, we have $I_i = 0$. Now consider the first coordinate, using the volume of the $d - 1$ sphere, we have:

$$I_1 = \int_{x \in [-R,R]} \frac{x^2 (R^2 - x^2)^{(d-1)/2} \pi^{(d-1)/2}}{\Gamma(\frac{d+1}{2})} dx$$

$$I_1 = \frac{2\pi^{(d-1)/2}}{\Gamma(\frac{d+1}{2})} \int_{x \in [0,R]} x^2 (R^2 - x^2)^{(d-1)/2} dx$$

We renormalize, let $v = \frac{x}{R}$, we have $dx = Rdv$ and

$$I_1 = \frac{2\pi^{(d-1)/2} R^{d+2}}{\Gamma(\frac{d-1}{2})} \int_{v \in [0,1]} v^2 (1 - v^2)^{(d-1)/2} dv$$

We change variables, let $u = (1 - v^2)$, we have $v = (1 - u)^{1/2}$ and $dv = \frac{-du}{2(1-u)^{1/2}}$, thus we have:

$$I_1 = \frac{\pi^{(d-1)/2} R^{d+2}}{\Gamma(\frac{d+1}{2})} \int_{u \in [0,1]} u^{(d-1)/2} (1 - u)^{1/2} du$$

Using the β function,

$$I_1 = \frac{\pi^{(d-1)/2} R^{d+2}}{\Gamma(\frac{d+1}{2})} \beta(\frac{d+1}{2}, \frac{3}{2})$$

Using the relationship between β and Γ,

$$I_1 = \frac{\pi^{(d-1)/2} R^{d+2}}{\Gamma(\frac{d+1}{2})} \frac{\Gamma(\frac{d+1}{2}) \Gamma(\frac{3}{2})}{\Gamma(1 + \frac{d+1}{2})}$$

$$I_1 = \frac{\pi^{d/2} R^{d+2}}{2\Gamma(1 + \frac{d+1}{2})}$$

□

Lemma 7. *Let* U *be a compact set such that* f *and* $\hat{f}_n$ *are bounded away from* 0 *on* U*. Let* $r > 0$*,* $m = \sup_{x \in U^r} f(x)^{-1}$*. Then, for any* $t, x \in U^r$*, we have:*

$$\begin{aligned} E(t, x) &= |(\frac{\hat{f}_n(t)}{\hat{f}_n(x)} - \frac{f(t)}{f(x)}) \frac{f(x)}{f(t)}| \\ &\leq 2m\|\hat{f}_n - f\|_\infty + o(m\|\hat{f}_n - f\|_\infty) \end{aligned}$$

Proof. We have:

$$\begin{aligned}
E &= |(\frac{\hat{f}_n(t)}{\hat{f}_n(x)} - \frac{f(t)}{f(x)})\frac{f(x)}{f(t)}| \\
&\leq |(1 + \frac{\hat{f}_n(t) - f(t)}{f(t)})(1 + \frac{\hat{f}_n(x) - f(x)}{f(x)})^{-1} - 1| \\
&\leq 2m\|\hat{f}_n - f\|_\infty + o(m\|\hat{f}_n - f\|_\infty)
\end{aligned}$$

□

Lemma 8. *Consider a compact set U such that f is bounded away from zero on U. Let $r > 0$. Let $m = \sup_{x \in U^r} f(x)^{-1}$ and $E = m\|\hat{f}_n - f\|_\infty$. Let $V_1 = \frac{2\Gamma(1+\frac{d+1}{2})}{\pi^{d/2} r^{d+2}}$ and $V_2 = \frac{\Gamma(1+\frac{d}{2})}{\pi^{d/2} r^d}$, for any $t \in U$ we note:*

$$\begin{aligned}
f_t^1(x) &= \frac{V_1}{f(t)} 1_{x \in \mathcal{B}(t,r)} (1 + (\beta - 1)\frac{\hat{f}_n(t)}{\hat{f}_n(x)})(x - t)(x - t)^T \\
f_t^2(x) &= \frac{V_1}{f(t)} 1_{x \in \mathcal{B}(t,r)} (1 + (\beta - 1)\frac{\hat{f}_n(t)}{\hat{f}_n(x)})(x - t) \\
f_t^3(x) &= \frac{V_2}{\beta f(t)} 1_{x \in \mathcal{B}(t,r)} (1 + (\beta - 1)\frac{\hat{f}_n(t)}{\hat{f}_n(x)})
\end{aligned}$$

We have that,

$$\begin{aligned}
&|\mathbb{E}[\frac{1}{n}\sum_{i \leq n} f_t^1(X_i))] - \beta I_d| \leq O(E) + O(r) \\
&|\mathbb{E}[\frac{1}{n}\sum_{i \leq n} f_t^2(X_i))] - \frac{\nabla f(t)}{f(t)}| \leq O(\frac{E}{r}) + O(r) \\
&|\mathbb{E}[\frac{1}{n}\sum_{i \leq n} f_t^3(X_i))] - 1| \leq O(E) + O(r)
\end{aligned}$$

Proof. We will prove the result for f^1, the proof is the same for the f^2 and f^3. Let $t \in U$, we have:

$$\begin{aligned}
\mathbb{E}[f_t^1(\mathbb{X})] = \int_{\mathbb{R}} f_t^1(x) f(x) dx & \\
&= \int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)} (1 + (\beta - 1)\frac{\hat{f}_n(t)}{\hat{f}_n(x)})(x - t)(x - t)^T f(x) dx
\end{aligned}$$

First, using Lemma 7:

$$
\begin{aligned}
&|\int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)}(\beta-1)(\frac{\hat{f}_n(t)}{\hat{f}_n(x)}-\frac{f(t)}{f(x)})f(x)(x-t)(x-t)^T dx| \\
&\qquad \leq \int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)} f(x)|\beta-1|(E+O(E^2))r^2 \\
&\qquad\qquad \leq (\frac{2\Gamma(1+\frac{d+1}{2})}{\Gamma(1+d/2)}+O(r))|\beta-1|(E+o(E))
\end{aligned}
$$

Now, using Taylor's formula,

$$
\begin{aligned}
&\int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)}(1+(\beta-1)\frac{f(t)}{f(x)})(x-t)(x-t)^T f(x)dx \\
&\qquad = \int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)}(\beta f(t)+\nabla f(t).(x-t)+O(\|x-t\|^2)) \\
&\qquad\qquad (x-t)(x-t)^T dx
\end{aligned}
$$

By symmetry, we have:

$$
\begin{aligned}
&\int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)}(1+(\beta-1)\frac{f(t)}{f(x)})(x-t)(x-t)^T f(x)X dx \\
&\qquad\qquad = \int_{\mathcal{B}(t,r)} V_1(\beta+O(\|x-t\|^2))(x-t)(x-t)^T dx
\end{aligned}
$$

Now, using (ii) from Lemma 6 for $\lambda=(x-t)$ and $u,v=e_i,e_j$ two elements of the canonical basis of $\mathbb{R}^d$, we obtain:

$$
\begin{aligned}
&\int_{\mathcal{B}(t,r)} \frac{V_1}{f(t)}(1+(\beta-1)\frac{f(t)}{f(x)})(x-t)(x-t)^T f(x)dx \\
&\qquad\qquad = \delta(i,j)\beta+O(r^2)
\end{aligned}
$$

Hence,

$$
|\mathbb{E}[f_t(\mathbb{X})]-\beta I_d| \leq \frac{2\Gamma(1+\frac{d+1}{2})}{\Gamma(1+d/2)}|\beta-1|(E+o(E))+O(r)
$$

□

B Proof of Corollary 3

In this proof, we only treat the case where Ω has a single connected component. The generalization to the case where Ω has multiple connected components is straightforward. Let ϵ be a strictly positive real and $U \subset \Omega$ be a compact set.

Consider $x \in \Omega$, let

- $\mu^+_{i,\delta}(x)$ be the probability that Y^x hits C_i^δ before any other $C_j^{-\delta}$,
- $\mu^-_{i,\delta}(x)$ be the probability that Y^x hits $C_i^{-\delta}$ before any other C_j^δ.

Let us show that for any i, if a trajectory enters C_i^δ then, it has a high probability to enter C_i if δ is small enough. Since the C_i are closed and disjoints there exists $\delta_0 > 0$ such that the $C_i^{\delta_0}$ are disjoints. Using the cone condition on C_i and considering truncated cone of length at most $\delta_0/2$, for any $z \in \partial C_i$ there exists a cone $O_{z,h_z,\alpha_z} \subset C_i$ with $h_z < \delta_0/2$. Using Lemma 9 there exist radii r_z such that if $x \in \cup_{z\in\partial C_i} \mathcal{B}(z, r_z)$ then the probability for Y^x to hit C_i before exiting $C_i^{\delta_0}$ is at least $1-\epsilon/8$. Since the closure of C_i is compact, there exists a single radius δ_i^+ such that if $d(x, C_i) \leq \delta_i^+$ then the probability for Y^x to hit C_i before exiting $C_i^{\delta_0}$ is at least $1 - \epsilon/8$.

Now, let us show the opposite: if a trajectory enters C_i, then it enters $C_i^{-\delta}$ in high probability. Using the previous notations for cones, for any $z \in \partial C_i$, there exists a cone $O_{z,h_z,\alpha_z} \subset C_i$. For $\delta < h_z$, this means that there exists z^δ in $C_i^{-\delta}$ such that there exists a cone $O_{s^\delta,h_z-\delta,\alpha_z}$. Thus by using a similar argument than in the previous case (with cones getting closer to points rather than points getting close to cones), for a set of radii δ_i^- such that if a trajectory hits C_i, then it hits $C_i^{-\delta}$ with probability at least $1 - \epsilon/8$.

Let $\delta = \min(\delta_j^+, \delta_j^-)$, by combining our results and using the strong Markov property of Y^x we obtain that:

- $\mu_{i,\delta}^+(x) - \mu_i(x) \leq \epsilon/4$,
- $\mu_i(x) - \mu_{i,\delta}^-(x) \leq \epsilon/4$.

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

$$B = \{w \in D([0,\infty], \mathbb{R}^d \mid \exists \tau \text{ such that } w(\tau) \in C_i^\delta \text{ and } \forall t < \tau \text{ we have } w(t) \in \Omega \setminus \cup_j C_j^{-\delta}\},$$

$$B_T = \{w \in D([0,T], \mathbb{R}^d \mid \exists \tau \text{ such that } w(\tau) \in C_i^\delta \text{ and } \forall t < \tau \text{ we have } w(t) \in \Omega \setminus \cup_j C_j^{-\delta}\}$$

We define the stopping time: $\tau(Y) = \inf_t Y \in C_i^\delta \cup_j C_j^{-\delta}$. Since $C_i \subset \Omega$ and Ω has a single connected component, we have that $\mathbb{P}(\tau(Y_t^x) < \infty) = 1$, in particular that means that there exists T_0 such that for any $T \geq T_0$, $\mathbb{P}(\tau(Y^x) \leq T) \geq 1 - \epsilon/6$. Using Theorem 2, we have that, almost surely

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

Hence, we have:

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

Using an argument similar to the one developed in Lemma 9, we can show that $\mathbb{P}(Y^x \in \partial B_T) = \mathbb{P}(Y^x \in \partial B) = 0$, hence by applying Theorem 2 on the set B_T, we obtain, in high probability,

$$\|\mathbb{P}(M^{x,h,n} \in B_T) - \mathbb{P}(Y^x \in B_T)\|_{\infty,U} \leq \epsilon/4$$

Combined with our previous result, we obtain:

$$\|\mathbb{P}(M^{x,h,n} \in B) - \mathbb{P}(Y^x \in B)\|_{\infty,U} \leq 3\epsilon/4$$

Using our assumption on $\hat{C}_{i,n}$, we have that, in high probability, $\mathbb{P}(M^{x,h,n} \in B) \geq \hat{\mu}_{i,h,n}$, therefore using our previous bound between μ and μ^+:

$$\hat{\mu}_{i,h,n}(x) - \mu_i(x) \leq \epsilon.$$

By applying the same arguments for μ^-,

$$\mu_i(x) - \hat{\mu}_{i,h,n}(x) \leq \epsilon,$$

concluding the proof.

Lemma 9. *Let $O_{z,h,\alpha}$ be a truncated cone. Let $\mathcal{B}_{z,h}$ be the ball of radius h centered in z. Let P^x ne the probability for Y^x to hit $O_{z,h,\alpha}$ before hitting $\mathcal{B}_{z,h}$.*

For any $\epsilon > 0$, there exists $r > 0$ such that if $\|x - z\|_2 \leq r$ then $P^x \geq 1 - \epsilon$.

Proof. The proof is composed of two steps, first we show that this result is true for the Brownian motion, then we use a change of measure to show that the result holds for the diffusion process.

Let us show that the result is true for a Brownian motion. Let $k \geq 0$ and let $x \in \mathcal{B}_{z,2^{-k-1}h}$, by the scaling invariance of the Brownian motion, there exists a such that the probability for a trajectory originated at x to hit $\mathcal{B}_{z,2^k h}$ before hitting the cone $O_{z,h,\alpha}$ is smaller than a Using the strong Markov property of the Brownian motion as well we have that, for $k \geq 0$ and $x_0 \in \mathcal{B}_{z,2^{-k-1}h}$, the probability for the Brownian motion to hit $O_{z,h,\alpha}$ before hitting $\mathcal{B}_{z,h}$ is $1 - a^k$. Moreover, as x gets closer to z the time necessary for the trajectories to hit the cone goes to 0.

We can then obtain the proof for the diffusion process using a change of measure, see [5], proof of Theorem 5.4 p.206. □