Report: Generalized Sliced Distances for Probability Distributions

Mathis Hardion

MVA, Télécom Paris

Abstract

Comparing probability distributions is of increasing relevance in many fields of statistics and machine learning, e.g. hypothesis testing and generative modeling. Finding computationally cheap, meaningful distances on the space of probability measures is therefore crucial to current tasks which often involve large, high dimensional datasets. Algorithms derived from such distances, for instance gradient flows, also need some convergence guarantees to be usable in practice and provide stable behaviors. The paper studied in this report introduces a class of probability metrics, named General Sliced Probabilty Metrics (GSPMs), shown to be proper distances. By studying a particular case, a link with Maximum Mean Discrepancy metrics is highlighted, and a corresponding gradient flow scheme is derived. The latter is shown under some regularity conditions to converge to the global optimum. This report studies the behavior of such metrics and corresponding flows on various synthetic examples to pinpoint its strengths and weaknesses, and a practical refinement of the gradient flow scheme is proposed.

I. Introduction

This report discusses the main contributions of [1], their implications, limitations and possible extensions. That paper aims to introduce a novel way to compute dissimilarity between probability distributions, i.e. introduce a family of metrics on the space of probability measures. The study and development of such distances is of high interest in machine learning and satistics, as it often deals with approximating the underlying distribution of data with a model, and optimize its parameters to get as "close" as possible to the target probability law, where the notion of "closeness" must be made objective adaptively to the problem as different applications may require different measures of distance. Examples include two-sample testing [2], generative modeling [3], [4], [5], clustering [6] etc. The computation of such metrics can be computationally demanding however, and considering the high dimensionality and large sample sizes of current tasks, it is crucial to consider cheaper yet still meaningful distances. In order to achieve this, [1] build upon previous frameworks which we now sketch.

I.1 Radon Transform

Denote $\mathcal{P}(\mathcal{X})$ for a measurable space $\mathcal{X}$ the set of probability measures on $\mathcal{X}$, and $C_b(\mathcal{X})$ the set of real-valued continuous bounded functions on $\mathcal{X}$. The "slice" of $\mu \in \mathcal{P}(\mathcal{X})$ with respect to $f \in C_b(\mathcal{X})$ is the pushforward $f_\sharp \mu$, corresponding to integrating μ along the

level sets of f, intuitively akin to slicing the space to observe a 1-dimensional distribution resulting from the "contraction" of μ along that slice. It is then of interest to find classes of functions $\mathcal{F} \subset C_b(\mathcal{X})$ such that the knowledge of the slices $f_\sharp\mu$ along all functions $f \in \mathcal{F}$ is sufficient to characterize μ, i.e. the map

$$\mathcal{R}_\mathcal{F} : \mu \mapsto \{f_\sharp\mu, f \in \mathcal{F}\} \tag{1}$$

is invertible. For the classical *Radon transform*, $\mathcal{X} = \left(\mathbb{R}^d, \langle\cdot,\cdot\rangle\right)$ and one considers a class of linear functions $F = \left\{x \mapsto \langle\theta, x\rangle, \theta \in \mathbb{S}^{d-1}\right\}$ where $\mathbb{S}^{d-1} \coloneqq \left\{\theta \in \mathbb{R}^d, \|\theta\|_2 = 1\right\}$ is the unit sphere. Such a Radon Transform is then invertible, see e.g. [7]. Since this transform corresponds to integration on hyperplanes, one can generalize it to hypersurfaces i.e. $(d-1)$-manifolds, resulting in the so-called Generalized Radon Transform (GRT). The classes of considered functiond are traditionally parametric, i.e. $\mathcal{F} = \{f_\theta \in C_b(\mathcal{X}), \theta \in \Omega\}$ for some $\Omega \subset \mathbb{R}^n \setminus \{0\}$. Some necessary regularity conditions on such functions are shown in [8], and some well-known examples of functions classes guaranteeing invertibility include circular slices $f_\theta : x \mapsto \|x - s\theta\|_2$ for some $s > 0$ and θ varying in $\mathbb{S}^{d-1}$ [9], and polynomials of odd degree homogenous in $\theta \in \mathbb{S}^{p-1}$ where p is the degree of said polynomials [10].

The knowledge of family of slices for which one can recover the original measure is of interest since it allows the definition of *sliced distances*. In the case of optimal transport, Wasserstein distances between 1-dimensional distributions are far easier to compute than in higher dimensions as they have an closed form. Thus, the idea of slicing probability measures to reduce complexity of barycenter computation is introduced in [4], under the classic Radon Transform framework. The resulting distance, coined *Sliced Wasserstein Distance*, has found successful extensions in generative imaging [11], [12], and was then generalized to the GRT case in [13], leading to the studied paper [1] aiming to explore the slicing of other distances on the space of probability measures.

I.2 Gradient Flows

Once a way to compare probability measures is provided, a problem of interest is then to minimize such a distance between a generated distribution and a target. To this end, gradient flows [14] have gained some attention, for instance in generative modeling [15], [16]. This method amounts to performing gradient descent in the space of probability distributions, often using particle systems and the Euler-Maruyama scheme [17]. Convergence towards minima of a distance can allow one to obtain a distribution "close" to the target yet distinct, hence its usefulness in generative purposes. Local minima could be sufficient in some cases, but it is difficult to explicit which ones, meaning that convergence towards a global minimum is a desirable property in general. Hence, though a particular case of interest, the paper builds upon previously introduced results for MMD distances [16] to derive a noisy gradient flow with respect to their newly introduced metrics which converges towards the target distribution.

II. Contributions

II.1 Generalized Sliced Probability Metrics (GSPMs)

The main objects introduced in the paper are the GSPMs, which we define as follows.

Definition 1. *Let ξ be a distance on the space of probability measures on $\mathbb{R}$, and $\mathcal{F} = \{f_\theta \in C_b(\mathcal{X}), \theta \in \Omega\} \subset C_b(\mathcal{X})$ a class of functions for which the GRT is invertible. Then*

the (r-)GSPM corresponding to ξ and $\mathcal{F}$ is defined for probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ as

$$\zeta_{\mathcal{F}}(\mu, \nu) = \left(\int_{\Omega} \xi \left(f_\theta \sharp \mu, f_\theta \sharp \nu \right)^r d\theta \right)^{\frac{1}{r}}. \tag{2}$$

Note that this is a slight reinterpretation of the original definition, as it was originally considered in the case of measures with density, but the definition is also valid in the general setting. It is easily shown that such a definition induces a proper distance [1]. Through studying a special case of GSPM, the authors highlight a link with Maximum Mean Discrepancy (MMD) norms introduced in [2], summarized in the following property.

Proposition 1. *Let A be positive definite linear operator on $L^2(\mathbb{R}, Leb)$ endowed with its usual norm $\|\cdot\|_2$, and consider the distance over the space of probability measures on $\mathbb{R}$ with density w.r.t. Lebesgue defined by $\xi(p, q) \coloneqq \|A(p - q)\|_2$ for two probabilty density functions p and q. Given observations $(x_i)_{i=1}^N \overset{i.i.d.}{\sim} \mu$ and $(y_j)_{j=1}^M \overset{i.i.d.}{\sim} \nu$, define the smoothened empirical sliced densities of μ as $\widehat{p}_\theta = \frac{1}{N} \sum_{i=1}^N \phi_\sigma(\cdot - f_\theta(x_i))$ where ϕ_σ is a radial basis function of radius σ i.e. smooth, L^2 converging to the dirac distribution as $\sigma \to 0$; and define analogously $\widehat{q}_\theta$ for ν, inducing measures on the whole space $\widehat{p}$, $\widehat{q}$ having corresponding slices. Then, the corresponding GSPM $\zeta(\widehat{p}, \widehat{q})$ is the empirical MMD associated to the PD kernel*

$$k : (x, y) \mapsto \int_{\Omega} \langle A\phi_\sigma(\cdot - f_\theta(y)), A\phi_\sigma(\cdot - f_\theta(x)) \rangle \, d\theta, \tag{3}$$

where the scalar product is taken in L^2.

The defined kernel is in practice approximated using a Monte-Carlo estimator as the integral is often intractable. We thus sample uniformly on Ω i.e. take $(\theta_\ell)_{\ell=1}^L \overset{\text{i.i.d}}{\sim} \mathcal{U}(\Omega)$.

II.1.1 Behavior of GSPM-MMDs

We aim to better understand the way GSPM-MMDs behave on some synthetic examples. We consider the We remind figure 1 the behavior of the Wasserstein distance between Gaussians as reference, and show figure 2 the correponding visualizations in the case of GSPM-MMD distances using different parameters. One observes that the latter plateaus quite quickly, and approaches the dirac distance as σ goes to 0. The fact that the empirical distance between two identical gaussians is nonzero simply comes from the stochasticity of the estimator, with the added effect of a decreasing σ resulting in higher distances as the set on which ϕ_σ is nonnegligible decreases. We explain the plateau with the following result.

Proposition 2. *Let ξ be the L^2 distance between densities (i.e. $A = id$ in the above), $\mathcal{F} = \{f_\theta \in C_b(\mathbb{R}^n), \theta \in \Omega\}$ for some $\Omega \subset \mathbb{R}^n \setminus \{0\}$ and ζ the corresponding GSPM(-MMD). Let μ, ν be two probability measures with densities, and assume these densities to be bounded above by respective constants B_μ, B_ν. Then, denoting Leb the Lebesgue measure, one has*

$$\zeta^2(\mu, \nu) \leq Leb\,(\Omega) \left(B_\mu^2 + B_\nu^2 \right). \tag{4}$$

The above proposition is straightforwardly proven using the triangular inequality. One can also observe figure 3 that the rate at which the distance plateaus is related to the thickness of the distributions' tail i.e. the rate at which the pdfs go to 0 at infinity, illustrated using Cauchy distributions.

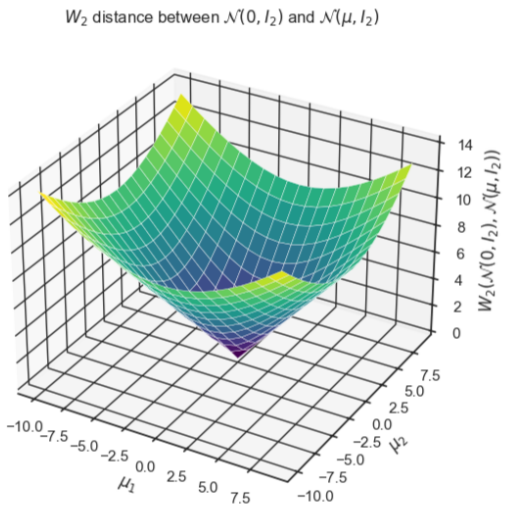

Figure 1: Wasserstein distance between gaussians ($L = 50$, $N = 200$, $M = 200$)

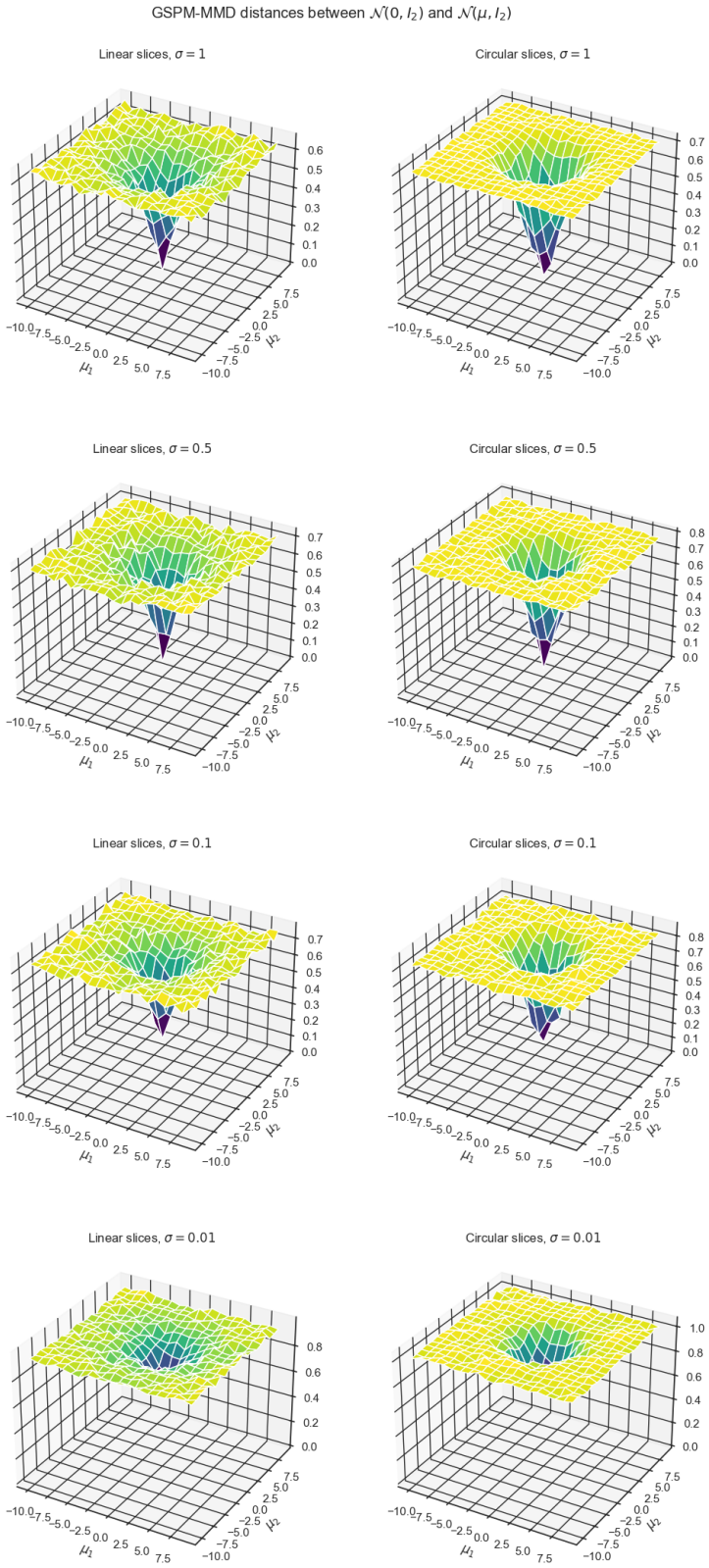

Figure 2: GSPM-MMD distance between gaussians for different slices and radii ($L = 50$, $N = 200$, $M = 200$, $A = \mathrm{id}$)

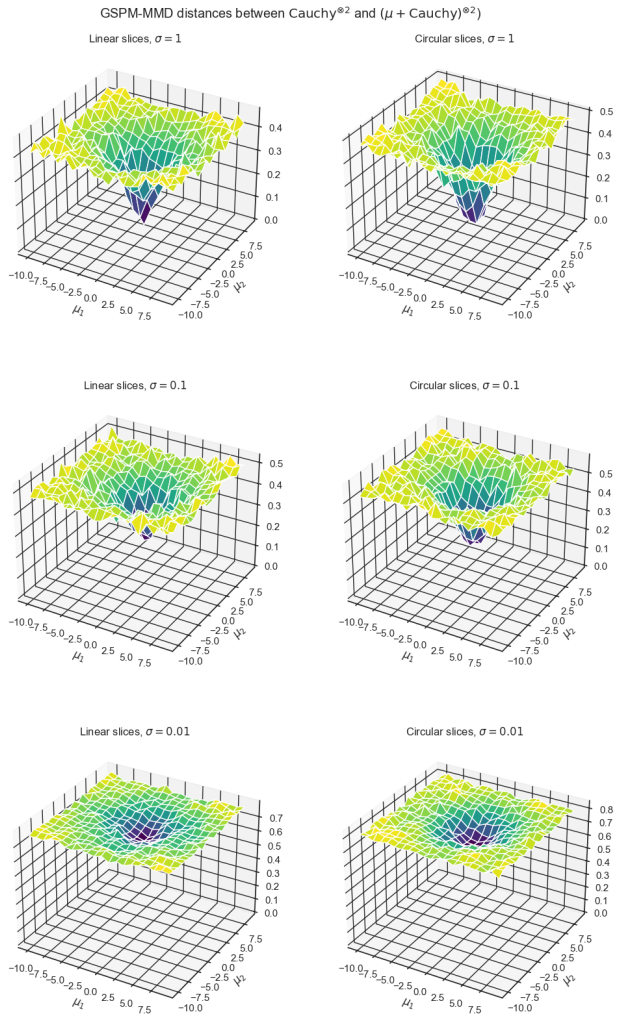

Figure 3: GSPM-MMD distance between Cauchy distributions for different slices and radii ($L = 50$, $N = 200$, $M = 200$, $A = \mathrm{id}$)

II.2 GSPM-MMD gradient flows

We now consider the gradient flows of GSPM-MMDs using the noisy Euler-Maruyama scheme explored in [1] following the work in [16]. For a target distribution μ, the goal is essentially to minimize $\nu \mapsto \zeta_{\mathcal{F}}(\nu, \mu)$ via gradient descent. The scheme writes as follows at iteration n:

$$X_{n+1} = X_n + \eta v(X_n + \beta_n U_n, \nu_n) \tag{5}$$

Where X_n denotes a particle of law ν_n, η is a step size, $(U_k)_k$ are i.i.d. standard normals, β_n is the noise standard deviation, and the vector field v is given by

$$\forall x \in \mathbb{R}^d, \forall \nu \in \mathcal{P}(\mathbb{R}^d), v(x, \nu) = -\nabla_x \left(\int k(\cdot, x) d\mu - \int k(\cdot, x) d\nu \right) \tag{6}$$

for the considered GSPM kernel defined in (3). The authors then show the following convergence result which we next discuss.

Theorem 1. *Assume ν_0 to have finite second order moment, A to be a linear bounded positive semi-definite operator, the gradients ∇f_θ to be bounded uniformly with respect to θ and Lipschitz for a constant independant of θ, and ϕ_σ to be bounded Lipschitz of bounded Lipschitz derivative. Further assume that $\sum_{i=0}^{\infty} \beta_i^2 = \infty$. Then, with the notations from (5), there exists constants L and λ verifying*

$$\zeta(\nu_n, \mu) \leq \zeta(\nu_n, \mu) e^{-2\lambda^2 \eta(1-3\eta L) \sum_{i=0}^{n} \beta_i^2}. \tag{7}$$

We refer to [1] following [16] for the proof and expression of constants L and λ, though the latter involve the operator norm of A and the Lipschitz constants and are thus intractable in practical cases. Considering the established bounds only goes to 0 if $\eta \leq \frac{1}{3L}$, this suggests the choice of "sufficiently small η" but essentially gives no information on when that is verified. It does give a guideline for the choice of the (β_n), however it is not used by the authors in their numerical experiments as they choose $\beta_n = \frac{\beta_0}{n+1}$ so that the distribution stabilizes quicker. We now illustrate the behavior of GSPM-MMD flows on our own synthetic distributions.

v is approximated using monte carlo estimation as before, requiring M samples of the target distribution and N evolving particles of law ν_n at each iteration, over n_{iter} iterations. The gaussian RBF is used, of standard deviation $\sigma/2$ so that the kernel involves a gaussian pdf of standard deviation σ, and the case $A = \mathrm{id}$ is considered for simplicity. Figure 4 illustrates flow between gaussians for linear and circular slices, with little notable difference between the two. In both cases, only a few particles actually make it to the target gaussian, which is due to the very thin tail of gaussian distributions resulting in a very flat landscape when far from the mode as previously discussed. Figure 5 illustrates the influence of the radius of the RBF on the flow for another pair of distributions, where we see that a higher σ results in smoother, initially faster trajectories but they are slower to explore the target distribution once it is reached, whereas a lower σ results in noisier paths but better exploring the target distribution. This is logical from the behavior seen figure 2, as a smaller sigma means a flatter landscape, meaning the gradients are smaller comparatively to the variance of the considered Monte Carlo estimators and thus the direction at each step is more erratic. We then study figure 6 a multimodal setting to observe whether the flow may get stuck in local minima, by looking at the flow from one gaussian to a gaussian mixture containing the initial gaussian. As one might expect, the flow indeed tends to get stuck in the potential well of the first gaussian, even with larger noise schemes. Larger values of σ slow down the convergence, which makes exploring other modes difficult. We now turn to a suggestion of solution to this issue.

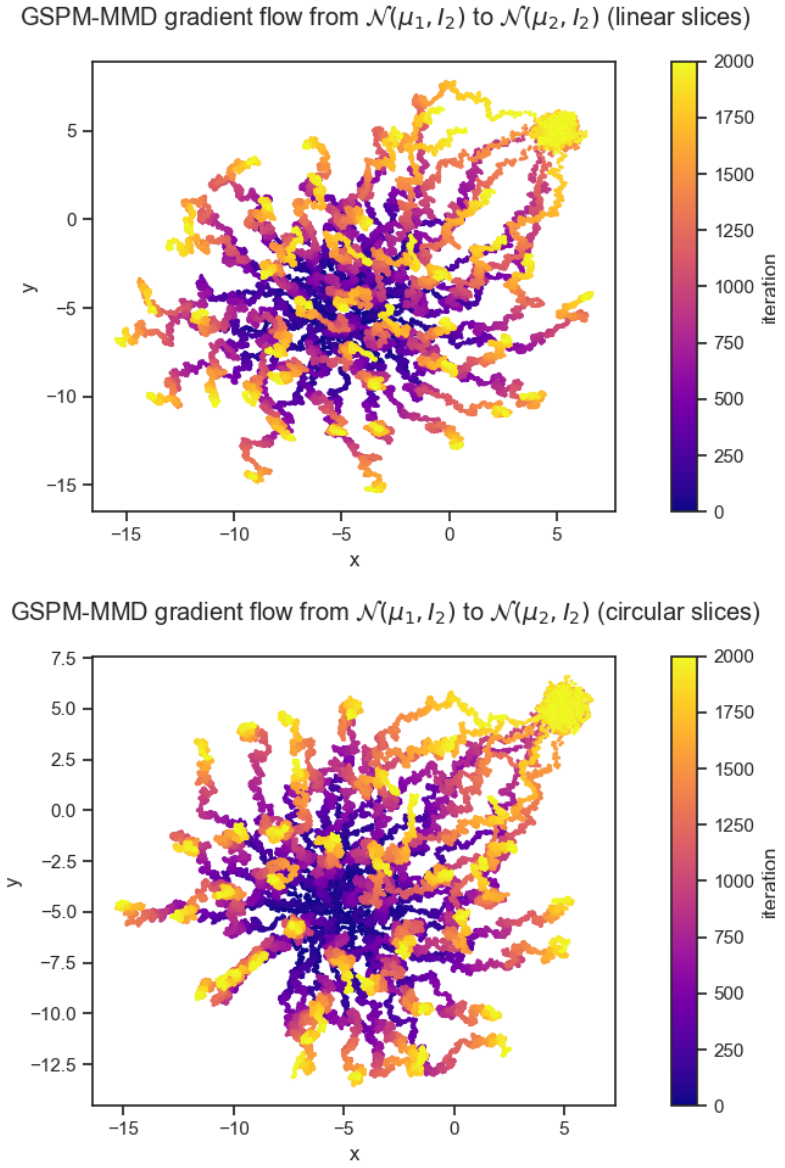

Figure 4: GSPM-MMD flow between two Gaussians with unit covariance, $\mu_1 = (-5,-5)^T, \mu_2 = (5,5)^T$, $N = 50$, $M = 50, L = 20, n_{iter} = 2000, \eta = 0.5, \beta_n = \frac{0.1}{(n+1)^2}, \sigma = 0.1$, linear slices (top), circular slices (bottom)

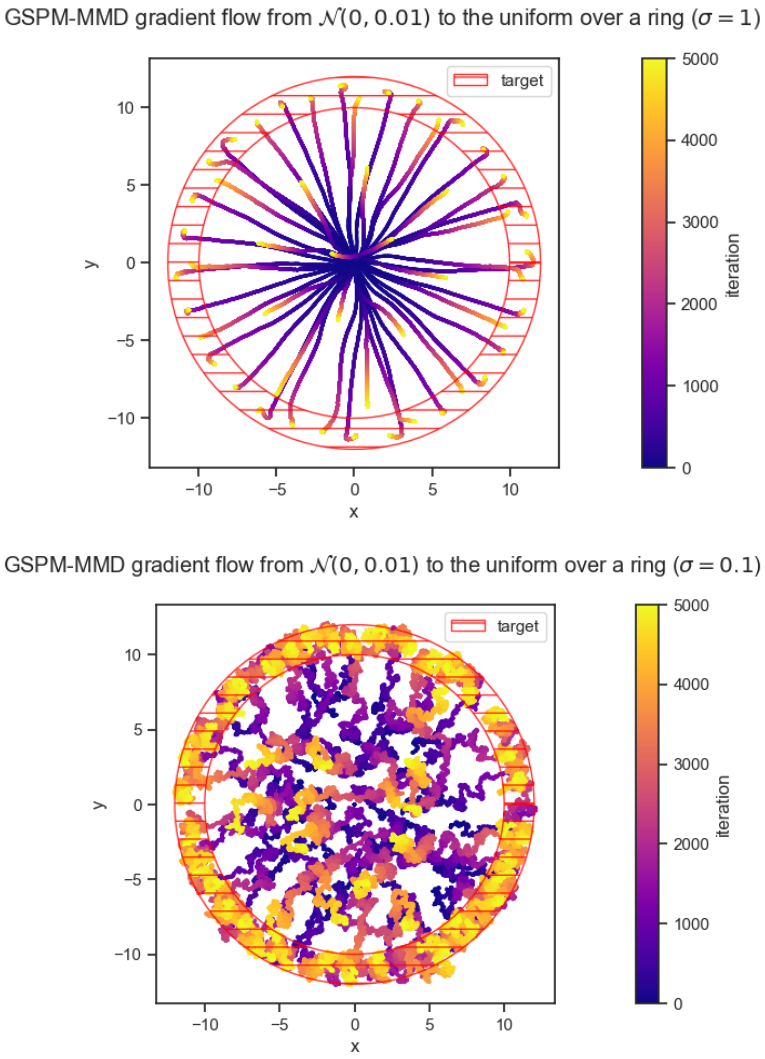

Figure 5: GSPM-MMD gradient flows from a $\mathcal{N}(0, 0.01I_2)$ to the uniform law over a ring of inner radius 10 and thickness 2, for RBF of radii $\sigma = 1$ (top) and 0.1 (bottom),$n_{iter} = 5000$, $N = M = 50$, $L = 20$, $\eta = .5$, $\beta_n = \frac{1}{(n+1)^2}$, linear slices

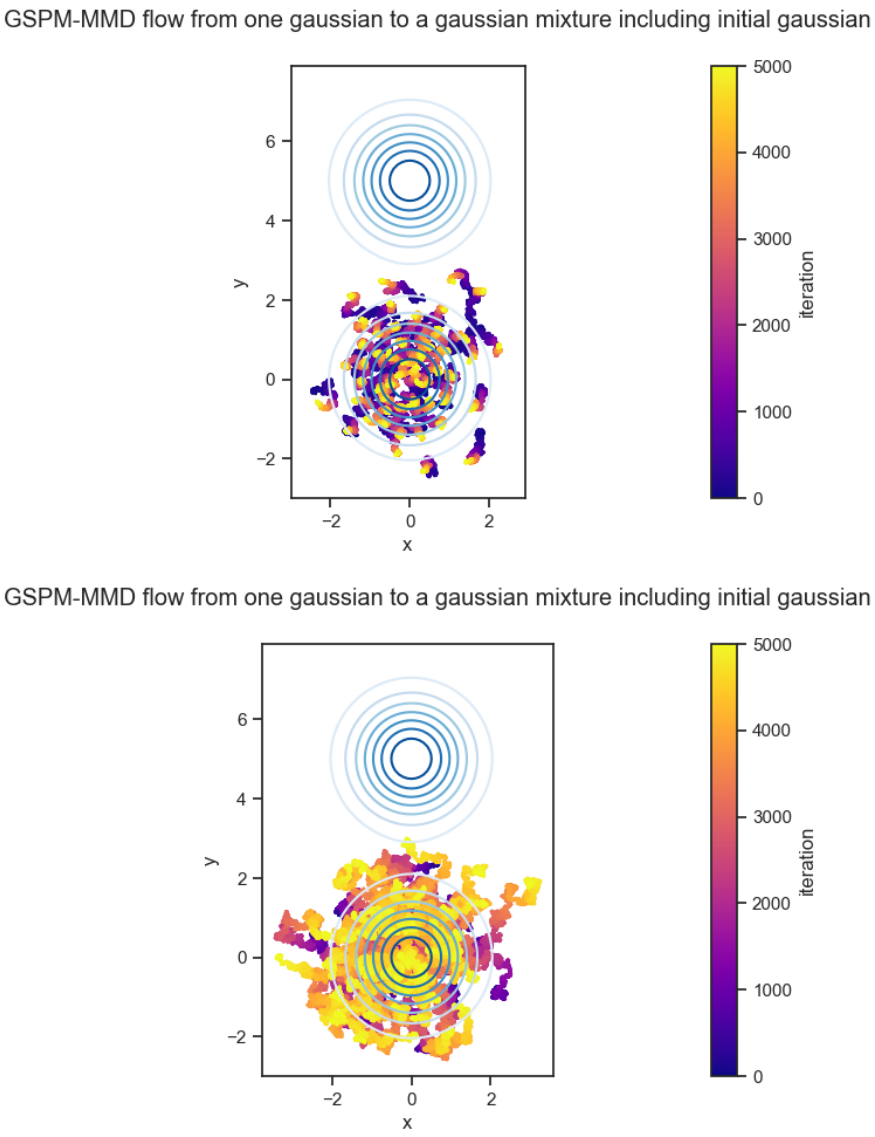

Figure 6: GSPM-MMD gradient flows in a multimodal setting, from $\mathcal{N}(0, I_2)$ to $\frac{1}{2}\left(\mathcal{N}(0, I_2) + \mathcal{N}((0, 5)^T, I_2)\right)$, $\beta_n = \frac{1}{n+1}$ (top) vs. $\frac{1}{\sqrt{n+1}}$ (bottom, $n_{iter} = 5000$, $N = M = 50$, $L = 20$, $\eta = .5$, linear slices

II.3 Extension of GSPM-MMD flow

With the behavior we have seen above in mind, it may be beneficial to change the value σ of the RBF's radius over iterations, so that it can initially benefit from larger gradients and gradually reduce them so that it explores the target distribution once close, staying above a minimal value $\sigma_{\min}$ so that the distance stays relevant. This is illustrated figure 7 and 8, on the distributions we have seen before. We qualitatively observe what was anticipated in that the trajectories are initially smoother and more explorative towards the end, and allow for seemingly better convergence than previously illustrated schemes. One also manages to explore different modes in the previous mixture case, as seen figure 9. The disadvantage is that this method requires extra parameter tuning, as the choice of the rate of decrease of σ_n may be a complex task in practical settings. From a theoretical standpoint, it is also complex to provide bounds for this scheme since the

considered kernel varies over iterations, although an extension of theorem 1 may be possible through a derivation akin to the one in [16] by carefully bounding distances at each iteration with respect to the distance of $\sigma_{\min}$, as distances appear to become larger as σ decreases. Verification of such properties is beyond the scope of this report.

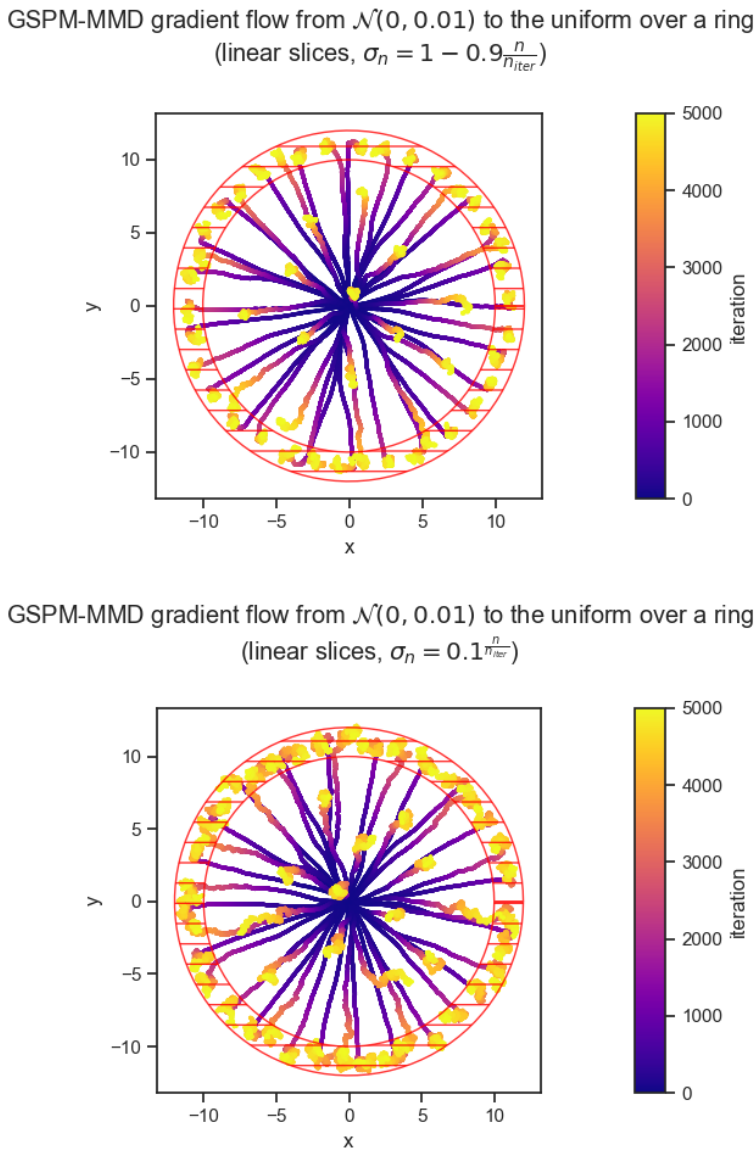

Figure 7: Changing radius GSPM-MMD flows with $\sigma_n = 1 - 0.9\frac{n}{n_{iter}}$ (top) and $0.1^{\frac{n}{n_{iter}}}$ (bottom), same other parameters as figure 5

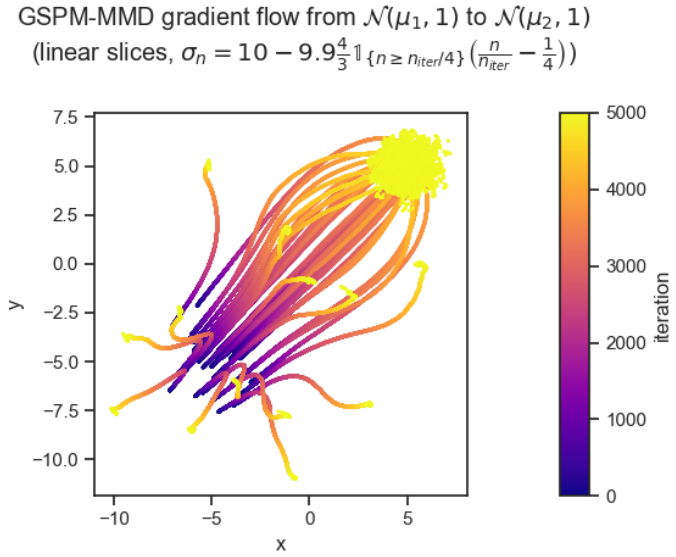

Figure 8: Changing radius GSPM-MMD flows with piecewise linear $\sigma_n = 10 - 9.9\frac{4}{3}\mathbb{1}_{\{n \geq n_{iter}/4\}}\left(\frac{n}{n_{iter}} - \frac{1}{4}\right)$ for the same other parameters as figure 4.

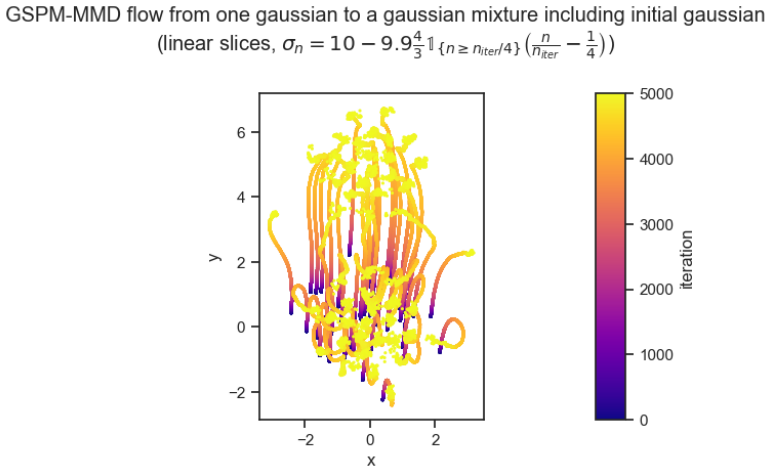

Figure 9: Changing radius GSPM-MMD flows with piecewise linear $\sigma_n = 10 - 9.9\frac{4}{3}\mathbb{1}_{\{n \geq n_{iter}/4\}}\left(\frac{n}{n_{iter}} - \frac{1}{4}\right)$ for the same other parameters as figure 6.

III. Conclusion

The studied paper [1] proposes new sliced metrics, allowing to lift a metric on 1-dimensional distribution to arbitrary spaces. It then explores a subcategory which is equivalent to MMD distances. The behavior of such metrics is seen to suffer from problems similar to that of Total Variation on most probability measures with density, which can be somewhat circumvented by tuning the RBF's radius but requires extra work as a result. The initial goal of generalizing sliced probability distributions to lower computational complexity is somewhat forgotten, as the introduced GSPM-MMDs require an extra Monte Carlo estimation because of the definition of the kernel as an integral. Moreover, for some of the most popular metrics outside of optimal transport, the mo-

tivation of slicing is not clear as if they already enjoy a closed form, for instance the Kullback Leibler divergence between discrete probability measures (which is essentially Monte Carlo estimation of the continuous case), adding an extra Monte Carlo step to integrate over slices could potentially add more computational cost as well as reduced accuracy. The idea of more general slices initially introduced in [13] may prove useful, but mostly in the case of Wasserstein distances.

As a second contribution, a gradient flow is proposed in the case of GSPM-MMDs, and the global convergence of the algorithm is proven, but is little more than the byproduct of propostions 1 and 8 of [16], i.e. the retained metrics are a particular case of well-studied objects. The articles' numerical experiments are only on very simple distributions, as even the MNIST dataset is far below the complexity of modern-day computational tasks. Hyperparameter tuning can be more comple as the introduction of an extra Monte-Carlo estimation of the kernel is needed and the RBF radius heavily influences the flows' behavior.

As an opening for future work, this report considers the case of a varying radius to obtain potentially more desirable flows which overcome local minima in the multimodal case. The suggested scheme could possibly be proven to also converge to the global optimum under regularity assumptions. From a theoretical standpoint, the relationship between sliced distributions and dual norms could be studied, since many popular metrics can be understood under this framework [18]. Some attempts made in the making of this report failed to highlight such a relationship if the 1D metric ξ is a dual norm, although it may still be feasible under more assumptions.

IV. Connexion with the course

The studied paper builds upon the notion of distances on the space of probability measures, first developed in optimal transport through the Wasserstein distances. The idea of slicing distances was first introduced in the Wasserstein case in [4]. As we have seen in the course, both Wasserstein and MMD (thus including GSPM-MMD) distances can be represented in the same framework as dual norms.

References

[1] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, and Shahin Shahrampour. *Generalized Sliced Distances for Probability Distributions*. 2020. arXiv: 2002.12537 [stat.ML].

[2] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. "A Kernel Method for the Two-Sample-Problem". In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, 2006. URL: https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.

[3] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. *Training generative neural networks via Maximum Mean Discrepancy optimization*. 2015. arXiv: 1505.03906 [stat.ML].

[4] Rabin Julien, Gabriel Peyré, Julie Delon, and Bernot Marc. "Wasserstein Barycenter and its Application to Texture Mixing". In: *SSVM'11*. Israel: Springer, 2011, pp. 435–446. URL: https://hal.science/hal-00476064.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[6] G. I. Papayiannis, G. N. Domazakis, D. Drivaliaris, S. Koukoulas, A. E. Tsekrekos, and A. N. Yannacopoulos. "On clustering uncertain and structured data with Wasserstein barycenters and a geodesic criterion for the number of clusters". In: *Journal of Statistical Computation and Simulation* 91.13 (Mar. 2021), pp. 2569–2594. ISSN: 1563-5163. DOI: 10.1080/00949655.2021.1903463. URL: http://dx.doi.org/10.1080/00949655.2021.1903463.

[7] Emmanuel Candes. *MATH 262/CME 372: Applied Fourier Analysis and Winter 2021 - Lecture 10*. Ed. by E. Bates. Elements of Modern Signal Processing. Stanford University. Feb. 2021. URL: https://candes.su.domains/teaching/math262/Lectures/Lecture10.pdf.

[8] Andrew Homan and Hanming Zhou. "Injectivity and Stability for a Generic Class of Generalized Radon Transforms". In: *The Journal of Geometric Analysis* 27.2 (Apr. 2017), pp. 1515–1529. ISSN: 1559-002X. DOI: 10.1007/s12220-016-9729-4. URL: https://doi.org/10.1007/s12220-016-9729-4.

[9] Peter Kuchment. "Generalized transforms of Radon type and their applications". In: *Proceedings of Symposia in Applied Mathematics* 63 (Jan. 2006). DOI: 10.1090/psapm/063/2208237.

[10] L. Ehrenpreis. "The universality of the radon transform". In: *Oxford University Press on Demand* (2003). DOI: 10.1017/S0013091505224823.

[11] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. "Sliced Wasserstein Auto-Encoders". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=H1xaJn05FQ.

[12] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. "Sliced and Radon Wasserstein Barycenters of Measures". In: *Journal of Mathematical Imaging and Vision* 1.51 (2015), pp. 22–45. DOI: 10.1007/s10851-014-0506-3. URL: https://hal.science/hal-00881872.

[13] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. "Generalized Sliced Wasserstein Distances". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/f0935e4cd5920aa6c7c996a5ee53a70f-Paper.pdf.

[14] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Birkhäuser, 2008.

[15] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. "Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 4104–4113. URL: https://proceedings.mlr.press/v97/liutkus19a.html.

[16] Michael Arbel, Anna Korba, Adil SALIM, and Arthur Gretton. "Maximum Mean Discrepancy Gradient Flow". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/944a5ae3483ed5c1e10bbccb7942a279-Paper.pdf.

[17] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin, Heidelberg, Aug. 1992. ISBN: 978-3-540-54062-5. DOI: 10.1007/978-3-662-12616-5.

[18] Jean Feydy. "Geometric Data Analysis, Beyond Convolutions". PhD thesis. Université Paris-Saclay, 2020.