## 8) IUF RESEARCH PROJECT 2022-2027

**Context: Optimal Transport**

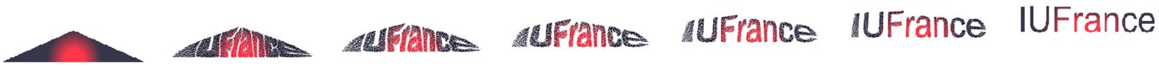

*Fig. 1. Top-view of optimally reshaping a source triangle-shaped mass into a target "IUFrance", with 5 interpolated/intermediate states. The red color is used to help following the trajectory of individual points.*

The domain of Optimal Transport (OT) stems from the problem that considers some mass (goods, units of material, people, etc) and that need to be redistributed from their current locations towards new locations. Intuitively, OT supposes that the cost of moving some material is proportional the mass and to a cost function $c$ (also called the *ground metric*) that depends on the source and target locations and which is often fixed for a problem, usually to the Euclidean distance.

### *Optimal Transport problem formulation*

In its probabilistic form, the optimal transport problem is formulated as finding a transport plan minimizing the total cost, i.e. a solution to the following OT problem:

$$OT(\mu, \nu, c) = \underset{\pi \in \Pi_{\mu,\nu}}{argmin} \int c(x,y) d\pi(x,y)$$

Where $\Pi_{\mu,\nu}$ is the set of admissible transport plans, i.e., the set of joint distributions on $X, Y$ that have marginals $\mu$ and $\nu$. A total cost is associated with each transport plan. Under conditions on the ground metric, the optimal total cost defines a distance between distributions ($\mu$ and $\nu$ here), namely the Wasserstein distance. Proving that a particular transport formulation yields a proper distance is one of the question around any new formulations in optimal transport, however, in the rest of this proposal, we will use the term "distance" loosely.

Formulated in terms of distributions, OT can be considered both between continuous distributions and between empirical ones (i.e., sums of Dirac-delta distribution, i.e, sets of points) but also between continuous and empirical ones. Most formulations can handle both continuous and empirical distributions, usually switching integrals for sums, but most algorithms are designed for empirical distributions (datasets).

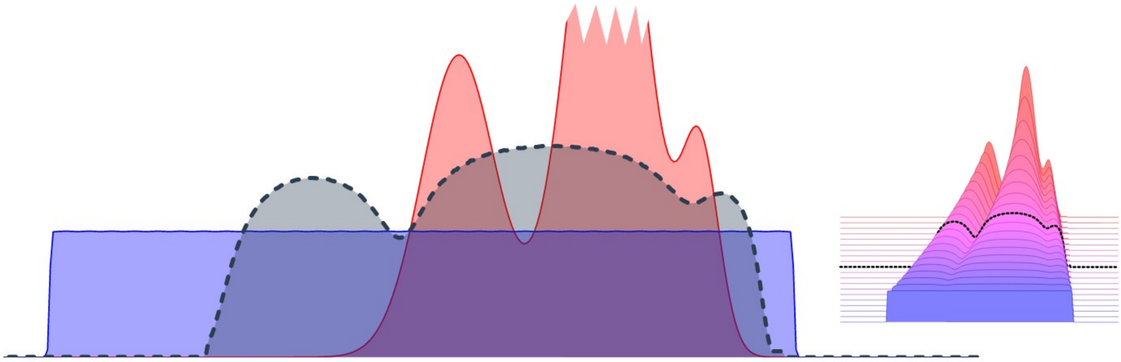

*Fig. 2. Illustration of 1D continuous optimal transport between the source distribution μ (blue) and the target distribution ν (red). The grey distribution shows an intermediate state during the transportation of mass and also corresponds to a Wasserstein barycenter of μ and ν. More steps are shown on the right.*

### *Importance of Optimal Transport in Machine Learning*

OT is widely used in statistical machine learning (ML) as it is a natural fit for many of the ML questions. Indeed, theoretical machine learning involves a lot reasoning about probability distributions, know (prior, dataset) or unknown (true distribution), continuous (prior, true distribution) or discrete (dataset, sampled distribution, stochastic approximations). In this context, the Wassertein distance is more and more used as a robust alternative to the KL divergence which might not be defined or might not provide good "differential/gradient" information. A lot remains to be done in this domain and we can expect a strong impact of OT formalism on ML.

In transfer learning, the formulation of the problems often involves a term of divergence between the source data distribution and the target one. As a mathematical tool to reason about such divergence, OT is pervasively now used in transfer learning: it has first been an alternate way to explain adversarial training approaches (e.g., for domain adaptation or generative adversarial networks) and has quickly hinted robust approaches to such problems.

OT is however mostly limited to traditional datasets where we need only to compare distributions on the same space. Some formulations have been introduced to go beyond this setting, making it possible to compare, e.g. distributions from different spaces, or graphs, but currently with major limitations on the scalability.

### *Extensions of Optimal Transport*

In this section, we focus on aspects related to modeling and mathematical formulations of some extensions of optimal transport beyond the simple distribution-to-distribution mapping. These extensions show that the formalism is flexible and can be adapted further to cover an ever-growing range of settings and applications. Considerations about scalability are paramount but they are pushed back to the next section.

A constraint of OT is that the sets of points (or distributions) to be matched, $\mu$ and $\nu$, should lie in the same space or, alternatively, that we provide a cost function that compares object from different spaces, which is usually not easy. This limits the applicability of OT and, thus, the Gromov Wasserstein (GW) problem has been proposed for distributions lying in different spaces. GW still aims at finding a transport between two sets of points but it aims at matching pairwise distances. The GW problem can be formulated as:

$$GW\left(\mathcal{L},C,C',\mu,\nu\right)=\underset{\pi\in\Pi_{\mu,\nu}}{argmin}\iint\mathcal{L}(C(x,x'),C'(y,y'))d\pi(x,y)d\pi(x',y')$$

or in its discrete form

$$GW_{discrete}\left(L,a,b\right)=\underset{T\in\Pi_{a,b}}{argmin}\sum_{i,j=1}^{I,I}\sum_{k,l=1}^{K,K}L_{ijkl}T_{ik}T_{jl}$$

As GW uses pairwise distances in each space, it can be applied on different spaces. It can also work with weighted graphs and defines a distance on them, which opens to a lot of applications that need a way to compare graphs.

Another OT extension to handle distributions in different spaces is Co-Optimal Transport (CO-OT) [1]. The principle is to find one transport between points, as in normal OT, but to also learn a second transport plan that aligns features between the two spaces.

Orthogonal to the previous extensions is the concept of multi-marginal optimal transport [2] in which more than two (e.g., $r=3$) marginal distributions need to be aligned. The problem is specified by $r$ marginals and a cost function with $r$ parameters; and a transport plan is an order $r$ tensor.

As OT defines a distance between distributions, it is also used to define a notion of barycenter (Fréchet mean) of distributions. This has been showcased for example for shape (seen as densities) interpolation [3] (and Fig. 1) or, with GW, for averaging contours [4] or graphs [5].

We proposed, in [6], the Optimal Tensor Transport (OTT) formulation which generalizes the problems of OT, GW, and CO-OT (see Fig. 3). OTT additionally allows to handle more complex structures such as datasets made of triplets (or any tuples), collections of weighted graphs, or points with several feature axes (generalization of CO-OT).

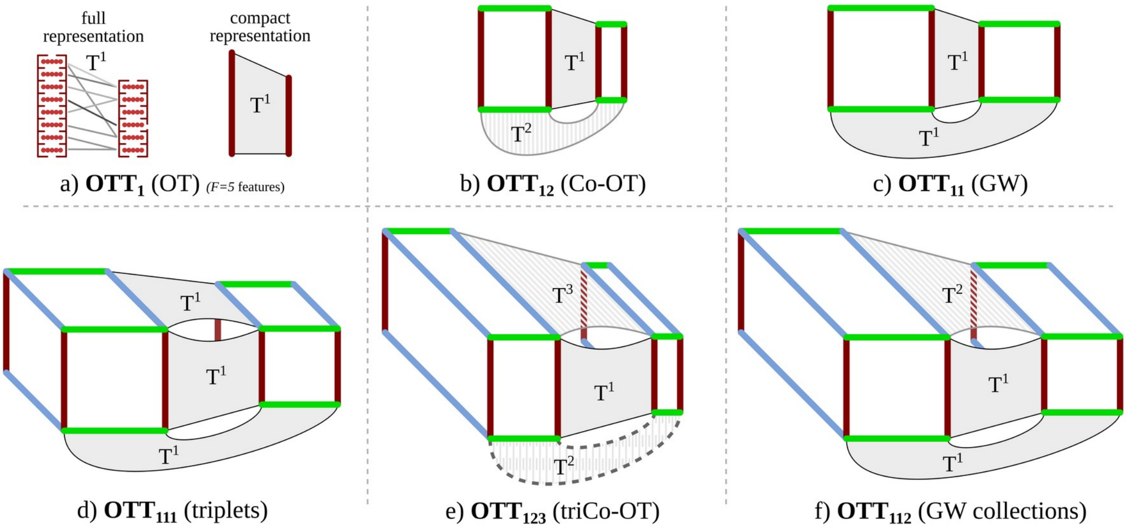

*Fig. 3. Visualization of Optimal Tensor Transport (OTT) with existing approaches (top) and new ones made possible by OTT (bottom).*

### *Complexity and Scalability of Optimal Transport*

Considering two empirical distributions each with $N$ points, the OT problem has a complexity that depends on the dimensionality of the space (and the ground metric used). In 1d, the optimal transport plan is the same for most relevant ground metrics and can be obtained by sorting each set of point and then matching them in order, yielding a $N.\log(N)$ time complexity. The general problem is a linear programming problem but it involves many variables and its complexity is $N^3.\log(N)$ in the worst case and super-cubic in practice [7].

Entropic reguralization [7] has been proposed to make the OT problem strongly convex and suggests the use of the Sinkhorn algorithm. For $P$ iterations of the algorithm, the time complexity is brought down to $P.N^2$, the number $P$ of iterations necessary to converge being mostly independent of $N$ [7].

Approaches like Sliced Wassertein [9] leverage random projections and the $N.\log(N)$ complexity of the one-dimensional OT solver to produce extremely fast algorithms. One limitation of most sliced methods is that they don't directly provide a transport plan, nor they estimate the Wassertein distance (they compute a different distance).

As for the GW problem, it is a generalization of the quadratic assignment problem (QAP) which is NP-hard, so all know algorithms are approximate ones. Entropic Gromov Wasserstein (EGW) is the most used algorithm for finding a solution to the GW problem. It runs $S$ iterations of a projected mirror descent where each iteration is itself an OT problem, solved with $P$ iterations of the Sinkhorn algorithm. Considering $P \lll N^2$, it yields a time complexity of $S.N^3$ for particular losses $L$ and $S.N^4$ in the general case (see [10] for details and a summary).

Hierarchical approaches such as [11] use a closed-form expression for barycenters, only available with certain loss functions, to hierarchically divide the GW problem, with a resulting $N^2.\log(N)$ complexity.

In [10], we identified a matrix-expectation in the GW formulation and proposed to sample it instead of using the full computation used in EGW. With $M$ samples this yields a $S.M.N^2$ time complexity, with any loss function. We also proposed a kind of sliced variation, by setting $M=1$ and leveraging the 1d OT solver. It brings down the time complexity to $S.N.\log(N)$, at the cost of accuracy in certain cases. Related to our sampling approach, a recent preprint [12] sparsifies the computations and brings down the complexity to $S.N^2$ for the general case.

As for the general OTT formulation, with order $D$ tensors (also an NP-hard problem), a typical non-stochastic algorithm would have a $S.N^{2D}$ or $S.N^{D+1}$ complexity with the square loss, which is prohibitive. Our sampling approach still displays a $S.M.N^2$ complexity.

### Project structure and planning

The project is articulated as 4 research directions (D1-D4). While D3 will be developed around a just-started Ph.D. Thesis, and D1 around a tentative funding for a Ph.D. Thesis (hopefully effective next year or the year after, or alternatively without Ph.D. student), I will develop D2 by working in collaboration with colleagues from the team but also internationally (starting right away in the context of the APRIORI ANR project). D4 will mainly be personal work at first (just started) but will most probably grow by creating a dedicated working group at the horizon of 1 or 2 years.

### *D1) Direction 1: scalable extensions of optimal transport*

This direction of work relies on 3 sub-axes that are well defined. These would ideally feed some collaboration around a thesis subject (or else internships), with applications on time series.

While OTT is a very generic formulation, it can be extended to include even more principles in a unified general framework. Among the orthogonal aspects that could benefit from being factored in the formulation are the multi-marginal setting , the marginal relaxation [13], the inclusion of class information, the inclusion of class shift [14] or the use of metric learning [15]. Another generalization, which is rather direct, is the inclusion of "fused" formulations where several OTT problems are optimized jointly, as done in Fused-GW which optimizes both (as a weighted sum) a OT loss and a GW loss. **Proposing an integrated generalized framework**, can transform optimal transport into a new way of modeling (like graphical models for probabilistic models), especially for transfer learning, and act as a pivot representation for novel orthogonal contributions like new types of transport plans (e.g., for times series), types of constraints (e.g., relaxed marginals with constraints), algorithms to solve the problem (see below).

One of the main challenges of OT, and especially of its generalization like GW and OTT is the scalability. It remains an open problem that prevents the use of these methods even on moderate-size datasets. Existing fast approaches can be classified in three categories: "sliced" approaches based on random projections [9][16][10], stochastic approaches [10][12] based on sampling, and hierarchical approaches [11][17]. **Designing hierarchical stochastic approaches with uncertainty quantification**, combined with the $N.\log(N)$ time complexity of slicing, can bring down the practical complexity of (approximate) generalized OT problems. Intuitively, a hierarchical approach allows to deal with huge datasets by simplifying them to an intermediate size. On this simplified form, the actual algorithms (like EGW) still have high complexity above $N^3$ for the simplest GW case and thus sampling is the way to reduce this complexity at the cost of higher uncertainty. By better modeling this uncertainty, for instance in a Bayesian formulation [18], we can expect to derive more robust fast algorithms. Slicing can play a role at different levels in a generalized setting, to bring down the complexity further: it can be used as an inner step as in PoGroW [10] but requires a selection (e.g., max-sliced) or voting scheme (to properly aggregate different slices), or it can be used as a randomized initialization method for an approach that would further refine it (e.g. to locally explore the polytope of admissible transport plans). A very recent preprint [19] explores hierarchical sliced approaches for the standard OT problem.

Another promising direction to improve scalability for problems such as OTT is the one proposed in DifFused Gromov Wasserstein (DFGW) [20]. Introduced in a fused-GW setting (working with labeled weighted graphs), DFGW leverages the edge information (graphs induced by the GW costs) to diffuse the node features (OT information) and then solves a traditional OT problem, which has much lower complexity than GW. **Generalizing DifFused-GW to OTT** opens many questions about how to generalize the graph diffusion to higher order tensors. Other interesting problems are suggested by DGFW and its potential application fields: how to use graphs with directional information, how to used it for sub-graphs matching, how to use faster approximate graph diffusion algorithms, how is it linked to graph neural networks combined with OT, etc. A natural way of working on this subject would be in a direct collaboration with the authors of DFGW [20].

### *D2) Direction 2: generalization guarantees and OT*

This direction of work considers mainly the properties that can be theoretically proven on the OT solutions or algorithms (e.g., from the D1 section). Theoretical guarantees for machine learning are a domain of expertise in the team (and collaborations).

As a distance between distributions, both continuous and empirical, the Wasserstein distance have been used in most situations where the Kullback-Leibler (KL) divergence usually appears. Indeed, the KL divergence is pervasive in statistical machine learning but may be not defined (e.g., infinite value, depending on the supports of the distributions). The Wasserstein distance gives meaningful values even for disjoint supports and thus provide better “gradient” information for optimization algorithms. As a robust distance, the Wasserstein distance has been used in many domains: domain adaptation, generative adversarial networks, auto-encoders (WAE), variational inference, etc. Theoretical work exist that derive generalization guarantees based on the Wasserstein distance (e.g., our work on metric learning [15]).

When designing new OT formulations and algorithms, in approaches like OTT (e.g., that use a stochastic Frank-Wolfe or projected mirror descent), we manage to prove the convergence of the algorithm, but no generalization aspects are taken into account. The standard (and entropic) optimal transport problem has been studied from this perspective (e.g., [21][22]). However, **deriving sample complexity and generalization bounds for structured transport (GW, OTT)** is an open and difficult problem.

Formulating the problem of optimal transport in a probabilistic manner opens the reasoning about uncertainty and generalization. In particular, a Bayesian formulation can be readily reused from [18] (which uses it to handle stochastic cost functions). Indeed, entropic optimal transport with empirical distributions naturally involves categorical (multinomial) distributions and the Bayesian formulation adds Dirichlet priors to it. Generally, **leveraging the PAC-Bayesian (PB) framework for optimal transport** is a very promising direction. Sliced Wasserstein, as an averaging method (across slices), lends itself quite directly to the voting view of PB and a recent preprint explores this [23]. The PB framework is not limited to this voting setting and it is able to produce extremely tight generalization bound in the case of Categorical/Dirichlet conjugacy as we have shown in [24] and it can most probably be adapted for the Bayesian OT formulation. A more exploratory direction of work around the PB framework is the use of “disintegrated PB bounds” that can be related to stochastic optimization methods such as the ones used for OTT.

### *D3) Direction 3: OT for structured latent representations*

This direction of work is driven by a particular application domain but might still suggest some developments for D1 and D2. This line of work is related to a starting Ph.D. thesis on *unsupervised object detection*, in which some auto-encoders with very structured latent spaces are used [25][26]: the latent space represents the object properties, the encoder is an object detector and the decoder a renderer. The idea behind these models is to automatically decompose a set of images into their constituent recurrent objects. The question of **using OT to regularize and transfer structured latent representations** to speed up and improve unsupervised and self supervised learning could be very interesting to explore. The Ph.D. student thesis will not explore this direction (the focus of the thesis is on some other aspects). With additional time, I could work in synergy with the Ph.D. student and leverage my background on motif mining, probabilistic models and optimal transport to explore this direction by myself, developing transfer or inference approaches e.g. in the vein of [27] but with structured representations.

### *D4) Direction 4: OT, partial differential equations (PDE) and diffusion processes*

This direction of work consists in **exploring and getting a deeper understanding of the link between OT and partial differential equations (PDE)**, which is a stimulating task involving a lot of literature review, analysis and synthesis. This kind of domain-broadening activity requires extensive research-focused time, that the IUF will provide. Having started to work, within the laboratory, with physicists working on the theory and applications of laser-mater interaction, PDE are pervasive: they may be known, they may need to learned from data or their general form may be supposed and used as a guide/regularization for learning. As such, bridging PDE and OT could be a way to unify the physical knowledge (side-information from a machine learning perspective) with the actual machine learning optimization problems. I now strongly believe that it is key and mutually beneficial for both domains (machine learning, and laser-mater interaction) to further understand the links between current learning methods and physical models of the mater. I am also convinced that my particular background (at the crossing of probabilistic modeling, theory of machine learning, deep learning, structured optimal transport, etc.) can give me an original viewpoint and instill a dynamic in the collaboration between ML and laser-matter interaction physics at the lab.

In parallel to purely understanding the inter-relation between OT and PDEs, there are already concrete directions that I will explore, both theoretically or empirically. Various OT research has already underlined several links with differential equations (e.g., [28][29][30]) but the relation to structured OT remains to be studied. Also, recently, very successful methods for density modeling (e.g. for image generation) have used diffusion models (as stochastic or deterministic PDE) to generate training samples for learning the reverse (denoising) diffusion process. As such, an interesting research direction consists in **expliciting the link between OT and diffusion models**. Diffusion approaches actually optimize the same objective as the score-based generative models and more precisely denoising score matching approaches. There is a parallel to be drawn between the score network that is learned in these methods and the (gradient) of the Kantorovich-Rubinstein duality function that yields the “critic” network used in Wasserstein GANs. While not formulated in these exact terms (on the OT side, it uses the Benamou-Brenier formulation), a recent preprint [31] shows an equivalence (mathematically proved for a special case and empirically verified in all tested cases) between the diffusion models and the optimal transport to a unit normal distribution. This link prompts for the question about whether the general OT problem (and further structured OT problems) between two distribution is also equivalent to two (independent) diffusion processes, and if not, whether bounds can be derived and can be used for initialization or as an elementary step in a iterative algorithm. The links with recent PAC-Bayesian bounds based on trajectories are also promising to explore [32] in relation with D2.

[1] T. Vayer, I. Redko, R. Flamary, and N. Courty, “CO-optimal transport,” in *NeurIPS*, 2020.

[2] G. Carlier, “On a class of multidimensional optimal transportation problems,” *Journal of convex analysis*, 2003.

[3] J. Solomon *et al.*, “Convolutional wasserstein distances: Efficient optimal transportation on geometric domains,” *ACM Trans. Graph.*, 2015,

[4] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-wasserstein averaging of kernel and distance matrices," in *ICML*, 2016.

[5] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty, "Fused gromov-wasserstein distance for structured objects," *Algorithms*, 2020,

[6] T. Kerdoncuff, R. Emonet, M. Perrot, and M. Sebban, "Optimal tensor transport," *AAAI*, 2022,

[7] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NeurIPS*, 2013.

[8] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via sinkhorn iteration," in *NeurIPS*, 2017.

[9] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and radon wasserstein barycenters of measures," *J Math Imaging Vis*, 2015,

[10] T. Kerdoncuff, R. Emonet, and M. Sebban, "Sampled gromov wasserstein," *Mach Learn*, 2021,

[11] H. Xu, D. Luo, and L. Carin, "Scalable gromov-wasserstein learning for graph partitioning and matching," in *NeurIPS*, 2019.

[12] M. Li, J. Yu, H. Xu, and C. Meng, "Efficient approximation of gromov-wasserstein distance using importance sparsification." 10.48550/arXiv.2205.13573.

[13] J. Li and L. Lin, "Optimal transport with relaxed marginal constraints," *IEEE Access*, 2021,

[14] I. Redko, N. Courty, R. Flamary, and D. Tuia, "Optimal transport for multi-source domain adaptation under target shift," in *AISTATS*, 2019.

[15] T. Kerdoncuff, R. Emonet, and M. Sebban, "Metric learning in optimal transport for domain adaptation," in *IJCAI*, 2021.

[16] T. Vayer, R. Flamary, N. Courty, R. Tavenard, and L. Chapel, "Sliced gromov-wasserstein," in *NeurIPS*, 2019.

[17] Q. Mérigot, "A multiscale approach to optimal transport," *Computer Graphics Forum*, 2011,

[18] A. Mallasto, M. Heinonen, and S. Kaski, "Bayesian inference for optimal transport with stochastic cost," in *ACML*, 2021.

[19] K. Nguyen, T. Ren, H. Nguyen, L. Rout, T. Nguyen, and N. Ho, "Hierarchical sliced wasserstein distance." 10.48550/arXiv.2209.13570.

[20] A. Barbe, M. Sebban, P. Gonçalves, P. Borgnat, and R. Gribonval, "Graph diffusion wasserstein distances," in *Machine learning and knowledge discovery in databases*, 2021.

[21] G. Mena and J. Niles-Weed, "Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem," in *NeurIPS*, 2019.

[22] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *AISTATS*, 2019.

[23] R. Ohana, K. Nadjahi, A. Rakotomamonjy, and L. Ralaivola, "Shedding a PAC-bayesian light on adaptive sliced-wasserstein distances." 10.48550/arXiv.2206.03230.

[24] V. Zantedeschi *et al.*, "Learning stochastic majority votes by minimizing a PAC-bayes generalization bound," in *NeurIPS*, 2021.

[25] Z. Lin *et al.*, "SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition," presented at the International conference on learning representations, 2022.

[26] W. Zhu, Y. Shen, M. Liu, and L. P. Aguirre Sanchez, "GMAIR: Unsupervised object detection based on spatial attention and gaussian mixture model," *Comput Intell Neurosci*, 2022,

[27] L. Ambrogioni, U. Güçlü, Y. Güçlütürk, M. Hinne, M. A. J. van Gerven, and E. Maris, "Wasserstein variational inference," in *NeurIPS*, 2018.

[28] L. C. Evans, "Partial differential equations and monge-kantorovich mass transfer," *Curr. Dev. in Math.*, 1997,

[29] Y. Brenier, "Extended monge-kantorovich theory," in *Optimal transportation and applications: Lectures given at the c.i.m.e. Summer school, held in martina franca, italy, september 2-8, 2001*, L. Ambrosio, L. A.

Caffarelli, Y. Brenier, G. Buttazzo, C. Villani, and S. Salsa, Eds. Berlin, Heidelberg: Springer, 2003.

[30] P. Mokrov, A. Korotin, L. Li, A. Genevay, J. M. Solomon, and E. Burnaev, "Large-scale wasserstein gradient flows," in *NeurIPS*, 2021.

[31] V. Khrulkov and I. Oseledets, "Understanding DDPM latent codes through optimal transport." 10.48550/arXiv.2202.07477.

[32] E. Clerico, G. Deligiannidis, B. Guedj, and A. Doucet, "A PAC-bayes bound for deterministic classifiers." 10.48550/arXiv.2209.02525.