PDE methods in machine learning: from continuum dynamics to algorithms

Katy Craig (UCSB),
Joan Bruna (NYU),
Qin Li (UW-Madison),
Lnac Chizat (EPFL)

June 10-14, 2024

1 Scientific context of the workshop

Over the past twenty years, machine learning (ML) algorithms have led to breakthroughs in our digital lives. However, theoretical understanding of key methods remains elusive. Traditional approaches from theoretical computer science, based on analyzing algorithms at the fully discrete level, are still far from explaining the mechanisms underlying modern methods. Instead, the recent years have seen a surge in interest in how the *continuum* perspective rooted in the study of *partial differential equations* (PDEs) can shed light on properties of algorithms, providing insight into properties such as convergence behavior, parameter selection, and the design of new algorithms. At the same time, the new equations arising from ML algorithms have attracted interest in the PDE community, as they often fail traditional criteria for well-posedness and evade simple characterizations of long time behavior, motivating the need to push the boundaries of existing theory and develop new mathematical tools for their study. This two-way interest has made it imperative to build new interfaces between these fields, to develop common vocabulary and techniques.

Motivated by this recently emerging interplay, **our workshop brought together international experts in ML and PDE** to examine how mathematical tools from optimal transport, the calculus of variations, parabolic and hyperbolic PDE, kinetic theory, and numerical analysis can be applied to solve open problems in ML. The workshop focused in particular on two areas of ML which lie at the heart of modern technology and are ripe for analysis via PDE methods: artificial neural networks and sampling.

1.1 Artificial Neural Networks

Artificial Neural Networks (NNs) are parameterized computer programs that transform input data to match a desired output, via adjusting the parameters during a *training phase*. For example, in a supervised learning task, a NN is a function $f(x, w)$ parametrized by weights w, and the training

phase uses a set of labeled data $\{(x_i, y_i)\}_{i=1}^n$ to adjust the weights in the parametrization so that $f(x_i, w) \approx y_i$ for all i.

A common choice of *architecture* f is the *multilayer perceptron*, which consists of a cascade of affine transformations, parameterized by weights $w = (W_1, \dots, W_L) \in \mathbb{R}^{d\times m} \times \mathbb{R}^{m\times m} \times \cdots \times \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times k}$, and intertwined with fixed element-wise non-linear maps $\sigma : \mathbb{R} \to \mathbb{R}$,

$$f(x, w) = W_L^\top \sigma(W_{L-1}^\top \sigma(\dots \sigma(W_1^\top x)) \dots)).$$

A typical choice of σ is the Rectified linear unit (ReLU) $\sigma(u) = \max\{0, u\}$. Given such an f and an initial random choice of weights w, the weights are then adjusted to improve the approximation $f(x_i, w) \approx y_i$ by applying gradient-based optimization methods to minimize the discrepancy between $f(x_i, w)$ and y_i for all $i = 1, \dots, n$. A major open problem in the theory of ML is to understand the *dynamics* by which the weights w change during the training phase: where do they converge? At which speed? How is this influenced by the structure of the data, the architecture, and the initialization of the training algorithm? Over the past five years, two distinct lines of research have emerged to apply PDE methods to analyze these problems, focusing on the *two-layer* ($L = 2$) and *multilayer* ($L > 2$) cases.

In the two layer case, a series of influential recent works discovered that, if the initial parameters are sampled with mean field variance scaling, the training dynamics converge to a *Wasserstein gradient flow* of the risk, which in the continuum limit as $m \to +\infty$ (the "infinite width limit") is a nonlinear, nonlocal transport PDE. This breakthrough opened up both an entirely new set of tools for studying the training dynamics as well as an entirely new class of equations to be studied via the Wasserstein gradient flow theory, equations which fail the traditional criteria for well-posedness and long time convergence to a global optimum. Open questions that appear ripe for analysis using these techniques include obtaining quantitative stability estimates for the training dynamics, understanding the interplay between regularity of solutions and convexity of the energy landscape, and analyzing the effect of regularization on convergence to equilibrium.

In the multilayer case, the Wasserstein gradient flow structure breaks down and there are many potential directions for studying the limiting behavior of dynamics, depending on the scale of the initial parameters In spite of the fact that research on the multilayer case has not yet coalesced around a single model, some questions still seem ripe for the study using existing tools. In particular, PDE based tools seem well-equipped to explain the phenomena of *neural collapse* [17], in which the optimal choice of parameters w found in practice leads to a function $x \mapsto f(x, w)$, the range of which concentrates around well-separated clusters.

1.2 Sampling

A second topic within ML that has emerged as a key point of intersection with PDE is sampling. Over the past decade, sampling has become one of the most important topics in theoretical ML. Indeed, sampling is the key computational bottleneck to train probabilistic models of high-dimensional data, and it is also fundamental in inverse problems, data assimilation, and non-parameterized modeling.

While the study of sampling witnessed several major advances in the 90s, through the design and study of MCMC solvers, with the rise of interest in ML, the main focus has shifted to non-asymptotic analysis and obtaining precise convergence rates, while preserving the underlying "dynamic" theoretical paradigm of MCMC. Based on these advances, new goals now seem achievable, including the acceleration of solvers, the design of gradient-free methods, and overcoming of the curse-of-dimensionality.

Among many possible mathematical techniques used in the study of sampling, PDEs stand out. This is due to the fact that passing to the continuum PDE formulation of an algorithm eliminates the complications of time-stepping constraints and discretization errors, simplifying the analysis and shifting the focus to *intrinsic dynamics*. For example, in methods using out-of-equilibrium dynamics and annealing schemes, the continuum perspective reveals an underlying system of reversible diffusion processes (the so-called score-based diffusion) [18], whose convergence properties are naturally addressed by studying their associated continuity equation. At the same time, sampling methods based on interacting particle systems have also benefited from the PDE perspective. Several recent works identified the continuum limit of Stein Variational Gradient Descent (SVGD) as a nonlinear, nonlocal transport equation and developed sufficient criteria for long time convergence to the global equilibrium. Interestingly, the PDE formally appears to have a gradient flow structure with respect to a *modified* Wasserstein distance; this has spurred significant interest in the PDE community, since the analysis of Wasserstein distances with *convex* mobilities, as arise in SVGD, seems completely inaccessible to existing mathematical tools. It has also spurred on alternative explanations of the success of SVGD, including recent work which viewed SVGD as a kernelized χ^2 gradient flow, relating the dynamics to the well-developed theory of porous medium equations.

A second reason for interest in PDE methods in the sampling context comes from the stunning success of the PDE approach in designing *probabilistic and generative* models, which are rooted in sampling. In particular, this is demonstrated by the DALL-E AI system, which uses transport and diffusion equations to create realistic works of art purely based on text input. This raises the important question of what other types of dynamics could be significant in designing generative, sampling-based models, as well as how existing dynamics can be implemented in a more efficient manner.

In spite of these recent advances, many open questions remain, which we believe are within reach of a PDE-based approach. First, how can our understanding of the *dynamics* of sampling algorithms lead to the development of new algorithms for sampling measures that fail traditional log-concavity assumptions? The Schrödinger bridge problem and analysis of the continuum limit of Föllner processes seem particularly promising here. Another important direction is to understand how the PDE perspective can shed light on the KLS conjecture—that the Cheeger constant of a non-log concave measure is independent of dimension. There is a long history of interplay between PDE and functional inequalities, and many of our participants on the PDE-side are international experts in the area. The PDE technique of *stochastic localization* also seems particularly promising here [9].

1.3 Impact on Partial Differential Equations

Finally, applications in NN and sampling also raise broader questions for PDE theory itself. When does a PDE have a gradient flow structure, and how can we find it? Given a functional inequality, what are the right PDE dynamics to perturb and probe its properties? Lastly, how can we connect the PDE perspective back to the discrete algorithms used in practice—via numerical analysis techniques or a redevelopment of PDE tools in the discrete setting?

Our workshop further developed these connections between ML and PDE. Due to the timeliness of the above problems, we believe that connections established at the workshop will lead directly to significant scientific breakthroughs.

2 Presentation Highlights

We highlight some of the talks given at the workshop, which connect to the above themes.

- Sinho Chewi: *Variational inference via Wasserstein gradient flows*[14, 19]

  By interpreting the variational inference problem as optimization over the Wasserstein space of probability measures, equipped with the optimal transport metric, one is led to new, principled algorithms, which exhibit strong practical performance and are backed by rigorous theory. Two particular cases of interest include Gaussian variational inference, in which one seeks to find the Gaussian which best approximates a given probability measures, and mean-field variational inference, in which one seeks to find best approximation by product measures.

- Nicolas Garcia Trillos: *Minimax rates for manifold learning*

  A novel formulation of the problem of learning a measure on the manifold was presented, along with optimal minimax rates.

- Jianfeng Lu: *Convergence analysis of classical and quantum dynamics via hypocoercivity* [10]

  This talk reviewed recent developments in the framework of hypocoervicity to obtain quantitative convergence estimate of classical and quantum dynamics, with focus on underdamped Langevin dynamics for sampling and Lindblad dynamics for open quantum systems.

- Jonathan Niles-Weed: *Central limit theorems for smooth optimal transport maps* [16]

  This talk considered plugin estimators of Brenier maps, which are defined as the Brenier map between density estimators of the underlying distributions, and showed that such estimators satisfy a pointwise central limit theorem when the underlying laws are supported on the flat torus of dimension. A key tool in the analysis was a new, quantitative linearization of the Monge-Ampére equation.

- Anna Korba: *Mirror Descent and Conditioning in Wasserstein space* [5]

  This talk presented new results on the convergence of the mirror descent algorithm, when optimizing a convex functional over the space of measures. Under appropriate smoothness and convexity assumptions (which are satisfied, for example, by the Kullback-Liebler divergence), it was shown that Sinhkorn's primal iterations for entropic optimal transport correspond to a mirror descent.

- Quentin Mrigot: *Uniform Quantization with (Sliced) Wasserstein Distances*

  This talk covered many recent results on optimal quantization of measures, with respect to both the classical Wasserstein metric and sliced Wasserstein metric. Despite the non-convexity of the corresponding optimization problem, it was shown that gradient descent strategies for the classical Wasserstein problem generally lead to low energy configurations. Similar properties were observed for the sliced Wasserstein problem, though rigorous theorems are more difficult to obtain in this setting.

- Boris Hanin: *Onset of Feature Learning in one hidden layer mean-field neural networks*

  This talk considered the process of feature learning in NN with one hidden layer, in the mean field regime. Key observations included that if all the directions of parameters are uncorrelated, the nonlinear model doesn't learn much more than the linear model.

- Gabriele Steidl: *Wasserstein gradient flows of maximum mean discrepancies* [11, 2]

  This talk presented recent work on inverse problems in imaging, in which one seeks to sample from the posterior given noisy measurement. The key strategy used was to do this via Wasserstein gradient flows maximum mean discrepancies defined with respect to appropriate kernels. In order to make this computationally feasible in high dimensions, approaches based on the Radon transform (slicing) and subsequent sorting were used. Applications to image generation and in inverse image restoration problems like computerized tomography were presented.

- Davide Carbone: *Jarzynski's idendity with energy based models*

  This talk presented connections between physical and generative diffusion models, with special attention given to energy-based models. New techniques, based on nonequilibrium statistical physics, were shown to improve the training of such models.

- Joan Bruna: *Learning Gaussian multi-index models with gradient flow* [4]

  This talk presented recent work studying gradient flow on the multi-index regression problem for high-dimensional Gaussian data, as a template for feature learning in neural networks. For the two-timescale algorithm, in which the low-dimensional link function is learnt with a non-parametric model infinitely faster than the subspace parametrizing the low-rank projection, it was proved that the resulting Grassmannian population gradient flow globally converges. A quantitative description of its associated saddle-to-saddle dynamics was also given.

- Franca Hoffmann: *Minmax games and Wasserstein gradient flows*

  This talk proposed a PDE framework for modeling the distribution shift of a strategic population interacting with a learning algorithm in two settings: one, where the objective of the algorithm and population are aligned, and two, where the algorithm and population have opposite goals. An analysis of convergence was presented in both settings, and the framework was applied to synthetic examples to illustrate its improved ability to capture distribution changes, when compared to simpler methods.

- Oliver Tse: Variational Acceleration Methods in the Space of Probability Measures [6]

  This talk reviewed various variational acceleration methods in Euclidean space and described one way of lifting these methods to the space of probability measures. Under suitable assumptions on the functional to be minimized, convergence rates were also given.

- Marco Mondelli: *Implicit Bias of Gradient Descent and Score-Based Generative Models*

  This talk presented two examples in which PDE methods have been useful in the the analysis of deep learning models. First, it was shown how PDE methods allowed the bias of two-layer

ReLU networks to be analyzed, and in the case of a univariate regression problem, ultimately implements a piecewise linear map of the inputs, where the number of points at which the tangent of the ReLU network estimator changes is at most three. Next, score-based generative models were analyzed, and convergence rates were given for a version of the popular predictor-corrector scheme.

- Jose A. Carrillo: *The Ensemble Kalman Filter in the Near-Gaussian Setting*

  This talk presented an analysis of the the accuracy of ensemble Kalman filter for problems where the filtering distribution is non-Gaussian, but can be characterized as close to Gaussian after appropriate lifting to the joint space of state and data, using the mean-field description.

- Stephan Wojtowytsch: *Acceleration under the PL inequality*

  This talk discussed how accelerated methods can be applied to non-convex optimization problems, with special attention given to the Modica-Mortola (or Cahn/Hilliard or Ginzburg-Landau) approximation to the perimeter functional.

- Theodor Misiakiewicz: *On the complexity of learning under group invariance*

  This talk connected learning with gradient descent to the computational complexity of general classes of learning algorithms. The talk introduced differentiable learning queries (DLQ) as a subclass of statistical query algorithms, and considered learning the orbit of a distribution under a group of symmetry, providing sharp upper and lower bounds for the computational complexity. Connections to training dynamics of neural networks in the mean-field regime were also presented.

- Anna Little: *Dimension reduction with path metrics: theory, applications, and challenges*

  This talk introduced a geometric framework for both unsupervised and supervised dimension reduction via the power-weighted path metric. Theoretical guarantees were presented, including convergence of graph Laplacian operators constructed with this random metric and applications include the analysis of single cell RNA data and multi-manifold clustering. Integration into Laplacian learning algorithms enabled accurate prediction of complex, nonlinear functions from sparse label information.

- Maria Bruna: *Macroscopic limits of systems of strongly interacting particles: from simple exclusion processes to interacting Brownian motions*

  This talk addressed the macroscopic dynamics of strongly interacting particles, which diverge significantly from weak (mean-field) or moderately interacting particles, especially in scenarios involving multiple species. Models based on strongly interacting particles arise in contexts where considerations like particle size (steric or excluded-volume effects) are paramount. While in simple cases, such as identical particles, the limit equations for the strongly interacting regime coincide with localising the mean-field limit, examples were also given showing how this correspondence breaks down for non-gradient type models, including multispecies systems.

- Matt Jacobs: *Lagrangian solutions to Porous medium equation and score matching*

This talk presented new results on constructing Lagrangian solutions to the porous medium equation and described how this analysis can be used to obtain convergence rates for certain deterministic versions of the score matching algorithm.

- Zhenjie Ren: *Approximations to the long time limit of Mean-field Langevin diffusion*

  This talk described how the training of two-layer neural network can be viewed as a convex mean-field optimization problem with entropy regularization, of which the minimizer distribution can be characterized as the invariant measure of the mean-field Langevin diffusion. Next, it explored recent progresses on different methods to approximate this target measure, including the uniform-in-time propagation of chaos results for the particle system and the ergodicity of the self-interaction diffusion.

- Claudia Totzeck: *Consensus-based optimization: multi-objective problems & gradient inference* [12]

  This talk presented how how ensemble-based gradient inference is able to improve the performance of particle methods for global optimization tasks, focusing especially on Consensus-based optimization (CBO) and sampling. The talk also presented an extension of CBO to approximate the Pareto front of multi-objective problems.

- Yulong Lu: *On the generalization of diffusion models in high dimensions*

  This talk considered the generalization theory of score-based generative models (SGMs) for learning high-dimensional distributions, showing that that SGMs achieve a dimension-free generation error bound when applied to a class of sub-Gaussian distributions characterized by certain low-complexity structures.

- Lukasz Szpruch: *FisherRao gradient flow for entropy regularised MDPs in Polish spaces* [15]

  This talk presented recent work studying the global convergence of a Fisher-Rao policy gradient flow for infinite-horizon entropy-regularised Markov decision processes with Polish state and action space. Global well-posedness of the gradient flow was established, and exponential convergence to the optimal policy was also shown. Furthermore, it was proved that the flow is stable with respect to gradient evaluation, offering insights into the performance of a natural policy gradient flow with log-linear policy parameterization.

- Daniel Sanz Alonso: *Structured Covariance Operator Estimation* [1]

  This talk introduced a notion of sparsity for infinite-dimensional covariance operators and a family of thresholded estimators which exploits it. In a small lengthscale regime, it was shown that thresholded estimators achieve an exponential improvement in sample complexity over the standard sample covariance estimator.

- Cristina Cipriani [8]: *An optimal control perspective on Neural ODEs and adversarial training*

  This talk showed how the training of Neural ODEs can be viewed as an optimal control problem, allowing for numerical approaches inspired by this control-oriented viewpoint. First-order optimality conditions in the form of a mean-field Pontryagin Maximum Principle were

derived and applied to different numerical examples. The case of adversarial training was formalized with perturbed data as a minimax optimal control problem, and first-order optimality conditions were again derived.

- Lisa Kreusser: *Fokker-Planck equations for score-based diffusion models*

  This talk analyzed the difference between the ODE and SDE dynamics of score-based diffusion models, which were linked to associated FokkerPlanck equations. The talk also provided a theoretical upper bound on the Wasserstein 2-distance between the ODE- and SDE-induced distributions in terms of a FokkerPlanck residual. Finally, it was numerically shown that reducing the FokkerPlanck residual by adding it as an additional regularisation term leads to closing the gap between ODE- and SDE-induced distributions. Numerical experiments suggested that this regularisation can improve the distribution generated by the ODE, at the cost of degraded SDE sample quality.

- Jaume de Dios Pont: *Complexity lower bounds for log-concave sampling* [7]

  This talk considered the problem of sampling a density, when the log-density is a strongly concave smooth function, and presented universal complexity bounds that no algorithm can beat.

- Raphal Barboni: *Understanding the training of infinitely deep and wide ResNets with Conditional Optimal Transport* [3]

  This talk presented recent work studying the convergence of gradient flow dynamics for training deep neural networks, in the mean-field regime. An approach to train the model with gradient flow with respect to the conditional Optimal Transport distance is considered, in which the conditional OT distance is a restriction of the classical Wasserstein distance which enforces the marginal condition. Well-posedness of the gradient flow equation and consistency with the training of ResNets at finite width was shown. Likewise, it was shown that gradient flows with well-chosen initializations converge towards a global minimizer.

- Nikola Kovachki: *Function Space Diffusion for Video Modeling*

  This talk presented a generalization of score-based diffusion models to function spaces by perturbing functional data via a Gaussian process at multiple scales. An appropriate notion of score was obtained by defining densities with respect to Guassian measures and generalizing denoising score matching. Likewise, the generative process could be then defined by integrating a function-valued Langevin dynamic. The talk showed that the corresponding discretized algorithm generates samples at a fixed cost that is independent of the data discretization. As an application for such a model, the talk formulated video generation as a sequence of joint inpainting and interpolation problems defined by frame deformations. Results were presented in which an image diffusion model was trained using Gaussian process inputs and used to solve the video generation problem by enforcing equivariance with respect to frame deformations. The results were state-of-the-art for video generation using models trained only on image data.

3 Open problem sessions

On days 3 and 4 of the workshop, we organized open problem sessions. In the first session, all participants discussed together in the main lecture hall and decided upon a list of five open problems of particular interest:

1. Can we develop better evaluation tools for generative models (beyond, say, FID). Given samples of two distributions in high dimension, what are the state-of-the-art theoretical tools to tell whether they are the same of not? (2-sample test).

2. What are the latest questions around mean-field limits of neural networks, for the two-layer case and beyond two-layer? Can we develop a sharp analysis of the particle discretization errors and propagation of chaos for these mean-field dynamics?

3. The space of probability measures can be endowed with a variety of structures, but a lot of existing efforts have been focused on the Wasserstein structure. Can we develop functional inequalities in other ("mirror") geometries, beyond log-Sobolev? Also what are the cases of benign non-convexity for measures?

4. Two topics that were discussed in the workshop – acceleration of Wasserstein gradient flows and minmax games in Wasserstein space – involve some sort of hypocoercivity (a term characterizing dynamics with a degenerate form of coercivity but that converge nontheless thanks to the "mixing effect" of a conservative component). What do these problems have in common, and can we transfer techniques from one case to the other?

5. What are the lower and upper bounds for optimization in the space of probability measures? For instance, is there a separation between mean-field accelerated and non-accelerated PDEs?

Sinho Chewi also presented, in detail, the following open problem for *sharp propagation of chaos for interacting particle systems*. Consider the optimization problem over the space of probability measures on $\mathbb{R}^d$

$$\min_{\mu} \int V d\mu + \frac{1}{2} \int W * \mu d\mu + \int \mu \log \mu$$

The stationary points of the Wasserstein gradient flow of this functional are of the form

$$\pi(x) \propto \exp\Big(- V(x) - \int W(x-y)\pi(dy)\Big)$$

and the stationnary points of the N-particle stochastic approximation of the dynamics are of the form

$$\hat{\pi}(x_1, \dots, x_N) \propto \exp\Big(- \sum_{i=1}^{N} V(x_i) - \frac{1}{2(N-1)} \sum_{i,j=1}^{N} W(x_i - x_j)\Big)$$

How far are these two distributions? For a while, it could be found in the literature the guarantee that $KL(\hat{\pi}|\pi^{\otimes N}) = O(1)$ which implies $KL(\hat{\pi}^1|\pi) = O(1/N)$. However, recent result by Daniel Lacker [13] have shown the surprising improvement that $KL(\hat{\pi}^1|\pi) = O(1/N^2)$, under the assumption that π satisfies a log-Sobolev inequality and that $\|\nabla^2 W\|$ is small enough. An open question is then the following: can we extend such improved guarantee under different assumptions, such as the geodesically convex case $\nabla^2 V \geq \alpha \geq 0$ and $\nabla^2 W \geq 0$?

In the second open problem session, participants broke into four small groups to discuss the above questions.

4 Outcome of the Meeting

The meeting succeeded in bringing together experts in ML and PDE to present recent progress and discuss open questions at the interface of the two fields. The workshop resulted in many new scientific connections, and we are confident that many interesting results and collaborations will emerge.

References

[1] Al-Ghattas, Omar, Jiaheng Chen, Daniel Sanz-Alonso, and Nathan Waniorek. "Optimal Estimation of Structured Covariance Operators." arXiv preprint arXiv:2408.02109 (2024).

[2] Altekrger, Fabian, Johannes Hertrich, and Gabriele Steidl. "Neural Wasserstein Gradient Flows for Discrepancies with Riesz Kernels." In *International Conference on Machine Learning*, pp. 664-690. PMLR, 2023.

[3] Barboni, Raphal, Gabriel Peyr, and Franois-Xavier Vialard. "Understanding the training of infinitely deep and wide ResNets with Conditional Optimal Transport." arXiv preprint arXiv:2403.12887 (2024).

[4] Bietti, Alberto, Joan Bruna, and Loucas Pillaud-Vivien. "On learning gaussian multi-index models with gradient flow." arXiv preprint arXiv:2310.19793 (2023).

[5] Bonet, Clment, Tho Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. "Mirror and Preconditioned Gradient Descent in Wasserstein Space." arXiv preprint arXiv:2406.08938 (2024).

[6] Chen, Shi, Qin Li, Oliver Tse, and Stephen J. Wright. "Accelerating optimization over the space of probability measures." arXiv preprint arXiv:2310.04006 (2023).

[7] Chewi, Sinho, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. "Query lower bounds for log-concave sampling." *Journal of the ACM* (2023).

[8] Cipriani, Cristina, Alessandro Scagliotti, and Tobias Whrer. "A minimax optimal control approach for robust neural ODEs." In *2024 European Control Conference (ECC)*, pp. 58-64. IEEE, 2024.

[9] R. Eldan, Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, (2013), 23(2), 532-569.

[10] Fang, Di, Jianfeng Lu, and Yu Tong. "Mixing time of open quantum systems via hypocoercivity." arXiv preprint arXiv:2404.11503 (2024).

[11] Hertrich, Johannes, Manuel Grf, Robert Beinert, and Gabriele Steidl. "Wasserstein steepest descent flows of discrepancies with Riesz kernels." *Journal of Mathematical Analysis and Applications* 531, no. 1 (2024): 127829.

[12] Klamroth, Kathrin, Michael Stiglmayr, and Claudia Totzeck. "Consensus-based optimization for multi-objective problems: a multi-swarm approach." *Journal of Global Optimization* (2024): 1-32.

[13] D. Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and Mathematical Physics*, 4.2 (2023): 377-432.

[14] Lambert, Marc, Sinho Chewi, Francis Bach, Silvre Bonnabel, and Philippe Rigollet. "Variational inference via Wasserstein gradient flows." *Advances in Neural Information Processing Systems* 35 (2022): 14434-14447.

[15] Lascu, Razvan-Andrei, Mateusz B. Majka, and ukasz Szpruch. "A Fisher-Rao gradient flow for entropic mean-field min-max games." arXiv preprint arXiv:2405.15834 (2024).

[16] Manole, Tudor, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. "Central Limit Theorems for Smooth Optimal Transport Maps." In *2023 IMS International Conference on Statistics and Data Science (ICSDS)*, p. 323. 2023.

[17] V. Papyan, X. Y. Han, D. L. Donoho, Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, (2020), 117(40), 24652-24663.

[18] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole, (2020) Score-Based Generative Modeling through Stochastic Differential Equations. *In International Conference on Learning Representations*, (2020).

[19] Jiang, Yiheng, Sinho Chewi, and Aram-Alexandre Pooladian. "Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space." In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2720-2721. PMLR, 2024.