Mathematisches Forschungsinstitut Oberwolfach

Report No. 13/2026

DOI: 10.4171/OWR/2026/13

Flows on Measure Spaces and Applications in Machine Learning

Organized by
Philippe Rigollet, Cambridge MA
Giuseppe Savaré, Milano
Gabriele Steidl, Berlin
François-Xavier Vialard, Champs-sur-Marne

22 March – 27 March 2026

Abstract. Flows on measure spaces have long been examined in stochastic analysis and have recently attracted significant interest in machine learning, leading to intriguing research questions that often fall outside the scope of existing theory. Normalizing flows, score-based diffusion, and flow matching models are among the most powerful generative neural methods and rely on the geometry of measure spaces. In particular, the Wasserstein metric and optimal transport techniques have advanced the field in recent years. However, involving different Riemannian-like metrics on measure spaces, e.g., by the framework of right-invariant metrics on the group of diffeomorphisms and their action on objects, e.g., densities, and designing transport inference functionals with advanced properties like equivariance led to new neural models. Generative models can be conditioned on (degraded) data, which leads to new developments in the solution of Bayesian inverse problems. Viewing transformers as interacting particle systems introduced a new mathematical perspective on these complex systems and shed light on their clustering behavior. Finally, learning neural models comes with new challenges in (stochastic) optimization, such as accelerated optimization, operator splitting, and mirror descent on measure spaces, ensemble filtering methods, the treatment of high dimensions via slicing or Fourier random features, as well as scalability questions and related lifting to infinite-dimensional spaces. The workshop will bring together scientists interested in different aspects of flows on measure spaces to further understand and develop their analysis, in particular to address questions in deep generative learning and to develop improved optimization methods for measure spaces.

Mathematics Subject Classification (2020): 49Q22, 35Q84, 37L05, 65C35, 68T05, 68T07.

Introduction by the Organizers

The aim of the workshop was to further understand and develop the analysis of flows on measure spaces, in particular facing questions in deep generative learning and to foster improved optimization methods on measure spaces.

Flow-based methods, such as normalizing flows, neural ODEs, diffusion models, and flow matching techniques, belong to the most successful techniques that aim to approximate an unknown distribution or to sample from an often high-dimensional distribution by pushing forward an easy-to-sample distribution, such as the Gaussian, by a learned (sequence of) neural network. The basic ideas stem from well-established results on flows in probability measure spaces, in particular, optimal transport in Wasserstein spaces. However, generative learning came with new questions on flow properties when using different geometries like Stein and Fisher-Rao metrics and combinations thereof of Hellinger–Kantorovich distances, including their linearized version. Important metrics on the space of densities are provided by the framework of right-invariant metrics on the group of diffeomorphisms and their action on objects such as densities. More recent instances of this framework include normalizing flows in machine learning and induced metrics on diffeomorphisms that arise naturally in applications such as Stein variational gradient descent. Therefore, right-invariant metrics on the group of diffeomorphisms and their connection to fluid flows provide a powerful framework for understanding various applications. A generative model can be conditioned on (degraded data) in order to solve Bayesian inverse problems, which raises robustness questions and also leads to the notion of conditional Wasserstein distances. Further, the Benamou-Brenier formulation of optimal transport was generalized in the study of mean field games and allowed the design of new transport inference functionals in probability spaces. Gromov-Wasserstein distances in network quantization, as well as learning of isometries, appear both practically and theoretically challenging. Finally, modern transformers can be seen as interacting particle systems, where a basic mathematical theory is available. As in neural ODEs, self-attention is viewed as a velocity field that evolves particles (tokens) toward a useful embedding, thereby raising entirely new research questions.

At the same time, efficient tools in optimization and numerical analysis have been addressed and are still an area of active research, including:

- accelerating optimization over probability measures,
- translation of operator splitting, mirror descent techniques from Hilbert to measure spaces,
- ensemble filtering methods,
- uncertainty quantification,
- sliced methods versus Fourier random features to tackle high-dimensional problems,
- scalability, modeling across different resolution levels to improve the efficiency of the training process, e.g., by lifting high-dimensional problems to an infinite-dimensional space and exploiting the "learn-then-discretise" paradigm.

The challenge and beauty of the field is that different kinds of mathematics can and should be applied from stochastic analysis and statistics over geometry and optimization toward neural modeling with quite different applications. We are far from capturing all the interesting activities in the field in this workshop, but we will focus on the above synergies.

Acknowledgements: The workshop organizers would like to thank the DFG for travel support within the SPP2298 "Theoretical Foundations of Deep Learning" and the Berlin MATH+ Cluster of Excellence. Moreover, the MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-2230648, "US Junior Oberwolfach Fellows".

Workshop: Flows on Measure Spaces and Applications in Machine Learning

Table of Contents

Abstracts

Introduce thermal noise before discretizing, not afterwards!

Felix Otto
(joint work with B. Gess, R. S. Gvalani, F. Kunick, and M. Sauerbrey)

1. The thin-film equation with thermal noise. The thin-film equation models the evolution of the height h of a liquid thin film over a planar d-dimensional[1] substrate, driven by surface tension γ and limited by (dynamic) viscosity η.

Starting from conservation of mass in form of a continuity equation, it expresses a quasi-stationary balance between viscous and capillary forces

$$\partial_t h + \nabla \cdot q = 0 \quad \text{where} \quad \frac{3\eta}{h^3} q = -\nabla p \quad \text{and} \quad p = -\gamma \Delta h;$$

the power 3 in the mobility $m(h) = \frac{h^3}{3\eta}$ that relates the flux q and the (potential) force is the only one that makes the model dimensionally correct. Following [1], by the fluctuation-dissipation principle, thermal effects at temperature T take the form (where k_B is the Boltzmann factor)

$$\begin{aligned} &q = -m\nabla p + f \quad \text{where} \\ &\mathbb{E} f_i(t,x) f_j(s,y) = 2k_B T\, m\, \delta_{ij}\, \delta(t-s)\, \delta(x-y); \end{aligned}$$

the fluctuation-dissipation principle passes through the lubrication approximation, see [5, Section 2.3]. Given an average film height H, the non-dimensional temperature is $\frac{1}{\beta} = \frac{\gamma H^2}{k_B T}$; hence for sufficiently thin films, thermal noise matters. We now non-dimensionalize to the effect of $\gamma = 3\eta = 1$. This results in the SPDE

$$\partial_t h - \nabla \cdot m \nabla p = \sqrt{\frac{2}{\beta}} \nabla \cdot \sqrt{m} \xi \quad \text{with} \quad p = -\nabla \cdot \nabla h, \tag{1}$$

where ξ is vectorial space-time white noise.

2. Naive discretization fails. The SPDE (1) is invariant in law under the change of variables $x = \lambda \hat{x}$, $t = \lambda^4 \hat{t}$, $\xi = \lambda^{-\frac{d+4}{2}} \hat{\xi}$, $h - H = \lambda^{1-\frac{d}{2}} \hat{h}$. This suggests that solutions are Hölder continuous in space with exponent $\alpha = 1 - \frac{d}{2}$ (and in time with exponent $\frac{\alpha}{4}$). This is too rough for the flux $-m(h)\nabla p$ to be given a sense so that the SPDE is singular (but subcritical for $d < 2$). It is therefore not surprising that one has to be careful when discretizing in space, cf. [8].

Considering a finite-volume discretization, the flux q lives on the (oriented) edges (of length $\frac{1}{N}$) of a graph, while the conserved quantity h lives at its vertices (which one should think of cells), and likewise the potential p. The continuity equation then involves a discrete divergence $\nabla_N \cdot q$; its transpose (up to the sign) yields a discrete gradient $\nabla_N p$. Hence, such a finite-volume discretization requires

[1] $d = 2$ is the physical dimension

a numerical mobility M that lives on (the un-oriented) edges, and thus is a (symmetric) function $M(h_0, h_1) > 0$ of the values of h at the two incident vertices, satisfying the consistency $M(h, h) = m(h)$. A naive Ansatz reads

$$\mathrm{d}h - \nabla_N \cdot M \nabla_N p \mathrm{d}t = \sqrt{\frac{2N}{\beta}}\, \nabla_N \cdot \sqrt{M} \mathrm{d}W \text{ with } p = -\nabla_N \cdot \nabla_N h, \tag{2}$$

where W stands for a Brownian motion, independent from edge to edge.

We find in [3] that one needs an additional term on the l. h. s. of (2), namely

$$\frac{N}{\beta}\, \nabla_N \cdot M'' \nabla_N h \, \mathrm{d}t, \tag{3}$$

where the new symmetric function M'' is induced by M through

$$\Big(\frac{\partial}{\partial h_1} - \frac{\partial}{\partial h_0}\Big) M(h_0, h_1) = (h_1 - h_0) M''(h_0, h_1). \tag{4}$$

Note that (3) acts as a – divergent – second-order term. We learn from (4) that $M'' \equiv 0$ iff $M = M_{\mathrm{flat}}$ where $M_{\mathrm{flat}}(h_0, h_1) \coloneqq m(\frac{h_0+h_1}{2})$, and it turns out that the correction term is there to emulate M_{flat}: It ensures on the level of the flux

$$-M \nabla_N p + \frac{N}{\beta} M'' \nabla_N h \approx -M_{\mathrm{flat}} \nabla_N p.$$

This expansion is (obviously) not classical and relies on the structure of the near-Gaussian equilibrium measure in form of $\mathbb{E} \nabla_N p \nabla_N h \approx \frac{2N^3}{\beta}$.

Discretizations with "diagonal" mobility have been proposed in [5, Section 2.3] and shown to not need a correction term in case of central differences in order to preserve (8) and to satisfy detailed balance in equilibrium [2, Section 4.1].

3. Gradient flow structure of thin-film equation. The (harmonic approximation of the) surface energy

$$E(h) \coloneqq \int \frac{1}{2} |\nabla h|^2 \tag{5}$$

defines a functional on configuration space. The minimal viscous dissipation required to generate an infinitesimal variation $\dot{h}$ of the film height h

$$g_h(\dot{h}, \dot{h}) \coloneqq \inf_q \Big\{ \int \frac{1}{m(h)} |q|^2 \;\Big|\; \dot{h} + \nabla \cdot q = 0 \Big\} \tag{6}$$

defines an inner product on the tangent space $\{\dot{h}\}$. By duality, this induces an inner product g^h on the co-tangent space, taking the form

$$g^h(p, p) = \int m(h) |\nabla p|^2,$$

where we identify co-tangent vectors with potentials p via $\dot{h} \mapsto \int p \dot{h}$.

It is easy to see that the deterministic thin-film evolution satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} F(h) = -g^h(\mathrm{diff} F_{|h}, \mathrm{diff} E_{|h}) \quad \text{for observables } F, \tag{7}$$

where the differential is defined via $\mathrm{diff} F_{|h}.\dot{h} = \frac{d}{ds}_{|s=0} F(h + s\dot{h})$. Statement (7) expresses that h evolves according to the gradient flow w. r. t. (E, g).

4. Natural incorporation of thermal noise. The gradient flow structure (E, g) allows for a canonical introduction of thermal noise: We postulate the equilibrium measure to be informally given by

$$\mu(\mathrm{d}h) = \frac{1}{Z} I(h > 0) \exp(-\beta E(h)) \mathrm{d}h, \tag{8}$$

which we interpret as the Gaussian free field conditioned on being non-negative, which is a regular conditioning provided $d < 2$.

In [4], we establish for $d < 2$ the existence of the corresponding overdamped Langevin dynamics: There exists a reversible stationary Markov process such that the transition probability μ_t is determined through

$$\frac{\mathrm{d}}{\mathrm{d}t} \int F \mathrm{d}\mu_t = -\int g^h \Big(\mathrm{diff} F, \mathrm{diff} \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu} \Big) \mathrm{d}\mu \quad \text{for observables } F. \tag{9}$$

This result relies on the theory of Dirichlet forms, and their closability.

5. Natural Galerkin discretization. A gradient flow structure (E, g) allows for natural discretizations, namely by a Galerkin Ansatz, i. e. by restricting to a submanifold of configuration space. In view of E being given by (5), the simplest energy-conformal Ansatz is to restrict to height functions h that are piecewise linear and continuous. We combine this with "mass lumping" in (6), for $d = 1$:

$$g_{h,N}(\dot{h}, \dot{h}) := \inf_q \Big\{ \int \frac{1}{m(h)} |q|^2 \;\Big|\; \frac{1}{N} \sum_{i=0}^{N} \dot{h}(i/N) \delta_{i/N} + \partial_x q = 0 \Big\},$$

which ensures that we obtain a numerical mobility M, given by

$$\frac{1}{M(h_0, h_1)} = \int_0^1 \frac{1}{m(s h_1 + (1-s) h_0)} \mathrm{d}s. \tag{10}$$

This coincides with the numerical mobility introduced in [6], designed such that the "entropy" $\frac{1}{N} \sum_{i=0}^{N} s(h(\frac{i}{N}))$ with $s'' = 1/m$ is a Lyapunov functional in the deterministic case, and thus promotes preservation of positivity. We note that for (10), we have for the coefficient in (3)

$$M''(h, h) = \frac{1}{6} \Big(m^2 \Big(-\frac{1}{m} \Big)'' \Big)(h),$$

so that the term (3) typically acts as a parabolic term, in particular, for $m(h) = h^3$.

6. Geometry of overdamped Langevin equations. In order to derive an SDE, one has to introduce coordinates $h^i = \Phi^i(h)$, $i = 0, \cdots, N$, for the manifold. In such coordinates, the metric tensor on co-tangent space is given by

$$g^{ij}(\Phi(h)) = g^h(\mathrm{diff} \Phi^i_{|h}, \mathrm{diff} \Phi^j_{|h}),$$

and the (free) energy functional E in coordinates is defined via push forward under the coordinate map:

$$\Phi\# \exp(-\beta E(h))\mathrm{d}h_0 \cdots \mathrm{d}h_N = \mathrm{d}\mu.$$

Writing $\Phi\#\rho(t,h)\mathrm{d}h_0 \cdots \mathrm{d}h_N = \mathrm{d}\mu_t$, we learn that (9) is the variational formulation of a Fokker–Planck equation that gives rise to a diffusion process. The latter may be written in Itô form[2]

$$\mathrm{d}h^i + (g^{ij}\partial_j E - \frac{1}{\beta}\partial_j g^{ij})\mathrm{d}t = \sqrt{\frac{2}{\beta}}\sigma^i_\alpha \mathrm{d}W^\alpha, \tag{11}$$

for independent Brownian motions $\{W^\alpha\}_\alpha$, provided the noise coefficients $\{\sigma^i_\alpha\}_{i,\alpha}$ satisfy the usual compatibility

$$g^{ij} = \sum_\alpha \sigma^i_\alpha \sigma^j_\alpha. \tag{12}$$

The term $-\frac{1}{\beta}\partial_j g^{ij}$ does not coincide with the Stratonovich-to-Itô correction $\frac{1}{\beta}\sum_\alpha \sigma^j_\alpha \partial_j \sigma^i_\alpha$. In fact, applying Leibniz' rule to (12) shows that (11) is equivalent to

$$\mathrm{d}h^i = \big(- g^{ij}\partial_j E + \frac{1}{\beta}\sum_\alpha \sigma^i_\alpha \partial_j \sigma^j_\alpha\big)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\sigma^i_\alpha \circ \mathrm{d}W^\alpha. \tag{13}$$

The remaining correction $\frac{1}{\beta}\sum_\alpha \sigma^i_\alpha \partial_j \sigma^j_\alpha$ is required for covariance, i. e. for (13) to define a process on the abstract manifold independent from the choice of coordinates. The treatment of (1) based on regularity structures from [7] does not yet reveal such a counter term.

References

[1] B. Davidovitch, E. Moro, H.A. Stone, *Spreading of viscous fluid drops on a solid substrate assisted by thermal fluctuations*, Phys. Rev. Lett. **95** (2005), p. 244505.

[2] M.A. Durán-Olivencia, R.S. Gvalani, S. Kalliadasis, G.A. Pavliotis, *Instability, rupture and fluctuations in thin liquid films: theory and computations*, J. Stat. Phys. **174** (2019), 579–604.

[3] B. Gess, R.S. Gvalani, F. Kunick, F. Otto, *Thermodynamically consistent and positivity-preserving discretization of the thin-film equation with thermal noise*, Math. Comp. **92** (2023), 1931–1976.

[4] B. Gess, F. Otto, M. Sauerbrey, in preparation.

[5] G. Grün, K.R. Mecke, M. Rauscher, *Thin-film flow influenced by thermal noise*, J. Stat. Phys. **122** (2006), 1261–1291.

[6] G. Grün, M. Rumpf, *Nonnegativity preserving convergent schemes for the thin film equation*, Numer. Math. **87** (2000), 113–152.

[7] R.S. Gvalani, M. Tempelmayr, *Stochastic estimates for the thin-film equation with thermal noise*, arXiv:2309.15829 (2023).

[8] M. Hairer, J. Maas, H. Weber, *Approximating rough stochastic PDEs*, Comm. Pure Appl. Math. **67** (2014), 776–870.

[2] We use Einstein's summation convention.

Kinetic optimal transport

Jan Maas

(joint work with Giovanni Brigati, Filippo Quattrocchi)

The 2-Wasserstein metric from optimal transport is *arguably* the natural metric on probability measures for the study of overdamped Langevin dynamics. Several arguments can be given to support this claim:

- the Fokker–Planck equation can be formulated as gradient flow of the relative entropy in the 2-Wasserstein space of probability measures [14, 13].
- the 2-Wasserstein geometry arises from a large deviation principle in the many-particle limit for independent Brownian particles [7, 1].
- absolutely continuous curves in the 2-Wasserstein space correspond to solutions of the continuity equation under natural moment assumptions [2].

In this talk, we discuss a natural geometry on probability measures for the description of *underdamped* Langevin dynamics. Let us motivate its definition by analogy with the Wasserstein geometry. For a curve $\gamma : [0,T] \to \mathbb{R}^d$ describing a particle trajectory, define its energy by $I(\gamma) \coloneqq \int_0^T |\gamma'(t)|^2 \,\mathrm{d}t$. This functional arises as the rate functional in Schilder's theorem, the large deviation principle for Brownian motion. The associated cost function c_T is defined by minimising I among curves with prescribed boundary values $x, y \in \mathbb{R}^d$:

$$c_T(x,y) \coloneqq \inf_\gamma \Big\{ I(\gamma) \;:\; \gamma(0) = x,\ \gamma(T) = y \Big\} = \frac{|x-y|^2}{2T}\,.$$

The 2-Wasserstein metric is then obtained by lifting the cost function to the space of probability measures.

To define the kinetic optimal transport problem, we consider the rate function $J(\gamma) \coloneqq \int_0^T |\gamma''(t)|^2 \,\mathrm{d}t$, which appears in the large deviation principle for a Brownian particle with inertia. Consider the phase-space $\mathbb{R}^{2d}$; a point $(x,v) \in \mathbb{R}^{2d}$ describes the position $x \in \mathbb{R}^d$ and the velocity $v \in \mathbb{R}^d$ of a particle. A natural kinetic cost function c_T^{kin} is then defined by

$$c_T^{\mathrm{kin}}\big((x,v),(y,w)\big) \coloneqq \inf_\gamma \Big\{ J(\gamma) \;:\; \big(\gamma(0),\gamma'(0)\big) = (x,v),\ \big(\gamma(T),\gamma'(T)\big) = (y,w) \Big\}.$$

A computation reveals that $c_T^{\mathrm{kin}}\big((x,v),(y,w)\big) = \frac{1}{2T}\Big(\frac{1}{12}\big|\frac{y-x}{T} - \frac{v+w}{2}\big|^2 + |v-w|^2\Big)$. This cost function appears in the heat kernel for the Kolmogorov diffusion $\partial_t u + v \cdot \nabla_x u = \Delta u$, just as the Euclidean cost function $\frac{1}{2T}|x-y|^2$ appears in the heat kernel for the diffusion equation $\partial_t u = \Delta u$.

It is then natural to consider the optimal transport problem with the cost function c_T^{kin}. Over the years, this transport problem has been considered by various authors in different contexts. Huang and Jordan constructed minimizing-movement schemes for kinetic problems [11, 12]. Later on, Duong, Peletier, and Zimmer developed the connection to large deviations [9], and they formulated the kinetic Fokker-Planck in terms of *GENERIC*, a framework to describe evolution equations with a dissipative part and a conservative part [8]. The kinetic optimal

transport problem has also been studied in the context of Wasserstein splines and related smooth interpolations of probability measures [3, 5, 6].

In this talk, we give an overview of these developments, and we present new results on the geometric structure on probability measures induced by the cost functions c_T^{kin}. While these cost functions do not induce a metric structure, they can still be used to define a concept of absolute continuity for curves in the space of probability measure [15]. One of our main results asserts that absolutely continuous curves in this geometry can be characterized as solutions to the Vlasov equation $\partial_t \mu + v \cdot \nabla_x \mu + \nabla_v \cdot (F\mu) = 0$ under natural moment assumptions on the acceleration field F. We also present a Benamou–Brenier formula, first stated in [5] and then proved in [10, 4], which expresses the equivalence of static and dynamical optimal transport in the kinetic setting. It is natural to expect that these results will have applications to the study of kinetic Fokker–Planck equations and related evolution equations.

References

[1] S. Adams, N. Dirr, M. A. Peletier, and J. Zimmer. From a large-deviations principle to the Wasserstein gradient flow: a new micro-macro passage. *Comm. Math. Phys.* **307** (2011), 791–815.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Birkhäuser Verlag, second edition (2008).

[3] J.-D. Benamou, T. O. Gallouët, and F.-X. Vialard. Second-order models for optimal transport and cubic splines on the Wasserstein space. *Found. Comput. Math.*, **19** (2019), 1113–1143.

[4] G. Brigati, J. Maas, and F. Quattrocchi. Kinetic optimal transport (OTIKIN) – Part 1: Second-order discrepancies between probability measures, *arXiv:2502.15665* (2025).

[5] Y. Chen, G. Conforti, and T. T. Georgiou. Measure-valued spline curves: An optimal transport viewpoint. *SIAM J. Math. Anal.*, **50** (2018), 5947–5968.

[6] S. Chewi, J. Clancy, T. Le Gouic, P. Rigollet, G. Stepaniants, and A. Stromme. Fast and smooth interpolation on Wasserstein space. *AISTATS, PMLR* (2021), 3061–3069.

[7] D.A. Dawson and J. Gärtner. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, **20**, 1987,247–308.

[8] M. H. Duong, M. A. Peletier, and J. Zimmer. GENERIC formalism of a Vlasov-Fokker-Planck equation and connection to large-deviation principles. *Nonlinearity* **26** (2013), 2951–2971.

[9] M. H. Duong, M. A. Peletier, and J. Zimmer. Conservative-dissipative approximation schemes for a generalized Kramers equation. *Math. Methods Appl. Sci.* **37** (2014), 2517–2540.

[10] K. Elamvazhuthi. Benamou–Brenier formulation of optimal transport for nonlinear control systems on $\mathbb{R}^d$. *arXiv:2407.16088* (2024).

[11] C. Huang. A variational principle for the Kramers equation with unbounded external forces. *J. Math. Anal. App* **250** (2000), 333–367.

[12] C. Huang and R. Jordan. Variational formulations for Vlasov–Poisson–Fokker–Planck systems. *Math. Methods Appl. Sci.* **23** (2000), 803–843.

[13] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.* **29** (1998), 1–17.

[14] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, **26** (2001), 101–174.

[15] R. Rossi and G. Savaré. Abstract action spaces and their topological and dynamic properties. *Discrete Contin. Dyn. Syst. Ser. S*, **17** (2024), 395–420.

Geometric calculations on probability manifolds from reciprocal relations in Master equations

Wuchen Li

Non-equilibrium thermodynamics [3, 5] studies dynamical systems that interact with reservoirs for thermal energy or volume. Examples are coupled physical irreversible processes in complex systems, which are modeled by probability functions over physical states. Typical dynamics include thermoelectric phenomena and heat conduction in an anisotropic medium. In the modeling of irreversible processes, Onsager introduced the reciprocal relationship [13], which describes the symmetric dissipation of the probability transition equations arising from physical processes, known as free energy dissipation or entropy production. Here, the free energy dissipation follows the second law of thermodynamics. Rather than addressing arbitrary physical mechanisms, one can focus on two canonical classes of stochastic models: overdamped Langevin dynamics in continuous state spaces, and detailed-balance Markov processes with transition rates on discrete domains. In both cases, entropy production is expressed as the time derivative of an entropy (or free energy) functional along the governing probability evolution equations, such as the Fokker–Planck equation in the continuous setting or the master equation in the discrete setting.

In recent decades, a particular type of gradient flow has been widely studied in optimal transport theory [1, 16]. They characterize a class of partial differential equations as gradient flows of energies on a probability space equipped with an infinite-dimensional Riemannian metric, namely the Wasserstein-2 space. Among these dynamics, the famous example is the probability density transition equation for overdamped Langevin dynamics, which is the gradient-drift Fokker-Planck equation. In this setting, the gradient drift Fokker-Planck equation satisfies the gradient flow of free energy in Wasserstein-2 space [7, 14]. From the angle of non-equilibrium thermodynamics [6], the gradient drift Fokker-Planck equation satisfies the Onsager reciprocal relation [13]. Along this direction, master equations with detailed balance conditions can also be formulated as gradient flows in discrete Wasserstein-2 spaces [4, 11, 12] for probability distributions on discrete states. It also satisfies Onsager reciprocal relations [13]. From now on, we mainly study these generalized Wasserstein-2 spaces on a simplex, called thermodynamical probability manifolds, in short *probability manifolds*.

Nowadays, geometric calculations in probability manifolds are essential for understanding fluctuation relations and thermodynamic properties [17]. E.g., [9, 15] introduces overdamped Langevin dynamics in the probability manifold, namely Wasserstein common noises. It is based on geometric quantities, such as second-order operators in finite-state probability manifolds. Along this direction, a natural question arises. *What are geometric quantities, including Riemannian and sectional curvatures, in probability manifolds? Can we use these computations to design algorithms in sampling distributions supported on simplex sets?*

In this talk, we study Riemannian calculations in a class of thermodynamical probability manifolds. This study is a generalization of Otto calculus on discrete

states with general mobility functions. We first compute the Levi-Civita connection, gradient, and Hessian operators on probability manifolds. We further derive Riemannian and sectional curvatures in the probability simplex set. Two examples of these computations are performed. One example is the Levi-Civita connection on a three-state probability simplex, where the discrete probability manifold is constructed from the master equation for a chemical monomolecular triangle reaction. The other example is the Ricci curvature on a probability manifold with a three-point lattice graph.

In the literature, geometric calculations in Wasserstein-2 spaces have been studied in [8, 10, 14], which show the link between Bakry-Émery Γ_1 operators [2] and Otto calculus. Geometric calculations in generalized Wasserstein-2 metric spaces from the Onsager reciprocal relation on discrete states are still partially known. The high-order derivative of the Onsager response matrices defines the Riemannian curvature tensors in probability manifolds. We leave the related network analysis of curvature tensors for future work. We expect this direction to benefit the design of machine learning architectures for sampling distributions on discrete states.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2006.

[2] D. Bakry and M. Émery. Diffusions hypercontractives. *Séminaire de Probabilités XIX*, volume 1123, 177–206, 1985.

[3] L. Bertini, A.D. Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim. Macroscopic fluctuation theory. *Rev. Mod. Phys.*, 87, 593, 2015.

[4] S.N. Chow, W. Huang, Y. Li, and H. Zhou. Fokker–Planck equations for a free energy functional or Markov process on a graph. *Archive for Rational Mechanics and Analysis*, volume 203, 969–1008, 2012.

[5] M. Grmela and H.C. Ottinger. Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. *Phys. Rev. E.*, 56 (6): 6620–6632, 1997.

[6] S. Ito. Geometric thermodynamics for the Fokker–Planck equation: stochastic thermodynamic links between information geometry and optimal transport. *Information Geometry*, 7, 441–483, 2024.

[7] R. Jordan, D. Kinderlehrer, and F. Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[8] J.D. Lafferty. The Density Manifold and Configuration Space Quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

[9] W. Li. Langevin dynamics for the probability of finite state Markov processes. *Information Geometry*, 2024.

[10] J. Lott. Some Geometric Calculations on Wasserstein Space. *Communications in Mathematical Physics*, 277(2):423–437, 2008.

[11] J. Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[12] A. Mielke. A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329, 2011.

[13] L. Onsager. Reciprocal Relations in Irreversible Processes. I. *Phys. Rev.* 37, 1931.

[14] F. Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[15] M.K. von Renesse, and K.T. Sturm. Entropic measure and Wasserstein diffusion. *The Annals of Probability*, 37(3):1114–1191, 2009.

[16] C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.
[17] M.T. Vrugt, H. Lowen, and R. Wittkowski. Classical dynamical density functional theory: from fundamentals to applications. *Adv. Phys.* 69, 2, 121–247, 2020.

Wasserstein gradient flows of functionals with entropic optimal transport

Théo Lacombe
(joint work with M. Hardion, G. Mordant, and F.-X. Vialard)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures, and $\epsilon > 0$. Define the entropic optimal transport problem (EOT) as

$$\text{(EOT)} \qquad \mathrm{OT}_\epsilon(\mu,\nu) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \langle c, \pi \rangle + \epsilon \mathrm{KL}(\pi | \mu \otimes \nu)$$

$$\text{(EOT-dual)} \qquad = \max_{f,g} \langle f, \mu \rangle + \langle g, \nu \rangle - \epsilon \left\langle e^{\frac{f \oplus g - c}{\epsilon}} - 1, \mu \otimes \nu \right\rangle,$$

where $c(x,y) \coloneqq \|x-y\|^2$ in this presentation, $\Pi(\mu,\nu)$ is the transportation polytope, KL is the relative entropy (i.e. Kullback-Leibler divergence), and $\langle f, \mu \rangle = \int f \mathrm{d}\mu$. Solutions of (EOT-dual), called Schrödinger potentials, are denoted by (f_μ^ν, g_μ^ν). When $\mu = \nu$, we denote by f_μ the (symmetrized) solution of the (self-)EOT problem. We recall that OT_ϵ presents the advantage of being easily computable when μ, ν are known through samples, and one has access to the gradients of the Schrödinger potentials ∇f_μ^ν at the same computational cost, which will play an important role when considering Wasserstein gradient flows of functionals involving OT_ϵ in their definition. We also recall that another convenient instance of this problem is that of Gaussian distributions. Indeed, when μ, ν are Gaussian distributions, both $\mathrm{OT}_\epsilon(\mu,\nu)$ and ∇f_μ^ν are known in closed form, the latter being in particular affine [3].

In this work, we study the Wasserstein gradient flow of two functionals. The first one is the Sinkhorn divergence, defined as

$$\text{(Sk div)} \qquad S_\epsilon \colon \mu \mapsto \mathrm{OT}_\epsilon(\mu,\mu_\star) - \frac{1}{2}\left(\mathrm{OT}_\epsilon(\mu,\mu) + \mathrm{OT}_\epsilon(\mu_\star,\mu_\star)\right),$$

where we restrict the analysis to the case where both the source μ_0 and the target $\mu_\star$ are Gaussian distributions.

The second functional is defined as

$$\text{(Sk. rel. entropy)} \qquad F_\epsilon \colon \mu \mapsto -\frac{1}{\epsilon}\mathrm{OT}_\epsilon(\mu,\mu) + \int V \mathrm{d}\mu,$$

where $V \colon \mathbb{R}^d \to \mathbb{R}$ is *analytic*, convex, with $\lambda_{\min} I_d \preceq \nabla^2 V \preceq \lambda_{\max} I_d$. We assume (without loss of generality) $V(0) = \nabla V(0) = 0$.

Wasserstein gradient flow of the Sinkhorn divergence between Gaussian distributions. [2] The Wasserstein gradient flow of (Sk div) is (expected to be) given by the PDE

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \left(\nabla f_\mu^{\mu_\star} - \nabla f_\mu\right)\right). \tag{1}$$

To the best of our knowledge, this expression is only derived in the compact case [1]. We prove that the Sinkhorn divergence (Sk div) restricted to (possibly singular) Gaussian distributions is convex along Wasserstein generalized geodesics and derive an expression for the corresponding Wasserstein derivative, yielding the existence of a unique solution of the flow (1) in that space, and for which we prove uniqueness in a larger class of measures using an EVI argument. Letting $(\mu_t)_t$ denote this flow, we prove in particular that

- There exists μ_∞ such that $\mu_t \to \mu_\infty$. Furthermore, $\mu_\infty = \mu_\star$ if and only if the support of $\mu_\star$ is included in the support of μ_0. In particular, if μ_0 is a non-singular Gaussian (i.e. admits a density with respect to Lebesgue), then $\mu_t \to \mu_\star$.
- If we furthermore assume that the covariance matrices of μ_0 and $\mu_\star$ commute, we obtain (i) exponential convergence rates when the support of the two measures coincides, but (ii) only a linear rate if the target $\mu_\star$ has a strictly smaller support than that of μ_0.

While our analysis is restricted to Gaussian distributions, we numerically observe that most of our conclusions hold beyond that setting. In particular, we observe that the gradient flow starting from a measure μ_0 which admits a density toward a target $\mu_\star$ with a lower-dimensional support seemingly converges globally, but at a slow rate. This suggests that using the Sinkhorn divergence as an objective function may not be adapted to such contexts (yet ubiquitous in generative modeling for instance).

The Sinkhorn relative entropy (Work in progress). The rescaled self-entropic term $-\frac{1}{\epsilon}\mathrm{OT}_\epsilon(\mu,\mu)$ is known to be—under some regularity assumption on μ—an approximation of the usual entropy $H(\mu) = \int \log(\mu)\mathrm{d}\mu$ when $\epsilon \to 0$, and the term $-\frac{2}{\epsilon}\nabla f_\mu^\mu$ converges toward the score $\nabla \log(\mu)$. This suggests that F_ϵ and its gradient flow when $\epsilon > 0$ can be interpreted as a variant of the relative entropy $\mathrm{KL}(\mu|e^{-V})$. Furthermore, in sharp contrast with the relative entropy, F_ϵ and its flow are well defined even for atomic measures, enabling a deterministic (and computationally efficient) implementation of the flow.

Restricting to sub-Gaussian distributions, we show that F_ϵ exhibits some (Wasserstein) convexity properties that guarantee the existence and uniqueness of the flow in that case. Regarding critical points, we show that there exists (at most) a unique stationary density $\mu_\star^\epsilon \propto e^{-V_\epsilon}$, where V_ϵ is an explicit shrunk version of V. Other critical points exist (e.g. δ_0 is always critical). We prove that when $\epsilon > \lambda_{\min}^{-1}$, *collapse* occurs, i.e. the flow $(\mu_t)_t$ starting from any μ_0 converges toward δ_0; conversely, this phenomenon does not occur if ϵ is smaller than $\lambda_{\max}^{-1}$, in the sense that $\inf_{t\geq 0}\int |x|^2 \mathrm{d}\mu_t > 0$.

When μ_0 admits a density, we expect that $\mu_t \to \mu_\star^\epsilon$ with exponential convergence rates (e.g. with respect to the relative entropy)—we can prove that this holds for Gaussian distributions. On the other hand, when μ_0 is discrete, the smoothness of the flow ensures that so is μ_t for all t. The analyticity of V can be used to prove that μ_t converges to some discrete (non-unique) critical point μ_∞^ϵ. Somewhat surprisingly, we prove that when $\epsilon \to 0$, $\mu_\infty^\epsilon \to \delta_0$. This suggests that the regularization parameter ϵ_{opt} should be chosen carefully in $(0, \lambda_{\max}^{-1})$ so that μ_∞^ϵ is a good (discrete) approximation of e^{-V}, which is also what we observe numerically.

References

[1] Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Displacement smoothness of entropic optimal transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:25, 2024.

[2] Mathis Hardion and Théo Lacombe. The Wasserstein gradient flow of the Sinkhorn divergence between Gaussian distributions. *arXiv preprint arXiv:2602.10726*, 2026.

[3] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced Gaussian measures has a closed form. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 10468–10479. Curran Associates Inc., 2020.

Telegrapher's Generative Model via Kac Flows

Richard Duong

(joint work with J. Chemseddine, P. K. Friz, and G. Steidl)

We break the mold in flow-based generative modeling by proposing a new model based on the damped wave equation, also known as telegrapher's equation. Similar to the diffusion equation and Brownian motion, there is a Feynman-Kac type relation between the telegrapher's equation and the stochastic Kac process in 1D. The Kac flow evolves stepwise linearly in time, so that the probability flow is Lipschitz continuous in the Wasserstein distance and, in contrast to diffusion flows, the norm of the velocity field is globally bounded by the finite speed of the Kac process. Furthermore, the Kac model has the diffusion model as its asymptotic limit. We extend these considerations to a multi-dimensional stochastic process. Since the damped wave equation is no longer mass-conserving in higher dimensions, we consider a componentwise construction which consists of independent 1D Kac processes in each spatial component. We show that this noising process gives rise to an absolutely continuous curve in the Wasserstein space and compute the conditional velocity field starting in a Dirac point analytically. Using the framework of flow matching, we train a neural network that approximates the velocity field and use it for sample generation. Our numerical experiments demonstrate the scalability of our approach, and show its advantages over diffusion models. Our componentwise construction of noises can be generalized to other one-dimensional processes, in particular, usual flow matching (via independent couplings) and diffusion models already employ this framework implicitly [4, 5], by adding isotropic

Gaussian noise to the data. This idea forms the basis of the author's subsequent work [2], where the 1D noising process is learned and adapted to the data. Further extensions of our model may include tasks in conditional generation, inverse problems, and sampling of Boltzmann distributions. An application toward model distillation, building on the finite-speed dynamics of the Kac framework, can be found in the work [3].

References

[1] R. Duong, J. Chemseddine, P.K. Friz, G. Steidl, *Telegrapher's Generative Model via Kac Flows*, arXiv preprint arXiv:2506.20641 (2025).
[2] J. Chemseddine, G. Kornhardt, R. Duong, G. Steidl, *Adapting Noise to Data: Generative Flows from 1D Processes*, arXiv preprint arXiv:2510.12636 (2025).
[3] W. Han, C. Meng, C.D. Manning, S. Ermon, *DistillKac: Few-Step Image Generation via Damped Wave Equations*, ICLR (2026).
[4] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le, *Flow Matching for Generative modeling*, ICLR (2023).
[5] M.S. Albergo, N.M. Boffi, E. Vanden-Eijnden, *Stochastic Interpolants: A Unifying Framework for Flows and Diffusions*, arXiv preprint arXiv:2303.08797 (2023).

Training Dynamics of Neural Networks in the Large-scale Limit

Lénaïc Chizat
(joint work with L.-P. Chaintron and J. Maass)

We study the training dynamics of neural networks in the regime where several architectural dimensions–depth, width, and embedding dimension–are simultaneously large. Empirically, increasing compute budget, in particular by scaling model size, consistently improves performance of neural networks, suggesting that optimal behavior emerges in suitable large-scale limits. This motivates the problem of identifying and analyzing such limits in a mathematically rigorous way. The talk is based on [1] and [2].

We begin with the classical mean-field limit of two-layer perceptrons, where the hidden width diverges $M \to \infty$ at fixed embedding dimension D (representing here the input and output dimension). In this regime, the dynamics converges to a deterministic evolution on probability measures, interpreted as a Wasserstein gradient flow. Building on earlier results on the 1D output case [3], we show that in the maximal update regime, the finite-width dynamics approximates this limit with error of order $\sqrt{D/M}$, under suitable assumptions. We also briefly mention some recent advances in the long-time behavior of such Wasserstein gradient flows for targets of Sobolev regularity [4].

We then turn to deep residual networks (ResNets), whose architecture involves three main scaling parameters: depth L, width M, and embedding dimension D. A central question is to understand the joint large-scale limit $L, M, D \to \infty$ and to identify the limiting dynamics.

Our main result establishes the existence of an explicit limiting dynamical system describing gradient descent. More precisely, for *clipped* gradient descent in

the maximal local update (MLU) regime, we prove a quantitative approximation result: the discrepancy Δ_k between the finite network and its limit after k steps satisfies, with high probability,

$$\Delta_k = O\Big(\frac{1}{L} + \frac{\sqrt{D}}{\sqrt{ML}} + \frac{1}{\sqrt{D}}\Big).$$

This bound is non-asymptotic and can be observed to be empirically tight when measured in embedding space. In particular, for a total parameter budget $P = \Theta(MLD)$, optimizing the architecture yields a convergence rate of order $P^{-1/6}$.

This analysis relies on several key structural insights. First, the general picture gains clarity and consistency once one considers the three free architectural parameters L, M, D rather than only two as was explored in prior works (e.g. by fixing $M = D$). Second, in the large-depth limit, ResNets behave as if they were also infinitely wide, even when the width M is fixed. This "hidden width" phenomenon shows that depth effectively increases the number of interacting units, leading to an effective width of order ML. Exploiting this phenomenon is crucial in order to identify the appropriate limit. Third, we identify the scaling of hyperparameters that lead to the maximal local update (MLU) regime, where features of the input are learned at a local level in the ResNet. In this regime, the system involves interactions in the CLT scaling, which are rarely studied in the context of mean-field systems. This led us to develop techniques of independent interest, such as a rigorous and quantitative version of the cavity method from statistical physics [5].

Extensions and perspectives. The framework extends, at least formally, to transformer architectures with bounded key/query dimension, leading to analogous limiting dynamics. We also discuss conjectured sharper rates in output space, suggesting an optimal scaling $L \sim P^{1/5}$ and $M, D \sim P^{2/5}$, with convergence rate $P^{-1/5}$, in agreement with numerical experiments.

References

[1] L. Chizat (2025). *The hidden width of deep ResNets: Tight error bounds and phase diagrams*, arXiv preprint arXiv:2509.10167.

[2] L.P. Chaintron, L. Chizat, J. Maas (2026). *ResNets of All Shapes and Sizes: Convergence of Training Dynamics in the Large-scale Limit*, arXiv preprint arXiv:2603.18168.

[3] S. Mei, T. Misiakiewicz, A. Montanari (2019). *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*, Conference on learning theory (pp. 2388-2464). PMLR.

[4] L. Chizat, M. Colombo, R. Colombo, X. Fernández-Real (2026). *Quantitative Convergence of Wasserstein Gradient Flows of Kernel Mean Discrepancies*, arXiv preprint arXiv:2603.01977.

[5] M. Mézard, G. Parisi, M.A. Virasoro, D.J. Thouless (1988). *Spin glass theory and beyond*.

On the Relation between Rectified Flows and Optimal Transport

Johannes Hertrich
(joint work with A. Chambolle and J. Delon)

In generative modeling, we aim to establish a sampling process from a probability distribution μ_1 out of a latent distribution μ_0. In this talk, we consider a generative model called rectified flows [1, 3, 5], which is also known as flow matching or stochastic interpolants. Starting with a coupling γ between μ_0 and μ_1, rectified flows define interpolations μ_t by $\int \psi(x)\mathrm{d}\mu_t(x) = \int \psi((1-t)x+ty)\mathrm{d}\gamma(x,y)$ and a velocity field v_t by minimizing the flow matching loss

$$\int_0^1 \int \|v_t((1-t)x+ty) - y + x\|^2 \mathrm{d}\gamma(x,y)\mathrm{d}t.$$

It can be shown that the pair (μ_t, v_t) satisfies the continuity equation $\partial_t \mu_t + \mathrm{div}(v_t\mu_t) = 0$. If v_t is sufficiently smooth, this yields that $\mu_1 = (\phi_1)_\# \mu_0$, where ϕ_t denotes the solution of the flow ODE $\partial_t \phi_t(x) = v_t(\phi_t(x))$ with $\phi_0(x) = x$. Consequently, one can generate samples from μ_1 by sampling $x \sim \mu_0$ and evaluating $\phi_1(x)$. Furthermore, one may introduce the "rectified" coupling $\mathcal{R}(\gamma) \coloneqq (I, \phi_1)_\# \mu_0$, which achieves a lower transport cost than the original coupling γ.

In order to generate couplings with straight trajectories in the flow ODE, Liu et al. [4, 5] propose to iterate this procedure and to consider the sequence $(\gamma_n)_n$ defined by $\gamma_{n+1} = \mathcal{R}(\gamma_n)$. In this talk, we consider limit points of reflow in the following settings:

- First, we consider a weak formulation of the rectified coupling $\mathcal{R}(\gamma)$ if the flow ODE does not admit a unique solution and show that limits of reflow admit straight trajectories in the flow ODE.
- If reflow is combined with minibatch OT with batch size N, we show that limits are rectifiable, N-monotone and straight couplings.
- If we in addition restrict the velocity v_t to (generalized) gradient fields, we show that under an additional assumption on the support the optimal transport plan is the unique fixed point of reflow. We also demonstrate that this support assumption is essential, in the sense that the result does not hold without it.

The talk is based on [2] and some work in progress.

References

[1] M. S. Albergo and E. Vanden-Eijnden, *Building normalizing flows with stochastic interpolants*, International Conference on Learning Representations, 2023.
[2] J. Hertrich, A. Chambolle, and J. Delon, *On the relation between rectified flows and optimal transport*, Conference on Neural Information Processing Systems, 2025.
[3] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow matching for generative modeling*, International Conference on Learning Representations, 2023.
[4] Q. Liu, *Rectified flow: A marginal preserving approach to optimal transport*, arXiv preprint arXiv:2209.14577, 2022.
[5] X. Liu, C. Gong, and Q. Liu, *Flow straight and fast: Learning to generate and transfer data with rectified flow*, International Conference on Learning Representations, 2023.

A Unifying View of Variational Generative Wasserstein Flows

Anna Korba
(joint work with P. Caucheteux and C. Bonet)

The recent success of generative modeling in machine learning has been driven by a variety of paradigms, including diffusion models, normalizing flows, generative adversarial networks (GANs), and optimal transport-based approaches. While these methods differ significantly in their formulations and algorithmic implementations, they all share a common objective: approximating an unknown target probability distribution from samples. Understanding these methods within a unified mathematical framework remains a central challenge, both for theoretical analysis and for the design of improved algorithms.

A natural and powerful perspective is provided by the theory of Wasserstein gradient flows, which describes the evolution of probability measures as steepest descent curves of functionals defined over the Wasserstein space. This viewpoint has deep roots in optimal transport and partial differential equations, and has recently gained increasing attention in machine learning. In this context, the Jordan–Kinderlehrer–Otto (JKO) scheme provides a principled time discretization of such flows, by iteratively solving proximal optimization problems in the space of probability measures.

In this work, we introduce a unified framework for generative modeling based on Wasserstein gradient flows, which we refer to as *Generative Wasserstein Flows* (GWF). Our approach is centered around the observation that many existing generative models can be interpreted as approximate implementations of JKO schemes associated with suitable discrepancy functionals. This perspective allows us to bridge a wide range of methods, revealing previously unnoticed equivalences and clarifying the role of geometric and variational structures in generative modeling.

We first focus on functionals defined by f-divergences, a broad class of discrepancy measures that includes the Kullback–Leibler divergence and the Jensen–Shannon divergence. By considering JKO schemes for such functionals, we derive generative algorithms that operate by iteratively transporting a source distribution toward the target distribution. A key feature of this formulation is that it naturally leads to optimization problems over transport maps, which can be parameterized using neural networks.

A central contribution of this work is to show that several recently proposed methods for generative modeling, which were introduced from different perspectives, can in fact be derived from a common JKO-based formulation. In particular, we establish the equivalence between variational Wasserstein gradient flow methods and scalable JKO schemes based on unbalanced optimal transport formulations. Despite their apparent differences–such as the order of minimization and maximization or the use of dual formulations–these methods are shown to correspond to the same underlying optimization principle.

Moreover, we demonstrate that adversarial training procedures, as used in GANs, arise naturally within this framework. Indeed, by exploiting variational

representations of f-divergences, the JKO step can be rewritten as a min-max optimization problem involving a generator and a discriminator. This establishes a direct connection between proximal schemes in Wasserstein space and adversarial learning, and shows that certain regularized GAN objectives can be interpreted as discretizations of Wasserstein gradient flows. In particular, relaxed Wasserstein proximal GANs are shown to coincide with JKO schemes up to the orientation of the divergence.

While f-divergences provide a natural starting point, they are limited in their ability to compare distributions with disjoint supports. To address this limitation, we extend the GWF framework to more general classes of discrepancies, including integral probability metrics (IPMs) such as the Wasserstein-1 distance, as well as kernel-based measures such as the Maximum Mean Discrepancy (MMD). Leveraging their variational formulations, we derive new JKO-based generative schemes that generalize existing approaches.

In the case of the squared MMD, we obtain a novel training procedure that can be interpreted as a JKO-regularized version of MMD-based generative models. Interestingly, unlike the case of f-divergences, the associated optimization problem admits a partially closed-form solution for the discriminator, leading to different algorithmic structures. We further show how to incorporate learned feature representations, thereby recovering adversarial formulations analogous to MMD-GANs within the JKO framework.

Beyond the formulation of new algorithms, our framework also sheds light on the role of regularization in generative modeling. The JKO scheme introduces a proximal term involving the Wasserstein distance, which acts as a form of temporal regularization and can significantly stabilize training. Through extensive numerical experiments, we investigate the impact of this regularization across a wide range of divergences and model architectures. Our results indicate that JKO regularization can substantially improve performance and stability, particularly for IPM-based objectives, while providing more moderate but consistent gains for f-divergences.

From a theoretical perspective, we further study the case of *parametric Wasserstein flows*, where the evolution of distributions is restricted to a family induced by parameterized maps. This setting is particularly relevant in practice, as generative models are typically defined through neural networks. We show that the induced dynamics can be interpreted as projected Wasserstein gradient flows, where the true Wasserstein gradient is approximated within a subspace determined by the parameterization. This leads to a connection between optimization in parameter space and gradient flows in measure space, and provides a principled interpretation of training dynamics in generative models.

Overall, the framework of Generative Wasserstein Flows provides a unifying lens through which a wide variety of generative modeling approaches can be understood and analyzed. By emphasizing the role of Wasserstein geometry and variational principles, it establishes deep connections between optimal transport and generative modeling. Our perspective opens new directions for the development

of generative models, both in terms of theoretical understanding and algorithmic design.

References

[1] P. Caucheteux, C. Bonet, A. Korba, *A Unifying View of Variational Generative Wasserstein Flows*, Submitted (2026).

A functional approach to differential programming in machine learning

Michel Arbel

(joint work with I. Petrulyonite, F. El Khoury, J. Mairal, E. Pawels, and S. Vaiter)

Machine learning has thrived by training expressive models on large datasets, but is now shifting toward integrating prior knowledge–like physical laws or structure–into the learning process. Many such problems can be formulated as an implicitly constrained learning (ICL) task, where a prediction model is subject to a partially known constraint whose unknown component is inferred by aligning model predictions with data. ICL problems arise across domains: enforcing physics in scientific models, planning from a learned world model, or accounting for hidden variables for causal effect estimation.

ICL formalizes problems where the goal is to find parameters θ, in a set Θ, that (locally) minimize a reduced objective $J(\theta)$ depending on a state variable, or *predictor*, u_θ, typically a function in an infinite-dimensional space $\mathcal{U}$, uniquely determined by an implicit constraint:

$$\min_{\theta\in\Theta} J(\theta) \coloneqq L(\theta, u_\theta), \qquad \text{s.t.} \quad C(\theta, u_\theta) = 0. \tag{1}$$

ICL connects to bilevel optimization when the constraint expresses an optimality condition [1], and to mathematical programs with equilibrium constraints (MPECs) [2]. However, unlike these formulations, ICL integrates statistical considerations—such as bias-variance trade-offs and generalization—absent from classical optimization. In practice, both the objective and constraint are estimated from finite data, which may be external or adaptively generated for the task. A learning theory of ICL must therefore analyze the generalization of both objective and constraint, while accounting for their intricate interdependence. **ICL lies at the intersection of learning and constrained optimization, yet falls outside the reach of existing theoretical frameworks for either.**

ICL possesses a *functional structure* that is often overlooked in deep learning. In other words, the constraint **uniquely specifies the state variable u_θ in a functional space $\mathcal{U}$**, so that u_θ becomes a well-defined *implicit function* of θ, ensuring that the reduced objective $J(\theta)$ is mathematically well-defined. The key to solving Eq. (1) with gradient-based methods is thus to account for this implicit dependence when computing $\nabla_\theta J(\theta)$. In practice, the predictor is often

approximated by deep neural networks f_ψ parameterized by ψ, given their expressivity and empirical success. When doing so, existing gradient-based approaches–such as those based on *differentiable programming* [3, 4]–replace the original ICL problem in Eq. (1) with a *surrogate* one, where the constrained variable is no longer the state u but the network's parameters ψ. These parameters are then treated as an implicit function of θ (i.e., $\theta \mapsto \psi^\star(\theta)$), whose dependence is accounted for when computing gradients. However, the symmetries and over-parameterization of neural networks [5, 6] imply that a single θ can correspond to **multiple solutions** $\psi^\star(\theta)$, rendering this surrogate problem *ill-defined*, as it depends on arbitrary choices of $\psi^\star(\theta)$ [7]. **This mismatch not only causes algorithmic instabilities** [8],[9, 7]**, but also precludes any rigorous theoretical analysis of these approaches.**

In this talk, I present a methodological framework for solving ICL, capable of handling complex implicit constraints–ranging from optimality conditions to partial differential equations and dynamical systems. The approach adopts a functional viewpoint that reframes ICL as a well-defined yet abstract problem, using deep networks only as approximation tools. We apply this framework to a class of bilevel optimization problems in Hilbert spaces covering several applications such as instrumental variable regression and model-based reinforcement learning. We expect this functional viewpoint on ICL to extend beyond Hilbert spaces to cover measure spaces, with possible applications to generative modeling, in particular for efficiently "guiding" diffusion models.

References

[1] S. Dempe, *Foundations of Bilevel Programming*. Kluwer Academic Publishers, 2002. doi: 10.1007/978-1-4615-1071-8.

[2] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996.

[3] M. Blondel and V. Roulet, "The elements of differentiable programming," *arXiv preprint*, 2024.

[4] Z. Hao, C. Ying, H. Su, J. Zhu, J. Song, and Z. Cheng, "Bi-level Physics-Informed Neural Networks for PDE Constrained Optimization using Broyden's Hypergradients," in *The Eleventh International Conference on Learning Representations*, 2023.

[5] B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, and J. Brea, "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances," in *International Conference on Machine Learning*, pp. 9722–9732, PMLR, 2021.

[6] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15849–15854, 2019.

[7] M. Arbel and J. Mairal, "Non-Convex Bilevel Games with Critical Point Selection Maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] P. Vicol, J. Lorraine, D. Duvenaud, and R. Grosse, "Implicit Regularization in Overparameterized Bilevel Optimization," in *ICML 2021 Beyond First Order Methods Workshop*, 2021.

[9] I. Petrulionyte, J. Mairal, and M. Arbel, "Functional Bilevel Optimization for Machine Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

The Stein–Log–Sobolev Inequality for the Continuous SVGD Flow

José A. Carrillo
(joint work with J. Skrzeczkowski and J. Warnett)

In this talk, I discussed the proof of this Stein–LSI relating the Stein discrepancy to the Kullback–Leibler divergence. We show that under very general assumptions on the potential V (merely bounded below by a parabola $|x|^2$) the inequality holds as long as the kernel is singular enough at 0. More precisely, as singular as the screened Poisson kernel.

The proof relies on a clever use of Fourier variables and realizing how to compensate the negative part with the positive contribution given by the dissipation. We rely on Fourier variables to connect the different weights in the inequality to conclude. We give several counterexamples to the Stein–LSI under several assumptions.

Effective fluctuating continuum models for stochastic gradient descent

Benjamin Gess
(joint work with V. Konarovskyi, R. Gvalani, and S. Kassing)

This report describes continuum limiting models for stochastic gradient descent (SGD) that incorporate training trajectories and fluctuations. The analysis was presented in the context of the MFO workshop Flows on Measure Spaces and Applications in Machine Learning, and is based on joint work with Vitalii Konarovskyi (Hamburg University), Rishabh Gvalani (Edinburgh University), and Sebastian Kassing (Wuppertal University). While deterministic limits describe the law of large numbers behavior of SGD in small learning rate or infinite-width regimes, these models do not account for the stochasticity of the training process. Two frameworks are discussed: Stochastic Modified Flows (SMF) and Conservative SPDEs for the mean-field limit of overparameterized networks in the small learning rate regime.

In the small learning rate regime ($\eta \to 0$), SGD trajectories are often approximated by Stochastic Modified Equations (SME). SMEs possess diffusion coefficients involving the square root of a degenerate covariance matrix and are designed to match single-point statistics. In joint work with S. Kassing and V. Konarovskyi [1], Stochastic Modified Flows (SMF) are introduced as an alternative approximation. These are stochastic differential equations driven by a cylindrical Wiener process W on a data-related Hilbert space $L_2(\Theta, \vartheta)$:

$$dX_t^\eta = -\nabla \left(R(X_t^\eta) + \frac{\eta}{4} |\nabla R(X_t^\eta)|^2 \right) dt + \sqrt{\eta} \int_\Theta G(X_t^\eta, \theta) W(d\theta, dt),$$

where G represents the noise in individual gradients. SMFs match the multi-point statistics of SGD to order η^2. This formulation allows for the analysis of the stability of SGD and its behavior near manifolds of minima as a consistent stochastic flow.

In the overparameterized regime where the number of parameters $M \to \infty$, and for shallow networks, the empirical measure of the network parameters is an interacting particle system. In joint work with R. Gvalani and V. Konarovskyi [2], the convergence of this empirical measure to solutions of a Stochastic Mean-Field Equation (SMFE) is established:

$$d\mu_t = -\nabla\cdot(V(\cdot,\mu_t)\mu_t)dt + \frac{\sigma}{2}D^2 : (A(\cdot,\mu_t)\mu_t)dt + \sigma^{1/2}\nabla\cdot\int_\Theta G(\cdot,\mu_t,\theta)\mu_t W(d\theta,dt).$$

Well-posedness for these conservative, nonlinear SPDEs is established via a superposition principle, where solutions are represented as superpositions of trajectories from an underlying SDE with interaction.

These fluctuating continuum models offer a mathematical basis for studying the implicit bias and stability properties of SGD.

References

[1] B. Gess, S. Kassing, and V. Konarovskyi, *Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent*, Journal of Machine Learning Research, 25(103):1–27, 2024.

[2] B. Gess, R. S. Gvalani, and V. Konarovskyi, *Conservative SPDEs as Fluctuating Mean-Field Limits of Stochastic Gradient Descent*, Probability Theory and Related Fields, 192(3-4):1447–1515, 2025.

Structure preservation in neural networks and for approximating dynamics

Elena Celledoni

(joint work with B. Owren, M. D. Hansen, M. Ghirardelli, and D. M. de Diego)

This work investigates a geometric and structure-preserving approach to deep learning, focusing on the relationship between neural networks, differential equations, and mechanical systems evolving on nonlinear manifolds. Contemporary architectures such as residual networks and neural ordinary differential equations (neural ODEs) admit rigorous interpretations as time discretizations of continuous dynamical systems. Adopting this dynamical systems viewpoint makes it possible to systematically incorporate techniques from numerical analysis, geometric integration, and mechanics into machine learning frameworks. As a result, the resulting models exhibit improved stability, robustness, and interpretability.

Deep neural networks can naturally be viewed through the lens of dynamical systems. For instance, residual networks can be interpreted as forward Euler discretizations of differential equations, while neural ODEs correspond to models with continuous depth. This interpretation motivates the analysis of properties such as stability, contractivity, and structure preservation using tools from geometric numerical integration. Furthermore, when the data evolve on nonlinear configuration spaces, it becomes advantageous to formulate learning problems directly on Lie groups and Riemannian manifolds.

Learning Mechanical Multibody Systems from Data. We introduce a structure-preserving framework for learning mechanical multibody systems from observational data. The proposed methodology is based on the Lagrange–d'Alembert principle together with ideas from discrete variational mechanics. By expressing the learning task through a discrete action principle, the resulting models inherently respect geometric constraints, balance laws, and the intrinsic structure of the configuration space.

An important feature of the method is that mechanical systems can be identified using position measurements alone, without requiring explicit velocity data. Variational integrators are employed to ensure consistency with the underlying mechanical principles while preserving key invariants such as momentum maps and geometric constraints. In this way, physical structure is incorporated directly into the learning architecture, producing interpretable models that remain stable even under long-time integration.

Potential applications include the analysis of human motion from motion-capture datasets (see Figure 1). In this context, complex multibody dynamics can be inferred while respecting the nonlinear manifold structure of joint configurations. The framework also enables tasks such as classification, dimensionality reduction, and physically consistent animation, thereby establishing a principled connection between geometric mechanics and data-driven modeling [4].

Stability on Riemannian Manifolds. A second focus of this work is the stability analysis of deep learning architectures defined on Riemannian manifolds. When neural networks are interpreted as discretizations of differential equations, questions of non-expansiveness, contractivity, and conditional stability naturally arise for the associated numerical schemes (Figure 2).

We extend classical stability results from Euclidean settings to the Riemannian manifold context by studying geometric integrators, including geodesic Euler methods. In particular, contractivity conditions are established in terms of curvature bounds and suitable step-size restrictions. These results provide rigorous

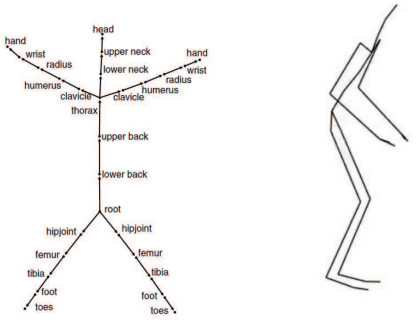

FIGURE 1. Human motion data.

guarantees for robustness in learning problems where the data take values on manifolds [1], [3].

The analysis offers theoretical guidance for designing deep learning architectures that operate on nonlinear configuration spaces such as Lie groups and shape manifolds. By enforcing geometric consistency and stability at the algorithmic level, the resulting models demonstrate improved robustness, stronger generalization properties, and enhanced numerical reliability.

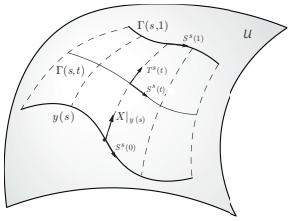

FIGURE 2. Construction for the proof of the main theorem.

Conclusion. In summary, this work promotes a unified perspective in which deep learning architectures are interpreted and designed as geometric numerical integrators. By combining structure-preserving mechanics with stability analysis on manifolds, the proposed framework connects machine learning, differential geometry, and computational mechanics within a coherent and mathematically principled setting.

REFERENCES

[1] Martin Arnold, Elena Celledoni, Ergys Çokaj, Brynjulf Owren, Denise Tumiotto, *B-stability of numerical integrators on Riemannian manifolds*, Journal of Computational Dynamics, 2024.

[2] E. Celledoni, M. Eslitzbichler and A. Schmeding, *Shape analysis on Lie groups with applications in computer animation*, Journal of Geometric Mechanics, 2016.

[3] M. Ghirardelli, B. Owren and E. Celledoni, *Conditional Stability of the Euler Method on Riemannian Manifolds*, arXiv:2503.09434.

[4] In preparation.

Distribution Matching with Sliced Optimal Transport

KIMIA NADJAHI

(joint work with G. Thurin and C. Boyer)

Many problems in modern machine learning require comparing and matching probability distributions, such as generative modeling, density estimation, or domain adaptation. The goal is typically to transform a source distribution in order to match a more complex target distribution. Distribution matching can be formalized through optimal transport (OT), which provides both a geometrically meaningful distance between probability measures and, when it exists, a transport map

pushing a source distribution σ to a target distribution μ. However, computing OT maps is in general expensive both computationally and statistically.

The high cost of OT has motivated alternative approaches that decompose the transport problem into simpler subproblems. A key idea is to build an interpolation between σ and μ through a sequence of elementary transformations, rather than estimating a single global transport map. This idea underlies many iterative correction schemes: although each step may only partially reduce the discrepancy between σ and μ, their composition is expected to gradually align them. Among all possible interpolations, McCann's interpolation plays a distinguished theoretical role, as it corresponds to geodesics in Wasserstein space, but it is rarely tractable. A generic iterative sequence of measures that mimics McCann's interpolation can be constructed through: $\sigma_0 = \sigma$, then

$$\forall k \geq 0, \qquad \sigma_{k+1} = \big((1-\gamma_k)\mathrm{Id} + \gamma_k \widehat{T}_k\big)_\sharp \sigma_k, \tag{1}$$

where $\widehat{T}_k$ is an approximate transport map from $\widehat{\sigma}_k$ to μ, $(\gamma_k)_k$ a sequence of step sizes, and $T_\sharp\sigma$ denotes the pushforward of σ by the function T: if $X \sim \sigma$, then $T(X) \sim T_\sharp\sigma$.

Different choices for $\widehat{T}_k$ have been proposed. In this presentation, we focus on *sliced optimal transport*, a computationally efficient alternative that leverages one-dimensional projections [1]. More precisely, we study the convergence of the *slice-matching scheme*, an efficient iterative method for distribution matching based on sliced OT, whose iterates are

$$\forall k \geq 0, \qquad \sigma_{k+1} = \big((1-\gamma_k)\mathrm{Id} + \gamma_k T_{\sigma_k, P_{k+1}}\big)_\sharp \sigma_k, \tag{2}$$

where $(P_k)_{k\geq 1}$ is a sequence of random orthonormal bases, and $T_{\sigma_k, P_{k+1}}$ the *slice-matching map* associated with matching one-dimensional projections along the sampled basis. This procedure admits a natural interpretation as a stochastic gradient descent method in Wasserstein space for the Sliced-Wasserstein objective [2].

Our main goal is to establish convergence rates for the slice-matching scheme. Our approach is based on identifying Polyak–Łojasiewicz (PL) inequalities for the Sliced-Wasserstein objective, which bound the loss by the squared norm of its Wasserstein gradient. These inequalities imply quantitative convergence rates to the target distribution. The main technical challenge is that the associated constants depend on lower and upper bounds on the density of the iterates, which are difficult to control along the trajectory.

We address this difficulty within the class of elliptic distributions, for which slice-matching maps are linear. In this regime, controlling the density of the iterates amounts to controlling the eigenvalues of their covariance matrices. When the target distribution is isotropic, we show that these eigenvalues can be controlled in expectation, which in turn yields explicit convergence rates. Such spectral control holds from the very first iteration when the updates use random orthonormal bases of directions. This stands in contrast with the single-direction setting, where the lack of orthogonality leads to larger fluctuations in the covariance structure before stabilization. We complement our theory with numerical experiments and

illustrate the predicted dependence on dimension and step-size, as well as the stabilizing effect of orthonormal-basis sampling [3].

References

[1] J. Rabin, G. Peyré, J. Delon and M. Bernot, *Wasserstein barycenter and its application to texture mixing*. International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), 2011.

[2] S. Li, C. Moosmueller and Y. Wang, *Measure transfer via stochastic slicing and matching*. 2023. https://arxiv.org/abs/2307.05705

[3] G. Thurin, C. Boyer and K. Nadjahi. *Convergence Rates for Distribution Matching with Sliced Optimal Transport*. 2026. https://arxiv.org/abs/2602.10691

The sliced Wasserstein flow and the sliced Wasserstein distance

Filippo Santambrogio

(joint work with G. Cozzi)

Let ρ and ν be probability measures on $\mathbb{R}^d$. For every direction $\theta \in \mathbb{S}^{d-1}$, we denote by $\pi_\theta(x) = x \cdot \theta$ the projection onto that direction and by $\rho_\theta := (\pi_\theta)_\# \rho$ and $\nu_\theta := (\pi_\theta)_\# \nu$ the push-forward measures. We then consider the map T_θ, which is the optimal transport map (i.e., the non-decreasing one) from ρ_θ to ν_θ, and define the vector field

$$v[\rho, \nu](x) := \int (T_\theta - id)(x \cdot \theta)\theta d\theta.$$

We consider the evolution equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v[\rho_t, \nu]) = 0.$$

This evolution is an isotropic version in continuous time of the one introduced in [1], and it is expected to converge in the long-time to ν. Indeed, it can be seen that this equation is the gradient flow in the space $W_2(\mathbb{R}^d)$ of the functional

$$F(\rho) := \frac{1}{2} SW_2^2(\rho, \nu) = \int W_2^2(\rho_\theta, \nu_\theta) d\theta,$$

and this functional is only minimized by ν. The functional F is the so-called sliced-Wasserstein distance, which has recently emerged as a computationally tractable alternative to the classical Wasserstein distance in high-dimensional problems. Its definition relies on reducing the transport problem to one-dimensional projections, where explicit formulas for the optimal map are available. The above equation is then called Sliced Wasserstein flow.

Besides studying the flow in terms of the density ρ, we can also consider the flow map. For each initial point x, one may define the trajectory $y_x(t)$ solving the characteristic system

$$\begin{cases} y_x'(t) = v_t(y_x(t)), \\ y_x(0) = x, \end{cases}$$

where $v_t(x) = v[\rho_t, \nu](x)$. Defining the map Y_t through $Y_t(x) := y_x(t)$, we can consider, if it exists, the limiting map Y_∞. A natural conjecture is that this map

coincides with the optimal transport map between the initial distribution and the target. This belief can be supported by the monotone nature of the ingredients defining this map, but we proved in [4] that in general it is not true that Y_∞ coincides with the optimal transport map from ρ_0 to ν, with a construction very similar to the one in [2] and [3].

On the other hand, the question of the asymptotic behavior of the evolving density ρ_t is much more challenging. The lack of geodesic convexity of F (in the W_2 geometry) prevents the use of standard tools guaranteeing convergence to a minimizer, and, despite empirical evidence, no general proof that ρ_t converges to ν is available. It is even possible to have a counterexample with atomic measures.

Example. Consider the following construction in $\mathbb{R}^2$: let the starting measure be

$$\rho_0 = \frac{\delta_{(-1,0)}}{2} + \frac{\delta_{(1,0)}}{2}$$

and the target measure

$$\nu = \frac{\delta_{(0,a)}}{2} + \frac{\delta_{(0,-a)}}{2}.$$

Then, if $a > 0$ is well-chosen (we need $a = \pi/2$ in dimension $d = 2$) the configuration ρ_0 is stationary, *i.e.* the couple $(\rho, v) \coloneqq (\rho_0, 0)$ is a solution of the continuity equation $\partial_t \rho + \nabla \cdot (\rho v) = 0$ satisfying $0 = v[\rho_0, \nu]$. Thus, in this case, the flow does not converge to the target.

On the other hand, it is possible to obtain a positive convergence result in the case where the target distribution is the standard Gaussian. This is the main result of [4]. In this setting, we can look at the evolution in time of the quantity

$$G(\rho) \coloneqq \mathrm{KL}(\rho|\nu)$$

and prove that we have

$$\frac{d}{dt}\mathrm{KL}(\rho_t|\nu) \leq -\int \mathrm{KL}((\rho_t)_\theta|\nu_\theta) - \frac{1}{2}SW_2^2(\rho_t,\nu) \leq -\frac{1}{2}SW_2^2(\rho_t,\nu). \tag{1}$$

This computation is done by differentiating the entropy of ρ_t and its second moment separately, and obtaining terms involving the entropy and the second moments of $(\rho_t)_\theta$, that we can recombine as above. Using the nonnegativity of the KL term, we obtain, if $G(\rho_0) < \infty$, the integrability in time of $F(\rho_t)$, which is also a nonincreasing quantity, so it tends to 0 (at least as $1/t$). Unfortunately, the magical computation leading to (1) only works when ν is of the form e^{-V} and $\int V d\rho_t$ is easy to differentiate along the flow, which in the end means only when V is quadratic, and thus only when ν is the standard Gaussian. Even Gaussians with other covariance matrices are not included in the result.

As a more recent development, still in collaboration with G. Cozzi (ongoing work), we considered the question whether the functional F can be λ-geodesically convex in W_2, possibly for a negative λ. It turns out that this question is intimately related to the Lipschitz behavior of the vector field $v[\rho, \nu]$. A simple computation shows that, whenever all the measures ν_θ have density bounded from below on

their supports, then we have

$$|Dv[\rho,\nu]| \leq \frac{1}{|x|} * \rho,$$

so that we only need $\rho \in L^p$ for $p > d'$ (since $1/|x|$ almost belongs to L^d; actually, the sharp assumption should be $\rho \in L^{d',1}$) so as to have an L^∞ bound on $|Dv[\rho,\nu]|$. Unfortunately, there are few measures, particularly in high dimensions, with lower bounds on projections. For instance, if one wants ν_θ to be constant, and not only lower bounded, there is a measure proportional to $(1-|x|^2)^{-1/2}$ in dimension $d = 2$, there is the uniform measure on the sphere in dimension $d = 3$, and no other one! We then turned to a more general result, which can be summarized as follows; suppose that ν (and not ν_θ) satisfies suitable lower bounds (for instance, measures bounded from below and above on a smooth and uniformly convex compact subset of $\mathbb{R}^d$ are accepted by the precise assumption), then F is λ-geodesically convex in W_2 on the (geodesically convex) set $\mathcal{M} = \{\rho : \rho \leq M \text{ a.e.}\}$, for a value of λ depending on the bounds on ν and on M.

Among the consequences of this result, considering that, as proven in [5], if $\rho_0 \in L^\infty$ then we can obtain quantitative L^∞ bounds for ρ_t (which, unfortunately, explodes exponentially as $t \to \infty$), we have the existence and - mainly - the uniqueness of the Sliced Wasserstein flow in the class L^∞as well as the rate of convergence of the JKO scheme.

References

[1] F. Pitié, A.C. Kokaram, R. Dahyot *Automated colour grading using colour distribution transfer* Computer Vision and Image Understanding **107** 1-2 (2007), 123–137

[2] A. Tanana *Comparison of transport map generated by heat flow interpolation and the optimal transport Brenier map* Communications in Contemporary Mathematics **23** 6, (2021) 2050025

[3] H. Lavenant, F. Santambrogio *The flow map of the Fokker-Planck equation does not provide optimal transport* Applied Mathematics Letters **133** (2022), 108225

[4] G. Cozzi, F. Santambrogio *Long-time asymptotics of the sliced-Wasserstein flow* SIAM Journal on Imaging Sciences **18** 1 (2025), 1–19

[5] N. Bonnotte *Unidimensional and Evolution Methods for Optimal Transportation* PhD thesis, Université Paris-Sud (2013)

Blind denoising diffusion models and the blessings of dimensionality

Sinho Chewi

(joint work with Z. Kadkhodaie, A.-A. Pooladian, and E. Simoncelli)

We develop a mathematical theory to explain the effectiveness of *blind denoising* in diffusion generative models. Recall that in generative modeling, the goal is to generate samples from a data distribution p_{data} by first evolving the distribution according to a stochastic process, and then learning the reverse transition kernels from data. In a standard setup, we let $p_\sigma = p_{\mathsf{data}} * \mathsf{N}(0, \sigma^2 I)$ denote the convolution of the data distribution with Gaussian noise, and we take the stochastic process to have marginal laws $(p_{\sigma_t})_{t\geq 0}$ for some choice of noise schedule $(\sigma_t)_{t\geq 0}$. Then,

the time reversal of this process can be implemented as a stochastic differential equation (SDE), whose drift coefficient depends on the score functions $\{\nabla \log p_\sigma : \sigma > 0\}$. These score functions are then learned from data (samples from p_{data}) with a neural network architecture through the use of a denoising training loss.

In practice, the neural network is a parametrized function $\widehat{f}\colon \mathbb{R}^d \times [0,\infty) \to \mathbb{R}^d$, which is trained so that $\widehat{f}(x_\sigma, \sigma) \approx x_0$, where $x_0 \sim p_{\mathsf{data}}$ and $x_\sigma = x_0 + \mathsf{N}(0, \sigma^2 I)$. In particular, the neural network takes as an input the noise level $\sigma > 0$, which poses difficulties for implementation: the standard prescription is to first embed the noise level as Fourier features—reminiscent of positional encoding in transformer architectures—and to pre-process these features before feeding them into other layers of the network in an ad hoc fashion.

Recently, in [1], it was shown that high-quality samples can still be generated using blind denoisers—that is, denoisers which do not take σ as an input. In our work [2], we provide a mathematical framework to justify these observations.

Specifically, the algorithm is described as follows.

- During training, one chooses a *prior* distribution Π_0 over $[\sigma_{\min}, \sigma_{\max}]$ and trains $\widehat{f}$ to minimize the loss $n^{-1}\sum_{i=1}^n \mathbb{E}[\|\widehat{f}(x^{(i)} + \sigma z) - x^{(i)}\|^2]$, where $\{x^{(i)}\}_{i=1}^n$ are the data samples and the expectation is taken over $\sigma \sim \Pi_0$ and $z \sim \mathsf{N}(0, I)$ independent.
- During inference, one chooses a *diffusion schedule* $(a_t)_{t\geq 0}$ and runs a discretization of the SDE $dY_t = (\widehat{f}(Y_t) - Y_t)\, dt + \sqrt{2a_t}\, dB_t$, initialized at $Y_0 \sim \mathsf{N}(0, \sigma_0^2 I)$ with $\sigma_0 \gg 1$.

We first show that the minimizer of the population loss can be expressed as $f^\star\colon y \mapsto y + \int \sigma^2 \nabla \log p_\sigma(y)\, \Pi(d\sigma \mid y)$, where $\Pi(\cdot \mid Y)$ is the *posterior* distribution of σ with prior Π_0 and observation $Y \sim p_\sigma$. Then, motivated by empirical evidence, we identify low intrinsic dimensionality as a key condition under which the posterior concentrates around a single point.

Namely, assume that p_{data} is compactly supported, and define the *intrinsic dimension* k of p_{data} to be the logarithm of the covering number of $\mathrm{supp}(p_{\mathsf{data}})$ at a certain small scale $r_0 \ll 1$ (we can take $r_0 \asymp \frac{\sigma_{\min}^2}{\sigma_{\max}\sqrt{d}}$). Then, for any "ground truth" $\sigma_\star \in [\sigma_{\min}, \sigma_{\max}]$ and for a broad class of priors (including power law priors $\Pi_0(\sigma) \propto \sigma^\alpha \mathbf{1}_{\sigma\in[\sigma_{\min},\sigma_{\max}]}$), when the observation Y is drawn from $p_{\sigma_\star}$, then $\int |\lambda(\sigma) - \lambda(\sigma_\star)|^2\, \Pi(d\sigma \mid Y) \ll_{\log} \lambda(\sigma_\star)^2\, (d^{-1} + \mathsf{k}^2 d^{-2})$ with high probability. Here, $\lambda(\sigma) = \sigma^{-2}$ and $\ll_{\log}$ suppresses logarithmic factors. Thus, the noise posterior $\Pi(\cdot \mid Y)$ concentrates on the ground truth $\sigma_\star$ provided that $d \gg_{\log} \mathsf{k}$.

The implication of this result is that under the low intrinsic dimensionality assumption, when the trained denoiser $\widehat{f}$ is close to the population minimizer $f^\star$, the dynamics are well-approximated by the idealized dynamics $dY_t^\star = \sigma_t^2 \nabla \log p_{\sigma_t}(Y_t^\star)\, dt + \sqrt{2a_t}\, dB_t$, where $Y_t^\star \sim p_{\sigma_t}$. In order for this equation to be self-consistent, it implies an ODE for $(\sigma_t)_{t\geq 0}$, which can be solved to yield $\sigma_t^2 = \sigma_0^2 e^{-2t} + 2\int_0^t a_s e^{-2(t-s)}\, ds$. Hence, we find that despite not being given an explicit noise schedule $(\sigma_t)_{t\geq 0}$, blind denoising diffusion models automatically and

implicitly track a noise schedule whose expression can be derived from the choice of diffusion schedule $(a_t)_{t\geq 0}$.

We formalize this reasoning into a precise error bound via Girsanov's theorem, and we also address the error incurred by discretization. In doing so, we show that (1) a particularly good choice of diffusion schedule for mitigating the discretization error is obtained by taking $a_t = \frac{1}{2}\sigma_t^2$, and (2) the discretization error only scales with the intrinsic dimension.

References

[1] Z. Kadkhodaie, E. Simoncelli (2021). *Stochastic solutions for linear inverse problems using the prior implicit in a denoiser*. Advances in Neural Information Processing Systems 34:13242–13254.

[2] Z. Kadkhodaie, A.-A. Pooladian, S. Chewi, E. Simoncelli (2026). *Blind denoising diffusion models and the blessings of dimensionality*. arXiv preprint arXiv:2602.09639.

Laplace Learning Gradient Flows

Matthew Thorpe
(joint work with M. C. A. Oliver and A. Esposito)

The *semi-supervised learning problem* is to find the missing labels from a partially labeled set of feature vectors, $\Omega_n = \{x_i\}_{i=1}^n$, with labels $\{\ell_i\}_{i\in\mathcal{I}_n}$. The set $\mathcal{I}_n \subseteq \{1,\dots,n\}$ indexes the labels, e.g. if $\mathcal{I}_n = \{1,\dots,n\}$ then every feature vector has a label. The objective is to estimate labels for $\{x_i\}_{i\in\{1,\dots,n\}\setminus\mathcal{I}_n}$.

The idea behind *Laplace Learning* is to assume similar feature vectors have similar labels. To define similarity between feature vectors we assume a graph structure where we have weights W_{ij} that represent how similar x_i is to x_j (the larger W_{ij} the more similar the feature vectors). This is represented as a variational problem corresponding to minimizing

$$\mathcal{E}_n^{(p)}(u_n) = \frac{1}{pn^2\varepsilon_n^p}\sum_{i,j=1}^n W_{ij}|u_n(x_i) - u_n(x_j)|^p$$

subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$ and where ε_n is a scaling constant we define shortly.

We can write

$$\mathcal{E}_n^{(p)}(u_n) = \frac{1}{2p}\langle \mathcal{L}_n^{(p)}u_n, u_n\rangle_{\mathrm{L}^p(\mu_n)}$$

where

$$\mathcal{L}_n^{(p)}(u_n)(x_i) = \frac{1}{n\varepsilon_n^p}\sum_{j=1}^n W_{ij}|u_n(x_i) - u_n(x_j)|^{p-2}(u_n(x_i) - u_n(x_j))$$

is the graph p-Laplacian and $\langle u_n, v_n\rangle_{\mathrm{L}^p(\mu_n)} = \frac{1}{n}\sum_{i=1}^n u_n(x_i)v_n(x_i)$ is the inner product with respect to the measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. It follows that the first variation of $\mathcal{E}_n^{(p)}$ is $\partial\mathcal{E}_n^{(p)}(u_n; v_n) = \langle \mathcal{L}_n^{(p)}(u_n), v_n\rangle_{\mathrm{L}^2(\mu_n)}$ and therefore the Fréchet

derivative (or more precisely, the Riesz representation theorem representative of the derivative) is $\mathcal{L}_n^{(p)}(u_n)$.

This talk is about the gradient flow associated with the above variational problem and, in particular, what happens when $n \to \infty$.

The energy dissipation balance formulation of the gradient flow is the equality

$$\mathcal{E}_n^{(p)}(u_n(t)) + \frac{1}{2}\int_0^t \|\mathcal{L}_n^{(p)} u_n(r)\|_{\mathrm{L}^p(\mu_n)}\,\mathrm{d}r + \frac{1}{2}\int_0^t \|\dot{u}_n(r)\|^2_{\mathrm{L}^p(\mu_n)}\,\mathrm{d}r = \mathcal{E}_n^{(p)}(u_n(0)).$$

That is $u_n : [0,T] \to \mathrm{L}^p(\mu_n)$ is a gradient flow if it satisfies the above equation.

To talk about asymptotics we use the following scaling on the edge weights: assuming $x_i \in \mathbb{R}^d$, we define $W_{ij} = \frac{1}{\varepsilon^d}\eta\left(\frac{\|x_i - x_j\|}{\varepsilon}\right)$ and choose $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$. It's an easy calculation to show that, if $u \in \mathrm{C}^2(\mathbb{R}^d)$ and $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(\mathbb{R}^d)$ that (almost surely)

$$\mathcal{E}_n^{(p)}(u\lfloor_{\Omega_n}) \to \mathcal{E}_\infty^{(p)}(u) \coloneqq \sigma_\eta \int_{\mathbb{R}^d} |\nabla u(x)|^p \rho^2(x)\,\mathrm{d}x$$

where ρ is the density of μ and $\sigma_\eta = \int_{\mathbb{R}^d} \eta(\|x\|)|x_1|^p\,\mathrm{d}x$ is a constant. The corresponding energy dissipation balance formulation of the continuum (limiting) gradient flow is

$$\mathcal{E}_\infty^{(p)}(u_\infty(t)) + \frac{1}{2}\int_0^t \|\mathcal{L}_\infty^{(p)} u_\infty(r)\|_{\mathrm{L}^p(\mu)}\,\mathrm{d}r + \frac{1}{2}\int_0^t \|\dot{u}_\infty(r)\|^2_{\mathrm{L}^p(\mu)}\,\mathrm{d}r = \mathcal{E}_\infty^{(p)}(u_\infty(0))$$

where

$$\mathcal{L}_\infty^{(p)}(u)(x) = -\frac{\sigma_\eta}{\rho(x)}\nabla\cdot\left(\rho^2|\nabla u|^{p-2}\nabla u\right)(x).$$

In this talk, we will discuss the convergence of the discrete gradient flow to the continuum one.

Finite-particle Rates for (Regularized) Stein Variational Gradient Descent

Krishnakumar Balasubramanian

(joint work with S. Banerjee, Y. He, and P. Ghosal)

Stein Variational Gradient Descent (SVGD) is a flexible particle-based method for approximate Bayesian inference that evolves an ensemble of particles through a kernelized interaction designed to decrease the Kullback–Leibler divergence to a target distribution. Despite its empirical success, SVGD is known to suffer from an intrinsic finite-particle bias. A regularized variant, regularized SVGD (R-SVGD), was introduced to mitigate this effect by applying a resolvent-type preconditioner to the kernelized Wasserstein gradient. This talk studies both algorithms from a finite-particle perspective and develops explicit, non-asymptotic convergence guarantees for their interacting particle systems, based on [1] and [2].

The analysis is based on an entropy method at the level of the joint particle law. A central ingredient is a differential inequality for the relative entropy between the

law of the full particle system and the product target measure. Its time derivative decomposes into a dominant dissipative term, proportional to the number of particles times the expected squared discrepancy, together with a smaller positive correction. For SVGD, the key control is in terms of the Kernelized Stein Discrepancy, while for R-SVGD the corresponding dissipation is expressed through a regularized Fisher-information-type quantity. This structure makes it possible to convert entropy dissipation into explicit quantitative rates.

As a consequence, one obtains finite-particle convergence bounds of order $N^{-1/2}$ for SVGD and R-SVGD in their natural discrepancy measures, with constants that grow only polynomially in the dimension under mild assumptions on the target potential and kernel. The results also imply Wasserstein convergence, marginal convergence of individual particles, and long-time propagation of chaos for the associated interacting particle systems. Taken together, these estimates provide a rigorous quantitative explanation for the large-particle behavior of both SVGD and its regularized counterpart, and clarify how regularization improves the approximation of the target law.

References

[1] Banerjee, Sayan, Krishnakumar Balasubramanian, and Promit Ghosal. "Improved finite-particle convergence rates for Stein variational gradient descent." arXiv preprint arXiv:2409.08469 (2024).

[2] He, Ye, Krishnakumar Balasubramanian, Sayan Banerjee, and Promit Ghosal. "Finite-Particle Rates for Regularized Stein Variational Gradient Descent." arXiv preprint arXiv:2602.05172 (2026).

From Geometric Hydrodynamics to Periodic Geodesics on Manifolds of Mappings

Levin Maier

Motivation. Since Arnold's interpretation of the Euler equation as the geodesic equation on the group of volume-preserving diffeomorphisms [2], many PDEs from mathematical physics have been understood geometrically as geodesic or Newton-type equations on infinite-dimensional manifolds; see, for example, [1, 6, 11]. This suggests studying the corresponding theory of charged particles in an external magnetic field in infinite dimensions.

In [9] we develop such a framework for strong Riemannian manifolds. A *magnetic system* consists of a Riemannian manifold (M, g) together with a closed 2-form $\sigma \in \Omega^2(M)$, called the *magnetic field*. The associated Lorentz force is the bundle endomorphism $Y \in \mathrm{End}(TM)$ defined by

$$g\big(Y(\xi), \eta\big) = \sigma(\xi, \eta), \qquad \xi, \eta \in TM.$$

The corresponding trajectories, called *magnetic geodesics*, satisfy

$$\nabla_{\dot\gamma}\dot\gamma = Y_\gamma(\dot\gamma).$$

As in the geodesic case, the kinetic energy is conserved.

Magnetic Euler–Arnold equations. If G is an infinite-dimensional Lie group with a right-invariant metric $\mathcal{G}$ and a right-invariant magnetic field σ, then the magnetic geodesic equation can be written in Eulerian variables as a nonlinear evolution equation on the Lie algebra. In [7] this leads to the notion of a *magnetic Euler–Arnold equation*, combining Arnold's hydrodynamical picture [2] with his description of magnetic flows [3]. This framework includes:

- the Korteweg–de Vries equation;
- generalized Camassa–Holm equations;
- the infinite conductivity equation;
- the global quasi-geostrophic equation on $\mathbb{S}^2$;
- the magnetic two-component Hunter–Saxton system [8];
- the magnetic EPDiff equation [10].

Thus several equations from fluid dynamics and dispersive PDE can be interpreted as the motion of a charged particle on an infinite-dimensional manifold in an external magnetic field.

Magnetic Hopf–Rinow theorem. The classical Hopf–Rinow theorem fails in infinite dimensions, so completeness does not automatically imply the existence of minimizing geodesics. A notable positive result of Bauer–Harms–Michor [5] shows that the full Hopf–Rinow theorem holds for half-Lie groups equipped with a right-invariant strong metric. Typical examples are Sobolev diffeomorphism groups

$$\mathrm{Diff}^{H^s}(M), \qquad s > \frac{\dim(M)}{2} + 1.$$

In [9] we prove the following magnetic analogue.

Theorem 1 (Magnetic Hopf–Rinow theorem, [9, Thm. 1.6])**.** *Let $(G, \mathcal{G}, \sigma)$ be a right-invariant weakly exact magnetic system on a half-Lie group G equipped with a strong right-invariant metric $\mathcal{G}$. Then:*

(1) *every magnetic geodesic exists for all time;*
(2) *under suitable regularity and completeness assumptions, there exists a critical value*
$$c(G, \mathcal{G}, d\alpha) \in (0, \infty)$$
such that for every $\kappa > c(G, \mathcal{G}, d\alpha)$ and every $x, y \in G$, there exists an action-minimizing magnetic geodesic of energy κ connecting x and y.

We refer to [9] for the precise meaning of weak exactness and the additional hypotheses. These assumptions are satisfied for Sobolev diffeomorphism groups with right-invariant strong Sobolev metrics.

Periodic geodesics on manifolds of mappings. A further basic question is the existence of periodic geodesics. Apart from a few explicit cases, this problem has remained essentially open in infinite-dimensional geometry. The main advance of the forthcoming work [4] is the following.

Theorem 2 (Periodic geodesic in each homotopy class, [4])**.** *Let $(G, \mathcal{G})$ be a half-Lie group equipped with a strong right-invariant metric $\mathcal{G}$ satisfying the assumptions before (e) in* [5, Thm. 7.7]. *Assume in addition that $\pi_1(G) \neq 0$. Then:*

(1) *for every nontrivial homotopy class* $[\eta] \in \pi_1(G)$*, there exists a closed geodesic* γ *representing* $[\eta]$*;*
(2) *if* $[\eta]$ *is a generator of* $\pi_1(G)$*, then* γ *is embedded.*

In particular, right-invariance implies the existence of infinitely many geometrically distinct periodic geodesics.

The theorem applies, for example, to $\left(\mathrm{Diff}^{H^s}(M), \mathcal{G}^s\right)$ whenever $\pi_1\left(\mathrm{Diff}^{H^s}(M)\right) \neq 0$. For the more restrictive family of $\mathcal{G}^k$ metrics one can strengthen the conclusion:

Proposition 3. [4] *Let* $G \in \left\{\mathrm{Diff}^{H^s}(M), \mathrm{Diff}^{H^s}_{\mathrm{vol}}(M)\right\}$ *be equipped with the* $\mathcal{G}^k$ *metric, where* $0 \leq k \leq s$*. Then:*

(1) *if* $\pi_1(\mathrm{Isom}(M,g)) \neq 0$*, there exist infinitely many periodic geodesics of* $(G, \mathcal{G}^k)$*;*
(2) *if* $\pi_q(\mathrm{Isom}(M,g)) \neq 0$ *for some* $q \geq 2$*, there exist infinitely many contractible periodic geodesics of* $(G, \mathcal{G}^k)$*.*

Consequently, if $\pi_q(\mathrm{Isom}(M,g)) \neq 0$ *for some* $q \geq 1$*, then for* $k = 0$ *there exist infinitely many periodic geodesics of* $(\mathrm{Diff}^{H^s}_{\mathrm{vol}}(M), \mathcal{G}^{L^2})$*.*

References

[1] V. Arnold and B. Khesin, *Topological Methods in Hydrodynamics*, Applied Mathematical Sciences, vol. 125, Springer, New York, 1998.
[2] V. I. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Ann. Inst. Fourier (Grenoble) **16** (1966), 319–361.
[3] V. I. Arnold, *Some remarks on flows of line elements and frames*, Soviet Math. Dokl. **2** (1961), 562–564.
[4] M. Bauer and L. Maier, *On Periodic Geodesics on Manifolds of Mappings*, in preparation.
[5] M. Bauer, P. Harms, and P. W. Michor, *Regularity and completeness of half–Lie groups*, Journal of the European Mathematical Society, 2025. doi: 10.4171/jems/1587.
[6] B. Khesin, G. Misiołek, and K. Modin, *Infinite-dimensional Newton's equations and geometric hydrodynamics*, Bull. Amer. Math. Soc. **58** (2021), 377–442.
[7] L. Maier, *On geometric hydrodynamics and infinite dimensional magnetic systems*, arXiv:2506.00544.
[8] L. Maier, *On Mañé's critical value for the two-component Hunter–Saxton system and an infinite-dimensional magnetic Hopf–Rinow theorem*, arXiv:2503.12901.
[9] L. Maier and F. Ruscelli, *The Hopf–Rinow Theorem and Mañé's Critical Value for Magnetic Geodesics on Half Lie groups*, arXiv:2510.19323.
[10] L. Maier and F. Ruscelli, *On Mañé's Critical Value for Tonelli Lagrangians on Half Lie groups*, arXiv:2511.13428.
[11] C. Vizman, *Geodesic equations on diffeomorphism groups*, SIGMA Symmetry Integrability Geom. Methods Appl. **4** (2008), Paper 030.

Optimal incompressible collective diffusion: a relaxation approach

Yann Brenier
(joint work with B. Geshkovski)

We consider phases indexed by $a \in \mathcal{A}$, with concentrations $c = c(t,x,a) \geq 0$ on $[0,T] \times \mathbb{T}^d \times \mathcal{A}$, where $\mathcal{A}$ is finite or continuous. The total density is constrained by $\int_{\mathcal{A}} c(t,x,a)\, da = 1$, so the phases rearrange while their sum stays pointwise constant. A common symmetric matrix field $B(t,x) \in \mathbb{R}^{d\times d}_{\rm sym}$, assumed row-wise divergence free, acts on every phase through

$$\partial_t c = \nabla_x \cdot (B \nabla_x c) = \nabla_x^2 : (Bc) = B : \nabla_x^2 c.$$

This may be viewed as an incompressible collective diffusion. It also appears as a large-β limit of continuum self-attention dynamics in which all data points $a \in \mathcal{A} = \{1, \cdots, N\}$ share the same attention field [5, 3].

We first consider the least-action problem

$$(1)\ \inf_{B,c}\Big\{ \int_0^T \int_{\mathbb{T}^d} \frac{|B|^2}{2}\, dxdt\ ;\ \nabla_x \cdot B = 0,\ \partial_t c = \nabla_x^2 : (Bc),\ c_{|t=0} = c_0,\ c_{|t=T} = c_T \Big\},$$

where the diffusion constraint holds for every $a \in \mathcal{A}$. This problem is nonconvex because of the product Bc. Introducing a scalar multiplier $\zeta = \zeta(t,x,a)$ and a vector multiplier $Q = Q(t,x)$, one formally obtains

$$(\partial_t + B : \nabla_x^2)\zeta = 0, \qquad \partial_t c = \nabla_x^2 : (Bc),$$

$$B = \nabla_{\rm sym} Q + \int_{\mathcal{A}} c\, \nabla_x^2 \zeta\, da, \qquad \nabla_x \cdot B = 0.$$

Since the optimal B need not be positive, the equation in ζ is of mixed type rather than a classical forward parabolic one.

The relaxation consists in replacing the nonlinear product Bc by a microscopic second-order momentum $b = b(t,x,a) \in \mathbb{R}^{d\times d}_{\rm sym}$, writing $B(t,x) = \int_{\mathcal{A}} b(t,x,a)\, da$, and requiring only $\nabla_x \cdot \int_{\mathcal{A}} b\, da = 0$. For a prescribed weight $w(a) \geq 0$, we substitute the macroscopic cost by the convex action

$$\int_0^T \int_{\mathbb{T}^d} \int_{\mathcal{A}} \frac{w(a)|b(t,x,a)|^2}{2c(t,x,a)}\, da\, dx\, dt,$$

which agrees with (1) whenever $b = Bc$ and $\int_{\mathcal{A}} c\, da = 1$. This leads to the relaxed convex problem

$$(2)\ \inf_{c,b}\Big\{ \int_0^T \int_{\mathbb{T}^d} \int_{\mathcal{A}} \frac{w(a)|b|^2}{2c}\ ;\ \partial_t c = \nabla_x^2 : b,\ \nabla_x \cdot \int_{\mathcal{A}} b\, da = 0,\ c_{|t=0} = c_0,\ c_{|t=T} = c_T \Big\}.$$

It is the analogue, for incompressible collective diffusion, of Brenier's relaxation of Arnold's least-action principle and of the multiphase transport framework of Brenier–Puel [1, 2].

By Fenchel–Rockafellar duality one finds the concave dual problem

$$\sup_{\zeta,Q}\Big\{\mathcal{L}(\zeta)\ ;\ \frac{1}{2}\big|\nabla_x^2\zeta(t,x,a)+\nabla_{\mathrm{sym}}Q(t,x)\big|^2+w(a)\,\partial_t\zeta(t,x,a)\le 0\Big\},\tag{3}$$

where

$$\mathcal{L}(\zeta)=\int_{\mathbb{T}^d\times\mathcal{A}} c_T(x,a)\zeta(T,x,a)-c_0(x,a)\zeta(0,x,a)\,dx\,da.$$

Hence the only nonlinearity is encoded in a 2nd order Hamilton–Jacobi type inequality.

Several special cases are instructive. If $\mathcal{A}=\{0,1\}$, $w(0)=1$, and $w(1)=0$, then incompressibility can be eliminated and one recovers the monophasic problem

$$\inf_{c,b}\Big\{\int_0^T\int_{\mathbb{T}^d}\frac{|b|^2}{2c}\ ;\ \partial_t c=\nabla_x^2:b,\ c_{|t=0}=c_0,\ c_{|t=T}=c_T\Big\}.\tag{4}$$

This is close in spirit to the Benamou–Brenier formulation of martingale optimal transport [6, 7]; in contrast with that setting, no convex-order condition is imposed on the endpoints. The formal optimality equations are

$$\partial_t c=\nabla_x^2:(c\beta),\qquad \beta=\nabla_x^2\zeta,\qquad \partial_t\beta+\nabla_x^2\Big(\frac{1}{2}|\beta|^2\Big)=0,$$

a matrix-valued porous-medium type equation whose spectrum is not constrained to be nonnegative. In one space dimension, the equal-weight two-phase problem reduces to

$$\inf\Big\{\int_0^T\int_{\mathbb{T}}\frac{b^2}{2c(1-c)}\ ;\ \partial_t c=\partial_{xx}b,\ c_{|t=0}=c_0,\ c_{|t=T}=c_T\Big\},$$

which is reminiscent of optimal transport with nonlinear mobility [4].

The slides also suggest an AGS-type gradient-flow interpretation of (4). Introducing an extra evolution variable $s\ge 0$ and a tensor $\Phi=\Phi(s,t,x)$ satisfying $\partial_s c=\nabla_x^2:\Phi$ and $\partial_s b=\partial_t\Phi$, with $\beta=b/c$, one gets formally

$$\frac{d}{ds}\int\frac{|b|^2}{2c}=-\int\Phi:\Big(\partial_t\beta+\nabla_x^2\Big(\frac{1}{2}|\beta|^2\Big)\Big).$$

Completing the square suggests the constitutive choice $\Phi=-c\big(\partial_t\beta+\nabla_x^2(\frac{1}{2}|\beta|^2)\big)$ and an associated evolution variational inequality. This provides a plausible metric-flow mechanism for the relaxed second-order action.

The main point is that the original least-action problem for incompressible collective diffusion is naturally nonconvex and only formally parabolic, whereas the relaxed formulation (2) is convex, has the dual description (3), and connects with generalized Eulerian, martingale, and mobility-based transport models.

References

[1] Y. Brenier, *Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations*, Comm. Pure Appl. Math. **52** (1999), 411–452.

[2] Y. Brenier and M. Puel, *Optimal multiphase transportation with prescribed momentum*, ESAIM Control Optim. Calc. Var. **8** (2002), 287–343.

[3] G. Bruno, F. Pasqualotto, and A. Agazzi, "A multiscale analysis of mean-field transformers in the moderate interaction regime," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[4] J. Dolbeault, B. Nazaret, and G. Savaré, *A new class of transport distances between measures*, Calc. Var. Partial Differential Equations **34** (2009), 193–231.

[5] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, Bull. Amer. Math. Soc. **62** (2025), 427–479.

[6] M. Huesmann and D. Trevisan, *A Benamou–Brenier formulation of martingale optimal transport*, Bernoulli **25** (2019), 2729–2757.

[7] D. Matthes, E.-M. Rott, and A. Schlichting, *Diffusive transport on the real line: semi-contractive gradient flows and their discretization*, arXiv:2501.14527.

A multiscale analysis of mean-field transformers in the moderate interaction regime

Giuseppe Bruno

(joint work with F. Pasqualotto and A. Agazzi)

The mean-field viewpoint on transformers introduced in [1, 2, 3] models tokens as interacting particles evolving through depth. In this framework, self-attention gives rise, in the large-N limit, to a transport equation on probability measures, and clustering phenomena can be studied with PDE and interacting-particle methods. We consider this dynamics on the sphere, induced by LayerNorm, in the *moderate interaction regime*, where both the number N of tokens and the inverse temperature $\beta = \beta_N$ tend to infinity, with $\beta_N \to \infty$ sufficiently slowly with respect to N. This scaling is motivated by long-context transformer regimes and leads to a genuinely multiscale asymptotic behavior.

More precisely, we study the continuous-depth self-attention system

$$\dot{x}_i(t) = P_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{N} e^{\beta \langle Q x_i(t), K x_j(t) \rangle} V x_j(t) \right), \qquad x_i(t) \in \mathbb{S}^{d-1},$$

with

$$Z_{\beta,i}(t) = \sum_{j=1}^{N} e^{\beta \langle Q x_i(t), K x_j(t) \rangle}.$$

Passing to the empirical measure $\mu_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i(t)}$, the large-$N$ limit is described by the continuity equation

$$\partial_t \mu + \operatorname{div}(\chi_\beta[\mu]\mu) = 0,$$

where

$$\chi_\beta[\mu](x) = P_x\left(\frac{\int_{\mathbb{S}^{d-1}} e^{\beta\langle Qx,Ky\rangle} Vy\, d\mu(y)}{\int_{\mathbb{S}^{d-1}} e^{\beta\langle Qx,Ky\rangle}\, d\mu(y)}\right).$$

A Dobrushin-type estimate shows that, on the relevant β-dependent timescales, the particle system is asymptotically described by this PDE, so the joint limit $N, \beta_N \to \infty$ reduces to the local limit $\beta \to \infty$ for the mean-field equation.

Our main result is that this regime exhibits three distinct dynamical phases.

The first one is the *alignment phase*, on timescales $O(1)$. Under mild assumptions on the parameters and on the initial density, we prove convergence to the limiting transport equation

$$\partial_t \mu = -\operatorname{div}\left(\mu\, P_x \frac{VK^\top Qx}{|K^\top Qx|}\right).$$

Hence, at leading order, the nonlinear interaction disappears and the dynamics is governed by an effective drift induced by the matrix $VK^\top Q$. As a consequence, the support of the limiting measure collapses onto the generalized eigenspace $E_{\max}$ associated with the eigenvalue of $VK^\top Q$ having maximal real part. This identifies a fast mechanism by which the token distribution concentrates onto a lower-dimensional subspace.

The second one is the *heat phase*, on timescales $O(\beta)$. Assuming that the dynamics has already concentrated on $E_{\max} \cap \mathbb{S}^{d-1}$, and imposing

$$Q^\top K|_{E_{\max}} = \lambda_1 I, \qquad V|_{E_{\max}} = \pm\lambda_2 I, \qquad \lambda_1, \lambda_2 > 0,$$

the leading-order drift vanishes and the next-order term becomes effective. After the rescaling $dt = \beta\, ds$, the limiting equation can be characterized as

$$\partial_s \mu = -\gamma\Delta\mu, \qquad \gamma = \pm 1,$$

on the lower-dimensional sphere $E_{\max} \cap \mathbb{S}^{d-1}$. If $\gamma < 0$, one obtains a forward heat equation; if $\gamma > 0$, one obtains a backward diffusion, leading to concentration and metastable clustering. Thus the second phase describes either smoothing or cluster formation inside the aligned subspace.

The third one is the *pairing phase*, on exponentially long timescales in β. Starting from atomic data

$$\mu_0 = \sum_{j=1}^m \alpha_j \delta_{x_j}, \qquad \alpha_j > 0, \qquad \sum_{j=1}^m \alpha_j = 1,$$

the interaction between clusters is exponentially weak, and the dominant contribution comes from the closest pair. After the rescaling

$$dt = e^{\beta(1-\langle x_i, x_j\rangle)}\, ds,$$

one recovers a finite-dimensional limiting dynamics in which only this pair moves, along the geodesic joining the two clusters, while the other ones remain frozen at leading order. This yields a sequential merging mechanism for the late-time dynamics.

In conclusion, in the moderate interaction regime, the transformer dynamics admits a multiscale description: rapid alignment onto a lower-dimensional subspace, heat-type evolution within that subspace, and exponentially slow pairwise merging of the resulting clusters. This provides a unified asymptotic picture for several phenomena previously observed separately in the literature. The later phases, however, still require stronger assumptions, and a full rigorous connection between them remains open.

References

[1] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *The emergence of clusters in self-attention dynamics*, in *Advances in Neural Information Processing Systems* 36 (2024).
[2] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, *Bulletin of the American Mathematical Society* 62 (2025), no. 3, 427–479.
[3] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, *Sinkformers: Transformers with doubly stochastic attention*, in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, PMLR 151 (2022), 3515–3530.

(Stochastic) Flows of Measures in Deep Transformers

Andrea Agazzi
(joint work with G. Bruno, E. M. Garcia, S. Saviozzi, and M. Romito)

Recent mathematical work has proposed a mean-field description of information flow in deep transformer architectures by interpreting the residual stream as a system of interacting particles [6, 8, 9]. In this picture, the tokens (e.g., words in the input sentence of the LLM) are viewed as particles moving on the sphere, which captures the effect of post-layer normalization, and the depth of the network plays the role of time. Passing to the large-depth limit turns the layerwise evolution into a continuous dynamics, while the large-token limit yields an evolution equation for the empirical distribution of tokens. This perspective translates a central mechanism of modern deep learning into the language of interacting particle systems and kinetic equations: rather than following each token separately, one studies how the distribution of tokens evolves under the collective action of self-attention.

From this viewpoint, self-attention induces a nonlinear transport of the token distribution. The interaction is of mean-field type, since each token, interpreted as a particle, is influenced by the empirical distribution of all the others through an attention kernel. In the spherical models introduced in [6, 8], the resulting dynamics can be interpreted as a Wasserstein gradient flow. The corresponding energy functional is minimized by fully synchronized configurations, namely by measures concentrated at a single point. In this sense, the model extends the classical synchronization mechanisms of Kuramoto-type systems.

At the same time, the long-time behavior of these deterministic mean-field transformer models is subtle. Although fully synchronized states are global minimizers of the energy, the landscape also contains saddle-type critical points corresponding to partially clustered configurations. These states can slow down the dynamics and generate metastable behavior, characterized, for example, in [1, 7].

As a consequence, establishing robust convergence to global minimizers for general initial data remains difficult. In particular, the presence of saddle points in the energy landscape prevents one from proving global exponential convergence estimates by a direct dissipation argument. Recent quantitative results in regimes where this obstruction can be controlled were obtained in [2].

In contrast to the deterministic picture discussed above, a stochastic picture emerges when considering the effect on the residual stream dynamics of the Multi-Layer Perceptron (MLP) layer at initialization. In the scaling regime we consider, the contribution of this layer does not disappear in the large-depth limit. Instead, after passing to the wide limit, it produces a common stochastic forcing acting simultaneously on all tokens. More precisely, in the limit, the MLP layer acts on the tokens as a Gaussian random field, with a covariance structure determined by the choice of activation function. Thus, the token dynamics converges in the deep and wide limit to a stochastic interacting particle system on the sphere, driven by a common noise reflecting the shared action of this random layer across all tokens.

At the level of empirical measures, this leads to a stochastic partial differential equation describing the evolution of the token distribution. The corresponding propagation-of-chaos statement follows the general roadmap developed for McKean–Vlasov systems with environmental noise [3]. This yields a natural stochastic counterpart of the deterministic mean-field transformer flow and provides, to our knowledge, one of the first rigorous noisy models for residual stream dynamics in deep transformers, in which the noise emerges directly from the architecture.

We then study the effect of stochasticity on the clustering phenomenon identified in the deterministic system. Under suitable nondegeneracy assumptions on the covariance induced by the activation, the stochastic dynamics forces trajectories to converge toward a random point attractor, in a way analogous to weak synchronization results for isotropic stochastic flows [4, 5]. In particular, one can prove exponential dissipation estimates in expectation for the deterministic attention energy. A notable consequence is that this noise-induced synchronization may persist even when the deterministic self-attention component is not purely attractive: sufficiently strong common noise can overcome repulsive effects in the attention parameters and still drive the system toward synchronization.

The picture that emerges is that deterministic and stochastic measure flows in deep transformers capture complementary aspects of the same architecture. The deterministic model explains how collective token dynamics can create clustered representations, but it also exhibits metastable partially synchronized states. The stochastic model, arising from the random MLP layer at initialization, suggests that common noise may regularize this landscape and restore quantitative convergence to synchronized states.

References

[1] G. Bruno, F. Pasqualotto, and A. Agazzi, *Emergence of meta-stable clustering in mean-field transformer models*, ICLR 2025; also available as arXiv:2410.23228.

[2] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet, *Quantitative Clustering in Mean-Field Transformer Models*, arXiv:2504.14697, 2025.
[3] M. Coghi and F. Flandoli, *Propagation of chaos for interacting particles subject to environmental noise*, Ann. Appl. Probab. **26** (2016), no. 3, 1407–1442.
[4] M. Cranston, B. Gess, and M. Scheutzow, *Weak synchronization for isotropic flows*, Discrete Contin. Dyn. Syst. Ser. B **21** (2016), no. 9, 3003–3014.
[5] M. Engel and A. Shalova, *Random Quadratic Form on a Sphere: Synchronization by Common Noise*, arXiv:2603.06187, 2026.
[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, Bull. Amer. Math. Soc. **62** (2025), no. 3, 353–417.
[7] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet, *Dynamic metastability in the self-attention model*, arXiv:2410.06833, 2024.
[8] P. Rigollet, *The Mean-Field Dynamics of Transformers*, arXiv:2512.01868, 2026.
[9] M. Sander, P. Ablin, M. Blondel, and G. Peyré, *Sinkformers: Transformers with doubly stochastic attention*. International Conference on Artificial Intelligence and Statistics (2022)

Lost in the Middle through Glauber Calculus

Borjan Geshkovski
(joint work with M. Duerinckx and S. Rossi)

Large language models based on causal transformers are routinely deployed in regimes where the available context is much longer than the part of the prompt that actually matters for the final prediction. A striking empirical observation is the so-called *lost-in-the-middle* effect: retrieval is typically best when the relevant information is placed near the beginning or the end of the context, and substantially worse when it is placed in the middle [6]. The aim of this work is to give a rigorous mathematical explanation of this phenomenon in a minimal continuous-time model for causal self-attention with recency bias.

Our starting point is the interacting particle system viewpoint on transformers developed in [3, 4, 5]. We consider a decoder-type dynamics on the circle $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, where the jth token is represented by an angle $\theta_j(t)$ and only attends to its predecessors. After reducing the original spherical model to dimension 2, and replacing the random softmax denominator by a deterministic normalization, we arrive at

$$\dot{\theta}_j(t) = \sum_{k=1}^{j} \frac{e^{-\frac{\lambda}{N}(j-k)}}{\sum_{\ell=1}^{j} e^{-\frac{\lambda}{N}(j-\ell)}} W'_\beta(\theta_j(t) - \theta_k(t)), \qquad W_\beta(x) = e^{\beta \cos x}.$$

The parameter $\beta > 0$ plays the role of inverse temperature, while $\lambda \in \mathbb{R}$ encodes an ALiBi-type positional bias toward recent tokens [7]. This model keeps the three mechanisms relevant for the phenomenon under study: one-sided causality, exponentially decaying memory, and nonlinear interaction through the kernel W'_β.

To probe memory, we fix a vocabulary of size M identified with the grid $\{2\pi m/M : 1 \le m \le M\} \subset \mathbb{T}$. For a distinguished site $i_* = \lfloor \sigma_0 N \rfloor$, with $\sigma_0 \in (0,1)$, we ask whether the last token $\theta_N(t)$ at time t decodes to the initial token at position i_*. Writing $|\cdot|_{\mathbb{T}}$ for the distance on the torus, the corresponding

accuracy is

$$\mathrm{Acc}_N(t,\sigma_0) \coloneqq \mathbb{P}\Big(|\theta_N(t)-\theta_{i_*}(0)|_{\mathbb{T}} \le \frac{\pi}{M}\Big).$$

A smoothed version of this observable admits the Fourier expansion

$$\mathrm{Acc}_N(t,\sigma_0) = \frac{\sqrt{\pi/2}}{M} + \frac{\sqrt{\pi/2}}{M}\sum_{n\neq 0} e^{-\frac{\pi^2 n^2}{2M^2}}\,\mathbb{E}\big[e^{in(\theta_N(t)-\theta_{i_*}(0))}\big],$$

so the whole question is reduced to understanding space-time correlations between the last particle at time t and the particle initially located at macroscopic position σ_0.

The first step is a mean-field description. Introducing the empirical measure

$$\mu_N(t) \coloneqq \frac{1}{N}\sum_{j=1}^{N} \delta_{(j/N,\theta_j(t))} \in \mathcal{P}((0,1)\times\mathbb{T}),$$

we prove that, for independent initial data with profile $f_\circ(\sigma,\theta)$, the sequence $\mu_N(t)$ converges to the unique weak solution of the causal kinetic equation

$$\partial_t f + \partial_\theta\left(f\int_0^\sigma Q_\lambda(\sigma,\sigma')\,(W'_\beta *_\theta f)(t,\sigma',\theta)\,d\sigma'\right) = 0,$$

with kernel

$$Q_\lambda(\sigma,\sigma') = \frac{\lambda e^{-\lambda(\sigma-\sigma')}}{1-e^{-\lambda\sigma}}\,\mathbf{1}_{\sigma'\le\sigma}, \qquad Q_0(\sigma,\sigma') = \frac{1}{\sigma}\,\mathbf{1}_{\sigma'\le\sigma}.$$

This limit already exhibits the main structural feature of the model: exchangeability is broken by causality, and the macroscopic variable σ records how much of the past is accessible to a given token. We also obtain a quantitative estimate on observables of $\mu_N(t)$, with error of order $N^{-2\delta\wedge 1}$ under suitable regularity assumptions on the initial Fourier modes.

Mean field, however, is not sufficient for the accuracy, because the relevant information is stored in correlations of order N^{-1}. We therefore analyze the next order in the propagation of chaos. For a smooth observable ϕ of the initial token, we introduce an autocorrelation field A^N_ϕ and a cross-correlation field C^N_ϕ, the latter encoding the covariance between a token at position σ and the initial token at position $\sigma_0<\sigma$. Our second main result shows that

$$A^N_\phi \to A_\phi, \qquad NC^N_\phi \to C_\phi,$$

locally uniformly in (t,σ,σ_0), with explicit polynomial control in the Fourier variable. The limit A_ϕ is transported by the mean-field flow, whereas C_ϕ solves a linear inhomogeneous equation forced at the source point σ_0:

$$\partial_t C_\phi + \partial_\theta\big(C_\phi\,\mathcal{V}[f] + f\,\mathcal{V}[C_\phi] + f\,Q_\lambda(\sigma,\sigma_0)(W'_\beta *_\theta A_\phi)\big) = 0, \qquad C_\phi(0)=0,$$

where $\mathcal{V}[g](t,\sigma,\theta) \coloneqq \int_0^\sigma Q_\lambda(\sigma,\sigma')(W'_\beta *_\theta g)(t,\sigma',\theta)\,d\sigma'$. This linearized equation is the continuum counterpart of the causal transport of information along the prompt.

The proof of this fluctuation result uses a non-hierarchical cumulant expansion adapted to the triangular structure of the dynamics. When differentiating a

covariance in time, one encounters third cumulants rather than closed two-point quantities. To control them, we use Glauber calculus with respect to the initial data, following the philosophy of [2, 1]. Causality implies that the Glauber derivative $D_k^\circ \theta_i(t)$ vanishes whenever $k > i$, and quantitative bounds on first and second Glauber derivatives then yield a sharp estimate on the third-cumulant remainder. This gives the desired truncation at the level of two-point correlations without resorting to a BBGKY hierarchy.

A particularly transparent regime is obtained for spatially homogeneous input data, namely $f_\circ \equiv 1$ on $\mathbb{T}$. In that case, the mean-field solution remains uniform, the autocorrelation is explicit, and the cross-correlation diagonalizes in Fourier:

$$\widehat{C}_\phi(t, \sigma, n; \sigma_0) = \widehat{\phi}(n)\, \widehat{G}_n(t, \sigma; \sigma_0).$$

The profile $\widehat{G}_n$ solves a Volterra–Hardy type equation that can be reduced, by a change of variables, to the Goursat problem $\partial_t \partial_y U - a_n U = 0$ with $a_n = n^2 I_n(\beta)$, where I_n denotes the modified Bessel function. This yields an explicit representation of $\widehat{G}_n$ in terms of I_1, and therefore an asymptotic expansion for the smoothed accuracy:

$$\mathrm{Acc}_N(t, \sigma_0) = \frac{\sqrt{\pi/2}}{M} + \frac{\sqrt{2\pi}}{MN} \sum_{n \geq 1} e^{-\frac{\pi^2 n^2}{2M^2}}\, \widehat{G}_n(t, 1; \sigma_0) + O(N^{-1-\delta'}),$$

uniformly for $\sigma_0 \in [\varepsilon, 1]$.

The problem is thus reduced to the profile

$$S(\sigma_0) \coloneqq \sum_{n \geq 1} e^{-\frac{\pi^2 n^2}{2M^2}}\, \widehat{G}_n(t, 1; \sigma_0).$$

Our final theorem shows that for every $\lambda > 0$ one has $S(\sigma_0) \to +\infty$ as $\sigma_0 \downarrow 0$. Moreover, if

$$t \sup_{n \geq 1} n^2 I_n(\beta) \leq \min\Big\{3 - \sqrt{3},\, 2(1 - e^{-\lambda})\Big\},$$

then $S'(1^-) > 0$ and S has a unique global minimum in $(0, 1)$. In other words, the retrieval profile is rigorously U-shaped: the earliest positions are privileged by repeated causal aggregation, the latest positions by recency, and the middle positions by neither. This provides a first derivation of a lost-in-the-middle law from a continuum interacting-particle model.

Beyond this theorem, the analysis shows that architectural position bias can already be read off from a deterministic kinetic description and its $1/N$ correlation corrections, before introducing training or data-dependent effects.

References

[1] M. Duerinckx, A. Gloria, and F. Otto, The structure of fluctuations in stochastic homogenization, *Communications in Mathematical Physics* **377** (2020), no. 1, 259–306.

[2] M. Duerinckx, On the size of chaos via Glauber calculus in the classical mean-field dynamics, *Communications in Mathematical Physics* **382** (2021), no. 1, 613–653.

[3] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, A mathematical perspective on transformers, *Bulletin of the American Mathematical Society* **62** (2025), no. 3, 427–479.

[4] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, The emergence of clusters in self-attention dynamics, *Advances in Neural Information Processing Systems* **36** (2023), 57026–57037.
[5] H. Koubbi, B. Geshkovski, and P. Rigollet, Homogenized Transformers, *arXiv preprint* arXiv:2604.01978, 2026.
[6] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, *Lost in the Middle: How Language Models Use Long Contexts*, Trans. Assoc. Comput. Linguistics **12** (2024), 157–173.
[7] O. Press, N. A. Smith, and M. Lewis, *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*, Proc. ICLR (2022).

Continuous transformations of probability distributions and their transport representations

Hugo Lavenant
(joint work with G. Savaré)

1. The question

For X and Y two Polish spaces, we consider $F\colon \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ transforming a probability distribution of X into a probability distribution over Y. We are interested in transport representations: can we find $f\colon \mathsf{X} \times \mathcal{P}(\mathsf{X}) \to \mathsf{Y}$ such that:

$$F(\mu) = f(\cdot, \mu)_\sharp \mu \qquad \text{for every } \mu \in \mathcal{P}(\mathsf{X}), \tag{1}$$

where $g_\sharp \mu$ is the push-forward of the measure μ by the map g. That is, we want $F(\mu)$ to be the push-forward of μ by a transport map, but the map $f(\cdot, \mu)$ is allowed to depend on μ.

While it is clear that, if a transport representation f exists, then F should inherit the regularity of f, in this talk we discuss the reverse problem: if F is given, can we find a f such that (1) holds, and how regular can f be found?

There is an obvious obstruction: not every map F admits a transport representative. On $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$, if $\gamma\colon \mathbb{R}^d \to \mathbb{R}_+$ is a continuous non-negative function which integrates to 1 then $F\colon \mu \mapsto \mu * \gamma$ cannot be represented as in (1), as $F(\delta_x) = \gamma(y-x)dy$, and no transport map can push a Dirac mass onto a measure having a density.

2. Motivation

This question was proposed to us by Gabriel Peyré. The motivation is in-context learning, as the recent work [1] showed that any jointly continuous function $f\colon \mathsf{X} \times \mathcal{P}(\mathsf{X}) \to \mathsf{Y}$ can be approximated by a transformer architecture. Thus, any continuous transformation F of probability distributions which can be written as in (1) with f continuous can be approximated by a transformer architecture. This can be of importance, as functions $F\colon \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ can correspond to very complex objects, while transformers architecture are now widely used.

3. Results

Our results hold for maps F which are defined on the whole set of probability measures $\mathcal{P}(\mathsf{X})$, sometimes with a restriction on the moments. It is crucial that F is defined on discrete measures, they are key in our analysis.

If $F\colon \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ has a transport representative, then necessarily we have a non-splitting property: tokens do not split, that is, a (finite) combination of Dirac masses cannot be split.

Specifically we say that $F\colon \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ is non-splitting on empirical measures if: when $x_1, \dots, x_m$ are distinct elements in X and $\mu = \sum_{i=1}^m a_i \delta_{x_i} \in \mathcal{P}(\mathsf{X})$ with $a_i \in \mathbb{Q}_+$ for all i, then there exists $y_1, \dots, y_m \in \mathsf{Y}$ (not necessarily distinct) such that

$$F(\mu) = F\left(\sum_{i=1}^m a_i \delta_{x_i}\right) = \sum_{i=1}^m a_i \delta_{y_i}.$$

Equivalently, for any empirical measure $\mu = \sum_{i=1}^m a_i \delta_{x_i}$, we assume that there exists $f\colon \mathsf{X} \to \mathsf{Y}$ such that $F(\mu) = f_\sharp \mu$.

We also need a topological setting. We assume that X, Y are Polish spaces endowed with their Borel σ-algebra. The spaces $\mathcal{P}(\mathsf{X})$ and $\mathcal{P}(\mathsf{Y})$ are endowed with the topology of narrow convergence, generated by the duality pairing with bounded continuous functions, and the associated Borel σ-algebra. In this context, our first main result reads as follows.

Theorem. *Assume $F\colon \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ is non-splitting on empirical measure and continuous. Then there exists $f\colon \mathsf{X} \times \mathcal{P}(\mathsf{X}) \to \mathsf{Y}$ measurable such that $F(\mu) = f(\cdot, \mu)_\sharp \mu$ for all $\mu \in \mathcal{P}(\mathsf{X})$.*

There are two notable consequences: that non-splitting on empirical measures guarantees existence of the transport representative f on *all* measures; and that it can be chosen in a measurable way. However, the transport representative $f(x, \mu)$ may not depend continuously on μ, or may not depend continuously on x.

Remarkably, we show that if we assume a more quantitative version of continuity, we can obtain much stronger consequences. For an exponent $p \geq 1$, we denote by W_p the p-Wasserstein distance, it is defined on $\mathcal{P}_p(\mathsf{X})$ and $\mathcal{P}_p(\mathsf{Y})$ the space of probability distributions with finite p moments.

Theorem. *For X a geodesic space, assume $F\colon \mathcal{P}_p(\mathsf{X}) \to \mathcal{P}_p(\mathsf{Y})$ is non-splitting on empirical measures and Lipschitz continuous for W_p. Then there exists $f\colon \mathsf{X} \times \mathcal{P}(\mathsf{X}) \to \mathsf{Y}$ measurable such that $F(\mu) = f(\cdot, \mu)_\sharp \mu$ for all $\mu \in \mathcal{P}(\mathsf{X})$, and which is jointly continuous in the sense $f(x_n, \mu_n) \to f(x, \mu)$ in Y if $x_n \to x$ in X and $\mu_n \to \mu$ in $\mathcal{P}_p(\mathsf{X})$ and x is in the support of μ.*

Compared to the previous result, we obtain a transport representative $f(x, \mu)$ which is continuous in both x and μ. While continuity in μ could be expected, continuity in x is more surprising as we did not assume it directly with the non-splitting property.

4. An example of a continuous F but discontinuous f

To justify that F being continuous is not enough to guarantee continuity of the transport representative, we consider the following example.

We take $\mathsf{X} = \mathsf{Y} = [0,1]$ and denote by λ the Lebesgue measure on $[0,1]$. Let $g\colon \mathbb{R} \to [0,1]$ be a 1-periodic and Lipschitz function such that $g_\sharp \lambda = \lambda$, for instance $g(x) = 2\min(x, 1-x)$. We also fix $\alpha \in (0,1)$. With W_p the p-Wasserstein distance, we define

$$f(x,\mu) = \begin{cases} g\left(\dfrac{x}{W_p(\mu,\lambda)^\alpha}\right) & \text{if } \mu \neq \lambda, \\ x & \text{if } \mu = \lambda. \end{cases}$$

We also define $F(\mu) = f(\cdot,\mu)_\sharp \mu$, in particular $F(\lambda) = \lambda$.

Then in this case we claim that we can prove that F is continuous, however that it cannot have a continuous transport representative. The transport representative is discontinuous for $\mu = \lambda$, and the whole technical aspect of the analysis of this example is showing that F is still continuous at $\mu = \lambda$.

References

[1] T. Furuya, M. V. de Hoop, and G. Peyré, Transformers are universal in-context learners, *arXiv preprint arXiv:2408.01367*, (2024).

Formation of clusters and coarsening in weakly interacting diffusions

André Schlichting

(joint work with N. J. Gerber, R. S. Gvalani, M. Hairer, and G. A. Pavliotis)

We study the formation and long-time evolution of clusters in a system of N interacting diffusions on the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. The key parameter regime is *strong* attraction ($\gamma \gg 1$) with *localised* attractive interaction ($\ell \ll 1$). The central question is: *To what extent is the mean-field description sufficient to capture the clustering dynamics?*

As it turns out, the answer is nuanced. While the mean-field PDE correctly predicts the existence and structure of clustered steady states, and correctly describes the initial linear instability of the uniform state, it fails to capture the coalescence dynamics of clusters. However, it is useful for describing dynamic metastability arising from mass exchange. This places the system within the framework of slow motion of gradient flows, as in [8], in the presence of random perturbations.

Setting. We consider N particles $X_t^i \in \mathbb{T} = \mathbb{R}/\mathbb{Z}$ evolving by the system of stochastic differential equations

$$dX_t^i = -\frac{\gamma}{N}\sum_{j=1}^N \nabla w\left(\frac{X_t^i - X_t^j}{\ell}\right) dt + \sqrt{2}\, dB_t^i, \qquad i = 1,\dots,N, \tag{1}$$

where B_t^i are independent standard Brownian motions on $\mathbb{T}$. The interaction potential $w\colon \mathbb{T} \to \mathbb{R}$ satisfies:

- $w \in C^2(\mathbb{T})$, even, with $w''(0) > 0$;
- w has a unique minimum at 0, is non-decreasing on $[0, s_w]$, and $s_w < \frac{1}{2}$;
- $w(0) < 0$ and $w(x) = 0$ for $|x| > s_w$ (localised).

The two key parameters are the *interaction strength* $\gamma \gg 1$ and the *interaction range* $\ell \ll 1$. The rescaled potential is

$$W_{\gamma,\ell}(x) \coloneqq \gamma\ell\, w\Big(\frac{x}{\ell}\Big).$$

Mean-field description and free energy. As $N \to \infty$, the empirical measure $\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$ converges [7] to the solution $\varrho_t \in \mathcal{P}(\mathbb{T})$ of the *McKean–Vlasov equation*

$$\partial_t \varrho_t = \Delta\varrho_t + \nabla\cdot(\varrho_t\, \nabla W_{\gamma,\ell} * \varrho_t) = \nabla\cdot(\varrho_t\, \nabla \mathcal{F}_{\gamma,\ell}(\varrho_t))\ . \tag{2}$$

The PDE (2) is the Otto-Wasserstein gradient flow [1] of the *free energy* $\mathcal{F}_{\gamma,\ell}$ given by

$$\mathcal{F}_{\gamma,\ell}(\varrho) \coloneqq \int_{\mathbb{T}} \varrho \ln \varrho\, dx + \frac{1}{2}\iint_{\mathbb{T}\times\mathbb{T}} W_{\gamma,\ell}(x-y)\, d\varrho(x)\, d\varrho(y). \tag{3}$$

The gradient flow description of (2) implies that stationary points are critical points of $\mathcal{F}_{\gamma,\ell}$ and the long-time behavior should be governed by the minimizers of the free energy.

Stability of clustered states. The following theorem characterises the global minimizers of $\mathcal{F}_{\gamma,\ell}$ for small ℓ and extends the previous studies from [3, 5].

Theorem (Gerber–Gvalani–Hairer–Pavliotis–S. [4])**.** *For $\ell > 0$ sufficiently small, there exists a critical strength $\gamma_c = \gamma_c(\ell) > 0$ such that for all $\gamma > \gamma_c$, all global minimizers $\varrho_\gamma^* \in \mathcal{P}(\mathbb{T})$ of $\mathcal{F}_{\gamma,\ell}$ over $\mathcal{P}(\mathbb{T})$ are* clustered states *in the sense that they are symmetric decreasing. Moreover, the critical strength satisfies*

$$\frac{1}{\ell} \lesssim \gamma_c(\ell) \lesssim \frac{1}{\ell}\log\frac{1}{\ell}.$$

The proof proceeds by observing that the fixed-point equation for ϱ^* reads

$$\varrho_\gamma^* = \frac{1}{Z}\exp(-W_{\gamma,\ell} * \varrho_\gamma^*) \approx \frac{1}{Z}\exp\Big(-\gamma\ell\, w(0) - \frac{\gamma w''(0)}{\ell}\, x^2 * \varrho_\gamma^*\Big),$$

which identifies ϱ_γ^* approximately as a Gaussian of variance $\ell/(\gamma w''(0))$. Global minimizers of the free energy are then shown to be symmetric decreasing using a new variant of the strict Riesz rearrangement inequality [2]. In particular, this inequality implies that among all probability measures on $\mathbb{T}$, the interaction energy $\frac{1}{2}\iint W_{\gamma,\ell}(x-y)\, d\varrho\, d\varrho$ is minimised by symmetric decreasing rearrangements.

Time-scales of the model. Beyond the mean-field limit, finite-N fluctuations drive the particle system through four distinct dynamical regimes. Their signatures can be tracked via the Dean–Kawasaki SPDE for the empirical measure μ_t^N, which contains a noise term of order $N^{-1/2}$ absent from the McKean–Vlasov PDE (2).

(E1) Initial clustering ($t_{\text{clust}} \sim \log N$). Linearising (2) around $\varrho \equiv 1$ yields the growth rates

$$\psi(k) = -k^2\big(1 + \hat{W}_{\gamma,\ell}(k)\big), \quad k \in \mathbb{Z},$$

where $\hat{W}_{\gamma,\ell}$ is the cosine transform of the rescaled potential. The uniform state is unstable for any $\gamma > \gamma_\sharp := -\big(\min_{k\in\mathbb{Z}\setminus\{0\}} \widehat{W}_{\gamma=1,\ell}(k)\big)^{-1}$. The fastest-growing mode $k_{\max} = \arg\min \widehat{W}_{\gamma=1,\ell}(k)$ determines the exponential instability of the uniform state provided that $\widehat{\rho}_{t=0}(k_{\max}) \neq 0$. For the finite N-particle system, if started from i.i.d. uniform initial states, the fluctuations are of order $1/\sqrt{N}$ and we expect that the first clusters appear at time

$$t_{\text{clust}} \approx \frac{\log N}{2\,\psi(k_{\max})}.$$

(E2) Coalescence of clusters ($t_{\text{coalesce}} \sim N$). After initial clustering, $K \approx k_{\max}$ clusters have formed with centers $\bar{X}_t^{(k)}$ and masses $m^{(k)} = |I_k|/N$. By symmetry of w, each center evolves approximately by $d\bar{X}_t^{(k)} = \sqrt{2/(Nm^{(k)})}\, dW_t^{(k)}$, so cluster centers undergo independent Brownian motions with diffusivity $(Nm^{(k)})^{-1}$. They merge once their separation drops to $\sim \ell$, yielding $t_{\text{coalesce}} \approx N$. This coalescence is a purely finite-N effect invisible to the mean-field PDE (2).

(E3) Dynamic metastability: mass exchange ($t_{\text{coarsen}} \sim e^{\gamma\ell\Delta_w/m_{\text{crit}}}$). Once the system settles into a K-cluster state, further coarsening proceeds via slow mass transfer between neighboring clusters. Approximating each cluster by a Gaussian of variance $\sigma_k^2 = \ell/(w''(0)m^{(k)}\gamma)$, the effective potential felt by particles in cluster k has barriers of height $\gamma\ell\,\Delta_w\, m^{(k)}$, where $\Delta_w = |w(0)|$. By Eyring–Kramers theory, mass leaves a cluster of mass $m^{(k)}$ on the Arrhenius timescale $e^{\gamma\ell\Delta_w m^{(k)}}$.

(E4) Microscopic reversibility ($t_{\text{reversibility}} \sim e^{N\Delta_{\mathcal{F}}}$). The particle system (1) is ergodic with respect to the Gibbs measure $d\pi^N \propto \exp\big(-\frac{\gamma}{2N}\sum_{i,j} W_{\gamma,\ell}(X^i - X^j)\big)\prod_i dX^i$. However, the presence of a discontinuous phase transition implies that the free energy (3) has a critical point. Since the clustered state is the global minimizer, its energy barrier $\Delta_{\mathcal{F}} > 0$ to this critical point determines the timescale of microscopic reversibility $t_{\text{reversibility}} \sim e^{N\Delta_{\mathcal{F}}}$ (see [5, Section 5]).

Conclusion. The identification of the timescales obtained raises several interesting questions: The rigorous proof of dynamical metastability of the mean-field PDE, in the sense of Otto–Reznikoff [8], is still open. Furthermore, the rigorous justification of the limit to the massive Arratia flow [6] in the scaling regime $N \to \infty, \gamma \to \infty, \ell \to 0 : \gamma\ell/\log N \to \infty$ is a challenging open question.

References

[1] L. Ambrosio, N. Gigli, G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed., Birkhäuser, Basel, 2008.

[2] A. Burchard. *Cases of equality in the Riesz rearrangement inequality.* Ann. of Math. 143.3 (1996), pp. 499–527.

[3] J. A. Carrillo, R. S. Gvalani, G. A. Pavliotis, A. Schlichting, *Long-Time behavior and Phase Transitions for the McKean–Vlasov Equation on the torus*. Arch. Ration. Mech. An. **235**, No. 1 (2020), pp. 635–690
[4] N. J. Gerber, R. S. Gvalani, M. Hairer, G. A. Pavliotis, A. Schlichting, *Formation of clusters and coarsening in weakly interacting diffusions*. Preprint arXiv:2510.17629 (2025).
[5] R. S. Gvalani, A. Schlichting, *Barriers of the McKean–Vlasov energy via a mountain pass theorem in the space of probability measures*, J. Funct. Anal. **279**, No. 11, (2020), 108720.
[6] V. Konarovskyi. "A system of coalescing heavy diffusion particles on the real line". Ann. Probab. 45.5 (2017), pp. 3293–3335.
[7] H. P. McKean, *A class of Markov processes associated with nonlinear parabolic equations*, Proc. Natl. Acad. Sci. USA **56** (1966), pp. 1907–1911.
[8] F. Otto, M. G. Reznikoff, *Slow motion of gradient flows*, J. Differential Equations **237** (2007), pp. 372–420.

Random Quadratic Form on a sphere: Synchronization by common noise

Anna Shalova
(joint work with M. Engel)

We study the Random Quadratic Form (RQF), the stochastic differential equation on a sphere $\mathbb{S}^{n-1}$ with multiplicative noise defined by the process Q_t, namely

$$\mathrm{d}X_t = -P_{X_t}\partial Q_t X_t, \tag{1}$$

where $P_X \coloneqq P_{T_X\mathbb{S}^{n-1}} = I - XX^T$ is the projection onto the tangent space of $\mathbb{S}^{n-1}$ at X and the noisy process Q_t is a stochastic matrix-valued process, given as

$$Q_t = \frac{1}{2}(B_t + B_t^T), \tag{2}$$

where $\{B_t^{ij} : i, j \in \{1 \dots n\}$ are independent Brownian motions. Finally, the notation ∂Q_t implies that the SDE (1) is understood in the Stratonovich sense.

The RQF (1) can be interpreted as a gradient flow of a *random quadratic functional* on a sphere; and the gradient structure of the dynamics provides important insights on the long-time behavior of the system. In particular, for a symmetric real matrix $M \in \mathrm{Sym}^n$ define the corresponding quadratic functional as

$$F_M(x) \coloneqq \frac{1}{2}x^T M x.$$

We call the gradient flow of F_M on $\mathbb{S}^{n-1}$, given by

$$\dot{x} \coloneqq -\nabla F_M(x) = -P_x M x, \qquad x(0) = x_0 \in \mathbb{S}^{n-1}, \tag{3}$$

the *deterministic quadratic form*. Rewriting (3) in integral form we thus obtain

$$x(t) = x_0 - \int_0^t \nabla F_M(x(s))\mathrm{d}s = x_0 - \int_0^t P_x M x(s)\mathrm{d}s.$$

Formally, replacing the deterministic functional F_M by the functional $F_{\partial Q_s}$ defined by the increments of Q_t, we obtain the gradient flow formulation of the RQF:

$$X_t = X_0 - \int_0^t \nabla F_{\partial Q_s}(X_s) \quad \text{where} \quad F_{\partial Q_s}(X_s) \coloneqq \frac{1}{2}X_s^T \partial Q_s X_s.$$

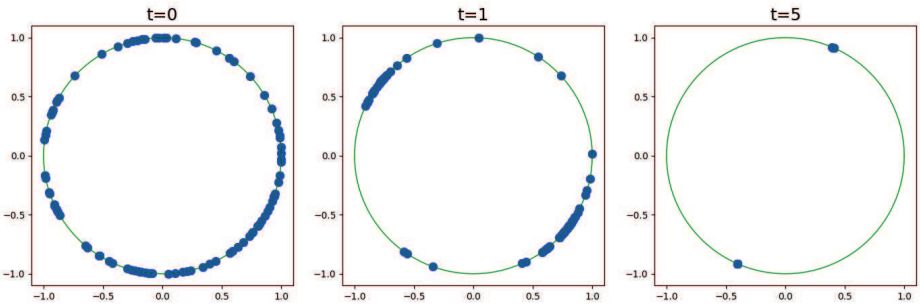

FIGURE 1. Ensemble of RQFs driven by the same process Q_t from different initial conditions. At time $t \sim 5$ the trajectories approach the *random attractor* consisting of two antipodal points that further move in time.

The standard (deterministic) gradient flow structure implies that the driving functional is the optimal Lyapunov function of the underlying dynamics, guaranteeing, under certain convexity assumptions, the convergence of the solutions to the minimizer of the driving functional. Remarkably, some of the properties of the deterministic gradient dynamics translate into the random case, which we illustrate through the example of RQF. In this work, we study distributional properties of the solutions to Eq. (1), i.e., *invariant measures*, and give a path-wise characterization, describing the *random attractor* of the RQF. We compare the long-time behavior of the RQF to the deterministic setting studied in [6]:

Theorem 1 (Deterministic Quadratic Form [6]). Let $M \in \mathrm{Sym}^n$ be a symmetric matrix, sampled from the Gaussian Orthogonal Ensemble. Then, w.p. 1, there exists $x^* \in \mathbb{S}^{n-1}$ such that the gradient flow of the quadratic form (3) satisfies

$$\lim_{t\to\infty} \min\left(\mathrm{dist}(x(t), x^*), \mathrm{dist}(x(t), -x^*)\right) = 0$$

for a.e. initial condition $x_0 \in \mathbb{S}^{n-1}$. In other words, almost every trajectory of the gradient flow $x(t)$ converges to either x^* or $-x^*$.

Analogously to the deterministic model, RQF synchronizes the points into the *random anti-polar configuration*, namely:

Theorem 2 (Random Quadratic Form [2]). Let Q_t be the driving process as described in (2). Let X_t and Y_t be the RQF processes driven by Q_t with (possibly) different initial conditions $X_0, Y_0 \in \mathbb{S}^{n-1}$, then

- the law of X_t (and Y_t) in the large time limit converges to the uniform measure on the sphere,
- for almost every ω, the two RQF processes X_t, Y_t satisfy

$$\lim_{t\to\infty} \min\left(\mathrm{dist}(X_t, Y_t), \mathrm{dist}(X_t, -Y_t)\right) = 0.$$

  In other words, X_t and Y_t either converge to each other (polar) or become opposite (anti-polar configuration).

The proof relies on the framework of random dynamical systems and the description of random attractors. In particular, we characterize the statistical equilibrium of the system as introduced in [1]. Remarkably, RQF is an example of an RDS exhibiting *partial* synchronization and therefore cannot be studied by the classical methods discussed in [4].

RQF and clustering in transformers. The RQF model illustrates the *synchronization by common noise* phenomena in the continuous-time models of transformers formulated e.g. in [5]. In particular, let $x_i : (0,t) \to \mathbb{S}^{n-1}$ be the trajectory of the i-th token in a continuous-time model of transformers, then the evolution of x_i is defined by two 'forces' and takes the form

$$\begin{aligned} \dot{x}_i &= P_{x_i}\left(\mathrm{FF}(x_i) + \mathrm{Attn}(x_i; x_1, x_2 \dots x_d)\right), \\ \mathrm{FF}(x_i) &= \sigma(Mx_i + B), \\ \mathrm{Attn}(x_i; x_1, x_2 \dots x_d) &= \frac{1}{\sum_j e^{x_i Q^T K x_j}} \sum_j e^{x_i Q^T K x_j} V x_j, \end{aligned}$$

where the self-attention mechanism (Attn) can be interpreted as an interaction force and the Feed-Forward layers (FF) play the role of the potential energy. One of the key features of such dynamics is the *clustering* phenomena, namely relative convergence of tokens in transformers in the long-time limit. Clustering in pure-attention transformers has been recently extensively studied, see [7, 3] and references therein.

In this work, we focus solely on the Feed-Forward layer and start with the following toy model:

$$\dot{x}_i = P_{x_i}\mathrm{FF}(x_i), \qquad \mathrm{FF}(x_i) = \sigma(M(t)x_i + B), \tag{4}$$

corresponding to the contribution of Feed-Forward layers in transformers. Under the additional structural assumptions $B \equiv 0$ and $\sigma(x) = x$, we conclude that every token follows the dynamics analogous to (3). Since the parameters in transformers are initialized randomly and are independent from layer to layer, we consider the white-noise structure of Q_t and obtain the RQF model

$$\mathrm{d}X_t = P_{X_t} \partial Q_t X_t,$$

as a simplified model of the dynamics of tokens driven by Feed-Forward layers.

Interpreting RQF as a model of the token dynamics, we provide an alternative (independent of Self-Attention) explanation of the clustering behavior in deep transformers and show that tokens cluster even in the absence of the Self-Attention mechanism.

References

[1] P. Baxendale. Statistical equilibrium and two-point motion for a stochastic flow of diffeomorphisms. In *Spatial Stochastic Processes: A Festschrift in Honor of Ted Harris on his Seventieth Birthday*, pages 189–218. Springer, 1991.

[2] M. Engel and A. Shalova, *Random Quadratic Form on a sphere: Synchronization by common noise*, *arXiv preprint arXiv:2603.06187*, 2026.

[3] L. Fedorov, M. Sander, R. Elie, P. Marion and M. Laurière. Clustering in deep stochastic transformers. *arXiv preprint arXiv:2601.21942*, 2026.
[4] F. Flandoli, B. Gess, and M. Scheutzow. Synchronization by noise. *Probab. Theory Related Fields*, 168(3-4):511–556, 2017.
[5] B. Geshkovski, C. Letrouit, Y. Polyanskiy and P. Rigollet, *A mathematical perspective on transformers*, Bulletin of the American Mathematical Society, 62(3):427–479, 2025.
[6] R. Mahony, U. Helmke, and J. Moore. Gradient algorithms for principal component analysis. *The ANZIAM Journal*, 37(4):430–450, 1996.
[7] P. Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.

Existence of Solution of Natural Gradient Flows in Neural Network Manifolds

Olga Mula
(joint work with D. Bon, B. Caris, and M. Peletier)

This work studies numerical methods for gradient flows in Hilbert spaces based on neural network approximations. The central idea is to represent the solution on a neural network manifold and evolve its parameters in time. This paradigm–appearing under names such as *natural gradient flow*, *parametric dynamical approximation*, or the *Dirac–Frenkel approach*–has recently attracted significant attention.

However, the standard formulation raises fundamental issues regarding the existence of solutions. While seemingly theoretical, these issues manifest in practice as strong numerical instabilities. A common remedy is regularization, which, however, breaks the connection with the underlying operator. Our contribution proposes a new viewpoint that restores existence without regularization and resolves these instabilities.

We consider the shallow neural network

$$\mathcal{U}_n(\theta)(x) = \sum_{i=1}^{n} a_i \, \varphi(x - b_i),$$

where $\varphi \in \mathcal{C}^\infty(\mathbb{R})$ and $\theta = (a_i, b_i)_{i=1}^n \in \mathbb{R}^{2n}$. The associated model class

$$\mathcal{M}_n \coloneqq \{\mathcal{U}_n(\theta) \colon \theta \in \mathbb{R}^{2n}\} \subset \mathcal{C}^\infty(\mathbb{R})$$

is shown to form a *singular manifold* in $L^2(\mathbb{R})$.

We analyze the limitations of using $\mathcal{M}_n$ as an ansatz space for evolution equations such as the Allen–Cahn equation, and show that classical approaches induce singularities in the dynamics. Exploiting the fact that Allen–Cahn is the L^2-gradient flow of an energy functional $\mathcal{E}$, we propose instead to consider the metric gradient flow of $\mathcal{E}$ in the metric space $(\mathcal{M}_n, d)$, where

$$d(u, v) = \|u - v\|_{L^2(\mathbb{R})}.$$

Under suitable assumptions, we prove that the corresponding minimizing-movement scheme converges, as $\Delta t \to 0$, to a curve of maximal slope for $\mathcal{E}$ on the completion $\overline{(\mathcal{M}_n, d)}^{L^2}$. We further show, by example, that parametrizations of

such curves may be discontinuous in time, and that this phenomenon is, in general, unavoidable.

The Hellinger–Kantorovich Metric Measure Geometry on Spaces of Measures

Lorenzo Dello Schiavo
(joint work with G. E. Sodini)

Motivations: large Lie groups and SPDEs. A *large Lie group* is a Lie group modeled on an infinite-dimensional Hilbert, Banach, or Fréchet space. Prototypical examples of such groups are *transformation groups*, as diffeomorphism groups of differential manifolds; *G-current groups*, i.e. groups of G-valued functions, for some Lie group G; *multiplier groups*, i.e. (Abelian) $(\mathbb{R}^+, \cdot)$-current groups, which are maximal toral subgroups in the corresponding SL_2-current groups.

A fruitful approach to large Lie groups via their *representations* has been the subject of a longstanding program initiated for diffeomorphism groups by Vershik, Gel'fand, and Graev in [14], and more recently by Kondratiev, Lytvynov, and Vershik for semidirect products of diffeomorphisms and multipliers in [7].

In order for the representations of these groups to be *faithful* —i.e. for them to retain sufficient information on the group— the representations need to be constructed on some 'large' Hilbert space. Especially in the case of diffeomorphisms and of multipliers, a concrete realization of such a Hilbert space is the space $L^2(\mathcal{Q})$ of some measure $\mathcal{Q}$ on a space of measures. Indeed, diffeomorphisms naturally act on measures by push-forward, and multipliers simply act on measures by multiplication by densities. When $\mathcal{Q}$ is a probability measure, it is usually regarded as (the law of) a *random measure*, typically, a random point process. This and similar constructions have appeared in [14] for *Poisson* processes; in [7] for *Gamma* processes; and in [3], for *Dirichlet–Ferguson* processes.

'Geometric' Brownian motions. As already noted in [3, 7], this action of diffeomorphisms, multipliers, or a combination thereof, on a space of measures induces an energy functional on $L^2(\mathcal{Q})$. As it turns out, the functional is, in many of these settings, a *Dirichlet form*, and it is therefore uniquely associated with a measure-valued Markov process. We call this process the *geometric measure-valued Brownian motion* induced by the group action.

On the one hand, it is one goal of the aforementioned program to study properties of the representation on $L^2(\mathcal{Q})$ of a given large Lie group via the corresponding geometric measure-valued Brownian motion. For instance, it is usually expected that invariant sets of this Brownian motion are in one-to-one correspondence with sub-spaces invariant under the group action.

On the other hand, geometric measure-valued Brownian motions are very interesting stochastic processes in their own right. This is readily seen from two important examples in the case when $\mathcal{Q}$ is the Dirichlet–Ferguson measure. In this case, one process induced by the action of multipliers is the *Fleming–Viot*

process with parent independent mutation, e.g. [11], while one process induced by the action of diffeomorphisms is the *Dirichlet–Ferguson diffusion* [3], the unique solution to the *Dean–Kawasaki stochastic partial differential equation with singular drift* [5, 6, 9].

Metric-measure Brownian motions. Another fundamental approach to the construction of energy functionals on spaces of measures is as follows. When the space of measures in question is endowed with some natural distance (e.g., Hellinger, Bhattacharyya, Kantorovich–Rubinstein, Hellinger–Kantorovich) and with a reference random measure $\mathcal{Q}$, we consider the *Cheeger energy* of the resulting *metric measure space*.

For Kantorovich–Rubinstein distances on spaces of probability measures, the study of these energy functionals has been undertaken in [4, 12]. Here, we rather consider the space of all non-negative finite measures with the *Hellinger–Kantorovich distance*. Also in this case, the Cheeger L^2-energy is a quadratic functional, and thus a Dirichlet form. We call the unique Markov process associated to it the *metric measure measure-valued Brownian motion* induced by the distance.

It is one main result of this work (see below) that, for a specific choice of $\mathcal{Q}$, the geometric point of view (group actions) and the metric-measure point of view (distances) are one and the same, i.e. that the geometric measure-valued Brownian motion coincides with the metric measure Brownian motion just described.

Not only does this provide an identification of the stochastic process in question; it will also grant us the possibility to import tools from metric measure geometry in the study of geometric Brownian motions and of the corresponding representations for a given group action, and vice versa to use the Lie-group construction for the study of the metric measure space arising from the Hellinger–Kantorovich distance and the reference measure $\mathcal{Q}$.

Main results. Let (M, g) be a closed Riemannian manifold with Riemannian distance d_g. We denote by $\mathcal{M}(M)$ the cone of all non-negative and finite Borel measures on M, endowed with the *Hellinger–Kantorovich* (or *Wasserstein–Fisher–Rao*) *distance* HK_g induced by d_g, [2, 8, 10].

Firstly, for a suitable algebra $\mathcal{A}$ of cylinder functions on $\mathcal{M}(M)$, we prove the following Myers–Serrin-type theorem. For every σ-finite Borel measure $\mathcal{Q}$ on $\mathcal{M}(M)$, the space $\mathcal{A}$ is dense in 2-energy in the metric Sobolev space, see [1],

$$H^{1,2}\left(\mathcal{M}(M), \mathsf{HK}_{\mathsf{d}_g}, \mathcal{Q}\right) ,$$

and the latter is a Hilbert space.

Secondly, we focus on a specific choice for $\mathcal{Q}$. For $\theta > 0$ we consider Vershik's *infinite-dimensional multiplicative Lebesgue measure* $\mathcal{L}_\theta$ on $\mathcal{M}(M)$ with intensity the *normalized* Riemannian volume, see [13]. We show that it is the unique natural measure for the Hellinger–Kantorovich geometry on $\mathcal{M}(M)$.

Further denote by $\boldsymbol{\nabla}$ the gradient for real-valued functions on $\mathcal{M}(M)$ associated to the Hellinger–Kantorovich geometry, and by $\langle \cdot \,|\, \cdot \rangle_\mu$ a suitably weighted scalar

product. We prove that the canonical energy form

$$\mathcal{E}(u,v) \coloneqq \int \langle (\boldsymbol{\nabla} u)_\mu | (\boldsymbol{\nabla} v)_\mu \rangle_\mu \, \mathrm{d}\mathcal{L}_\theta(\mu)\,, \qquad u, v \in \mathcal{A}\,,$$

is closable on $L^2(\mathcal{L}_\theta)$. Its closure is a conservative quasi-regular strongly local Dirichlet form on $L^2(\mathcal{L}_\theta)$; it is further identical with the Cheeger energy of the metric measure space $(\mathcal{M}(M), \mathsf{HK}_g, \mathcal{L}_\theta)$, and properly associated with a Hunt diffusion with state space $\mathcal{M}(M)$, the '*Brownian motion*' of the Hellinger–Kantorovich geometry on $\mathcal{M}(M)$. The latter process is recurrent if $\theta \in (0, 1]$, and transient otherwise.

Acknowledgments. This research was funded in part by the Austrian Science Fund (FWF) projects ESP208 and F65. This research was funded in part by PRIN Department of Excellence MatMod@TOV (CUP: E83C23000330006).

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. *Invent. Math.*, 195(2):289–391, 2014. doi:10.1007/s00222-013-0456-1.

[2] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An Interpolating Distance Between Optimal Transport and Fisher–Rao Metrics. *Found. Computat. Math.*, 18(1):1–44, October 2016. doi:10.1007/s10208-016-9331-y.

[3] L. Dello Schiavo. The Dirichlet–Ferguson Diffusion on the Space of Probability Measures over a Closed Riemannian Manifold. *Ann. Probab.*, 50(2):591–648, 2022. doi:10.1214/21-AOP1541.

[4] M. Fornasier, G. Savaré, and G. E. Sodini. Density of subalgebras of Lipschitz functions in metric Sobolev spaces and applications to Wasserstein Sobolev spaces. *J. Funct. Anal.*, 285(11):110153, dec 2023. doi:10.1016/j.jfa.2023.110153.

[5] V. V. Konarovskyi and M.-K. von Renesse. Modified Massive Arratia flow and Wasserstein diffusion. *Comm. Pure Appl. Math.*, 72(4):764–800, 2019. doi:10.1002/cpa.21758.

[6] V. V. Konarovskyi and M.-K. von Renesse. Reversible coalescing-fragmentating Wasserstein dynamics on the real line. *J. Funct. Anal.*, 286(8):110342, April 2024. doi:10.1016/j.jfa.2024.110342.

[7] Yu. G. Kondratiev, E. W. Lytvynov, and A. M. Vershik. Laplace operators on the cone of Radon measures. *J. Funct. Anal.*, 269(9):2947–2976, 2015. doi:10.1016/j.jfa.2015.06.007.

[8] S. Kondratyev, L. Monsaingeon, and D. Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11-12):1117–1164, 2016. doi:10.57262/ade/1476369298.

[9] L. Dello Schiavo. Massive Particle Systems, Wasserstein Brownian Motions, and the Dean–Kawasaki Equation. *arXiv:2411.14936*, pages 1–103, 2024. doi:10.48550/arXiv.2411.14936.

[10] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117, 2018. doi:10.1007/s00222-017-0759-8.

[11] L. Overbeck, M. Röckner, and B. Schmuland. An analytic approach to Fleming–Viot processes with interactive selection. *Ann. Probab.*, 23(1):1–36, 1995.

[12] G. E. Sodini. The general class of Wasserstein Sobolev spaces: density of cylinder functions, reflexivity, uniform convexity and Clarkson's inequalities. *Calc. Var. PDE*, 62(212), 2023.

[13] N. V. Tsilevich, A. M. Vershik, and M. Yor. An Infinite-Dimensional Analogue of the Lebesgue Measure and Distinguished Properties of the Gamma Process. *J. Funct. Anal.*, 185(1):274–296, 2001. doi:10.1006/jfan.2001.3767.
[14] A. M. Vershik, I. M. Gel'fand, and M. I. Graev. Representations of the Group of Diffeomorphisms. *Russ. Math. Surv.+*, 30(6):1–50, 1975.

Evolution of Gaussians in the Spherical HK–Boltzmann Gradient Flow

Oliver Tse
(joint work with J.-J. Zhu, M. Liero, and A. Mielke)

We present the study of the gradient flow of the KL-divergence functional and its long-time behavior within the family of Gaussian measures, where the underlying geometry is obtained by projecting the Spherical Hellinger–Kantorovich (SHK) metric onto this family.

1. Motivation: variational inference and three approaches

Given a reference measure $\pi \in \mathcal{P}(\mathbb{R}^d)$ with $\pi \propto e^{-V}$ for a λ-convex potential V ($\lambda > 0$, but $\lambda \ll 1$), the variational inference problem is

$$\min_{\mu\in\mathcal{G}} \mathcal{E}(\mu) := \int \log\frac{d\mu}{d\pi}\, d\mu,$$

for some subset $\mathcal{G} \subset \mathcal{P}(\mathbb{R}^d)$. Three natural approaches are available:

(1) Otto–Wasserstein flow. The Fokker–Planck equation

$$\partial_t \mu = \Delta\mu + \operatorname{div}(\mu\nabla V) =: -\mathbb{K}_{\mathrm{Otto}}(\mu) D\mathcal{E}(\mu)$$

admits an Otto–Wasserstein gradient structure [3], with convergence rates [1]

$$\begin{cases} \mathcal{E}(\mu_t) \le O(e^{-2\lambda t}) & \text{for } \lambda > 0, \\ \mathcal{E}(\mu_t) \le O(t^{-1}) & \text{for } \lambda = 0. \end{cases}$$

(2) Accelerated methods. Polyak [8] and Nesterov [9] inspired kinetic PDEs that achieve $\mathcal{E}(\mu_t) \le O(e^{-\beta t})$ with $\beta_t \le 1$ via a doubling of variables [10, 11].

(3) SHK (or Wasserstein–Fisher–Rao) flow. Developed by Liero–Mielke–Savaré [4] (see also [7, 12]), the combined flow reads

$$\partial_t \mu = -\alpha\, \mathbb{K}_{\mathrm{Otto}}(\mu)\, D\mathcal{E}(\mu) - \beta\, \mathbb{K}_{\mathrm{SHe}}(\mu)\, D\mathcal{E}(\mu) =: -\mathbb{K}_{\alpha,\beta}(\mu)\, D\mathcal{E}(\mu),$$

where $\mathbb{K}_{\mathrm{SHe}}(\mu)D\mathcal{E}(\mu) = \mu(D\mathcal{E}(\mu) - \mathcal{E}(\mu))$. Lu–Lu–Nolen [5] (see also [2]) showed that, when $\mathcal{G} = \{\text{Gaussians}\}$ and initial data are close enough, the decay rate of the SHK flow is *independent* of $\lambda > 0$.

2. Gaussian SHK flow and gradient structure

Restrict to $\mathcal{G} = \{G(\Sigma, m)\} : \Sigma^\top = \Sigma,\ \Sigma \succ 0,\ m \in \mathbb{R}^d\}$, and set $\pi = G(\Gamma, n)$. Projecting the SHK flow onto $\mathcal{G}$ gives the Bures–Wasserstein–Fisher–Rao system

$$\text{(BWFR)} \qquad \begin{aligned} \dot\Sigma &= \alpha\big(2I - \Gamma^{-1}\Sigma - \Sigma\Gamma^{-1}\big) + \beta\big(\Sigma - \Sigma\Gamma^{-1}\Sigma\big), \\ \dot m &= -(\alpha + \beta\Sigma)\,\Gamma^{-1}(m - n). \end{aligned}$$

A general principle due to Maas–Mielke [6] guarantees a gradient structure for the projected flow with *reduced driving energy* $\mathcal{E}(\Sigma, m) = \frac{1}{2}|m - n|^2_{\Gamma^{-1}} + \mathcal{F}(\Sigma)$ with

$$\mathcal{F}(\Sigma) \coloneqq \frac{1}{2}\Big(\operatorname{tr}[\Gamma^{-1}\Sigma - I] - \log\det(\Gamma^{-1}\Sigma)\Big).$$

In this case, the equation for the covariance Σ in (BWFR) takes the form

$$\dot\Sigma = -\big(\alpha\mathbb{A}(\Sigma) + \beta\mathbb{B}(\Sigma)\big)D\mathcal{F}(\Sigma) =: -\mathbb{K}^\Sigma_{\alpha,\beta}D\mathcal{F}(\Sigma),$$

where $\mathbb{A}(\Sigma)\Lambda = 2(\Lambda\Sigma + \Sigma\Lambda)$, $\mathbb{B}(\Sigma)\Lambda = 2\Sigma\Lambda\Sigma$, and $D\mathcal{F}(\Sigma) = \frac{1}{2}(\Gamma^{-1} - \Sigma^{-1})$.

Note that the equation for Σ in (BWFR) decouples from m, so decay estimates for $\mathcal{F}(\Sigma)$ can be established independently.

3. Decay estimates for Gaussian targets

3.1. Geodesic convexity.

The first result states a negative result concerning the λ-convexity of $\mathcal{F}$ in the presence of the SHe-term $\mathbb{B}$.

Theorem 1. *$\mathcal{F}$ is λ-convex w.r.t. $\mathbb{K}^\Sigma_{\alpha,\beta}$ if and only if $\beta = 0$ and $\lambda \le \alpha\lambda_{\min}(\Gamma^{-1})$. In particular, the SHe-term $\mathbb{B}$ destroys geodesic convexity.*

3.2. Gradient dominance (PL-condition).

Define the dissipation

$$\mathcal{D}_{\alpha,\beta}(\Sigma) \coloneqq \langle D\mathcal{F}(\Sigma), (\alpha\mathbb{A}(\Sigma) + \beta\mathbb{B}(\Sigma))D\mathcal{F}(\Sigma)\rangle.$$

The situation improves when considering gradient dominance.

Theorem 2 (Global PL)**.** *$\exists\, \mathsf{c}_{\mathrm{PL}} > 0$ such that*

$$\mathcal{D}_{\alpha,\beta}(\Sigma) \ge \mathsf{c}_{\mathrm{PL}}\mathcal{F}(\Sigma) \ \text{ for all } \Sigma \iff \alpha\lambda_{\min}(\Gamma^{-1}) > 0.$$

The necessary condition $\alpha\lambda_{\min}(\Gamma^{-1}) > 0$ can be removed if the requirement for a global PL-condition is relaxed.

Theorem 3 (Sublevel PL)**.** *For every $E > 0$, $\exists\, \mathsf{c}_{\mathrm{PL}}(E) > 0$ such that*

$$\mathcal{D}_{\alpha,\beta}(\Sigma) \ge \mathsf{c}_{\mathrm{PL}}(E)\mathcal{F}(\Sigma) \quad \text{for every } \Sigma : \mathcal{F}(\Sigma) \le E.$$

In fact, the sublevel PL-condition holds even when $\alpha = 0$ and the constant $\mathsf{c}_{\mathrm{PL}}(E)$ improves when $E \to 0$.

3.3. **Refined estimates.** Since $\mathcal{F}(\Sigma)$ decreases along the evolution, we hope to refine the decay estimates. Indeed, we obtain

Theorem 4. *For any $\Sigma_0 \in \operatorname{dom}\mathcal{F}$ there exists $\mathsf{c}_\beta = \mathsf{c}_\beta(\Sigma_0)$ such that*

$$\mathcal{F}(\Sigma(t)) \le \mathsf{c}_\beta\, e^{-\nu t}\, \mathcal{F}(\Sigma_0), \qquad \nu = 2\alpha\lambda_{\min}(\Gamma^{-1}) + \beta.$$

Idea of proof. Set $B \coloneqq \Gamma^{-1/2}\Sigma\Gamma^{-1/2}$ with eigenvalues (b_i). Hellmann–Feynman gives two-sided bounds on $b_i(t)$, from which we obtain

$$\frac{d}{dt}\mathcal{F}(\Sigma_t) \le -\big(2\alpha\lambda_{\min}(\Gamma^{-1}) + \beta\, b_{\min}(t)\big)\mathcal{F}(\Sigma_t).$$

The proof then concludes after a careful Gronwall argument. □

4. Non-Gaussian targets

Now let $\pi \propto e^{-V}$ with V λ-convex ($\lambda > 0$), which is not assumed to be Gaussian. The projected dynamics on Gaussians become

$$\dot\Sigma = \alpha(2I - \Gamma^{-1}\Sigma - \Sigma\Gamma^{-1}) + \beta(\Sigma - \Sigma\Gamma^{-1}\Sigma),$$
$$\dot m = -(\alpha + \beta\Sigma)\int \nabla V(x)\, G(\Sigma, m)(dx), \qquad \Gamma^{-1} \coloneqq \int \nabla^2 V(x)\, G(\Sigma, m)(dx),$$

and one still has a projected gradient structure on $\mathcal{G}$.

Theorem 5 (Unique minimizer)**.** *$\mathcal{E}$ attains a unique minimizer $(\Sigma^{\mathrm{opt}}, m^{\mathrm{opt}}) \in \mathcal{G}$.*

Theorem 6 (Convergence)**.** *The following decay estimate holds with $\gamma \propto \lambda$:*

$$\mathcal{E}(\Sigma(t), m(t)) \le e^{-\gamma t}\,\mathcal{E}(\Sigma_0, m_0) + (1 - e^{-\gamma t}) \min_{(\Lambda,n)\in\mathcal{G}} \mathcal{E}(\Lambda, n).$$

We conjecture that the convergence result above is far from optimal, since, unlike when π is Gaussian, it depends on the parameter $\lambda > 0$. A better decay estimate remains open, as does a good (provably convergent) numerical scheme for non-Gaussian targets.

References

[1] L. Ambrosio, N. Gigli, G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Birkhäuser, 2nd ed., 2008.

[2] Y. Chen, D.Z. Huang, J. Huang, S. Reich, A.M. Stuart, *Sampling via gradient flows in the space of probability measures*, arXiv:2310.03597 (2023).

[3] R. Jordan, D. Kinderlehrer, F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal. **29** (1998), 1–17.

[4] M. Liero, A. Mielke, G. Savaré, *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures*, Invent. Math. **211** (2018), 969–1117.

[5] Y. Lu, J. Lu, J. Nolen, *Accelerate Langevin Sampling with Birth-Death Process*, arXiv:1905.09863 (2019).

[6] J. Maas, A. Mielke, *modeling of chemical reaction systems with detailed balance using gradient structures*, J. Stat. Phys. **181**(6) (2020), 2257–2303.

[7] L. Monsaingeon, S. Kondratyev, D. Vorotnikov. *A new optimal transport distance on the space of finite Radon measures*, Adv. Differ. Equ. **21** (2016), 1117–1164.

[8] B.T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys. **4** (1964), 1–17.

[9] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Sov. Math. Dokl. **27** (1983), 372–376.
[10] A. Wibisono, A.C. Wilson, M.I. Jordan, *A variational perspective on accelerated methods in optimization*, PNAS **113** (2016), E7351–E7358.
[11] R. Zhang, S. Chewi, M. Li, A. Krishnamurthy, K. Balasubramanian, M.A. Erdogdu, *Improved analysis of score-based generative modeling*, 2023.
[12] L. Chizat, G. Peyré, B. Schmitzer, F.X. Vialard (2018), *An interpolating distance between optimal transport and Fisher–Rao metrics*, Found. Comp. Math. **18**(1) (2018), 1–44.

A Network-Simplex implementation of unbalanced optimal transport problems

Antonin Chambolle
(joint work with G. Agazzotti and C. Royer)

The computation of discrete optimal transport problems, which is useful for many applications in engineering and computer science (see for instance [12, 4]), requires the resolution of a (usually high-dimensional) linear program, of the form:

$$\text{(OT)} \qquad \min_X \left\{ C : X \ : \ X \geq 0, X\mathbf{1} = a, X^T\mathbf{1} = b \right\}.$$

Here, $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ are non-negative weights representing measures over n and m points, $C = (C_{i,j})_{i,j} \in \mathbb{R}^{n\times m}$ is a cost-matrix (which may be assumed to be non-negative as well), $X \in \mathbb{R}^{n\times m}$ an assignment matrix (or discrete transport plan) and $C : X = \sum_{i,j} C_{i,j}X_{i,j}$ is the Frobenius scalar product. The constraints on X express that the mass $X_{i,j}$ transported from i to j is non-negative ($X \geq 0$ is understood componentwise) and that the first and second marginals of X are a and b, respectively. Here, $\mathbf{1}$ denotes vectors with all components equal to 1 (by an abuse of notation, we omit their dimension, which is obvious from the context). Observe that one must have $\sum_i a_i = \sum_{i,j} X_{i,j} = \sum_j b_j$, otherwise the problem has no feasible points. Among the numerous numerical methods implemented to solve (OT), the simplex method is one of the most efficient (for sizes $m, n \sim 10^3 - 10^4$), when implemented as a "network-simplex" algorithm [2]: this is for instance the implementation in the Python library `POT` [8], based on the `C++` library "LEMON" [7].

This approach relies on the observation that (OT) is a "minimal cost flow" problem, for which the support $\mathcal{T}_X \coloneqq \{(i,j) : X_{i,j} > 0\}$ of the primal variable X can be represented as a tree on a graph, with at most $n + m - 1$ branches, as we further explain. In practice, one builds a bipartite graph with n points on the left, connected (densely) to m points on the right, and i and j are connected whenever $X_{i,j} > 0$. Given such a tree, the variables $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ of the dual problem

$$\text{(OT}_{\text{dual}}\text{)} \qquad \max_{f,g} \{ f \cdot a + g \cdot b : C_{i,j} \geq f_i + g_j \ \forall i, j \}$$

can be immediately computed from the complementary conditions $C_{i,j} = f_i + g_j$ when $X_{i,j} > 0$. Note that the dual variables (f, g) belong to a dual space $(\mathbb{R}^n \times \mathbb{R}^m)/\sim$ where $(f, g) \sim (f', g')$ if and only if $f - f' = c\mathbf{1}$ and $g - g' = -c\mathbf{1}$ for some $c \in \mathbb{R}$, so that the dual space has dimension $n + m - 1$. Hence, the vertices of the

polytope $\{f_i + g_j \leq C_{i,j}\}$ are points where $n+m-1$ equalities $f_i + g_j = C_{i,j}$ hold, which gives the size of the support of the corresponding primal variables.

The simplex algorithm runs as follows: starting from an admissible primal X, the associated tree $\mathcal{T}_X$, and the corresponding duals (f, g), one identifies $(i, j) \notin \mathcal{T}_X$ where $f_i + g_j > C_{i,j}$, so that adding the edge (i, j) to the current tree creates a cycle. Then, one modifies X along this cycle to reduce the cost in (OT), until X vanishes on another edge of the cycle. One obtains a new value X and a new tree $\mathcal{T}_X$, on which one may recompute the dual variables. The process stops after finitely many steps, and even if the worst-case complexity of the method is that of a simplex method (which could be exponential if the polytope is in a very bad position), the method is usually very efficient.

The talk addressed the extension of this method to *unbalanced* optimal transport problems, of the form

$$\text{(uOT)} \qquad \min_X \left\{ C : X + \psi_1(X\mathbf{1}) + \psi_2(X^T\mathbf{1}) \,:\, X \geq 0 \right\},$$

for two convex functions ψ_1, ψ_2. Although we considered in [1] quite general costs, the case of quadratic (and, actually, entropic) costs is a bit easier, and the talk was focusing on quadratic problems:

$$\text{(quOT)} \qquad \min_X \left\{ C : X + \tfrac{1}{2}\|X\mathbf{1} - a\|^2 + \tfrac{1}{2}\|X^T\mathbf{1} - b\|^2, X \geq 0 \right\}.$$

Observe that now, the problem makes sense for any pair of vectors a and b. In that case, the problem can be cast as a "quadratic minimal cost-flow" problem, for which, again, efficient methods exist: in particular, a strongly polynomial algorithm [14], accelerated first-order iterative methods [9] or other methods designed for the similar "Lasso" problem [5], approaches based on entropic regularization and matrix scaling [6], see also [3, 13, 11]. Of particular interest are the semi-unbalanced variants (when one marginal remains fixed, e.g. $X\mathbf{1} = a$), which correspond to computing the "proximal operator" of the OT distance to the given measure a. This is the basic brick in many splitting algorithms to tackle discretized "JKO" gradient flows problems, see for instance [10].

Yet, it seems the idea of extending the network-simplex method to this problem had not been addressed before. Again, a basic remark here is that once $X\mathbf{1}$ and $X^T\mathbf{1}$ are known, the problem boils down to (OT) so that all the remarks which are valid for the linear program (size of the support, fact that it is acyclic) remain also valid for this problem, and the solutions are represented with the same tree structure. One important difference is that the dual variables (hence the support $\mathcal{T}_X$) now determine the marginals. Indeed, the dual problem for (quOT) is now

$$\text{(quOT}_{\text{dual}}\text{)} \qquad \max_{f,g} \left\{ f \cdot a + g \cdot b - \tfrac{1}{2}(\|f\|^2 + \|g\|^2) : C_{i,j} \geq f_i + g_j \ \forall i, j \right\}$$

and at optimality, one has

$$\text{(Margin.)} \qquad X\mathbf{1} = a - f\,, \quad X^T\mathbf{1} = b - g.$$

As before, given an admissible primal variable X, admissible duals (f, g) which satisfy the complementary condition on $\mathcal{T}_X$ and (Margin.), the algorithm identifies a pair (i, j) for which $\Delta := f_i + g_j - C_{i,j} > 0$, and the corresponding loop in the

graph $\mathcal{T}_X \cup \{(i,j)\}$. Flow is first pushed through this loop, as for solving (OT), in order to lower the primal value $C:X$ without affecting the marginals. Yet, contrary to the standard network simplex, one cannot update the dual variables as easily as before, as this requires updating simultaneously the value X to satisfy (Margin.). We described a method to compute the optimal X, f, g on the new tree (or forest, as the support of X usually becomes disconnected), with worst-case complexity the square of the size of the initial tree. In practice, this step gets slower whenever the connected component of $\mathcal{T}_X \cup \{(i,j)\}$ containing (i,j) is split into many subtrees, making the subsequent iterations possibly faster. Overall, the complexity of this step does not slow down the method too much compared to the standard network simplex: indeed the bottleneck for these algorithms is the search of a new "pivot" (i,j) which violates the dual constraints (we use a block pivot search rule, as implemented in `POT`).

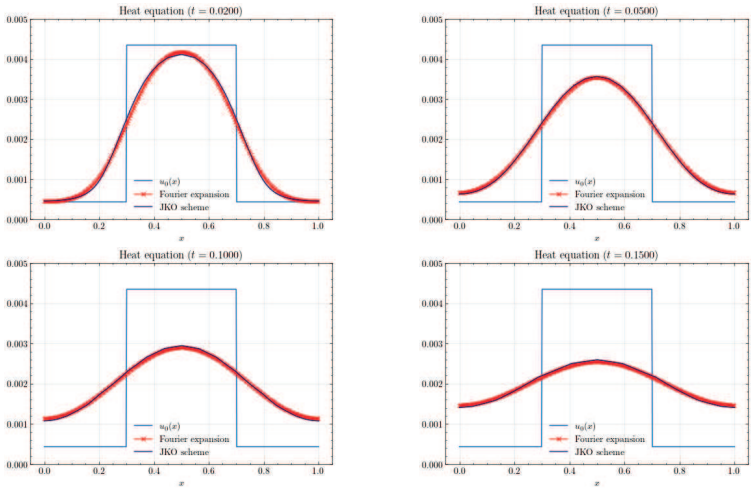

FIGURE 1. Heat equation as a JKO flow of the entropy

Complexity. Heuristically, the (worst case) complexity of the method is at worst of order $\max\{n,m\}$ times the runtime of a network simplex algorithm for (OT). Yet, in the quadratic case, one can exploit a simplification, due to the strong concavity of the dual problem ($\text{quOT}_{\text{dual}}$). One shows that the (primal) energy which is obtained after an update is of order Δ^2 smaller than the energy prior to the update. If Δ may be chosen of order $\max_{i,j} f_i + g_j - C_{i,j}$, this is also a bound for the primal optimality gap. This allows us to derive strategies which ensure that the algorithm can terminate, in case C, a, b are integer-valued and positive, with at most $O(n^2m^2(\|a\|_1 + \|b\|_1))$ operations. This is of the same order in n, m as the algorithm of [14], yet of course not strongly polynomial as it also depends on the amplitude of the values of a, b. In practice, the running time seems excellent,

even if we still need to compare with a few other algorithms for convex minimal cost flows.

Example. The analysis can be modified to solve semi-unbalanced optimal transport problems, as well as problems where the marginals are penalized with a Kullback-Leibler divergence. In particular, one can implement a standard "JKO" scheme to approximate the Wasserstein gradient flow of the entropy. This is probably not a recommended approach to solving the heat equation, yet it delivers very precise results, see Figure 1.

References

[1] Gaetano Agazzotti, Antonin Chambolle, Clément Royer, *Adapted network-simplex algorithms for unbalanced optimal transport*, in preparation.

[2] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin, *Network flows: theory, algorithms, and applications*, Prentice-Hall, Inc., USA, 1993.

[3] Jean-David Benamou, *Numerical resolution of an "unbalanced" mass transport problem*, ESAIM: Mathematical Modelling and Numerical Analysis, **37**(5) (2003), 851–868.

[4] N. Bonneel and J. Digne, *A survey of optimal transport for computer graphics and computer vision*, *Computer Graphics Forum*, **42**(2) (2023) 439–460.

[5] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso, *Unbalanced optimal transport through non-negative penalized linear regression*, Advances in Neural Information Processing Systems, **34** (2021), 23270–23282.

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *Scaling algorithms for unbalanced transport problems*, Math. Comp. **87** (2018), 2563-2609.

[7] Balázs Dezs, Alpár Jüttner, and Péter Kovács, *Lemon – an open source c++ graph template library*, Electron. Notes Theor. Comput. Sci., 264(5) (2011), 23–45.

[8] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer, *Pot: Python optimal transport*, Journal of Machine Learning Research, **22**(78) (2012), 1–8.

[9] Yurii Nesterov, *Lectures on Convex Optimization*, Springer Optimization and Its Applications **137** (2018).

[10] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet, *Optimal Transport with Proximal Splitting*, SIAM J. Imaging Science, **7**:1 (2014).

[11] Gabriel Peyré, Marco Cuturi, et al, *Computational optimal transport: With applications to data science.* Foundations and Trends® in Machine Learning, **11**(5-6) (2019), 355–607.

[12] F. Santambrogio, *Optimal Transport for Applied Mathematicians.* Birkäuser Cham, 2015.

[13] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré, *Sinkhorn divergences for unbalanced optimal transport. arXiv preprint arXiv:1910.12958*, 2019.

[14] László A. Végh, *A strongly polynomial algorithm for a class of minimum-cost flow problems with separable convex objectives*, SIAM J. Comput., **45**(5) (2016), 1729–1761.

Approximating the optimal transport map with flows of control-linear Neural ODEs

Alessandro Scagliotti
(joint work with S. Farinelli)

In this report, we address the problem of approximating optimal transport maps between probability measures on $\mathbb{R}^n$ by means of flows generated by controlled dynamical systems. More precisely, we consider systems of the form

$$\dot{x}(t) = F(x(t))u(t) = \sum_{i=1}^{k} F_i(x(t))u_i(t), \qquad \text{for a.e. } t \in [0,1],$$

where $F_1, \dots, F_k$ are given vector fields and the control u belongs to $L^2([0,1];\mathbb{R}^k)$ and depends only on time. The corresponding flow at final time $t = 1$ defines a map $\Phi_u \colon \mathbb{R}^n \to \mathbb{R}^n$. The family of such maps provides a class of diffeomorphisms that we use to approximate optimal transport maps.

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ be compactly supported probability measures, with μ absolutely continuous. Denote by T the optimal transport map pushing μ forward to ν with respect to the quadratic cost. Our first goal is to investigate whether T can be approximated by maps of the form Φ_u. Under suitable assumptions on the measures and on the vector fields, one can show that T belongs to the closure (in a suitable topology) of the set of flows generated by the system. This relies on controllability properties of the dynamics [1, 2] and on regularity results ensuring that the optimal transport map is a diffeomorphism isotopic to the identity.

A central motivation for this problem comes from applications where the measures μ and ν are not explicitly known. In a data-driven setting, one typically has access only to discrete approximations μ_N and ν_N, for instance, given by empirical samples. In this case, the optimal transport map is not directly available, and one must rely on approximate procedures. A natural strategy consists in first computing an optimal coupling γ_N between μ_N and ν_N, and then reconstructing a transport map from this discrete information.

In this direction, in [5], we introduce a family of optimal control problems defined, for $\beta > 0$, by

$$\mathcal{F}^{N,\beta}(u) \coloneqq \int_{\mathbb{R}^n \times \mathbb{R}^n} |\Phi_u(x) - y|^2 \, d\gamma_N(x,y) + \frac{\beta}{2}\|u\|_{L^2}^2. \tag{1}$$

The first term enforces consistency between the flow Φ_u and the coupling γ_N, while the second term provides regularization and ensures coercivity. minimizers of $\mathcal{F}^{N,\beta}$ yield flows that can be interpreted as approximations of the optimal transport map.

We show that, as $N \to \infty$, the functionals $\mathcal{F}^{N,\beta}$ Γ-converge (with respect to the weak topology of L^2) to a limiting functional of the form

$$\mathcal{F}^{\infty,\beta}(u) \coloneqq \int_{\mathbb{R}^n} |\Phi_u(x) - T(x)|^2 \, d\mu(x) + \frac{\beta}{2}\|u\|_{L^2}^2, \tag{2}$$

where T is the optimal transport map between μ and ν. This result ensures that minimizers of the discrete problems converge to minimizers of the limiting one. Moreover, under the assumption that T can be approximated by admissible flows, the minimizers generate maps that are close to T in $L^2(\mu)$, provided that the regularization parameter β is sufficiently small.

The approach can be naturally interpreted within the framework of generative models. Indeed, the flows Φ_u can be seen as instances of normalizing flows, i.e., invertible transformations used to map a simple reference distribution into a more complex target distribution. In contrast with standard constructions based on deep neural networks, here the transformations are generated by controlled differential equations with controls taking values in a finite-dimensional space. This viewpoint connects optimal transport, control theory, and machine learning, and provides a structured way to design transport maps with desirable properties.

From a computational perspective, the minimization of $\mathcal{F}^{N,\beta}$ can be addressed using tools from optimal control. In particular, one can derive first-order optimality conditions via the Pontryagin Maximum Principle, leading to an iterative numerical scheme for approximating minimizers. This yields a practical method for reconstructing transport maps starting from discrete data. Indeed, we remark that approximating optimal transport maps is particularly relevant in situations where sampling from the target measure ν is difficult, while sampling from μ is inexpensive. In such cases, an approximate transport map allows one to generate samples from ν by pushing forward samples from μ, thus providing an efficient mechanism for generating new observations distributed as ν.

Finally, the approximation of the optimal transport map provided by the minimization of $\mathcal{F}^{\infty,\beta}$ is achieved in the L^2_μ sense. However, the controllability results in [1, 2] ensure that the optimal transport map can in fact be approximated in the stronger C^0_c topology. Bridging this gap is an interesting direction for future research. In this perspective, we plan to exploit risk measures, in particular, Conditional Value-at-Risk (CVaR) [3], which are known to interpolate between risk-neutral formulations (such as mean squared error minimization) and the worst-case-oriented criteria studied in [4].

References

[1] A. Agrachev, A. Sarychev. Control in the spaces of ensembles of points. *SIAM J. Control Optim.*, 58(3): 1579–1596 (2020). doi: 10.1137/19M127304

[2] A. Agrachev, A. Sarychev. Control on the Manifolds of Mappings with a View to the Deep Learning. *J. Dyn. Control Syst.*, 28: 989–1008 (2022). doi: 10.1007/s10883-021-09561-2

[3] D.P. Kouri, T.M. Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM J. Control Optim.*, 26(1): 365–396 (2016). doi: 10.1137/140954556

[4] A. Scagliotti. Minimax problems for ensembles of control-affine systems. *SIAM J. Control Optim.*, 63(1) (2025). doi: 10.1137/24M167531X

[5] A. Scagliotti, S. Farinelli. Normalizing flows as approximations of optimal transport maps via linear-control neural ODEs. *Nonlinear Analysis*, 257: 113811 (2025). doi: 10.1016/j.na.2025.113811

SympFormer: accelerated attention blocks via inertial dynamics on density manifolds

Viktor Stein
(joint work with W. Li and G. Steidl)

Recent work has revealed that the attention mechanism of transformers admits a mean-field interpretation: token updates can be viewed as interacting particle dynamics, and in suitable scaling limits the empirical token distribution evolves according to a nonlinear continuity equation on a space of probability measures [1, 2, 3]. This point of view suggests that attention layers should be understood as particular discretizations of certain variational flows on measure spaces. The goal of the present work is to push this principle from first-order to *accelerated* dynamics.

We start from the continuous-time limit of Nesterov's accelerated gradient descent [4] for minimizing a convex function $F\colon \mathbb{R}^d \to [0,\infty)$,

$$\ddot{x}_t + \alpha_t \dot{x}_t + \nabla F(x_t) = 0.$$

We interpret this ODE as a damped Hamiltonian system, and transfer this idea to the density manifold of smooth probability densities endowed with Wasserstein- or Stein-type metrics, in the spirit of accelerated information gradient flows [5, 6]. This yields a second-order dynamics for a density ρ_t and a cotangent variable Φ_t, with Hamiltonian

$$\mathcal{H}(\rho,\Phi) = \frac{1}{2}\int \Phi\, G_\rho^{-1}[\Phi]\,\mathrm{d}x + \mathcal{F}(\rho),$$

where G_ρ is the metric tensor and $\mathcal{F}$ is the energy induced by the attention mechanism. The resulting evolution provides a geometric notion of accelerated attention on probability spaces.

For *linear attention*, the transformer PDE takes the form of a Stein variational gradient flow associated with the bilinear kernel $k(x,y) = y^\top Ax$ and a quadratic potential energy. In this case, the accelerated flow remains sufficiently explicit to permit a structural analysis. In particular, elliptically contoured distributions are invariant under the dynamics; e.g., for Gaussian or Student-t input laws, the infinite-dimensional PDE reduces to a closed finite-dimensional system for the mean, covariance, and a quadratic momentum potential. This provides one of the few situations where the accelerated mean-field attention dynamics can be described by a tractable closed ODE system.

For *softmax attention*, one recovers the transformer PDE as a generalized Wasserstein gradient flow with a nonlinear mobility depending on the softmax normalization [2]. The accelerated counterpart again leads to a Hamiltonian particle system, but now with a genuinely *non-separable* Hamiltonian: the kinetic term depends on positions and momenta through the attention matrix. This is the main conceptual difference from Euclidean Nesterov-type transformer variants, such as Yuriiformer [7]. In particular, the exact expression for the acceleration is not appended externally to the architecture, but is derived from the geometry of the density evolution underlying self-attention.

A particle discretization of the accelerated PDE yields token dynamics in position–momentum variables. We discuss several low-oracle time discretizations of the damped Hamiltonian system, including explicit Euler, conformally symplectic Euler, exponential Euler, and Adams–Bashforth–2 schemes. Since the dominant computational cost lies in evaluating the attention kernel, the implementation is designed to preserve oracle complexity per layer. The resulting architecture, termed *SympFormer*, combines an accelerated attention step with an accelerated MLP step, together with learnable step sizes and a learnable log-linear damping schedule.

The present analysis is partly formal on $\mathbb{R}^d$, but it identifies a mathematically coherent route from attention PDEs to accelerated transformer blocks. Small-scale experiments on TinyStories indicate that these inertial attention blocks can improve validation loss over both vanilla transformers and Yuriiformer-style baselines at fixed oracle complexity, albeit presently at a higher wall-clock cost. This suggests that geometric acceleration on measure spaces may provide a useful design principle for transformer architectures beyond first-order gradient-flow models.

References

[1] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré, *A unified perspective on the dynamics of deep transformers*, preprint, arXiv:2501.18322, 2025.

[2] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, *A mathematical perspective on transformers*, *Bull. Amer. Math. Soc.* **62** (2025), no. 3, 427–479.

[3] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, *Understanding and improving transformer from a multi-particle dynamic system point of view*, preprint, arXiv:1906.02762, 2019.

[4] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, *J. Mach. Learn. Res.* **17** (2016), Paper No. 153, 1–43.

[5] A. Taghvaei and P. Mehta, *Accelerated flow for probability distributions*, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97, 6076–6085, 2019.

[6] Y. Wang and W. Li, *Accelerated information gradient flow*, *J. Sci. Comput.* **90** (2022), Paper No. 12.

[7] A. Zimin, Y. Polyanskiy, and P. Rigollet, *Yuriiformer: A suite of Nesterov-accelerated transformers*, preprint, arXiv:2601.23236, 2026.

[8] V. Stein, W. Li, and G. Steidl, *SympFormer: Accelerated attention blocks via inertial dynamics on density manifolds*, preprint, arXiv:2603.16535, 2026.

The Wasserstein geometry of random measures

Alessandro Pinzi

(joint work with G. Savaré)

One of the features of the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is its Riemannian-like structure. Indeed, 2-absolutely continuous curves $(\mu_t)_{t\in[0,1]} \in AC^2([0,1], \mathcal{P}_2(\mathbb{R}^d))$ can be characterized as solution to the *continuity equation* $\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0$, intended in the distributional sense, for some Borel measurable vector field $v_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ that is in $L^2(\mu_t \otimes dt)$. This characterization leads to the celebrated Benamou–Brenier formula and to the definition of a tangent bundle,

which can be characterized as vector fields of minimal velocity, representing the infinitesimal behavior of absolutely continuous curves passing through μ. In [2], we reproduced these kinds of results in the Wasserstein space of random measures $\big(\mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d)), \mathcal{W}_p\big)$, for $p \in (1, +\infty)$.

Random measures. By random measure, we mean a probability measure over probability measures, that is $M \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$, where the space $\mathcal{P}(\mathbb{R}^d)$ is endowed with the narrow topology and its induced σ-algebra. The Wasserstein space $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is a Polish metric space, so that the Wasserstein space of random measures $\big(\mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d)), \mathcal{W}_p\big)$ is defined simply by iterating the Wasserstein construction (indeed, it is usually called Wasserstein-on-Wasserstein distance).

Continuity equation for random measures. Given a continuous (with respect to the narrow topology) curve of random measures $(M_t)_{t\in[0,1]} \subset \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$, to give a meaning to the continuity equation

$$\partial_t M_t + \operatorname{div}(b_t M_t) = 0, \tag{1}$$

we need to understand two main things: what is a vector field in this setting? In which sense is the equation intended? The natural vector fields to consider here are the *non-local* ones, that have the form $b : [0,1] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$, for which the mild integrability assumption (and thus measurable)

$$\int_0^1 \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)| d\mu(x) dM_t(\mu) < +\infty \tag{2}$$

is always assumed (for a slightly more general integrability assumption, see [3]). Then, the equation must be tested against *cylinder functions*.

Definition. We say that $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is a cylinder function, we write $F \in \mathrm{Cyl}$, if there exists $k \geq 1$, $\phi_1, \dots, \phi_k \in C_c^\infty(\mathbb{R}^d)$, $\Psi \in C_c^\infty(\mathbb{R}^k)$ such that

$$F(\mu) = \Psi\left(\int \phi_1(x) d\mu(x), \dots, \int \phi_k(x) d\mu(x)\right).$$

Its Wasserstein gradient is given by

$$\nabla_W F(x,\mu) = \sum_{i=1}^k \partial_i \Psi\left(\int \phi_1(x) d\mu(x), \dots, \int \phi_k(x) d\mu(x)\right) \nabla \phi_i(x),$$

for all $(x,\mu) \in \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$. Then (1) is satisfied if for all $F \in \mathrm{Cyl}$ it holds

$$\frac{d}{dt} \int_{\mathcal{P}(\mathbb{R}^d)} F(\mu) dM_t(\mu) = \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \nabla_W F(x,\mu) \cdot b_t(x,\mu) d\mu(x) dM_t(\mu),$$

in the sense of distributions of $[0,1]$.

Superposition principles for random measures. The characterization of the geometry of the Wasserstein space of random measures is a consequence of two *nested superposition principles* for curves of random measures: one for absolutely continuous curves (in the spirit of the result by S. Lisini [4]) and one for curves that solve the continuity equation (see [1, Theorem 8.2.1] for the classic result).

Theorem. For $p \in (1, +\infty)$ and $(M_t)_{t\in[0,1]} \in AC^p([0,1], \mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d)))$, it holds:

(1) there exists $\Lambda \in \mathcal{P}(AC^p([0,1], \mathcal{P}_p(\mathbb{R}^d)))$ with time-marginals M_t such that

$$|\dot{M}|^p_{\mathcal{W}_p}(t) = \int |\dot{\mu}|^p_{W_p}(t)\, d\Lambda\big((\mu_s)_{s\in[0,1]}\big) \quad \text{for a.e. } t;$$

(2) there exists $\mathfrak{L} \in \mathcal{P}(\mathcal{P}(AC^p([0,1], \mathbb{R}^d)))$ with time-marginals M_t, that is $(e_t)_{\sharp\sharp}\mathfrak{L} = M_t$ for all $t \in [0,1]$, and such that

$$|\dot{M}|^p_{\mathcal{W}_p}(t) = \int\int |\dot{\boldsymbol{x}}|^p(t)\, d\lambda\big((\boldsymbol{x}_s)_{s\in[0,1]}\big)\, d\mathfrak{L}(\lambda) \quad \text{for a.e. } t.$$

Theorem. Let $b : [0,1] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$. For any curve $(M_t)_{t\in[0,1]} \in C([0,1], \mathcal{P}(\mathcal{P}(\mathbb{R}^d)))$ that solves $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(b_t M_t) = 0$, and (2), we have

(1) there exists $\Lambda \in \mathcal{P}(C([0,1], \mathcal{P}(\mathbb{R}^d)))$ with time-marginals M_t and such that it is concentrated over curves of measures $(\mu_t)_{t\in[0,1]}$ that are solutions of the **non-local continuity equation**

$$\partial_t \mu_t + \mathrm{div}(b_t(\cdot, \mu_t)\mu_t) = 0;$$

(2) there exists $\mathfrak{L} \in \mathcal{P}(\mathcal{P}(C([0,1], \mathbb{R}^d)))$ with time-marginals M_t and such that it is concentrated over laws of processes $\lambda \in \mathcal{P}(C([0,1], \mathbb{R}^d))$ which, in turn, are concentrated over absolutely continuous curves that solves the **interacting particle system**

$$\dot{\boldsymbol{x}}_t = b_t(\boldsymbol{x}_t, (e_t)_\sharp \lambda).$$

Geometry of the Wasserstein space of random measures. The previous superposition results allow us to prove the wanted geometric features of the L^p-Wasserstein space of random measures for $p \in (1, +\infty)$:

- (**Characterization of absolutely continuous curves**) Assume that $M_0 \in \mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d))$ and $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(b_t M_t) = 0$ with

$$\int_0^1 \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^p d\mu(x) dM_t(\mu) dt < +\infty.$$

Then $(M_t)_{t\in[0,1]} \in AC^p([0,1], \mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d)))$ and

$$|\dot{M}|^p_{\mathcal{W}_p}(t) \le \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^p d\mu(x) dM_t(\mu) \quad \text{for a.e. } t.$$

Vice versa, if $(M_t)_{t\in[0,1]} \in AC^p([0,1], \mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d)))$, there exists $b : [0,1] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ for which it holds $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(b_t M_t) = 0$ and

$$\int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^p d\mu(x) dM_t(\mu) \le |\dot{M}|^p_{\mathcal{W}_p}(t) \quad \text{for a.e. } t.$$

- (**Benamou–Brenier formula**) For all $M_0, M_1 \in \mathcal{P}_p(\mathcal{P}_p(\mathbb{R}^d))$ it holds

$$\mathcal{W}_p^p(M_0, M_1) = \min\Big\{ \int_0^1 \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^p d\mu(x) dM_t(\mu) dt \ : \\ \partial_t M_t + \mathrm{div}_{\mathcal{P}}(b_t M_t) = 0 \Big\}.$$

- (**Tangent space**) Let $p = 2$ for simplicity, and $M \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$:

$$\mathrm{Tan}_M \, \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) := \overline{\Big\{ \nabla_W F(x,\mu) \ : \ F \in \mathrm{Cyl}(\mathcal{P}(\mathbb{R}^d)) \Big\}}^{L^2(\widetilde{M};\mathbb{R}^d)} \subseteq L^2(\widetilde{M};\mathbb{R}^d);$$

where $\widetilde{M} \in \mathcal{P}(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d))$ is defined as $\widetilde{M} := \int_{\mathcal{P}(\mathbb{R}^d)} \mu \otimes \delta_\mu \, dM(\mu)$.

Let $b : [0,1] \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ and $(M_t) \in C([0,1]; \mathcal{P}(\mathcal{P}(\mathbb{R}^d)))$ such that

$$\int_0^1 \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^2 d\mu(x) dM_t(\mu) dt < +\infty \quad \text{and} \quad \partial_t M_t + \mathrm{div}_{\mathcal{P}}(b_t M_t) = 0.$$

Then, for a.e. $t \in [0,1]$

$$b_t \in \mathrm{Tan}_{M_t} \, \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) \iff |\dot{M}|^2_{\mathcal{W}_2}(t) = \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} |b_t(x,\mu)|^2 d\mu(x) dM_t(\mu).$$

References

[1] L. Ambrosio, N. Gigli, G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Birkhäuser Verlag, Basel, 2nd edition (2008).

[2] A. Pinzi, G. Savaré, *Nested superposition principle for random measures and the geometry of the Wasserstein on Wasserstein space*, arXiv preprint (2025).

[3] A. Pinzi, *First order equation on random measures as superposition of weak solutions to the McKean–Vlasov equation*, arXiv preprint (2025).

[4] S. Lisini, *Characterization of absolutely continuous curves in Wasserstein spaces*, Calc. Var. Partial Differential Equations **28** (2008).

Stochastic Interpolants and Generalizations

Eric Vanden-Eijnden

(joint work with M. Albergo and N. M. Boffi)

Generative modeling asks for a principled way to transport a tractable reference distribution μ_0 (e.g. a Gaussian) to a complex target μ_1 available only through data by learning a map or a flow. We described three approaches to this problem: the *stochastic interpolant* framework [1, 2], a unified construction that subsumes score-based diffusion models and flow matching as special cases; *equilibrium matching* [4], which replaces the time-dependent interpolant flow by a single time-independent velocity satisfying a divergence condition, enabling both a training-free instantiation (Poisson flow) and a learned one, and particularly suited to Boltzmann sampling; and *drifting* [5], which learns a generative map by functional descent on a McKean–Vlasov equation whose drift evolves the pushforward toward the target distribution.

Stochastic interpolants. Given samples $x_0 \sim \mu_0$ and $x_1 \sim \mu_1$, the stochastic interpolant is the process

$$I_t \;=\; \alpha_t\, x_0 + \beta_t\, x_1, \qquad t \in [0,1],$$

where (α_t, β_t) are smooth interpolating coefficients satisfying $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = 0$. The marginal law μ_t of I_t continuously deforms μ_0 into μ_1 and satisfies the continuity equation $\partial_t \mu_t + \nabla \cdot (b_t \mu_t) = 0$ for a velocity field b_t that can be expressed as a conditional expectation and learned by minimizing the quadratic loss

$$\mathcal{L}(b) \;=\; \int_0^1 \mathbb{E}\big[\, |b_t(I_t) - \dot{I}_t|^2 \,\big]\, dt.$$

The two generalizations below exploit this structure in different ways.

Equilibrium matching and the divergence condition. Suppose that the target μ_1 may be singular (e.g. supported on a data manifold or a set of atoms). Then any *time-independent* velocity field b satisfying the distributional divergence condition

$$\nabla \cdot (\nu\, b) \;=\; \mu_0 - \mu_1$$

for some positive weight $\nu > 0$ transports μ_0 onto μ_1: the flow $\dot{X}_t = b(X_t)$ with $X_0 \sim \mu_0$ satisfies $(X_\tau)_{\#}\mu_0 = \mu_1$, where τ is the hitting time of $\mathrm{supp}(\mu_1)$. This follows from the divergence theorem applied to each basin of attraction and does not require the velocity to be a gradient or to have any special structure beyond the divergence condition.

This unifies two apparently different methods as instantiations for specific choices of ν:

- **Poisson Flow ($\nu = 1$) [3]:** $b = \nabla\phi$ where $\Delta\phi = \mu_0 - \mu_1$ is the Poisson equation. The resulting b is the electric field generated by μ_1 in the background of μ_0, computable without any training from fresh mini-batches of both distributions.
- **Flow-matching-derived velocity ($\nu = \nu_{\mathrm{FM}}$) [4]:** Integrating the FM continuity equation $\partial_t \mu_t^{\mathrm{FM}} + \nabla \cdot (b_t^{\mathrm{FM}} \mu_t^{\mathrm{FM}}) = 0$ over $t \in [0,1]$ yields a time-independent velocity $b_{\mathrm{FM}} = j/\nu_{\mathrm{FM}}$ where $j = \int_0^1 b_t^{\mathrm{FM}} \mu_t^{\mathrm{FM}}\, dt$ and $\nu_{\mathrm{FM}} = \int_0^1 \mu_t^{\mathrm{FM}}\, dt$. This b_{FM} automatically satisfies $\nabla \cdot (\nu_{\mathrm{FM}}\, b_{\mathrm{FM}}) = \mu_0 - \mu_1$, and can be learned from data by a standard regression loss.

Drifting via McKean–Vlasov flows. This method was proposed in [5] for learning a generative map G_θ. Here we give drifting a theoretical foundation via *functional descent*. The starting point is the McKean–Vlasov equation

$$\partial_t \mu_t + \nabla \cdot \big(v[\mu_t, \mu_*]\, \mu_t\big) = 0, \qquad \mu_0 \text{ given},$$

where the drift $v[\mu_t, \mu_*]$ *depends on the current law* μ_t and is designed so that $\mu_t \to \mu_*$ as $t \to \infty$; it is estimable from fresh samples of μ_t and μ_* without access to densities, using e.g. kernel or optimal-transport estimators. Rather than simulating this self-interacting particle system and then distilling its trajectories

into a map via flow matching — which incurs an unavoidable approximation floor from finite particle counts — we learn G_θ directly via the fixed-point loss:

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{x_0\sim\mu_0}\Big[\ell\Big(G_\theta(x_0),\ \mathrm{stopgrad}\big(\Phi[\hat{\mu}_\theta^N,\hat{\mu}_*^M](G_\theta(x_0))\big)\Big)\Big],$$

where $\hat{\mu}_\theta^N = \frac{1}{N}\sum_i \delta_{G_\theta(x_0^i)}$ is the empirical pushforward, and $\Phi[\mu,\mu_*](x) = x + h\,v[\mu,\mu_*](x)$ is one Euler step of the McKean–Vlasov drift, and $\ell(x,y)$ is a discrepancy appropriate to the geometry of the output space (e.g. $\|x-y\|^2$ for Euclidean outputs, cross-entropy for discrete targets), and both batches are redrawn at every gradient step. The stopgrad means Φ is never differentiated through, and the self-consistency condition $\mu_\theta = (G_\theta)_\#\mu_0$ is enforced implicitly at each step: the drift is always evaluated at the *current* generated distribution, so the map and its pushforward co-evolve without requiring an offline particle simulation.

Under mild assumptions on Φ (a fixed-point condition and contractivity toward μ_* in the ℓ-transport cost), the iteration converges with no hard approximation floor: resampling both batches at every step causes estimation errors to self-average, so accuracy improves indefinitely with training regardless of batch size. Concrete instantiations include MMD gradient flows, Nadaraya–Watson kernel transport, and entropic optimal transport, each verified to satisfy the fixed-point and attractivity conditions.

References

[1] M. Albergo, E. Vanden-Eijnden, *Building Normalizing Flows with Stochastic Interpolants*, International Conference on Learning Representations (ICLR), 2023. arXiv:2209.15571.

[2] M. Albergo, N. M. Boffi, E. Vanden-Eijnden, *Stochastic Interpolants: A Unifying Framework for Flows and Diffusions*, preprint, 2023. arXiv:2303.08797.

[3] Y. Xu, Z. Liu, M. Tegmark, T. Jaakkola, *Poisson Flow Generative Models*, Advances in Neural Information Processing Systems (NeurIPS) 35, 2022. arXiv:2209.11178.

[4] R. Wang, Y. Du, *Equilibrium Matching: Generative modeling with Implicit Energy-Based Models*, preprint, 2025. arXiv:2510.02300.

[5] M. Deng, H. Li, T. Li, Y. Du, K. He, *Generative modeling via Drifting*, preprint, 2026. arXiv:2602.04770.

A McKean-Pontryagin formulation for entropic-regularized optimal transport

Sebastian Reich

1. Problem statement

We wish to bridge two given distributions π_0 and π_T in a variable $x \in \mathbb{R}^d$ along the controlled stochastic differential equation

$$\mathrm{d}\tilde{X}_t = U_t\mathrm{d}t + \sqrt{2}\Sigma^{1/2}\mathrm{d}\tilde{B}_t, \quad \tilde{X}_0 \sim \pi_0, \quad \tilde{X}_t \sim \pi_T, \tag{1}$$

subject to minimizing the cost

$$\mathcal{J}(U) = \frac{1}{2}\int_0^T \mathbb{E}_{\tilde{X}_t}\|U_t\|_R^2\,\mathrm{d}t \tag{2}$$

in the control $U = U_{[0,T]}$ for weighted norm

$$\|u\|_R^2 = u^{\mathrm{T}} R^{-1} u = R^{-1} : uu^{\mathrm{T}} \tag{3}$$

and given a symmetric positive definite matrix R. The problem reduces to the classical optimal transport problem for $\Sigma = 0$ and $R = I$ and to the Schrödinger bridge problem for $\Sigma = R = I$ [1].

While the Schrödinger bridge problem is well investigated [1], this presentation summarizes an alternative mean-field formulation for the general coupling problem (1)-(2) based on the classical Pontryagin maximum principle [2]. The proposed formulation extends previous joint work with Manfred Opper [3].

2. McKean–Pontryagin variational formulation

We apply the McKean–Pontryagin variational formulation from [3] with states $X_t(a) \in \mathbb{R}^d$ and co-states $P_t(a) \in \mathbb{R}^d$. The distribution of $X_t(a)$ is denoted by ρ_t. Here, $a \in \mathbb{R}^d$ are labels that remain constant and serve as independent variables. The proposed variational formulation will ensure that X_t and $\tilde{X}_t$ agree in law for all $t \in [0, T]$.

In a first step, we introduce the action functional $\mathcal{S} = \mathcal{S}(X, P, U, \beta, \psi)$ defined by

$$\mathcal{S} = \int_0^T \int_{\mathbb{R}^d} \left(\langle P_t, \dot{X}_t - U_t \rangle - \Sigma : D_x^2 \psi_t(X_t)) + \frac{1}{2} \|U_t\|_R^2 \right) \rho_0(a) \mathrm{d}a \, \mathrm{d}t \tag{4a}$$

$$- \int_0^T \int_{\mathbb{R}^d} \langle P_t - \nabla_x \psi_t(X_t), \beta_t \rangle \, \rho_0 \, \mathrm{d}a \, \mathrm{d}t \tag{4b}$$

$$- \int_{\mathbb{R}^d} \psi_T(X_T) \, \rho_0(a) \mathrm{d}a + \pi_T[\psi_T] - \int_{\mathbb{R}^d} \psi_0(X_0) \, \rho_0(a) \mathrm{d}a + \pi_0[\psi_0]. \tag{4c}$$

with $\langle a, b \rangle = a^{\mathrm{T}} b$. We also introduce the Hamiltonian $\mathcal{H} = \mathcal{H}(X, P, U, \beta, \psi)$

$$\mathcal{H} = \int_{\mathbb{R}^d} \left(\langle P, U \rangle + \Sigma : D_x^2 \psi(X) - \frac{1}{2} \|U\|_R^2 + \langle \beta, P - \nabla_x \psi(X) \rangle \right) \rho_0(a) \mathrm{d}a. \tag{5}$$

In a second step, taking variations of (4) with respect to P_t, X_t, β_t, ψ_t, and U_t, the following constrained evolution equations arise:

$$\dot{X}_t = \frac{\delta \mathcal{H}}{\delta P_t} = U_t + \beta_t, \tag{6a}$$

$$-\dot{P}_t = \frac{\delta \mathcal{H}}{\delta X_t} = \nabla_x (\Sigma : D_x^2 \psi_t(X_t)) - D_x^2 \psi_t(X_t) \, \beta_t, \tag{6b}$$

$$0 = \frac{\delta \mathcal{H}}{\delta \beta_t} = P_t - \nabla_x \psi_t(X_t), \tag{6c}$$

$$0 = \frac{\delta \mathcal{H}}{\delta \psi_t} = \nabla_x \cdot \left(\rho_t(X_t) \left\{ \beta_t + \Sigma \nabla_x \log \rho_t(X_t) \right\} \right), \tag{6d}$$

$$0 = \frac{\delta \mathcal{H}}{\delta U_t} = P_t - R^{-1} U_t \tag{6e}$$

subject to the boundary conditions

$$\rho_0 = \pi_0, \qquad \rho_T = \pi_T, \qquad P_0 = \nabla_x \psi_0, \qquad P_T = \nabla_x \psi_T, \tag{7}$$

which arise from variations with respect to X_0, ψ_0 and X_T, ψ_T, respectively. Here we have assumed for simplicity that $\rho_0(x) = \pi_0(x) > 0$ everywhere. The potentials ψ_0 and ψ_T are unknown and arise from the boundary constraints on the density ρ_t. We also find that β_t is not uniquely determined by (6d); but a natural choice is

$$\beta_t = -\Sigma \nabla_x \log \rho_t(X_t). \tag{8}$$

We recall from [3] that the time evolution of $\psi_t(x)$ is independent of the choice of β_t. The gauge freedom in the choice of β_t arising from (6d) is a consequence of the relabelling symmetry of the McKean–Pontryagin formulation.

The Hamiltonian (5) is preserved and takes the constant value

$$\mathcal{E} = \int_{\mathbb{R}^d} \left(\frac{1}{2} \|U_t\|_R^2 + \Sigma : D_x^2 \psi_t(X_t) \right) \rho_0(a) \mathrm{d}a \tag{9}$$

along solutions of (6).

The optimal transport formulation of Benamou–Brenier is recovered for $\Sigma \to 0$ and $R = I$ [1]. Furthermore, for $\Sigma = R = I$, we have found a (deterministic) mean-field formulation of the dynamic Schrödinger bridge problem. Indeed, the law, $\rho_t(x)$, of X_t satisfies the Fokker–Planck equation

$$\partial_t \rho_t = \mathcal{L}_{u_t}^\dagger \rho_t, \tag{10}$$

while $\psi_t(x)$ satisfies the Hamilton–Jacobi–Bellman equation

$$-\partial_t \psi_t = \mathcal{L}_{u_t} \psi_t - \frac{1}{2} \|u_t\|^2 \tag{11}$$

with generator $\mathcal{L}_{u_t} f = \langle \nabla_x f, u_t \rangle + \Sigma : D_x^2 f$ and control law $u_t(x) = R \nabla_x \psi_t(x)$ [5]. Also note that the potentials ψ_t are directly related to the Schrödinger potentials ϕ_t and $\hat{\phi}_t$ [1] via

$$\hat{\phi}_t = e^{\psi_t}, \qquad \rho_t = \phi_t \hat{\phi}_t. \tag{12}$$

for all $t \in [0, T]$.

3. Closing remarks

We emphasize that the proposed formulation can be easily extended to stochastic differential equations of the form

$$\mathrm{d}\tilde{X}_t = b(\tilde{X}_t)\mathrm{d}t + G U_t \mathrm{d}t + \sqrt{2}\Sigma^{1/2} \mathrm{d}\tilde{B}_t \tag{13}$$

for given $b(x)$ and G and provides an alternative to approaches based on forward-backward stochastic differential equations [5].

One can also extend the variational formulation to multiplicative noise $\Sigma(x)$ and mean-field formulations such as the Kalman–Wasserstein gradient flow [4], which corresponds to $R = \Sigma = C(\rho_t)$ with $C(\rho_t)$ the covariance matrix of $X_t \sim \rho_t$.

Inspired by a talk by Yann Brenier at this workshop on a closely related problem, it would also be of interest to treat Σ as an additional control term with reward $-\gamma\Sigma : \Sigma$, $\gamma > 0$. Our variational approach does not require the symmetric matrix Σ to be positive definite; however, the connection to a stochastic differential equation of the form (13) is no longer straightforward and would result in a potentially highly non-trivial combination of forward and backward stochastic differential equations [5]. We finally mention that a related approach based on replacing Brownian motion by score functions has been proposed in [6]. The main differences are that instead of co-states $P_t = \nabla_x \psi_t(X_t)$, one considers $Y_t = \psi_t(X_t)$ as a dual variable, and that the resulting mean-field evolution equations are therefore not of Hamiltonian nature.

Acknowledgments. This work has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294.

References

[1] Sophia Tang. Foundation of Schrödinger bridges for generative modeling. Technical report, arXiv:2603.18992, 2026.
[2] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko. *The Mathematical Theory of Optimal Processes*. John Wiley & Sons, New York, 1962.
[3] Manfred Opper and Sebastian Reich. On a mean-field Pontryagin minimum principle for stochastic optimal control. Technical report, arXiv:2506.10506, 2025.
[4] Edoardo Calvello, Sebastian Reich, and Andrew M. Stuart. Ensemble Kalman methods: A mean field perspective. *Acta Numerica*, 34:123–291, 2025.
[5] René Carmona. *Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications*. SIAM, Philadelphia, 2016.
[6] M. Zhou, S. Osher, and W. Li. A deep learning algorithm for computing mean field control problems via forward-backward score dynamics. Technical report, arXiv:2401.09547, 2024.

Games & Gradient Flows: Modeling Strategic Behavior

Lauren Conger

(joint work with F. Hoffmann, E. Mazumdar, L. Ratliff, and G. Savaré)

Agents interacting with machine learning algorithms report data that shifts over time; in parallel, algorithms train on this evolving data, resulting in a coupled gradient flow system [2, 3]. We represent this as a two-species system: one species is the algorithm, which updates by training, and the other is a distribution of strategic agents. This is a specific example of the more general n-species setting, where each species aims to minimize its own cost functional and evolves according to gradient descent,

$$\dot{x}_i(t) = -\nabla_{d_i, x_i} F_i(x_1, \dots, x_n), \quad i = 1, \dots, n, \tag{1}$$

where (X_i, d_i) is the metric space for the i-th species, $F_i : X \to \mathbb{R}$ is its cost functional, and $x = [x_1, \dots, x_n] \in X$ is the joint action in metric space (X, d), where $X = \prod_{i=1}^n X_i$ and $d(x, y) = \sqrt{\sum_{i=1}^n d_i^2(x_i, y_i)}$. The notation ∇_{d_i, x_i} is the direction of steepest descent in the d_i metric with respect to x_i. Since the resulting

coupled system is not necessarily itself a gradient flow, new analysis tools are needed. We ask:

1. When does a system of n energies $(F_i)_{i=1}^n$ have a Nash equilibrium?
2. When is the coupled gradient flow system (1) well-posed?
3. When does the solution to (1) converge to the Nash equilibrium of $(F_i)_{i=1}^n$?

We present classical results from the Euclidean setting, followed by new results in the Wasserstein-2 space [3, 4] and in general metric spaces. A key condition for all three questions is *monotonicity*. To build intuition, consider finite dimensions where $X_i = \mathbb{R}^{d_i}$ and $X = \mathbb{R}^d$ equipped with the 2-norm. The cost functionals $(F_i)_{i=1}^n$ are λ-*monotone* if

$$\sum_{i=1}^n \langle \nabla_{x_i} F_i(x) - \nabla_{x_i} F_i(y),\, x_i - y_i \rangle \geq \lambda \|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^d .$$

When $\lambda > 0$, the Lyapunov functions $V_1(x) = \frac{1}{2}\sum_{i=1}^n \|\nabla_{x_i} F_i(x)\|^2$ and, if a Nash equilibrium x^* of $(F_i)_{i=1}^n$ exists, $V_2(x) = \frac{1}{2}\|x - x^*\|^2$, both decay exponentially with rate 2λ along solutions to (1).

In the Wasserstein-2 setting, $X_i = \mathcal{P}_2(\mathbb{R}^{d_i})$, $X = \overline{\mathcal{P}}_2$, $d_i = \mathscr{W}_2$ the Wasserstein-2 metric, $d = \overline{\mathscr{W}}$ and the dynamics (1) are

$$\dot{\rho}_i(t) = -\nabla_{W_2,\rho_i} F_i(\rho_1, \ldots, \rho_n) = \operatorname{div}(\rho_i \nabla_{x_i} \delta_{\rho_i} F_i(\rho)) , \quad i = 1, \ldots, n . \tag{2}$$

In [4], we extend monotonicity to this setting: the functionals $\{F_i : \mathcal{A} \subset \overline{\mathcal{P}}_2 \to \mathbb{R}\}_{i=1}^n$ are λ-*monotone* in $\mathcal{A}$ if, for any $\rho, \mu \in \mathcal{A}$ and all Wasserstein-2 optimal couplings $\gamma \in \Gamma^*(\rho, \mu)$,

$$\sum_{i=1}^n \int \langle x_i - y_i,\, \nabla_{x_i} \delta_{\rho_i} F_i[\rho](x_i) - \nabla_{x_i} \delta_{\rho_i} F_i[\mu](y_i) \rangle \, d\gamma_i(x_i, y_i) \geq \lambda\, \overline{\mathscr{W}}^2(\rho, \mu) .$$

This definition of monotonicity enables us to prove results about the long-time behavior of (2), summarized in the following theorem.

Theorem 1 (Convergence [4])**.** *Let (F_i) be λ-monotone in $\mathcal{A}$ with $\lambda > 0$, and let $\rho(t) \in \mathcal{A}$ evolve according to (2). Then*

(a) *there exists a unique steady state ρ^∞ of (2), and $\rho(t)$ converges exponentially to ρ^∞ in $\overline{\mathscr{W}}$ with rate λ; and*
(b) *if the energies F_i are lower semicontinuous, then ρ^∞ is the unique Nash equilibrium in $\mathcal{A}$.*

This brings us to the second question: when are solutions to (1) well-posed? The rest of the results are from current work in collaboration with F. Hoffmann and G. Savaré. We construct solutions via a time-discrete scheme. Two natural options arise for how each species updates toward the steepest descent of F_i. The *partially implicit* scheme minimizes F_i against the previous step of all other species:

$$X_{i,\tau}^k = \operatorname{argmin}_{x_i \in X_i} \left\{ F_i\big(x_i, X_{-i,\tau}^{k-1}\big) + \tfrac{1}{2\tau}\, d_i^2\big(x_i, X_{i,\tau}^{k-1}\big) \right\},$$

where $x_{-i} \coloneqq [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$. This scheme requires upper Lipschitz bounds on the energy gradients for stability, as in [5]. The *variational movement*

scheme (VMS) instead finds, at each step, a Nash equilibrium of the penalized energies:

$$\text{(VMS)} \qquad X_\tau^k \text{ is a Nash eq. of } \left\{ x \mapsto F_i(x_i, x_{-i}) + \tfrac{1}{2\tau}\, d_i^2\big(x_i, X_{i,\tau}^{k-1}\big) \right\}_{i=1}^n .$$

The VMS generalizes the minimizing movement scheme [1] to the multispecies setting: choosing $n = 1$ reduces the Nash equilibrium problem to a single minimization, recovering the standard JKO argmin scheme.

The sequence (X_τ^k) generated by (VMS) satisfies a discrete evolution variational inequality (EVI). To pass to continuous time to obtain a continuous EVI, we construct the piecewise constant interpolant $\overline{X}_\tau(t) = X_\tau^k$ for $t \in ((k-1)\tau, k\tau]$ and establish a Cauchy estimate $d(\overline{X}_{\tau_1}(t), \overline{X}_{\tau_2}(t)) \to 0$ as $|\tau_1 - \tau_2| \to 0$. This guarantees the existence of a limit curve $x(t) = \lim_{\tau \to 0} \overline{X}_\tau(t)$, which satisfies the *evolution variational inequality*

$$\text{(EVI)} \qquad \tfrac{1}{2}\tfrac{d}{dt}\, d^2(x(t), y) + \tfrac{\kappa}{2}\, d^2(x(t), y) \le b(y, x(t)) , \quad \text{for all } y \in X ,\ t \ge 0 ,$$

with the bifunction $b(y, x) = \sum_{i=1}^n F_i(y_i, y_{-i}) - F_i(x_i, y_{-i})$. The single-species choice $b(y, x) = \varphi(y) - \varphi(x)$ reduces (EVI) to the classical gradient flow EVI of [1]. Existence of solutions to (VMS) and existence of such a curve $x(t)$ follow from two conditions on b, along with appropriate continuity and metric space properties:

- *Barycentric κ-convexity*: $y \mapsto b(y, x)$ is κ-convex along some class of multipoint interpolations in X for some $\kappa \in \mathbb{R}$.
- *η-interaction dissipativity*: $b(x, y) + b(y, x) \le \eta\, d^2(x, y)$ for all $x, y \in X$.

The barycentric κ-convexity and η-interaction dissipativity together imply $\lambda = \kappa - \eta$ monotonicity of $(F_i)_{i=1}^n$, and any two solutions to (1) satisfy

$$d(x(t), y(t)) \le e^{-\lambda(t-s)}\, d(x(s), y(s)) \quad \text{for all } t \ge s > 0 .$$

Thus, we have established conditions under which solutions to (1) exist and we characterize the long-time behavior with the distance estimate. When additionally (X, d) is complete and $\lambda > 0$, a unique steady state exists and coincides with the Nash equilibrium of $(F_i)_{i=1}^n$.

References

[1] L. Ambrosio, N. Gigli, G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Birkhäuser, Basel, 2008.

[2] L. Conger, F. Hoffmann, E. Mazumdar, L. Ratliff, *Strategic Distribution Shift of Interacting Agents via Coupled Gradient Flows*, Advances in Neural Information Processing Systems **36** (2023), pp. 45971–46006.

[3] L. Conger, F. Hoffmann, E. Mazumdar, L. J. Ratliff, *Coupled Wasserstein Gradient Flows for Min-Max and Cooperative Games*, arXiv:2411.07403 (2024).

[4] L. Conger, F. Hoffmann, E. Mazumdar, L. J. Ratliff, *Monotone Multispecies Flows*, arXiv:2506.22947 (2025).

[5] M. Di Francesco, S. Fagioli, *Measure Solutions for Non-Local Interaction PDEs with Two Species*, Nonlinearity **26** (2013), pp. 2777–2808.

Mean-Field Limits for Interacting Particle Systems on Weighted Graphs

Nathalie Ayi
(joint work with N. P. Duteil and D. Poyato)

Transformers have become the dominant architecture in modern machine learning. They can be reinterpreted as interacting particle systems, where the successive layers of a neural network correspond to time discretizations of an underlying dynamical system [3]. Many models rely on full attention mechanisms, in which each token interacts with every other token. In contrast, causal attention mechanisms restrict each token to interact only with preceding tokens [6]. In this setting, the interacting particle system can be naturally formulated on a graph. In this talk, we explain how concepts from graph theory can be used to extend classical mean-field limit results, originally developed for systems of indistinguishable particles, to non-exchangeable settings. This approach paves the way for rigorous derivations of mean-field limits in the context of causal attention, extending existing results obtained for full attention mechanisms.

The framework we adopt originates from social dynamics for which a general model can be written as

$$\frac{d}{dt}x_i^N(t) = \frac{1}{N}\sum_{j=1}^{N} w_{ij}^N\, \phi(x_i^N(t), x_j^N(t)),$$

where $x_i^N \in \mathbb{R}^d$ is the state variable, which will represent the opinion, the position of agent i among a population of N agents and $w_{ij}^N \in \mathbb{R}$ is the interaction coefficient. These equations can be reinterpreted as a system of ODEs posed on the weighted graph $G_N =< V(G_N), E(G_N), w^N >$ where the vertices $V(G_N)$ are $\{1, \dots, N\}$, the edges $E(G_N)$ are $\{1, \dots, N\}^2$ and $w_N = (w_{ij}^N)_{1\leq i,j\leq N}$ is the matrix which attributes to each edge (i, j) the weight w_{ij}^N.

Graph theory has undergone significant progress in recent years, see [9], and has proven to be particularly useful for addressing mean-field limit problems for non-exchangeable particle systems. Indeed, it provides a natural framework to define the convergence of a sequence of graphs by relating it to a fundamental object known as a graphon, that is, a symmetric function $w \in L_+^\infty([0, 1]^2)$. A number of works have built upon this approach in various settings, see [2, 4, 5, 7, 8] for instance, and established the rigorous derivation of the mean-field limit toward the following Vlasov-type equation:

$$\partial_t \mu_t^\xi(x) + \nabla_x \cdot \left(\left(\int_I \int_{\mathbb{R}^d} w(\xi, \zeta)\phi(x, y)\mu_t^\zeta(dy)d\zeta \right) \mu_t^\xi(x) \right) = 0$$

where $\mu_t^\xi(x)$ is the probability of finding an agent with identity ξ and opinion x at time t.

In this talk, we focus on a variant of this setting where higher-order interactions are taken into account. This may open new perspectives for studying variants

of causal attention in which a token is compared simultaneously with multiple preceding tokens. In that case, the general model can be written as

$$\begin{cases} \dfrac{dX_i^N(t)}{dt} = \displaystyle\sum_{\ell=1}^{N-1} \sum_{j_1,\dots,j_\ell=1}^{N} w_{ij_1\cdots j_\ell}^{\ell,N} K_\ell(X_i^N(t), X_{j_1}^N(t), \dots, X_{j_\ell}^N(t)), \\ X_i^N(0) = X_{i,0}^N, \qquad i \in \{1, \cdots, N\}. \end{cases}$$

Here, interactions among agents are given as the superposition of all possible $(\ell+1)$-body interactions. The functions $K_\ell = K_\ell(x, x_1, \dots, x_\ell)$ represent the $(\ell+1)$-body interaction kernels, and each weight $w_{ij_1\cdots j_\ell}^{\ell,N}$ describes the underlying $(\ell+1)$-body couplings or connections among agents. In [1], we show how, by exploiting the space of hypergraphons of unbounded rank endowed with the cut distance, we can rigorously establish the mean-field limit and prove the convergence to the corresponding Vlasov-type equation:

$$\begin{cases} \partial_t \mu_t^\xi + \operatorname{div}_x(F_w[\mu_t](\cdot,\xi)\,\mu_t^\xi) = 0, \quad t \geq 0,\, x \in \mathbb{R}^d,\, \xi \in [0,1], \\ \mu_{t=0}^\xi = \mu_0^\xi. \end{cases}$$

where $F_w[\mu_t](x,\xi)$ is defined as

$$\sum_{\ell=1}^{\infty} \int_{[0,1]^\ell} w_\ell(\xi, \xi_1, \dots, \xi_\ell) \left(\int_{\mathbb{R}^{d\ell}} K_\ell(x, x_1, \dots, x_\ell)\, d\mu_t^{\xi_1}(x_1) \cdots d\mu_t^{\xi_\ell}(x_\ell) \right) d\xi_1 \dots d\xi_\ell.$$

References

[1] N. Ayi, N. Pouradier Duteil, D. Poyato, Mean-field limit of non-exchangeable multi-agent systems over hypergraphs with unbounded rank, arXiv:2406.04691 (2024).

[2] H. Chiba, G. S. Medvedev, The mean field analysis of the Kuramoto model on graphs I. The mean field equation and transition point formulas, Discrete Contin. Dyn. Syst. 39(1): 131-155 (2019).

[3] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet, A mathematical perspective on transformers, Bulletin of the American Mathematical Society, 62:427–479 (2025).

[4] P.-E Jabin, D. Poyato, J. Soler, Mean-field limit of non-exchangeable systems, Comm. Pure Appl. Math., (2025).

[5] D. Kaliuzhnyi-Verbovetskyi, G.S. Medvedev, The mean field equation for the Kuramoto model on graph sequences with non-Lipschitz limit, SIAM J. Math. Anal. 50(3): 2441–2465 (2018).

[6] N. Karagodin, Y. Polyanskiy, and P. Rigollet, Clustering in causal attention masking, Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24), 37:115652-115681 (2024).

[7] C. Kuehn, Network dynamics on graphops, New J. Phys., 22(5) (2020).

[8] C. Kuehn, C. Xu, Vlasov equations on digraph measures, J. Differ. Equ., 339: 261–349 (2022).

[9] L. Lovász, Large Networks and Graph Limits, Colloquium Publications (2012).

Accelerated Fixed-point Iteration over Spaces of Probability Measures

Vitalii Aksenov
(joint work with M. Eigel and M. Oster)

Various statistical tasks, such as sampling or computing Wasserstein barycenters, can be reformulated as fixed-point problems for operators on probability distributions. Accelerating standard fixed-point iteration schemes provides a promising novel approach to the design of efficient numerical methods for these problems. The Wasserstein geometry on the space of probability measures allows us to define various useful Riemannian notions, such as tangent spaces, exponential maps, and parallel transport [1], motivating the adaptation of Riemannian numerical methods. This, however, is not straightforward, because the Wasserstein space is not exactly an infinite-dimensional manifold, thus such properties as the isomorphism of tangent spaces or injectivity of the exponential map, do not hold. To sidestep this problem, we propose to consider a more regular subset of measures or tangent vectors.

As a proof of concept, we have studied the Anderson mixing algorithm on the Bures-Wasserstein space of Gaussian measures [2]. We show that in a small enough ball around a measure with a nondegenerate covariance matrix Σ, the space behaves as a closed Riemannian manifold with bounded sectional curvature. We investigate the Riemannian Anderson Mixing (RAM), a well-known fixed-point acceleration algorithm, in this setting. Using the properties of the manifold, we improve the estimates for the radius of local convergence of the method, as given in [3]. We also perform extensive numerical evaluation of the method on various problems, such as Wasserstein Barycenter or Median problems, and report significant acceleration of convergence compared to the previously used Riemannian Gradient Descent [4] and comparable or superior performance in comparison with other methods, such as Riemannian Conjugate Gradient. We also explore different options for the vector transport map, used in the algorithm. We argue that the trivial transport map, i.e., using the same coordinates for each tangent vector in every tangent space, is valid in the sense of the convergence theory of the method, and does not significantly hinder the iteration cost of the method.

As for the general smooth probability densities, we work with Stein geometry, as introduced in [5]. Constraining the tangent vectors to a Reproducible Kernel Hilbert Space (RKHS) yields tractable formulas for approximating scalar products, retraction, and vector transport, using samples from the current measure. Interpreting the Stein Variational Gradient Descent (SVGD) method as a Riemannian gradient descent with respect to the Stein geometry, we can use our approach to construct accelerated sampling algorithms. Numerically, we report accelerated performance in sampling from various model distributions, as well as in the problem of Bayesian regression for neural networks. The theoretical proof of convergence, however, remains an open question.

References

[1] Ambrosio, Luigi, and Gigli, Nicola. *Construction of the parallel transport in the Wasserstein space*, (2008): 1-30.

[2] Aksenov, Vitalii, Martin Eigel, and Mathias Oster. *Anderson Mixing in Bures Wasserstein Space of Gaussian Measures*, arXiv preprint arXiv:2601.22038 (2026).

[3] Li, Zanyu, and Chenglong Bao. *Riemannian Anderson Mixing Methods for Minimizing C 2 Functions on Riemannian Manifolds*, Mathematics of Operations Research (2025).

[4] Altschuler, Jason, et al. *Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent*, Advances in Neural Information Processing Systems 34 (2021): 22132-22145.

[5] Duncan, Andrew, Nikolas Nüsken, and Lukasz Szpruch. *On the geometry of Stein variational gradient descent*, Journal of Machine Learning Research 24.56 (2023): 1–39.

Reporter: Viktor Stein

Participants

Prof. Dr. Andrea Agazzi
Institute for Mathematical Statistics and Actuarial Sciences,
University of Bern
Alpeneggstrasse 22
3012 Bern
SWITZERLAND

Vitalii Aksenov
Weierstraß-Institut für Angewandte Analysis und Stochastik
Anton-Wilhelm-Amo-str. 39
10117 Berlin
GERMANY

Dr. Michael Arbel
INRIA Rhône-Alpes
655 avenue de l'Europe
38334 Montbonnot, St. Ismier, Cedex
FRANCE

Dr. Nathalie Ayi
Laboratoire Jacques Louis Lions
UMR 7598, Campus Jussieu
Sorbonne Université
Institut Universitaire de France
4 Place Jussieu
75005 Paris Cedex
FRANCE

Dr. Krishnakumar Balasubramanian
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
UNITED STATES

Prof. Dr. Yann Brenier
CNRS, Laboratoire de mathématiques d'Orsay (LMO), Université Paris-Saclay
Bâtiment 307, rue Michel Magat
91405 Orsay
FRANCE

Giuseppe Bruno
Institute for Mathematical Statistics and Actuarial Sciences,
University of Bern
Alpeneggstrasse 22
3012 Bern
SWITZERLAND

Prof. Dr. José Antonio Carrillo de la Plata
Mathematical Institute
University of Oxford
Andrew Wiles Building
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG
UNITED KINGDOM

Prof. Dr. Elena Celledoni
Department of Mathematical Sciences
Norwegian University of Science and Technology
A. Getz vei 1
7491 Trondheim
NORWAY

Prof. Dr. Antonin Chambolle
CEREMADE, CNRS
Université Paris-Dauphine,
PSL Research University
Place de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

Dr. Jannis Chemseddine
Institut für Mathematik
Technische Universität Berlin
Sekretariat MA 4-3
Straße des 17. Juni 136
10623 Berlin
GERMANY

Dr. Sinho Chewi
Department of Statistics and Data Science
Yale University
New Haven, CT 06520-8285
UNITED STATES

Lénaïc Chizat
Mathematics institute
EPFL Lausanne
Station 8
1015 Lausanne
SWITZERLAND

Lauren Conger
California Institute of Technology
1200 E California Blvd
91125 Pasadena
UNITED STATES

Prof. Dr. Lorenzo Dello Schiavo
Università degli Studi di Roma
"Tor Vergata"
Viale della Ricerca Scientifica 1
00133 Roma
ITALY

Dr. Richard Duong
Fachbereich Mathematik
TU Berlin
Sekr. Ma 6-4
Straße des 17. Juni 136
10623 Berlin
GERMANY

Dr. Borjan Geshkovski
Inria & Laboratoire Jacques-Louis Lions
Sorbonne Université
4 Rue Jussieu
P.O. Box 15-25-320
Paris 75005
FRANCE

Prof. Dr. Benjamin Gess
Institut für Mathematik
TU Berlin
Str. des 17. Juni 136
10587 Berlin
GERMANY

Dr. Johannes Hertrich
ENS Paris
45 Rue d'Ulm
75005 Paris
FRANCE

Dr. Anna Korba
Department of Statistics
ENSAE/CREST/IP Paris
5 Avenue Henry Le Chatelier
91120 Palaiseau
FRANCE

Hugo Koubbi
CEREMADE
Université Paris Dauphine
Place du Marechal de Lattre deTassigny
75775 Paris Cedex 16
FRANCE

Dr. Théo Lacombe
Laboratoire d'Informatique Gaspard Monge,
Université Gustave Eiffel, CNRS
5, Boulevard Descartes,
Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2
FRANCE

Dr. Hugo Lavenant
Department of Decision Sciences
Universita Bocconi
Via Sarfatti 25
20100 Milano
ITALY

Prof. Dr. Wuchen Li
Department of Mathematics
University of South Carolina
586 Eagles Rest DR, Chapin
Columbia, SC 29208
UNITED STATES

Prof. Dr. Jan Maas
IST Austria
Am Campus 1
3400 Klosterneuburg
AUSTRIA

Levin Maier
Mathematisches Institut
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Hrushikesh N. Mhaskar
Institute of Mathematical Sciences
Claremont Graduate University
1232 N. Dartmouth Avenue
Claremont, CA 91711
UNITED STATES

Prof. Dr. Olga Mula
Institut für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
AUSTRIA

Dr. Kimia Nadjahi
CNRS – ENS Paris
45, rue d'Ulm
75005 Paris
FRANCE

Prof. Dr. Felix Otto
Max-Planck-Institut für Mathematik
in den Naturwissenschaften
Inselstraße 22 - 26
04103 Leipzig
GERMANY

Dr. Alessandro Pinzi
Universita Bocconi
Via Roentgen 1
20136 Milano
ITALY

Prof. Dr. Sebastian Reich
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam
GERMANY

Prof. Dr. Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge MA 02139-4307
UNITED STATES

Prof. Dr. Filippo Santambrogio
Institut Camille Jordan
Université Claude Bernard Lyon 1
43 blvd. du 11 novembre 1918
69622 Villeurbanne Cedex
FRANCE

Prof. Dr. Giuseppe Savaré
Via Roentgen 1
Department of Decision Sciences,
Bocconi University
Via Roentgen 1
Milano 20136
ITALY

Dr. Alessandro Scagliotti
Munich Center for Machine Learning (MCML)
& Technical University of Munich
Boltzmannstr. 3
85748 Garching bei München
GERMANY

Prof. Dr. André Schlichting
Institute of Applied Analysis
Universität Ulm
Helmholtzstr. 18
89081 Ulm
GERMANY

Dr. Anna Shalova
Korteweg-de Vries Instituut
Universiteit van Amsterdam
Postbus 94248
1090 GE Amsterdam
NETHERLANDS

Prof. Dr. Gabriele Steidl
TU Berlin
Institute of Mathematics
Straße des 17. Juni
10623 Berlin
GERMANY

Viktor Stein
Technische Universität Berlin
Institut für Mathematik
FG Angewandte Mathematik
Straße des 17. Juni
10623 Berlin 10623
GERMANY

Dr. Matthew Thorpe
Department of Statistics
University of Warwick
Gibbet Hill Road
Coventry CV4 7AL
UNITED KINGDOM

Dr. Oliver Tse
Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
NETHERLANDS

Prof. Dr. Eric Vanden-Eijnden
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York NY 10012-1110
UNITED STATES

Prof. Dr. François-Xavier Vialard
Laboratoire d'Informatique Gaspard Monge
Université Gustave Eiffel
UMR 8049
77420 Champs-sur-Marne
FRANCE