Mathematisches Forschungsinstitut Oberwolfach

Report No. 8/2025

DOI: 10.4171/OWR/2025/8

Mini-Workshop: Statistical Challenges for Deep Generative Models

Organized by
Sören Christensen, Kiel
Alain Oliviero-Durmus, Palaiseau
Claudia Strauch, Heidelberg
Lukas Trottner, Birmingham

16 February – 21 February 2025

Abstract. Over the last decade, deep generative modelling has emerged as a powerful probabilistic tool in machine learning. The idea behind generative modelling is simple: transform noise to create new data that matches a given training data set. Such transformations must adapt to the information contained in the training data, which is high-dimensional in typical machine learning applications. Generative models, which have demonstrated outstanding empirical generation capabilities for images, videos, text, and many others, have in common that they train deep neural networks to either approximate the transformation directly (e.g., Generative Adversarial Networks) or to approximate the characteristics of a stochastic process that dynamically evolves noise into data (e.g., diffusion models). To explain this empirical success mathematically, we face the statistical task of identifying scenarios in which the distance between the target and generated distributions converges with minimax optimal rate in terms of the sample size as well as the intrinsic dimension and smoothness of the data distribution. While there has been significant progress on this question in rather idealised settings, existing statistical theory is still far from providing a convincing mathematical explanation for why deep generative models perform so well for very different tasks. Due to the complex nature of the field, answering such questions requires a concerted effort from a diverse group of researchers working in probability, nonparametric statistics, functional analysis and optimisation. The aim of this Mini-Workshop was therefore to bring these experts together to foster intensive interactions and to address the statistical challenges posed by generative modelling.

Mathematics Subject Classification (2020): 62G05, 62G07, 62E17, 65C05.

Introduction by the Organizers

The Mini-Workshop *Statistical Challenges for Deep Generative Models*, organized by Sören Christensen (Kiel), Alain Oliviero-Durmus (Palaiseau), Claudia Strauch (Heidelberg) and Lukas Trottner (Birmingham), brought together 18 researchers from across Europe, America and Asia, reflecting a wide geographical and mathematical diversity. The workshop was held in a hybrid format, with two participants joining online and the rest coming to Oberwolfach in person. Over the course of the week, 16 talks covered a broad spectrum of topics related to generative models, with a particular focus on their statistical analysis and sampling guarantees as well as novel probabilistic modelling approaches.

Deep generative models (DGMs) is an umbrella term for a wide variety of model classes following a common underlying generative principle: training deep neural networks to learn a data-driven transformation of easy-to-sample-from noise to generate synthetic data samples for a target distribution that is typically not analytically available but can only be accessed indirectly via a given training sample. Driven by their natural applicability in machine learning, the last decade has produced a huge surge of interest in such models, resulting in a still ongoing process of developing new approaches and refining existing models for specific tasks.

The development of mathematical theory to help explain the empirical success of DGMs has been a comparatively much slower process. Even for the most fundamental classes, such as deep Generative Adversarial Networks (GANs), we lack a definitive mathematical understanding that goes beyond rather idealised settings and that can provide a unified perspective on statistical, probabilistic, and numerical issues arising from such models. A distinctive feature of this workshop was the novel combination of researchers from different but overlapping mathematical disciplines, fostering cross-disciplinary discussions that had not taken place in this form before. This unique constellation provided a fertile ground for exploring state-of-the-art methods and laying the groundwork for future research collaborations. The open and flexible structure of the mini-workshop, which allowed ample time for discussion, proved to be an ideal format for this emerging area of research, where there is not yet a well-established community. As a result, the workshop may well serve as a first step towards the formation of a dedicated research network in this area. Participants, including a significant number of early career researchers at PhD and postdoctoral level, actively engaged in discussions, contributing to the inclusive and dynamic atmosphere of the workshop. The positive feedback from participants repeatedly underlined the enlightening value of the event and its effectiveness in stimulating new knowledge and collaborations. In addition, the MFO's exceptional facilities provided the perfect environment for in-depth discussions and focused research efforts, further enhancing the impact of the workshop.

The central workshop topics can be broadly summarised as follows:

Statistical convergence rates of deep generative models: Minimax optimality of deep generative models under mild assumptions on the data density is a central

statistical question that has been the subject of intense research efforts over the last couple of years. In particular, the statistical implications of the so-called *manifold hypothesis* – which imposes lower-dimensional support assumptions on the data distribution and, at a heuristic level, commonly serves to explain the empirical performance of generative models – has been of great interest. In this context, Judith Rousseau presented recent results on almost minimax optimal convergence rates for score-based diffusion generative models in Wasserstein-1 distance and Kullback–Leibler divergence, in particular improving previous work in this direction by establishing convergence bounds that are independent of the ambient dimension. Eddie Aamari presented related work explaining how Wasserstein GANs can overcome the curse of dimensionality by adapting to the intrinsic dimension of the target data. Mathias Trabs showed how such results for Wasserstein GANs can be extended to Vanilla GANs by establishing compatibility results between the Vanilla GAN and the Wasserstein-1 distance. Lea Kunkel described in her talk how Wasserstein-1 convergence rates for Flow Matching that do not rely on imposing early stopping assumptions on the generative process can be obtained. Claudia Strauch and Lukas Trottner presented first statistical results on reflected generative diffusion models, a class of score-based generative models that allows to incorporate domain constraints by exploiting boundary reflections for both the noising as well as the denoising process.

Sampling guarantees for score-based diffusion generative models: A primary objective of the workshop was to create synergies between researchers working mainly in statistics and experts in sampling methods and their iteration complexities. Regarding the latter, Alain Oliviero-Durmus showed how, given an L^2-approximation error (which is one of the primary concerns of the statistical analysis alluded to above), only assuming finite relative Fisher information of the data distribution with respect to a Gaussian is sufficient to obtain sharp sampling guarantees in Kullback–Leibler divergence in both overdamped and kinetic settings, while also avoiding early stopping restrictions. Yuting Wei discussed discrete-time convergence rates of (deterministic) probability flow ODE and (stochastic) Denoising DDPM type samplers, as well as their accelerated variants. Her results rely on a given control of the ℓ^2-score approximation error and on the Jacobian of the score estimator, but not on additional smoothness assumptions on the absolutely continuous data density. Iskander Azangulov complemented Judith Rousseau's talk on statistical aspects of diffusion generative models under the submanifold hypothesis by showing that their iteration complexity in Kullback–Leibler divergence scales linearly in the *intrinsic* dimension d up to logarithmic factors.

Optimal transport and Schrödinger bridge problems: Several talks were devoted to (entropic) optimal transport problems, especially the Schrödinger Bridge problem (SBP), and their role in generative modelling. Stefano Peluchetti presented his work on Bridge Matching and showed that an iterative bridge matching procedure converges to a solution of the SBP. Denis Belomestny proposed a numerical solution to the SBP for general reference processes based on nonparametric kernel

regression techniques. Giovanni Conforti presented fundamental results on exponential convergence of the Sinkhorn algorithm to the solution of an empirical optimal transport problem. Arthur Stéphanovitch discussed higher order regularity of Langevin transport maps and statistical implications of his results for density estimation with Wasserstein GANs and diffusion generative models.

New directions in denoising generative models: The final main pillar of the workshop was concerned with more fundamental ideas for the design of generative algorithms that build on the concept of time reversal. Sören Christensen presented a novel generative model that preserves the time-homogeneous nature of the forward model and is thus able to dynamically adapt to the noise level present in the generation step. This is achieved via an application of Doob's h-transform to some reference process, which allows conditioning the forward process to be terminated in a suitably chosen sampling distribution at a *random* time. Christian A. Naesseth extended the basic ideas behind diffusion generative modelling in a different direction by introducing Neural Flow Diffusion Models and SDE Matching as a general framework that does not treat the forward process as fixed but as a learnable object. Dario Shariatian discussed in his talk the benefits of using heavy-tailed α-stable noise instead of the usual Gaussian noise in the Denoising Probabilistic Model setup, thus establishing the novel class of Denoising Lévy Probabilistic Models. Focusing on diffusion guidance, Yazid Janati discussed the limitations of existing methods that use diffusion generative models for inverse problems, and proposed instead a new sampling procedure that runs through weighted mixture approximations of intermediate posteriors.

Mini-Workshop: Statistical Challenges for Deep Generative Models

Table of Contents

Alain Oliviero-Durmus (joint with Giovanni Conforti and Marta Gentiloni Silveri)
KL convergence guarantees for score diffusion models under minimal data assumptions 381

Claudia Strauch and Lukas Trottner (joint with Asbjørn Holk)
Statistical guarantees for reflecting denoising diffusion models 382

Sören Christensen (joint with Claudia Strauch and Lukas Trottner)
Against the Flow – Time-Reversed Stochastic Processes and Their Role in Generative Models 384

Denis Belomestny (joint with John Schoenmakers)
Kernel Forward Reverse Regression for Schrödinger bridge problem 385

Giovanni Conforti (joint with Alberto Chiarini, Giacomo Greco and Luca Tamanini)
Semiconcavity of entropic potentials and exponential convergence of Sinkhorn algorithm 386

Judith Rousseau (joint with Iskander Azangulov, G. Deligiannidis)
Convergence rate of diffusion generative models under the manifold hypothesis: the impact of the ambiant dimension 388

Mathias Trabs (joint with Lea Kunkel)
The role of probability distances in generative models: A Vanilla GAN case study 389

Arthur Stéphanovitch
Regularity of (non-optimal) transport maps and applications to SGMs . . 391

Yazid Janati (joint with Badr Moufad, Mehdi Abou El Qassime, Alain Durmus, Eric Moulines and Jimmy Olsson)
A mixture-based framework for guiding diffusion models 392

Christian A. Naesseth (joint with Grigory Bartosh and Dmitry Vetrov)
Neural Flow Diffusion Models and SDE Matching 393

Stefano Peluchetti
Bridge Matching Schrödinger Bridges 394

Dario Shariatian (joint with Alain Durmus, Umut Simsekli)
Denoising Lévy Probabilistic Models: A Heavy-Tailed Diffusion Approach for Generative Modeling 395

Eddie Aamari (joint with Arthur Stéphanovitch and Clément Levrard)
Minimax Optimality of Wasserstein GAN Estimators 397

Yuting Wei (joint with Gen Li, Yu Huang, Timofey Efimov, Yuxin Chen, Yuejie Chi)
Towards Faster Non-asymptotic Convergence for Diffusion-based Generative Models 398

Lea Kunkel (joint with Mathias Trabs)
Flow Matching from a KDE perspective 399

Iskander Azangulov (joint with Peter Potaptchik, George Deligiannidis, Judith Rousseau)
Iteration Complexity of Diffusion Models under the Manifold Assumption 400

Abstracts

KL convergence guarantees for score diffusion models under minimal data assumptions

Alain Oliviero-Durmus
(joint work with Giovanni Conforti and Marta Gentiloni Silveri)

In recent years, deep generative models (DGMs) have emerged as a key area of research in artificial intelligence due to their impressive capabilities. These models aim to learn mappings that generate realistic new data points from a simple prior distribution, typically a standard Gaussian. Well-trained DGMs can approximate complex, high-dimensional probability distributions and serve as proxies for data likelihood estimation.

Among generative models, score-based diffusion generative models (SGMs) have become one of the most influential approaches. SGMs leverage the time-reversal of a diffusion process to transform noise into structured data samples. The first step involves estimating the score function of an ergodic forward diffusion process over a fixed time window $[0, T]$. Once the score function is learned, the second step consists of simulating the time-reversal of the diffusion process to generate samples. To make this step computationally feasible, a time-discretization scheme is introduced, and the backward process is initialized at the invariant distribution of the forward process. Since this distribution is easier to sample from than intermediate steps of the forward process, it facilitates efficient approximation of the target data distribution.

The empirical success of SGMs has fueled significant research efforts to analyze how various sources of error-score approximation, time-discretization, and initialization-impact the quality of generated samples. This has led to a growing body of work aimed at providing theoretical guarantees for SGMs. In particular, there is increasing interest in understanding the sampling phase, which corresponds to the second step of the process, and in deriving rigorous bounds on its effectiveness. Despite recent progress in developing a mathematical theory for diffusion models, there is still no comprehensive quantitative result that provides a priori error bounds on the discrepancy between generated and true data distributions without assuming smoothness conditions (e.g., Lipschitz continuity of the score function or its estimator).

One way to circumvent this limitation is to introduce an early stopping rule, where the backward process is only run until time $T - \delta$. Some recent studies have shown that under minimal assumptions on the data, this approach allows for a quantitative analysis of the gap between the generated samples and the law of the forward process at time δ. This can be seen as a noised (smoothed) version of the target distribution. In this talk, we present a new analysis of the performance of two widely used families of score-based diffusion models, where the forward process is either the Ornstein–Uhlenbeck (OU) diffusion or its kinetic counterpart (kOU), under various assumptions on the data distribution. To approximate the

time-reversed diffusion, we employ the exponential integrator Euler–Maruyama discretization with a constant step size, a widely adopted scheme in prior research on the subject.

Our main result establishes explicit, simple, and sharp bounds on the Kullback–Leibler (KL) divergence between the data distribution and the law of the score-based generative model (SGM) in both overdamped and kinetic settings. These bounds hold under minimal assumptions: (i) an L^2 score approximation error and (ii) the finite relative Fisher information of the data distribution with respect to a standard Gaussian. Unlike previous works that avoid early stopping or exponentially decreasing step sizes [2, 3, 1], our results do not require the data distribution to be supported on a bounded manifold or assume the Lipschitz continuity of the score (or its approximation) over the sampling interval $[0, T]$. Instead, requiring finite Fisher information imposes only a mild integrability condition on the score function.

Moreover, our bounds are sharp: when the data distribution is a standard Gaussian, the only remaining term corresponds to the approximation error of the score function. Additionally, our results match or surpass the accuracy of previously established bounds, providing a significant improvement in the theoretical understanding of score-based diffusion models.

References

[1] H. Chen, H. Lee, and J. Lu. *Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions.* International Conference on Machine Learning, pp. 4735–4763, PMLR, 2023.

[2] H. Lee, J. Lu, and Y. Tan. *Convergence for score-based generative modeling with polynomial complexity.* Advances in Neural Information Processing Systems **35**, pp. 22870–22882, 2022.

[3] H. Lee, J. Lu, and Y. Tan. *Convergence of score-based generative modeling for general data distributions.* International Conference on Algorithmic Learning Theory, pp. 946–985. PMLR, 2023.

Statistical guarantees for reflecting denoising diffusion models

Claudia Strauch and Lukas Trottner
(joint work with Asbjørn Holk)

In the past two years, much progress has been made in the statistics literature in explaining the generative ability of unconstrained denoising diffusion models that use an Ornstein–Uhlenbeck forward process [1, 4, 5]. In practice, however, implementations introduce thresholding procedures for the generative process to overcome performance issues arising from the unbounded state space of such models. To overcome this mismatch between theoretical design and implementation of diffusion models, *reflected* diffusion processes as the driver of noise have instead been suggested and empirically tested in the literature [2, 3].

In this talk, we have presented a first statistical analysis of such reflected diffusion models, focusing explicitly on a forward model governed by the SDE

$$\mathrm{d}X_t = \nabla f(X_t)\,\mathrm{d}t + \sqrt{2f(X_t)}\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t^D,$$

where W is a d-dimensional Brownian motion, $f\colon \mathbb{R}^d \to [f_{\min}, \infty) \subset (0, \infty)$ is a smooth potential, D is an open and bounded domain with smooth boundary, ν is the inward-pointing normal vector field on ∂D, and $\ell_\cdot^D$ is the local time at the boundary. This process is constrained to $\overline{D}$ through normal reflections at the boundary and is analytically characterised by the divergence form generator $\mathcal{A} = \nabla \cdot f\nabla$ subject to Neumann boundary conditions. If is f is constant, this is nothing else but a reflected scaled Brownian motion in $\overline{D}$.

The specific form of the generator implies time-reversibility of the process w.r.t. its uniform stationary distribution and the spectral gap of $\mathcal{A}$ yields an exponential convergence rate. This combination of a fast speed of convergence and an easy-to-sample-from limiting distribution makes this class of processes particularly suitable for generative modelling purposes. Indeed, as with unconstrained SDEs, such reflected SDEs are stable under time reversal in the sense that for some forward running time $\overline{T}$, the process $(X_{\overline{T}-t})_{t\in[0,\overline{T})}$ is a normally reflected diffusion as well. The backward drift is entirely characterised by the potential f and the *score* $s^\circ(x,t) = \nabla \log p_t(x)$, where for the forward transition densities $q_t(x,y)$ and the underlying data distribution p_0, the forward density $p_t(x)$ is given by $p_t(x) = \int_0^t q_t(y,x)\,p_0(\mathrm{d}y)$.

Since the data distribution p_0 is generally unknown, we don't have direct access to the score and therefore need to estimate this space-time function based on the data $(X_{0,i})_{i=1}^n$, which can then be used to simulate the backward generative process initialised in the uniform distribution on $\overline{D}$. As for unconstrained models, training is based on empirical risk minimisation of the *denoising score matching loss* over a suitable class of sparse deep neural networks. The central mathematical challenge is then of an analytical nature and requires an efficient construction of a neural network approximation of the score $s^\circ(x,t)$, which is particularly demanding in our reflected diffusion setting due to the absence of explicit formulae for the transition densities $q_t(x,y)$. For this reason, we discussed how the spectral decomposition of the generator, which for an orthonormal eigensystem $(\lambda_j, e_j)_{j=0}^\infty$ of $\mathcal{A}$ and an absolutely continuous data density yields the score representation

$$s^\circ(x,t) = \frac{\sum_{j=1}^\infty \mathrm{e}^{-\lambda_j t}\langle e_j, p_0\rangle_{L^2}\nabla e_j(x)}{\sum_{j=0}^\infty \mathrm{e}^{-\lambda_j t}\langle e_j, p_0\rangle_{L^2} e_j(x)},$$

can be exploited in the approximation analysis to obtain a neural network class with optimised size constraints that allow to balance the score approximation bias and the stochastic error of the optimisation procedure, encoded in the metric entropy of the neural network class. As our main result, we showed that the combination of a strictly positive lower bound and a Sobolev smoothness assumption on the data density leads to minimax optimal convergence rates up to small log-factors in terms of the expected deviation in total variation of the true and generated data distributions.

References

[1] I. Azangulov, G. Deligiannidis, and J. Rousseau, *Convergence of Diffusion Models Under the Manifold Hypothesis in High-Dimensions*, arXiv preprint arXiv:2409.18804, 2024.

[2] N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. Hutchinson, *Diffusion Models for Constrained Domains*, Transactions on Machine Learning Research, PMLR, 2023.

[3] A. Lou, and S. Ermon, *Reflected Diffusion Models*, International Conference on Machine Learning, pp. 22675–22701, PMLR, 2023.

[4] K. Oko, S. Akiyama, and T. Suzuki, *Diffusion Models are Minimax Optimal Distribution Estimators*, International Conference on Machine Learning, pp. 26517–26582, PMLR, 2023.

[5] R. Tang, and Y. Yang, *Adaptivity of Diffusion Models to Manifold Structures*, International Conference on Artificial Intelligence and Statistics, pp. 1648–1656, PMLR, 2024.

Against the Flow – Time-Reversed Stochastic Processes and Their Role in Generative Models

Sören Christensen

(joint work with Claudia Strauch and Lukas Trottner)

The stochastic theory of time-reversed Markov processes is a well-established area within classical probability theory. While it is well known that a time-reversed Markov process inherits the Markov property, the resulting process is typically time-inhomogeneous when a fixed time horizon is considered. In this talk, we emphasize the advantages of employing random time horizons instead of fixed ones to circumvent this.

A key feature of this framework is its connection to Doob's h-transform, which facilitates process termination based on a suitable sampling distribution at a potentially random time. Although the general theory has been developed for Markov processes in broad terms, as demonstrated in [2], we argue that there is a lack of references applying this theory to specific classes of processes. This talk aims to address this gap by focusing on diffusion processes. Specifically, we conduct an h-transform with respect to a function h, which is an r-potential of the form $h(x) = \int G_r(x,y)\,\kappa(dy)$, where $G_r(x,y)$ is the r-Green kernel and r is a fixed constant. This approach leads to natural results that generalize the well-known formulas for fixed time horizons in a meaningful way.

In particular, we show that the h-transformation of a symmetric diffusion process Z results in another diffusion process Z^h of the form

$$dZ_t^h = b^h(Z_t^h)\,dt + \sigma(Z_t^h)\,dW_t, b^h(y) = b(y) + \sigma(y)\sigma(y)^\top \nabla \log h(y),$$

killed at a finite lifetime ζ. When considering the time-reversed process from the random lifetime ζ, for Z starting from the distribution α, this becomes an $\tilde{h}$-transformation with $\tilde{h}(x) = \int \frac{G_r(x,y)}{h(y)}\,\alpha(dy)$. From a mathematical standpoint, this implies that there are no structural advantages in using fixed time horizons over random ones.

Building on these theoretical insights, we introduce a novel approach to diffusion models by proposing a new class of generative diffusion models. Unlike conventional denoising diffusion models, our model preserves time-homogeneity in both the noising and denoising processes. This time-homogeneous structure allows for adaptive adjustment of the number of steps based on the noise level, leading to a more efficient sampling procedure. Moreover, our model is particularly well-suited for data with lower intrinsic dimensionality, as the termination criterion simplifies to a first-hitting rule. This finding provides a fresh perspective on the manifold hypothesis.

A crucial feature of our model is its adaptability to target data, enabling a variety of downstream tasks with a pre-trained unconditional generative model. These tasks include natural conditioning through the appropriate initialization of the denoising process and classification of noisy data. Our findings have significant potential to impact applications in machine learning and statistics, particularly in scenarios that require efficient and adaptable generative models.

References

[1] S. Christensen, C. Strauch, and L. Trottner, *Beyond Fixed Horizons: A Theoretical Framework for Adaptive Denoising Diffusions*, arXiv preprint arXiv:2501.19373, 2025.

[2] K. L. Chung and J. B. Walsh, *Markov Processes, Brownian Motion, and Time Symmetry*, Grundlehren der mathematischen Wissenschaften **249**, Springer, 2005.

Kernel Forward Reverse Regression for Schrödinger bridge problem

Denis Belomestny

(joint work with John Schoenmakers)

This work deals with the Schrödinger Bridge Problem (SBP), which seeks to construct a Markov process that evolves a given initial distribution into a given terminal distribution in a way that minimizes the relative entropy with respect to some "reference" (or prior) Markov process. The SBP has classical connections to optimal transport, stochastic control, and entropy-regularized probability matching. A key point in SBP is to identify Markov processes that fit the prescribed begin-end distributions. In general, bridging those distributions with minimal relative entropy leads to entropic optimal transport ideas. For a reference Markov process X with transition density $q(0,x;T,z)$, the SBP solution requires finding two "potentials" ν_0 and ν_T such that

$$\rho_0(x) \;=\; \nu_0(x) \int q(0,x;T,z)\,\nu_T(z)\,dz, \quad \rho_T(z) \;=\; \nu_T(z) \int q(0,x;T,z)\,\nu_0(x)\,dx.$$

From these potentials, one builds a Markov measure on paths that achieves the desired boundary conditions and solves the SBP. Existence and uniqueness (up to a scalar factor) of these potentials are well-known from the 1930s (Fortet, Beurling) and by modern fixed-point arguments (e.g., via the Hilbert metric). A popular way to solve the Schrödinger system is to alternate between updating ν_0 and ν_T in a Picard iteration (analogous to the "Sinkhorn algorithm" in discrete entropic

OT). In discrete spaces, this is known as Iterative Proportional Fitting consisting of two steps

- Forward step: updates ν_0 using the current guess of ν_T.
- Reverse step: updates ν_T using the new ν_0.

This iteration is a contraction in the Hilbert projective metric, ensuring convergence. The main innovation of our work is a nonparametric way to solve the SBP numerically in higher dimensions and general reference process X. It is only assumed that one can sample from a reverse process to compute integrals in a backward-time manner (e.g., to evaluate functionals of the form $\int g(x)\, q(0,x;T,\cdot)\, dx$).

(1) Monte Carlo Samples: simulate from the reference forward process (to gather (X_0, X_T) pairs) and from a suitably chosen "reverse" process (to gather analogous backward-time data).
(2) Kernel Regression: approximate the integrals $\mathbb{E}[\rho_T(X_T)\nu_T(X_T) \mid X_0 = x]$ and similar expressions by kernel smoothing. Essentially, one uses Nadaraya–Watson estimators to learn the maps
$$x \;\mapsto\; \int q(0,x;T,z)\, \rho_T(z)\, \nu_T(z)\, dz$$
from samples.
(3) Iterative Updates: after each forward regression, update ν_0. Then perform a reverse regression to update ν_T. Repeat until convergence.

We derive (under regularity assumptions, such as Hölder continuity and boundedness) Nonparametric Convergence Rate in terms of sample size N and bandwidth δ. Specifically, we show an $O\big(N^{-\frac{1+\alpha}{2(1+\alpha)+d}}\big)$ type rate (where α is a Hölder exponent, d the dimension), matching standard kernel regression theory. We also show that these rates are essentially unimprovable under standard nonparametric assumptions.

Semiconcavity of entropic potentials and exponential convergence of Sinkhorn algorithm

Giovanni Conforti

(joint work with Alberto Chiarini, Giacomo Greco, Luca Tamanini)

Given two Polish spaces $\mathcal{X}, \mathcal{Y}$, marginal probability distributions $\rho \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the entropic optimal transport problem (EOT) reads as

$$\text{(EOT)} \qquad \inf_{\pi \in \Pi(\rho,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi + \varepsilon\, H(\pi|\rho \otimes \nu),$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν, H denotes the Kullback–Leibler divergence (also known as relative entropy), and $\varepsilon > 0$ is a regularization parameter. The study of EOT has greatly intensified since the observation [2] that adding an entropic penalty in the objective function of the Monge–Kantorovich problem

(corresponding to $\varepsilon = 0$ in (EOT)) leads to a more convex, more regular, and numerically more tractable optimization task, thus opening new perspectives for the computation of transport distances in machine learning and beyond, see [1]. Much of the success of entropic regularization techniques in applications can be attributed to the fact that EOT can be solved by means of an exponentially-fast matrix scaling algorithm, Sinkhorn's algorithm, and to the fact that EOT is more stable than the Monge–Kantorovich problem with respect to variations in the cost or marginals. Because of this, considerable efforts have been made over the last decade to turn these intuitions into sound mathematical statements. This has produced many important contributions. Nonetheless, several open questions remain. For example, exponential convergence of Sinkhorn algorithm is not well understood when both the marginals' support and the ground cost are unbounded, as it is the case in the landmark example of the quadratic cost with Gaussian marginals. In this talk we present the results of [3], where it is shown that if one is able to show semiconcavity of Sinkhorn potentials, then Sinkhorn's algorithm converges exponentially fast. To introduce Sinkhorn algorithm, we recall that under mild integrability conditions on the cost c, there exist two functions φ^ν, ψ^ν, called entropic potentials, such that the unique optimal plan π^ν for (EOT) admits the Radon–Nikodým density

$$\frac{d\pi^\nu}{d(\rho\otimes\nu)}(x,y) = \exp\left(-\frac{c(x,y)+\varphi^\nu(x)+\psi^\nu(y)}{\varepsilon}\right), \quad \rho\otimes\nu \text{ a.e.} \tag{1}$$

Defining for any $\varphi : \mathcal{X} \to \overline{\mathbb{R}}$ and $\psi : \mathcal{Y} \to \overline{\mathbb{R}}$ the maps

$$\Phi_0^\rho(\varphi)(y) := -\varepsilon \log \int_{\mathcal{X}} \exp\left(-\frac{c(x,y)+\varphi(x)}{\varepsilon}\right)\rho(dx)\,,$$
$$\Psi_0^\nu(\psi)(x) := -\varepsilon \log \int_{\mathcal{Y}} \exp\left(-\frac{c(x,y)+\psi(y)}{\varepsilon}\right)\nu(dy)\,.$$

and imposing that a probability measure of the form (1) belongs to $\Pi(\rho,\nu)$ yields the *Schrödinger system*:

$$\begin{cases} \varphi^\nu = -\Psi_0^\nu(\psi^\nu)\,, \\ \psi^\nu = -\Phi_0^\rho(\varphi^\nu)\,. \end{cases} \tag{2}$$

Sinkhorn's algorithm solves (2) as a fixed point problem. That is, it constructs two sequences of potentials $(\varphi_\varepsilon^n, \psi_\varepsilon^n)$ defined through the iterations

$$\begin{cases} \varphi_\varepsilon^{n+1} &= -\Psi_0^\nu(\psi_\varepsilon^n)\,, \\ \psi_\varepsilon^{n+1} &= -\Phi_0^\rho(\varphi_\varepsilon^{n+1})\,. \end{cases}$$

Typically, the initialization is $\varphi_\varepsilon^0 = 0$, but other choices are possible. We can also associate to Sinkhorn's potentials a sequence of plans $\pi^{n,n}$ as follows

$$\frac{d\pi^\nu}{d(\rho\otimes\nu)}(x,y) = \exp\left(-\frac{c(x,y)+\varphi^n(x)+\psi^n(y)}{\varepsilon}\right), \quad \rho\otimes\nu \text{ a.e.}$$

We establish exponential convergence of the algorithm in the form

$$H(\pi^\nu|\pi^{n,n}) \le \exp(-\lambda n) H(\pi^\nu|\pi^{0,0}) \tag{3}$$

for some $\lambda > 0$, provided Sinkhorn potentials ψ^n and the cost are uniformly semiconcave, and ν satisfies a Talagrand inequality, also known as transport-entropy inequality. Our exponential convergence results are deduced from the following stability results for optimal plans, which is of independent interest.

Theorem (KL stability of optimal plans). *Let π^ν, π^μ denote the unique optimizers in* (EOT) *for the set of marginals (ρ,ν) and (ρ,μ). If there exists $\Lambda > 0$ such that*

$$y \mapsto c(x,y) + \psi^\nu(y)$$

is Λ-semiconcave uniformly in x in the support of ρ, then

$$H(\pi^\mu|\pi^\nu) \le H(\mu|\nu) + \frac{\Lambda}{2\varepsilon} W_2^2(\mu,\nu)\,.$$

For example, if ν is a standard Gaussian distribution and ρ a strongly log-concave probability measure, the convergence rate in (3) can be taken to be

$$\lambda = \frac{\varepsilon}{\varepsilon + \alpha_\rho^{-1/2}}.$$

Moreover, it can be shown that such rate has optimal dependence on ε.

References

[1] M. Cuturi, G. Peyré. *Computational optimal transport: with applications to data science*, Foundations and trends in Machine Learning **11**(5-6), pp. 355–607. , 2019.

[2] M. Cuturi. *Advances in Neural Information Processing Systems*, Advances in Neural Information Processing Systems **26**, 2013.

[3] A. Chiarini, G. Conforti, G. Greco, and L. Tamanini (2024). *A semiconcavity approach to stability of entropic plans and exponential convergence of Sinkhorn's algorithm.* arXiv preprint arXiv:2412.09235, 2024.

Convergence rate of diffusion generative models under the manifold hypothesis: the impact of the ambiant dimension

Judith Rousseau

(joint work with Iskander Azangulov, G. Deligiannidis)

Denoising Diffusion Probabilistic Models (DDPM) are powerful state-of-the-art methods used to generate synthetic data from high-dimensional data distributions and are widely used for image, audio, and video generation as well as many more applications in science and beyond. The *manifold hypothesis* states that high-dimensional data often lie on lower-dimensional manifolds within the ambient space, and is widely believed to hold in provided examples. While recent results have provided invaluable insight into how diffusion models adapt to the manifold hypothesis, they do not capture the great empirical success of these models, making this a very fruitful research direction.

In this work, we study DDPMs under the manifold hypothesis and prove that they achieve rates independent of the ambient dimension in terms of score learning. More precisely assuming that the true generative process μ lives on a smooth manifold M of dimension d, which is unknown. Consider observations $X_1, \cdots, X_n$ according to μ, which is assumed to have a density with respect to the volume measure of the supporting manifold M, which is α holder. Let $X(t), t \leq \bar{T}$ be a realisation of an Ornstein–Uhlenbeck process with initial distribution μ and $Y(t), t \leq \bar{T}$ be the backward process. We also denote the score $s^o(x,t) = \nabla \log p(x,t)$ where $p(x,t)$ is the marginal density of $X(t)$. We show that for well chosen architecture of the deep neural network used to approximate the score, for $d \geq 3$

$$\mathbb{E}\left(\int_{\underline{T}}^{\bar{T}} \sigma_t^2 \|\hat{s}(X(t),t) - s^o(X(t),t)\|^2 dt\right) = O(n^{\gamma\alpha} n^{-2\frac{(\alpha+1)}{2\alpha+d}}), \quad \underline{T} = n^{\gamma - \frac{2(\alpha+1)}{2\alpha+d}}$$

independently of the ambient dimension D, as long as $D \leq n^H$ for some $H > 0$. As a consequence, we can control the Kullback–Leibler divergence between the generated density of $\hat{Y}(\underline{T})$ and the perturbed density of $X(\underline{T})$ with a rate of order

$$KL(X(\underline{T}), Y(\underline{T})) = O(n^{\gamma\alpha} n^{-\frac{2\alpha}{2\alpha+d}})$$

while the Wasserstein distance between the distribution of $\hat{Y}(\underline{T})$ and the true generative process is bounded by

$$W_1(\mu, Y(\underline{T})) = O(\sqrt{D} n^{\gamma\alpha} n^{-2\frac{(\alpha+1)}{2\alpha+d}}).$$

We do this by developing a new framework connecting diffusion models to the well-studied theory of extrema of Gaussian Processes.

It is still unclear if the term $\sqrt{D}$ in the bound on the Wasserstein distance is sharp.

The role of probability distances in generative models: A Vanilla GAN case study

Mathias Trabs

(joint work with Lea Kunkel)

The empirical success of Generative Adversarial Networks (GANs) caused an increasing interest in theoretical research. The statistical literature is mainly focused on Wasserstein GANs and generalizations thereof, see e.g. [3]. Statistical results for Vanilla GANs, the original optimization problem, are still rather limited and require assumptions such as smooth activation functions and equal dimensions of the latent space and the ambient space, see [2]. A main reason is that Vanilla GANs are by construction linked to the Jensen–Shannon distance which is not compatible with dimension reduction settings and the manifold hypothesis. To bridge this gap, we draw a connection from Vanilla GANs to the Wasserstein distance in [1]. By doing so, existing results for Wasserstein GANs can be extended to Vanilla GANs.

Let $X_1, \dots, X_n \sim \mathbb{P}^*$ be an i.i.d. sample of training data in $[0,1]^d$ with unknown distribution $\mathbb{P}^*$ and empirical measure $\mathbb{P}_n$. The Vanilla GAN estimator for $\mathbb{P}^*$ can be defined as the empirical risk minimizer

$$\widehat{G}_n \in \operatorname*{arg\,min}_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})$$

where $Z \in [0,1]^{d^*}$ is a latent easy to sample random variable, $\mathcal{G}$ is the class of generators $G\colon [0,1]^{d^*} \to [0,1]^d$ and the risk is given by

$$V_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) := \sup_{W \in \mathcal{W}} \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \Big[-\log\Big(\frac{1 + e^{-W(X)}}{2}\Big) - \log\Big(\frac{1 + e^{W(Y)}}{2}\Big)\Big]$$

for the class of discriminators $\mathcal{D} = \{\frac{1}{1+e^{-W(\cdot)}} : W \in \mathcal{W}\}$ for some set $\mathcal{W}$ of functions $W\colon \mathbb{R}^d \to \mathbb{R}^d$.

By proving compatibility of $V_{\mathcal{W}}(\mathbb{P}, \mathbb{Q})$ with the Wasserstein distance for appropriate choices of $\mathcal{W}$, we obtain an oracle inequality for Vanilla GANs in Wasserstein distance. Choosing $\mathcal{W} = \mathrm{Lip}(L)$ as the set of all Lipschitz functions with Lipschitz constant $L > 2$, $d^* > 2$ and $\mathcal{G}$ be compact. The empirical risk minimizer satisfies

$$\begin{aligned}&\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\widehat{G}_n}(Z))] \\ &\quad = \mathcal{O}\Big(\inf_{G^* \in \mathrm{Lip}(M)} \Big\{ \mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)^{1/2} + \inf_{G \in \mathcal{G}} \|G^* - G\|_\infty^{1/2} \Big\} + n^{-\frac{1}{2d^*}} \Big).\end{aligned}$$

This oracle inequality can be generalized to allow for neural network classes $\mathcal{W}$. The assumptions of this oracle inequality are designed to be satisfied by network architectures commonly used in practice, such as feedforward ReLU networks and d^* can be chosen independently of d. By providing a quantitative result for the approximation of a Lipschitz function by a feedforward ReLU network with bounded Hölder norm, we conclude a rate of convergence for Vanilla GANs as well as Wasserstein GANs as estimators of the unknown probability distribution.

At the end of the talk we discuss with all mini-workshop participants possible alternatives to measure the quality of generative models beyond standard choices like the total-variation distance, Kullback–Leibler or more general f-divergence, Jensen–Shannon distance, or Wasserstein distances. Suggestions are the following:

- Proper scoring rules, e.g. energy scores,
- Maximum Mean Discrepancies,
- different cost functions in optimal transport,
- Fréchet inception distance.

Finally, it was noticed that the generated distribution should not be too close the empirical measure as discussed in [4].

References

[1] L. Kunkel and M. Trabs, *A Wasserstein perspective of Vanilla GANs*, Neural Networks **181**, 106770, 2025.

[2] N. Puchkin, S. Samsonov, D. Belomestny, E. Moulines and A. Naumov, *Rates of convergence for density estimation with generative adversarial networks*, Journal of Machine Learning Research **25**(29), pp. 1–47, 2024.

[3] A. Stéphanovitch, E. Aamari and C. Levrard, *Wasserstein generative adversarial networks are minimax optimal distribution estimators*, Annals of Statistics **52**(5), pp. 2167-2193, 2024.

[4] E. Vardanyan, S. Hunanyan, T. Galstyan, A. Minasyan and A. S. Dalalyan, *Statistically Optimal Generative Modeling with Maximum Deviation from the Empirical Distribution*, International Conference on Machine Learning pp. 49203–49225, PMLR, 2024.

Regularity of (non-optimal) transport maps and applications to SGMs

Arthur Stéphanovitch

In this work, we explore the regularity of transport maps constructed via diffusion processes, extending the classical theory beyond Lipschitz continuity. While Caffarelli's seminal regularity results for Monge–Ampère equations have provided foundational insights into optimal transport maps, obtaining regularity in unbounded settings remains a significant challenge. Here, we address this gap by demonstrating that transport maps generated through diffusions can indeed exhibit higher-order regularity, leading to the construction of the first smooth transport map within this framework. These results open new avenues for understanding the structure and properties of transport maps in broader settings. We develop the framework in full generality, where we emphasize the following key contributions:

1. Higher order regularity of transport maps. Mirroring the classical regularity theory of optimal transport (Theorem 12.50 in [1]), we prove that the Langevin transport map between the d-dimensional Gaussian distribution and a log β-Hölder perturbation, is of Hölder regularity $\beta + 1$. Futhermore, we obtain a Lusin-type result for the transport of the Gaussian to a class of measures supported on the ball. It is shown that a set of mass $1 - \epsilon$ is transported by a $(\beta + 1)$-Hölder map having a norm controlled by a logarithmic power of ϵ.

2. Applications. The existence of smooth transport maps enables a range of applications, which we present in the subsequent sections.

Transfer of functional inequalities The regularity result allows to transfer the functional inequalities involving higher order derivatives from the Gaussian measure to the transported one. As applications we extend the class of measures satisfying the generalized Sobolev inequality [2].

Applications to generative models We provide applications of our regularity result to the estimation of densities by Wasserstein Generative Adversarial Networks (WGAN) and score-based generative models. Firstly, the existence of smooth transports maps allows to show the optimality of the WGAN estimator within the (widely used) Gaussian setting. Secondly, given the strong correlation between the score function in diffusion models and the transport map's velocity field, we show that higher regularity of the target distribution transfers to higher regularity of the score function.

References

[1] C. Villani, *Optimal transport: old and new*, Grundlehren der mathematischen Wissenschaften **338**, Springer, 2009.

[2] J. Rosen, *Sobolev inequalities for weight spaces and supercontractivity*, Transactions of the American Mathematical Society **222**, pp. 367–376, 1976.

A mixture-based framework for guiding diffusion models

Yazid Janati

(joint work with Badr Moufad, Mehdi Abou El Qassime, Alain Durmus, Eric Moulines and Jimmy Olsson)

Inverse problems–such as reconstructing images from partial or noisy measurements, or separating individual sources from mixed signals–are inherently challenging due to their ill-posed nature. In such settings, Bayesian inference, when combined with generative modeling, provides a systematic and principled approach. By using generative models trained on representative data distributions, these methods incorporate meaningful prior knowledge, which can then be integrated with the likelihood function describing the observed data. This leads to a posterior distribution, whose samples represent plausible solutions that harmonize both the observed data and prior assumptions.

In recent developments, diffusion models have emerged as state-of-the-art generative models, demonstrating exceptional capabilities in image and audio generation tasks. Diffusion models function by first progressively adding noise to data samples through a forward diffusion process, ultimately converting them into pure Gaussian noise. The generative model is then trained to reverse this noising process, effectively learning to reconstruct original data from noise. While diffusion models provide powerful priors, directly using them for inverse problems typically requires constructing a posterior denoiser that blends this learned prior with the gradient of the log-likelihood function derived from the observations. However, existing posterior sampling methods for diffusion models often rely on crude approximations of the likelihood gradient and require significant heuristic tuning and adjustments specific to each task.

In this talk, I will introduce a novel principled approach specifically designed to overcome these limitations. The core contribution of this approach is the construction of a mixture approximation of intermediate posterior distributions defined by the diffusion model. The sampling is carried out sequentially via Gibbs sampling, a Markov Chain Monte Carlo method, using a careful data augmentation scheme. Gibbs sampling is employed here due to its simplicity and theoretical guarantees, allowing for exact conditional updates at each iteration, thus ensuring stability and efficiency.

One key advantage of the presented algorithm is its flexibility: it adapts to varying levels of computational resources by adjusting the number of Gibbs iterations. Consequently, substantial performance gains can be achieved by increasing inference-time computational effort. I will present extensive experimental results

demonstrating strong empirical performance across ten diverse image restoration tasks, involving both pixel-space and latent-space diffusion models, and showcase its successful application in musical source separation.

Neural Flow Diffusion Models and SDE Matching

Christian A. Naesseth
(joint work with Grigory Bartosh and Dmitry Vetrov)

In this talk I introduced Neural Flow Diffusion Models (NFDMs) [1, 2] and SDE Matching [3]. These frameworks rely on a variational perspective of Diffusion Models, Schrödinger Bridges, and Latent Stochastic Differential Equations (SDEs) that lets us directly derive objective functions as standard variational evidence lower bounds (ELBOs).

From the variational perspective, Diffusion Models relies on a *fixed* forward process for inference. This can often complicate the reverse process' task in learning generative trajectories, and results in costly inference for diffusion models. To address this limitation, NFDM introduces a *learnable* forward process with a corresponding variational bound that can be estimated *simulation-free*.

The NFDM posterior process, also known as the forward process, is constructively defined by following a three-step procedure:

(1) Define the sequence of marginal distributions through a normalizing flow;
(2) Construct an ODE with random initial condition with matched marginals;
(3) Extend the ODE to an SDE that preserves the marginal distributions.

First, the sequence of marginal distributions $q_t(z_t|x)$ for latent variable z_t and data-point x is defined by a normalizing flow

$$z_t = F(\varepsilon, t, x), \quad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

where F is a diffeomorphism parameterized by a neural network. This induces a sequence of distributions with density that can be evaluated through the standard change-of-variables formula.

Then, an ordinary differential equation (ODE) with random initial condition that matches the marginals of (1) is given by

$$\mathrm{d}z_t = \partial_t F(\varepsilon, t, x)\Big|_{\varepsilon = F^{-1}(z_t, t, x)} \mathrm{d}t, \quad z_0 \sim q_0(z_0|x). \tag{2}$$

Finally, the conditionally deterministic ODE trajectories are extended to a distribution over stochastic trajectories $(z_t)_t$ that preserves the marginal distributions

$$\mathrm{d}z_t = \left[\partial_t F(\varepsilon, t, x)\Big|_{\varepsilon = F^{-1}(z_t, t, x)} + \frac{1}{2} g(t) g(t)^\top \nabla_{z_t} \log q_t(z_t|x)\right] \mathrm{d}t + g(t)\mathrm{d}w_t. \tag{3}$$

Using (3) as a posterior process approximation lets the user estimate the corresponding ELBO, which only depends on the marginal distributions, in a simulation-free manner by directly sampling as in (1).

In the talk I illustrated how NFDM and SDE Matching improves on previous state-of-the-art for image generation, molecular generation, straightening generative flows, bridge matching, and time series applications.

References

[1] G. Bartosh, D. Vetrov, and C. A. Naesseth, *Neural Diffusion Models*, International Conference on Machine Learning, PMLR, pp. 3073-3095, 2024.
[2] G. Bartosh, D. Vetrov, and C. A. Naesseth, *Neural Flow Diffusion Models: Learnable Forward Process for Improved Diffusion Modelling*, Advances in Neural Information Processing Systems **37**, 2024.
[3] G. Bartosh, D. Vetrov, and C. A. Naesseth, *SDE Matching: Scalable and Simulation-Free Training of Latent Stochastic Differential Equations*, arXiv:2502.02472, 2025.

Bridge Matching Schrödinger Bridges

Stefano Peluchetti

This talk presents a review of the Bridge Matching framework and illustrates its application in approximating Schrödinger Bridges and Entropic Optimal Transport plans.

Bridge Matching (BM, [1]) establishes an exact transport in finite time between two distributions Ψ_0, Ψ_1. This stands in contrast to Denoising Diffusion models (DDM, [2]), where an exact transport is achieved only asymptotically as the integration time approaches infinity. In its simplest implementation, the BM transport is characterized by the solution to the following d-dimensional stochastic differential equation (SDE):

$$X_0 \sim \Psi_0, \quad dX_t = \frac{\mathbb{E}_\Pi[X_1|X_t] - X_t}{1-t}dt + dW_t, \quad t \in [0,1], \tag{1}$$

where the expectation is taken under the measure Π corresponding to $(X_0, X_1) \sim \Psi_{0,1}$, for some coupling $\Psi_{0,1}$ of Ψ_0, Ψ_1, and $X_t|X_0, X_1 \sim \mathcal{BB}$, representing the transition law of the standard Brownian Bridge from X_0 to X_1.

A crucial property of the BM transport is that the marginal distribution of X_t from the solution to (1) matches the marginal distribution of X_t from Π. Consequently, (1) defines a transport from Ψ_0 to Ψ_1. Analogous to DDM, BM can be learned through a straightforward regression objective:

$$\mathbb{E}_\Pi[X_1|X_t] = \operatorname*{argmin}_{\alpha(x,t)} \mathbb{E}_\Pi[\|\alpha(X_t,t) - X_1\|^2],$$

where $\alpha(x,t)$ is implemented via a neural network and optimization is carried out by stochastic gradient descent (SGD).

Originally introduced as an alternative to DDM for generative applications, in this context BM considers Ψ_1 as the data distribution (from which samples are available), Ψ_0 as a simple distribution (typically $\mathcal{N}(0, I_d)$), and employs the independent coupling $\Psi_0 \times \Psi_1$.

While the BM transport (1) preserves the marginal distributions of Π, it does not preserve the coupling $\Psi_{0,1}$. Indeed, the findings of [3] demonstrates that

iterative computation of the BM transport (1), using the coupling produced by (X_0, X_1) from the previous iteration as input, converges to the Schrödinger Bridge with law $\mathcal{S}$, where:

$$\mathcal{S} = \operatorname*{argmin}_{\mathcal{P}} \mathbb{KL}(\mathcal{P} \mid \mathcal{W}) \quad \text{s.t.} \quad \mathcal{P}_0 = \Psi_0, \mathcal{P}_1 = \Psi_1,$$

and $\mathcal{W}$ represents the Brownian law. Schrödinger Bridges play a fundamental role in measure transport theory, providing solutions to Entropic Optimal Transport and Stochastic Optimal Control problems [4].

The iterated BM procedure of [3] and the Diffusion Iterative Proportional Fitting procedure [5] share a common limitation: they require solving multiple optimization problems, one per iteration. To address this limitation, [6] proposes a forward-backward SDE approach, where forward and backward SDEs perform BM based on couplings produced by their counterparts. This approach yields a loss function optimizable in a single SGD loop. Initial theoretical results from [6] indicate that this method successfully recovers the Schrödinger Bridge $\mathcal{S}$ at convergence.

References

[1] S. Peluchetti, *Non-Denoising Forward-Time Diffusions*, 2021.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, International Conference on Learning Representations, 2021.

[3] S. Peluchetti, *Diffusion Bridge Mixture Transports, Schrödinger Bridge Problems and Generative Modeling*, Journal of Machine Learning Research, 2023.

[4] C. Léonard, *A survey of the Schrödinger problem and some of its connections with optimal transport*, Discrete & Continuous Dynamical Systems, 2014.

[5] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, *Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling*, Advances in Neural Information Processing Systems **34**, 2021.

[6] S. Peluchetti, *BM^2: Coupled Schrödinger Bridge Matching*, Transactions on Machine Learning Research, 2024.

Denoising Lévy Probabilistic Models: A Heavy-Tailed Diffusion Approach for Generative Modeling

Dario Shariatian

(joint work with Alain Durmus, Umut Simsekli)

In this talk, I introduce Denoising Lévy Probabilistic Models (DLPM), a novel class of generative models that extend the diffusion framework by replacing the standard Gaussian noise with heavy-tailed α-stable noise. While classical diffusion models–such as DDPM [1] and its score-based SDE reformulation [2] have achieved impressive results in image or audio synthesis, they face challenges when dealing with heavy-tailed or imbalanced datasets. In contrast, DLPM leverages

the stability property of α-stable distributions to define a discrete-time noising process

$$X_t = \gamma_t X_{t-1} + \sigma_t \epsilon_t^{(\alpha)} , \quad X_0 \sim p_0 , \tag{1}$$

where p_0 is the data distribution, $(\gamma_t, \sigma_t)_{t=1}^T$ is the noising schedule, and $(\epsilon_t^{(\alpha)})_{t=1}^T$ are distributed as the centered, unit-scale, symmetric α-stable distribution $\mathcal{S}\alpha\mathcal{S}$. This recovers DDPM when $\alpha = 2$ and introduces heavy tails for $\alpha < 2$. A key technical contribution is a transformation that decomposes the α-stable noise into a product of a one-dimensional random variable and a Gaussian vector:

$$\epsilon_t^{(\alpha)} \sim A_t^{1/2} G_t , \tag{2}$$

where A_t is distributed as a totally skewed to the right $\alpha/2$ -stable distribution of scale $\cos^{2/\alpha}(\pi\alpha/4)$, which we denote by $\mathcal{S}_{\alpha/2}^{\text{skewed}}$, and $G_t \sim \mathcal{N}(0, \mathrm{I}_d)$. By conditioning on the sequence $A_{1:T}$, we recover Gaussian transitions and thus re-employ the classical DDPM machinery. We introduce the following probabilistic model:

$$p^\theta(x_0, \cdots, x_t, a_{1:T}) = \underbrace{p_T^\theta(x_T)}_{\text{noise}} \cdot \prod_{t=1}^{T} \underbrace{p_{t-1|t,a}^\theta(x_{t-1}|x_t, a_{1:t})}_{\text{Gaussian transitions}} \Psi^{\otimes T}(a_{1:T}) , \tag{3}$$

where Ψ is the density of $\mathcal{S}_{\alpha/2}^{\text{skewed}}$, and

$$p_{t-1|t,a}^\theta(\cdot|x_t, a_{1:t}) = \mathcal{N}(\cdot \ ; \ \mathrm{m}_t^\theta(x_t, a_{1:t}), \Sigma_t^\theta(a_{1:t})) . \tag{4}$$

Analogous to DDPM, we fit our parametric family of distribution with a variational inference scheme, based on the classical ELBO trick applied to

$$\mathcal{L}(\theta) \mapsto \mathbb{E}\left[\mathrm{KL}(p_0(\cdot) \| p_{0|a}^\theta(\cdot|A_{1:T}))^{1/2}\right] . \tag{5}$$

Ultimately, this yields a modified denoising loss – a non-squared ℓ_2 loss, that ensures finite expectations despite the infinite variance of the heavy-tailed noise. I further demonstrate that DLPM is not only mathematically simpler than its continuous-time counterpart, the Lévy-Itô Model (LIM) ([3]), but also more flexible. In particular, DLPM:

- Maintains compatibility with standard DDPM implementations, requiring only minimal changes to existing codebase
- Admits a deterministic sampler, dubbed Denoising Lévy Implicit Models (DLIM)
- Enables more design choices, as compared to its continuous counterpart LIM, thanks to the use of elementary mathematics instead of fractional calculus (e.g., use alternative noising schedule, learn the optimal variance of the Gaussian transitions in the generative process, etc.)
- Provides improved tail coverage and better performance on unbalanced datasets, while offering faster computation times.

I concluded the talk with a discussion on experimental results.

References

[1] J. Ho, A. Jain, and P. Abbeel, *Denoising Diffusion Probabilistic Models*, Advances in Neural Information Processing Systems **33**, pp. 6840–6851, 2020.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, International Conference on Learning Representations, 2021.

[3] E. Yoon, K. Park, S. Kim, and S. Lim, *Score-based Generative Models with Lévy Processes*, Advances in Neural Information Processing Systems **37**, 2023.

Minimax Optimality of Wasserstein GAN Estimators

Eddie Aamari

(joint work with Arthur Stéphanovitch and Clément Levrard)

I have presented an analysis of the minimax optimality of the Wasserstein Generative Adversarial Networks (WGAN) estimator for generative modeling. Given a probability measure μ supported in $\mathbb{R}^p$, the goal is to approximate μ and sample from a distribution close to it.

I reviewed the theoretical foundations of Vanilla GANs and Wasserstein GANs, formulating the generative adversarial problem as minimizing an Integral Probability Metric (IPM) over a suitable class of generator functions $\mathcal{G}$ and discriminator functions $\mathcal{D}$

$$\hat{g} \in \operatorname{argmin}_{g\in\mathcal{G}} \sup_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} D(X_i) - D(g(U_i))$$

To address the curse of dimensionality, it is assumed that some mapping $g^\star$ from a lower-dimensional torus $\mathbb{T}^d$ to $\mathbb{R}^p$ is such that μ can be expressed as the pushforward measure $g^\star_{\#U}$ of a uniform distribution U on $[0,1]^d$. Unlike classical nonparametric density estimators relying on discrepancy measures such as the Kullback–Leibler divergence, this high-dimensional setting makes conventional density-based approaches ineffective.

In this context, I presented a general oracle inequality, together with the two main ingredient leading the construction of tractable classes of neural networks $\mathcal{G}$ and $\mathcal{D}$, such that

$$\sup_{g^\star\in\mathcal{H}^{\beta+1}} \mathbb{E}_{X_i\sim g^\star_{\#U}}[d_{\mathcal{H}^\gamma}(g^\star_{\#U}, \hat{g}_{\#U})] \le C(\log n)^{C'}\left(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}}\right),$$

Namely, we insisted on:

- How wavelets allow to understand finely the bias of the method ;
- How functional interpolation inequalities *à-la* Gagliardo-Nirenberg allow to only analyze the case of large γ values, and then transfer the optimal rates to stronger norms with smaller γ.

References

[1] A. Stéphanovitch, E. Aamari, and C. Levrard, *Wasserstein generative adversarial networks are minimax optimal distribution estimators*, The Annals of Statistics, **52**(5), pp. 2167–2193, 2024.

[2] N. Puchkin, S. Samsonov, D. Belomestny, E. Moulines, and A. Naumov, *Rates of convergence for density estimation with generative adversarial networks*, Journal of Machine Learning Research, **25**(29), pp. 1–47, 2024.

Towards Faster Non-asymptotic Convergence for Diffusion-based Generative Models

Yuting Wei

(joint work with Gen Li, Yu Huang, Timofey Efimov, Yuxin Chen, Yuejie Chi)

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative artificial intelligence. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. Given the complexity of developing a full-fledged end-to-end theory, a divide-and-conquer approach has been advertised, decoupling the score learning phase (i.e., how to estimate score functions from training data) and the generative sampling phase (i.e., how to generate new data given the score estimates). In particular, the past two years have witnessed growing interest and remarkable progress from the theoretical community towards understanding the sampling phase. Our works described in this talk contribute to this growing list of theoretical endeavors by developing a new suite of non-asymptotic theory for several score-based generative modeling algorithms.

In the first part of the talk, we concentrate on two types of samplers in discrete time: (i) a deterministic sampler based on a sort of ordinary differential equations (ODEs) called probability flow ODEs (which is closely related to the DDIM); and (ii) a DDPM-type stochastic sampler motivated by reverse-time SDEs. For the deterministic sampler [2], we establish a convergence rate proportional to $1/T$ (with T the total number of steps). As far as we know, this is the first result for this deterministic sampler that accounts for score estimation errors in discrete time. In comparison, other theoretical results that accommodate score errors for the probability flow ODE approach either study certain stochastic variations of this deterministic sampler or fall short of accommodating discretization errors. For the DDPM-type sampler [1], we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results characterize how ℓ_2 score estimation errors affect the quality of the data generation processes.

In the second part, we design accelerated variants for both the deterministic and the stochastic samplers [3]. In the deterministic setting, we demonstrate how to speed up the ODE-based sampler by exploiting some sort of momentum term to

adjust the update rule, leverages insights from higher-order ODE approximation in discrete time with an improved convergence rate $1/T^2$. In the stochastic setting, we propose a novel sampling procedure to accelarate the SDE-based sampler and establish a rate of $1/T$, thus unveiling the superiority of the proposed sampler compared to the original DDPM sampler. A series of numerical experiments have also been conducted to illustrate the effectiveness of the accelerated samplers.

References

[1] G. Li, Y. Wei, Y. Chen, and Y. Chi, *Towards faster non-asymptotic convergence for diffusion-based generative models*, International Conference on Learning Representations, 2024.

[2] G. Li, Y. Wei, Y. Chi, and Y. Chen, *A sharp convergence theory for the probability flow odes of diffusion models*, arXiv:2408.02320, 2024.

[3] G. Li, Y. Huang, T. Efimov, Y. Wei, Y. Chi, and Y. Chen, *Accelerating convergence of score-based diffusion models, provably*, International Conference on Machine Learning, PMLR, 2024.

Flow Matching from a KDE perspective

Lea Kunkel

(joint work with Mathias Trabs)

Flow Matching, a generative model introduced by [1], has recently attained significant interest due to its considerably more straightforward simulation process in comparison to diffusions, which have been regarded as the state-of-the-art generative method.

Let $\mathbb{U}$ be a distribution on $\mathbb{R}^d$, $p\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}_{>0}$ be a time dependent probability density path and $v\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ be a time dependent vector field. If $\psi\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ solves the ODE

$$\frac{d}{dt}\psi_t(x) = v_t(\psi_t(x)), \quad \psi_0(x) = x$$

and

$$p_t = [\psi_t]_{\#}p_0$$

then we say that v_t generates p_t. Approximating a vector field that generates a certain density p_t using a function $\tilde{v}$ out of a class of neural networks leads to the Flow Matching Objective ([1])

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ X_t\sim p_t}}\big[\|v_t(X_t) - \tilde{v}_t(X_t)\|^2\big]. \tag{1}$$

[1] have constructed for $t\in[0,1]$ the function p_t in terms of the marginal probability path $p_t(\cdot|y)\colon \mathbb{R}^d \to \mathbb{R}$ as

$$p_t(x) = \int p_t(x|y)p^*(y)\,\mathrm{d}y,$$

and v_t in terms of the marginal vector fields as

$$v_t(x) = \int v_t(x|y)\frac{p_t(x|y)p^*(y)}{p_t(x)}dy, \tag{2}$$

where $v_t(\cdot|y)\colon \mathbb{R}^d \to \mathbb{R}^d$ is a vector field that generates $p_t(\cdot|y)\colon \mathbb{R}^d \to \mathbb{R}$. [1] show that in this case, v_t generates p_t and that the minimizing arguments of (1) and

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1],\\ Y\sim p^*,\\ X_t\sim p_t(\cdot|Y)}} \left[\|\tilde{v}_t(X_t) - v_t(X_t \mid Y)\|^2 \right], \tag{3}$$

are the same.

The statistical properties have only recently been studied by [2] and [3] in the Wasserstein 2 distance. Both do not use the exact setting of [1], but rather use stopping times that are adaptive to the number of samples.

The talk aimed to shorten this gap between practice and theory. First, we showed that the motivation of Flow Matching also applies to its empirical counterparts. We then built on the connection between flow matching and kernel density estimation by extending our analysis to latent distributions, which are often used as kernels in nonparametric statistics. In doing so, we proposed moment conditions for a proper choice of the latent distribution. We then showed convergence rates in Wasserstein 1 distance in the case of perfect approximation of the empirical counterpart of the vector field (2), as well as in the case where neural networks are used for the vector field. For the latter, we exploited the approximations properties of ReQU networks, which have been studied by [4].

References

[1] Y. Lipman, R. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow Matching for Generative Modeling*, International Conference on Learning Representations, 2023.

[2] K. Fukumizu, T. Suzuki, N. Isobem, K. Oko, and M. Koyama, *Flow matching achieves minimax optimal convergence*, arXiv preprint arXiv:2405.20879, 2024.

[3] Y. Gao, J. Huang, Y. Jiao, and S. Zheng, *Convergence of Continuous Normalizing Flows for Learning Probability Distributions*, arXiv preprint arXiv:2404.00551, 2024.

[4] D. Belomestny, A. Naumov, N. Puchkin, and S. Samsonov, *Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations*, Neural Networks **161**, pp. 242–253, 2023.

Iteration Complexity of Diffusion Models under the Manifold Assumption

Iskander Azangulov

(joint work with Peter Potaptchik, George Deligiannidis, Judith Rousseau)

Score-matching generative models, also known as diffusion models, have proven highly effective at sampling from complex, high-dimensional unknown distributions. The core idea behind these models is to consider a forward process, X_t, that progressively adds Gaussian noise to the original complex distribution over the interval $[0, T]$. This forward process can be modeled using a standard Ornstein-Uhlenbeck (OU) process.

The backward process, denoted as $Y_{T-t} = X_t$, is a diffusion process where the drift function is adjusted by the *score function* $s_t(x)$, which is proportional to the conditional expectation of the noise at a given point. Diffusion models exploit this

framework by learning the score function $s_t(x)$, and then discretizing it to model the backward process effectively.

In their work, Benton et al. [1] studied the relationship between iteration complexity (the number of discretization steps) and the quality of the generated samples. They demonstrated that for a distribution μ in $\mathbb{R}^D$, assuming a finite second moment, the number of steps required scales at most as $O(D)$. Importantly, this bound is tight unless additional assumptions are made.

However, in many real-world applications, it is believed that the distribution of interest concentrates on a much lower-dimensional manifold embedded within the D-dimensional space. This phenomenon is referred to as the manifold hypothesis.

In this talk, we demonstrate that the number of steps required by diffusion models to converge in Kullback–Leibler (KL) divergence is linear–up to logarithmic factors–in the intrinsic dimension d of the underlying manifold. Moreover, we show that this linear dependency is tight, meaning the scaling with d is optimal.

This result helps explain why diffusion models excel in tasks such as synthetic image generation. While the extrinsic dimensionality of image datasets is large (e.g., approximately 1.5×10^5 for ImageNet), research by Pope et al. [2] suggests that the true intrinsic dimension is much lower, on the order of around 50 for ImageNet. Our findings imply that the number of diffusion steps needed to generate high-quality samples scales with this intrinsic dimension rather than the much larger extrinsic dimension. This explains why diffusion models can generate sharp image samples in fewer than 1000 iterations, despite the large dimensionality of the input data.

References

[1] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis, *Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization*, International Conference on Learning Representations, 2024.

[2] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein *The intrinsic dimension of images and its impact on learning*. arXiv preprint arXiv:2104.08894, 2021.

[3] P. Potaptchik, I. Azangulov, and G. Deligiannidis, *Linear convergence of diffusion models under the manifold hypothesis*. arXiv preprint arXiv:2410.09046, 2024.

Reporter: Lukas Trottner

Participants

Dr. Eddie Aamari
Département de mathématiques et applications
École normale supérieure, Université PSL, CNRS
45 rue d'Ulm
75230 Paris Cedex 05
FRANCE

Iskander Azangulov
Department of Statistics
Oxford University
24-29 St Giles
Oxford OX1 3LB
UNITED KINGDOM

Dr. Denis Belomestny
Fakultät für Mathematik
Universität Duisburg-Essen
45117 Essen
GERMANY

Prof. Dr. Sören Christensen
Mathematisches Seminar
Christian-Albrechts-Universität Kiel
Heinrich-Hecht-Platz 6
24118 Kiel
GERMANY

Dr. Giovani Conforti
Dipartimento di Matematica
Universita di Padova
Via Trieste, 63
35121 Padova
ITALY

Yazid Janati
Ecole polytechnique
Paris 75003
FRANCE

Lea Kunkel
Institut für Stochastik
Karlsruher Institut f. Technologie (KIT)
76131 Karlsruhe
GERMANY

Dr. Christian A. Naesseth
Informatics Institute
University of Amsterdam
1081 HV Amsterdam
NETHERLANDS

Prof. Dr. Alain Oliviero-Durmus
Centre de Mathématiques Appliquées
École Polytechnique
91128 Palaiseau Cedex
FRANCE

Dr. Stefano Peluchetti
Sakana AI
Toranomon Hills Business Tower 15F, Minato City
1 Chome-17-1 Toranomon
105-6490 Tokyo
JAPAN

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Judith Rousseau
Department of Statistics
University of Oxford
24-29 St Giles'
Oxford OX1 3LB
UNITED KINGDOM

Dario Shariatian
INRIA
48 Rue Barrault
75013 Paris
FRANCE

Arthur Stephanovitch
Ecole Normale Supérieure
45 rue d'ulm
75005 Paris Cedex
FRANCE

Prof. Dr. Claudia Strauch
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Mathias Trabs
Institut für Stochastik
Karlsruher Institut f. Technologie (KIT)
76128 Karlsruhe
GERMANY

Dr. Lukas Trottner
School of Mathematics
University of Birmingham
Birmingham B15 2TT
UNITED KINGDOM

Dr. Yuting Wei
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES