Mathematisches Forschungsinstitut Oberwolfach

Report No. 50/2025

DOI: 10.4171/OWR/2025/50

Mini-Workshop: Probabilistic Perspectives in Neural Network-Based Machine Learning

Organized by
Steffen Dereich, Münster,
Aymeric Dieuleveut, Palaiseau
Sebastian Kassing, Berlin
Sophie Langer, Bochum

26 October – 31 October 2025

Abstract. Artificial neural networks (ANNs) have emerged as a powerful tool in modern machine learning, yet their mathematical foundations remain only partially understood. A key challenge is the inherently stochastic nature of ANN training: optimization occurs in high-dimensional parameter spaces with complex loss landscapes, influenced by stochastic initialization and noisy gradient updates. Understanding these dynamics requires probabilistic methods and asymptotic frameworks. This workshop explored recent advances in stochastic training dynamics, emphasizing probabilistic techniques and limit theorems. By bringing together researchers from probability, optimization, and deep learning theory, this workshop laid the groundwork for new directions in understanding neural network training from a stochastic perspective.

Mathematics Subject Classification (2020): 62M45, 60Gxx, 62G20, 90C30.

Introduction by the Organizers

The workshop *Probabilistic Perspectives in Neural Network-Based Machine Learning*, organized by Steffen Dereich (Universität Münster), Aymeric Dieuleveut (École Polytechnique, Palaiseau), Sebastian Kassing (Bergische Universität Wuppertal), and Sophie Langer (Ruhr Universität Bochum), was attended by 16 participants from Germany, Switzerland, France, Italy, and the Netherlands. The program included 20 talks (45 minutes each) and two open problem sessions (90 minutes), providing ample time for discussions. Interactions during talks, breaks,

and problem sessions fostered new collaborations, strengthened existing connections, and enabled participants to exchange ideas across disciplines. The workshop brought together researchers working on probability theory, stochastic processes, interacting particle systems, and optimization. Several talks highlighted recent advances in limit theorems, stochastic dynamics, random structures, and applications to statistical physics and algorithmics. Substantial interaction between communities allowed the blending of analytical, probabilistic, and geometric techniques.

The scientific highlights can be summarized as follows:

Analysis and Design of Optimization Algorithms: Several talks of this workshop focused on providing rigorous theoretical guarantees, such as limit theorems, and developing systematic methods for fundamental optimization algorithms used in training machine learning models. S. Dereich established a Central Limit Theorem (CLT) for the averaged Adam algorithm, while J. Schmidt-Hieber provided a quenched CLT and convergence rates under noisy SGD (such as with dropout). S. Weissmann provided almost sure convergence guarantees for the last iterate of SGD under a gradient domination condition. S. Kassing analyzed convergence of SGD schemes for Lojasiewicz-landscapes (a weak non-convex condition), providing bounds on step-sizes and perturbation size to guarantee almost sure convergence of the iterates. A. Dieuleveut used the PEPIT framework to provably show that the classical Heavy Ball method does not achieve accelerated convergence on smooth, strongly convex functions, resolving a long-standing question. A. Taylor overviewed systematic methods for algorithm analysis, including the primal and dual Performance Estimation Problems (PEP), which facilitate constructive worst-case analysis and the design of optimal first-order algorithms.

Continuum and Fluctuation Models of Training Dynamics Several talks analyzed discrete training dynamics, such as SGD, via continuous limits (SDEs/SPDEs) to study noise and implicit bias. B. Gess introduced Stochastic Modified Flows, which are high-order SDEs that provide a more accurate, fluctuation-aware approximation of SGD. V. Konarovskyi derived a nonlinear, measure-valued Stochastic PDE (SPDE) to capture the fluctuations around the deterministic mean-field limit in overparameterized shallow neural networks, offering a higher-order continuum model of training noise. A. Shalova applied gradient flow formalism to analyze noisy gradient descent, revealing multiscale dynamics where the structure of noise (e.g., in SGD or dropout) determines the time scale of evolution. S. Pesme used simple diagonal linear networks to analyze the implicit bias of SGD, showing it favors ℓ_1-norm solutions, and characterized saddle-to-saddle dynamics in the training landscape.

Architecture Theory and Probabilistic Effects: Another line of talks focused on how architectural choices (depth, weight sharing) and parameterizations fundamentally shape the signal propagation, training trajectory, and final model properties. M. Seleznova demonstrated a qualitative failure of the infinite-width limit for recurrent

networks with shared weights. K. Papagianoulli explained the observed incremental rank increase in Transformers through saddle-to-saddle training dynamics in a low-rank surrogate model, where optimization proceeds through near-stationary low-rank phases followed by rapid rank-increase events. A. Shalova characterized the stationary measures and bifurcation branches of the McKean-Vlasov equation, which models the mean-field dynamics of noisy Transformer tokens on manifolds.

Generative Models and High-Dimensional Problems: Some talks further focused on the theoretical foundations of modern generative models and the computational tractability of high-dimensional problems in numerical analysis. L. Trottner developed a novel class of generative diffusion models to achieve adaptive and time-homogeneous denoising. S. Pesme provided a scheme for using pre-trained denoisers as surrogates for the intractable proximal operator in Maximum a Posteriori (MAP) estimation for inverse problems. J. Ackermann proved the capability of DNNs to bypass the Curse of Dimensionality (CoD) for solving high-dimensional semilinear PDEs by constructing networks that mimic Multilevel Picard (MLP) approximations.

Bridging Statistics and Optimization S. Wang and S. Langer discussed the existing gap between statistical and computational-related results. S. Wang reviewed the parameter identification problem in PDEs and SDEs used to tackle non-linear inverse problems. He reviewed the computational burden of those methods and presented a new theory proving the globally polynomial-time computability for corresponding estimators. S. Langer reviewed existing statistical results of neural network estimators within nonparametric regression problems, highlighted the need to embed the optimization procedure in the analysis and presented a first setting in which this is possible.

Mini-Workshop: Probabilistic Perspectives in Neural Network-Based Machine Learning

Table of Contents

Julia Ackermann (joint with Arnulf Jentzen, Thomas Kruse, Benno Kuckuck, and Joshua Lee Padgett)
Deep neural networks overcome the curse of dimensionality for space-time solutions of semilinear PDEs 2679

Steffen Dereich (joint with Arnulf Jentzen, Sebastian Kassing, Adrian Riekert)
Asymptotic analysis of the Adam algorithm 2680

Aymeric Dieuleveut
Systematic proofs and performance estimation: Provable non acceleration of Heavy Ball method and Application to communication-constrained algorithms 2681

Benjamin Gess (joint with Sebastian Kassing, Vitalii Konarovskyi, Nimit Ran)
Effective Fluctuating Continuum Models for Stochastic Gradient Descent 2682

Sebastian Kassing (joint with Benjamin Gess, Steffen Dereich, Simon Weissmann)
Beyond Strong Convexity: Geometry and Optimization under the Polyak–Łojasiewicz Condition 2684

Vitalii Konarovskyi
Fluctuation Analysis of Mean-Field Limits in Overparameterized SGD . 2685

Sophie Langer (joint with Adam Krzyzak, Michael Kohler, Alina Braun)
Deep Learning Theory: Statistics, Optimization and the Space Between . 2686

Johannes Schmidt-Hieber (joint with Jiaqi Li, Wei-Biao Wu)
CLTs for noisy SGD via geometric moment contraction 2687

Scott Pesme (joint with Nicolas Flammarion and Loucas Pillaud-Vivien)
Deep Learning Theory Through the Lens of Diagonal Linear Networks . 2687

Scott Pesme (joint with Giacomo Meanti, Michael Arbel and Julien Mairal)
MAP Estimation with Denoisers: Convergence Rates and Guarantees . . 2688

Anna Shalova
Noisy gradient flows in machine learning: two examples 2689

Mariia Seleznova
The probabilistic effects of depth in deep learning 2691

Lukas Trottner (joint with Sören Christensen, Jan Kallsen, Claudia Strauch)
Adaptive denoising diffusion modelling via random time reversal 2692

Katerina Papagianoulli (joint with Hana Tseran, Federico Pasqualotto)
Low-Rank Bias in Transformers: Saddle-to-Saddle Dynamics, Symmetries, and a Surrogate Model 2693

Adrien Taylor
Towards principled and systematic approaches to the analysis and design of first-order optimization algorithms 2694

Sven Wang
On global polynomial-time computable estimators in non-linear inverse problems 2695

Simon Weissmann (joint with Waïss Azizian, Leif Döring, Sara Klein)
Convergence analysis of stochastic gradient methods under gradient domination 2695

Abstracts

Deep neural networks overcome the curse of dimensionality for space-time solutions of semilinear PDEs

Julia Ackermann
(joint work with Arnulf Jentzen, Thomas Kruse, Benno Kuckuck, and Joshua Lee Padgett)

Approximating high-dimensional partial differential equations (PDEs) is a challenging task due to the curse of dimensionality (COD). In the last few years, it has been shown that deep neural networks (DNNs) have the expressive power to overcome the COD for several PDE approximation tasks, in the sense that there exists an approximating sequence of DNNs such that the number of DNN parameters grows at most polynomially in the PDE dimension and the reciprocal of the prescribed approximation accuracy.

My talk was based on [1, 2] where we establish that DNNs with the rectified linear unit (ReLU), leaky ReLU and softplus activation function can approximate solutions of semilinear heat PDEs with Lipschitz-continuous nonlinearities in the L^p-sense in space-time without the COD.

The general approach to prove such results is of a probabilistic nature, starting from a stochastic representation of the PDE solution. It employs approximation results of the PDE solution by nonlinear Monte Carlo methods called multilevel Picard (MLP) approximations, which have been shown to overcome the COD (see, e.g., [3]).

The aim is then to construct DNNs that mimic the MLP approximations. To this end, we replace the terminal condition and the nonlinearity of the PDE by DNNs. Indeed, we can show that the nonlinearity can be suitably approximated by DNNs for the above activation functions. Existence of a suitable DNN approximation of the terminal condition becomes, as in related work, an assumption. Furthermore, we translate all mathematical operations used in the MLP approximations to the DNN framework while keeping track of the number of DNN parameters. To obtain approximation results of the PDE in the space-time sense, we additionally interpolate the MLP approximations in time and show that also the interpolation procedure admits a suitable DNN approximation.

The resulting DNNs inherit randomness from the MLP approximations. By showing that the expected approximation error of the PDE solution by these DNNs does not exceed the prescribed accuracy, we can conclude that there exists a sequence of deterministic DNNs with the desired properties.

References

[1] J. Ackermann, A. Jentzen, T. Kruse, B. Kuckuck, and J.L. Padgett, *Deep neural networks with ReLU, leaky ReLU, and softplus activation provably overcome the curse of dimensionality for Kolmogorov partial differential equations with Lipschitz nonlinearities in the L^p-sense.*, arXiv:2309.13722, 2023.

[2] J. Ackermann, A. Jentzen, B. Kuckuck, and J.L. Padgett, *Deep neural networks with ReLU, leaky ReLU, and softplus activation provably overcome the curse of dimensionality for space-time solutions of semilinear partial differential equations.*, arXiv:2406.10876, 2024.
[3] M. Hutzenthaler, A. Jentzen, B. Kuckuck, and J.L. Padgett, *Strong L^p-error analysis of nonlinear Monte Carlo approximations for high-dimensional semilinear partial differential equations.*, arXiv:2110.08297, 2021.

Asymptotic analysis of the Adam algorithm

Steffen Dereich
(joint work with Arnulf Jentzen, Sebastian Kassing, Adrian Riekert)

In this talk, we will analyse the Adam algorithm, introduced by Kingma and Ba in 2014, for fixed parameters α and β and step sizes (γ_n) that decay to zero. I will present new error estimates for the Adam algorithm, developed jointly with Arnulf Jentzen. Specifically, we show that the algorithm's effective behaviour is closely related to a particular vector field, which we refer to as the Adam field. If this field satisfies a local coercivity condition around one of its zeros, we can prove convergence of order $\sqrt{\gamma_n}$ towards that zero, provided the iterates remain within a suitable neighbourhood [1].

Based on the new techniques it is also possible to prove an ODE approximation for the Adam algorithm, where the driving vector field is not the classical gradient but the Adam field. In the general setting, zeros of this vector field do typically not coincide with the zeros of the gradients of the loss and in this case limit points of the Adam algorithm are not local minima of the loss. However, in an overparametrised setting, the algorithm is shown to converge to the set of global minima if it enters a neighbourhood of these minima infinitely often, see [2].

In another related project, we analyse the *averaged* Adam algorithm. We prove a central limit theorem for the averaged Adam algorithm in [3] under an additional non-degeneracy condition on the innovation Numerical results for the averaged algorithm for deep neural network approximations for partial differential equation and optimal control problems are provided, see [4].

References

[1] S. Dereich and A. Jentzen. *Convergence rates for the Adam optimizer*. 2024. (arXiv:2407.21078).
[2] S. Dereich, A. Jentzen, and S. Kassing. *ODE approximation for the Adam algorithm: General and overparametrized setting*. 2025. (arXiv:2511.04622).
[3] S. Dereich and A. Jentzen. *The averaged Adam algorithm: a central limit theorem*. In preperation.
[4] S. Dereich, A. Jentzen, and A. Riekert *Averaged Adam accelerates stochastic optimization in the training of deep neural network approximations for partial differential equation and optimal control problems*. 2025. (arXiv:2501.06081).

Systematic proofs and performance estimation: Provable non acceleration of Heavy Ball method and Application to communication-constrained algorithms

Aymeric Dieuleveut

We present two independent applications of systematic approaches for the analysis and design of first-order optimization algorithms. Although motivated by distinct questions, both rely on constructive, worst-case reasoning and leverage performance estimation techniques to obtain tight guarantees.

First we revisit the long-standing question of whether the classical heavy-ball (HB) method can achieve acceleration beyond quadratic objectives. Despite its widespread use and strong empirical behavior, its theoretical status on the standard class of L-smooth and μ-strongly convex functions had remained unclear. Building on constructive cycling arguments, it is shown in [1, 2, 3] that for any condition number and any choice of parameters, HB either fails to converge on a function of $\mathcal{F}_{\mu,L}$ or converges at a non-accelerated rate. This closes an open question on one of the most iconic momentum-based schemes and highlights a striking contrast between its practical success and its worst-case guarantees.

Second, we consider learning with compressed information, where error-feedback mechanisms are commonly used to mitigate the loss induced by communication operators. A tight Lyapunov-based analysis of EF and EF21 is provided in [4], including explicit optimal rates and matching lower bounds. Interestingly, this study reveals that both mechanisms share identical worst-case performance in the single-agent smooth strongly convex setting, and that compressed gradient descent may outperform them in both achievable rates and convergence domains. These results enable an "apples-to-apples" comparison between compressed methods using the same Lyapunov framework.

Together, these two applications illustrate how constructive worst-case analyses and systematic Lyapunov search tools provide qualitative insights into fundamental limits of first-order algorithms, from momentum acceleration to communication-efficient learning. They also complement a growing body of work advocating principled methods for certifying tight guarantees in optimization [5, 6, 7, 8].

References

[1] B. Goujaud, A. Taylor, A. Dieuleveut. *Provable non-accelerations of the heavy-ball method*. Mathematical Programming B, 2025.(arXiv:2307.11291).

[2] B. Goujaud, A. Dieuleveut, A. Taylor *Counter-examples in first-order optimization: a constructive approach* IEEE Control Systems Letters 7, 2485–2490

[3] B. Goujaud, A. Taylor, A. Dieuleveut. Open Problem: Two Riddles in Heavy-Ball Dynamics arXiv preprint arXiv:2502.19916

[4] D. B. Thomsen, A. Taylor, A. Dieuleveut. *Tight analyses of first-order methods with error feedback*. 2025.(arXiv:2506.05271).

[5] A. Taylor, J. Hendrickx, and F. Glineur. *Smooth strongly convex interpolation and exact worst-case performance of first-order methods*. Mathematical Programming 161(1), 2017.

[6] A. Taylor, J. Hendrickx, and F. Glineur. *Exact worst-case performance of first-order methods for composite convex optimization*. SIAM J. Optimization 27(3), 2017.

[7] B. Goujaud, A. Dieuleveut, and A. Taylor. *On fundamental proof structures in first-order optimization*. Conference on Decision and Control (CDC), 2023.
[8] B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, and A. Dieuleveut. *PEPit: computer-assisted worst-case analyses of first-order methods in Python*. Mathematical Programming Computation 16(3), 2024.

Effective Fluctuating Continuum Models for Stochastic Gradient Descent

Benjamin Gess
(joint work with Sebastian Kassing, Vitalii Konarovskyi, Nimit Ran)

Stochastic gradient descent (SGD) is the workhorse of modern machine learning, yet its discrete and noisy dynamics remain poorly understood from a stochastic-analytic perspective. A central challenge is to derive continuum models that capture not only the mean behavior of SGD but also its inherent fluctuations, thereby linking discrete optimization algorithms with continuous stochastic dynamical systems. This talk presents recent progress on the rigorous derivation and analysis of such *effective fluctuating continuum models* for Euclidean and Riemannian variants of SGD, in regimes of small learning rates and for large, shallow networks.

Diffusion approximations and their limitations. Classical diffusion approximations replace the SGD iteration

$$Z_{n+1} = Z_n - \eta \nabla \widetilde{R}(Z_n, \xi_n)$$

by a continuous-time stochastic differential equation of the form

$$dY_t = -\nabla R(Y_t)\, dt - \tfrac{\eta}{4}\nabla |\nabla R(Y_t)|^2\, dt + \sqrt{\eta}\, \Sigma^{1/2}(Y_t)\, dW_t,$$

the so-called *stochastic modified equation* (Li, Tai, E 2019). Such equations capture the first-order behavior of SGD and approximate expectations to order $\mathcal{O}(\eta^2)$, but they require strong regularity of the matrix square root $\Sigma^{1/2}$. In practical machine learning settings, in particular in overparameterized settings, where the covariance Σ is typically degenerate, this assumption fails. Moreover, the SME only captures one-point marginals of SGD and thus misses the multi-point correlations essential for dynamical properties such as stochastic synchronization.

Stochastic modified flows. To overcome these limitations, we introduced the notion of *stochastic modified flows (SMFs)*—infinite-dimensional stochastic differential equations driven by cylindrical Wiener processes:

$$dX_t = -\nabla R(X_t)\, dt - \tfrac{\eta}{4}\nabla |\nabla R(X_t)|^2\, dt + \sqrt{\eta} \int_{\Xi} G(X_t, \xi)\, W(d\xi, dt),$$

where $G(x, \xi) = \nabla \widetilde{R}(x, \xi) - \nabla R(x)$. The SMF solves the same martingale problem as the SME while avoiding the irregular square root, requiring only regularity of the individual losses $\widetilde{R}(\cdot, \xi)$. We proved quantitative weak-error bounds showing that the SMF approximates the SGD dynamics to order $\mathcal{O}(\eta^2)$ and, crucially,

matches the multi-point distributions of SGD. This enables the analysis of dynamical features such as stochastic synchronization and Lyapunov stability at the level of the time continuous flow.

In the context of wide neural networks, combining the small-learning-rate limit with the infinite-width (mean-field) scaling yields *distribution-dependent stochastic modified flows (DDSMFs)*. These measure-valued SDEs describe the evolution of the empirical particle distribution of network parameters and provide a quantified continuum description of SGD in large, shallow networks. The DDSMFs thus extend classical McKean–Vlasov and Wasserstein gradient flow limits by incorporating fluctuation effects.

Riemannian extensions. Many optimization tasks in modern learning are naturally posed on non-Euclidean manifolds—examples include principal component analysis on the Grassmann manifold, weight normalization on spheres, and information-geometric optimization on the Fisher–Rao manifold. For such settings we developed the *Riemannian stochastic modified flow (RSMF)*, a diffusion model that approximates Riemannian stochastic gradient descent (RSGD)

$$Z_{n+1} = \mathrm{retr}_{Z_n}\big(-\eta \,\mathrm{grad}\, \widetilde{f}(Z_n, \xi_n)\big)$$

to higher order in the learning rate η. The RSGD dynamics are first approximated by the deterministic Riemannian gradient flow $\dot{z}_t = -\mathrm{grad}\, f(z_t)$ to weak order $\mathcal{O}(\eta)$, and then by the RSMF

$$dX_t = B(X_t)\, dt + G(X_t, \cdot) \circ dW_t,$$

to order $\mathcal{O}(\eta^2)$, where W is a cylindrical Wiener process on $L^2((\Xi, \vartheta); \mathbb{R})$. The drift B includes geometric correction terms accounting for curvature and the Stratonovich interpretation on manifolds. The RSMF provides a geometrically consistent stochastic flow capturing the fluctuations of RSGD.

Outlook. The resulting framework unifies diffusion approximations, mean-field limits, and Riemannian optimization within a single stochastic–geometric theory. The proposed fluctuating continuum models match SGD dynamics to higher order, remain well-posed under minimal regularity, and reproduce the correct multi-point and distributional statistics. They thereby open the way for a systematic analysis of the stochastic dynamics and implicit regularization properties of SGD in high-dimensional and geometrically structured settings by means of analyzing continuous time SDEs.

References

[1] B. Gess, S. Kassing, and V. Konarovskyi, *Stochastic modified flows, mean-field limits and dynamics of stochastic gradient descent*, *J. Mach. Learn. Res.* **25** (2024), Paper No. 103, 1–54; `arXiv:2302.07125`.

[2] B. Gess, S. Kassing, and N. Rana, *Stochastic modified flows for Riemannian stochastic gradient descent*, *SIAM J. Control Optim.* (to appear, 2025); `arXiv:2402.03467`.

Beyond Strong Convexity: Geometry and Optimization under the Polyak–Łojasiewicz Condition

Sebastian Kassing

(joint work with Benjamin Gess, Steffen Dereich, Simon Weissmann)

Many theoretical results in (stochastic) optimization have been derived under strong convexity assumptions or even for quadratic objective functions. However, such assumptions often fail to hold in modern machine learning applications, where objectives are typically non-convex. This talk explores a recent line of research that extends classical results in stochastic gradient-based optimization to broader classes of functions satisfying the Polyak–Łojasiewicz (PL) inequality, a condition that is significantly more relevant for practical deep learning models.

We analyze the heavy ball (HB) method, introduced by Polyak [1], which is one of the earliest momentum-based acceleration techniques in first-order optimization. The heavy ball method can be defined vie the iteration scheme

$$\begin{aligned} X_{n+1} &= X_n + \gamma_{n+1} V_{n+1}, \\ V_{n+1} &= V_n - \gamma_{n+1}\mu V_n - \gamma_{n+1}(\nabla f(X_n) + D_{n+1}), \end{aligned}$$

where (γ_n) is a sequence of step-sizes, $\mu > 0$ is the friction parameter, $f : \mathbb{R}^d \to \mathbb{R}$ is the objective function we aim to minimize and (D_n) is an additional noise term.

It is known that HB may exhibit oscillatory behavior and even divergence for non-quadratic problems. In this talk, we provide a concise account of the local behavior of the heavy ball method in the neighborhood of local minima. We consider the following noise assumptions

- Bounded variance, i.e. $\mathbb{E}[|D_{n+1}|^2 \mid \mathcal{F}_n] \leq C$,
- Overparametrized setting, i.e. $\mathbb{E}[|D_{n+1}|^2 \mid \mathcal{F}_n] \leq C(F(X_n - F(x_*)))$, and,
- Noise-free setting, i.e. $D_n \equiv 0$.

We show almost sure convergence of $(X_n)_{n\in\mathbb{N}}$ and derive convergence rates. In the noise-free setting and under suitable smoothness assumptions, we show that HB indeed accelerates the local linear rate of convergence compared to standard gradient descent. The generalizes a known result for strongly-convex objective functions, to the class of functions satisfying the PL-inequality. The proof is based on a geometric interpretation of the PL-inequality which was recently presented in [2].

References

[1] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, *U.S.S.R. Comput. Math. Math. Phys.* **4** (1964), no. 5, 1–17.

[2] Q. Rebjock and N. Boumal, *Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions*, *Math. Program.* (2024), 1–49.

[3] S. Dereich and S. Kassing, *Convergence of Stochastic Gradient Descent Schemes for Łojasiewicz-Landscapes*, *Journal of Machine Learning* (2024), 245–281.

[4] S. Kassing and S. Weissmann, *Polyak's heavy ball method achieves accelerated local rate of convergence under the Polyak–Łojasiewicz inequality*, `arXiv:2410.16849` (2024).

[5] B. Goujaud, A. Taylor, and A. Dieuleveut, *Provable non-accelerations of the heavy-ball method*, *Math. Program.* (to appear, 2025), 1–59.

[6] B. Gess and S. Kassing, *Exponential convergence rates for momentum stochastic gradient descent in the overparametrized setting*, *Math. Program.* (to appear, 2025).

Fluctuation Analysis of Mean-Field Limits in Overparameterized SGD

Vitalii Konarovskyi

In this talk I reported on the joint work with Benjamin Gess and Rishabh Gvalani [2] concerning the emergence of *conservative stochastic partial differential equations* (SPDEs) as fluctuating mean–field limits of stochastic gradient descent (SGD) in overparametrized shallow neural networks. We view the network parameters as an interacting particle system whose empirical measure evolves under the stochastic updates induced by SGD. In the classical mean–field regime $M \to \infty$ (where M is the network width), one obtains a deterministic transport-type PDE describing the law-of-large-numbers limit of this empirical distribution [4, 5]. Our focus is on the next-order description: the fluctuations around this deterministic limit.

We show that, under natural assumptions on the activation σ and the data distribution, these fluctuations are captured by a nonlinear *conservative measure-valued SPDE*. The equation displays nonlocal and potentially degenerate coefficients reflecting the structure of the neural network and the SGD noise. A substantial part of the work is devoted to developing a well-posedness theory for this class of SPDEs, including strong uniqueness, a superposition principle, and stability estimates in Wasserstein metrics. These analytic tools then allow us to prove quantitative convergence of the empirical measure $\mu_{M,\alpha}$ of the SGD dynamics to the SPDE solution μ_α at the higher rate $o(M^{-1/2})$.

Beyond the mean-field limit, the fluctuation SPDE provides a refined continuum approximation of SGD that captures stochastic effects invisible at the deterministic level. Such higher-order descriptions are central for understanding the role of noise in training dynamics and for connecting SGD with models of fluctuating hydrodynamics. The resulting equations share structural similarities with the Dean–Kawasaki equation [1, 3] but avoid some of its singular features due to the nonlocality induced by the neural network structure.

References

[1] David S. Dean, *Langevin equation for the density of a system of interacting Langevin processes*, J. Phys. A **29** (1996), no. 24, L613–L617. MR 1446882

[2] Benjamin Gess, Rishabh S. Gvalani, and Vitalii Konarovskyi, *Conservative spdes as fluctuating mean field limits of stochastic gradient descent*, Probab. Theory Related Fields **192** (2025), 1447–1515.

[3] Vitalii Konarovskyi, Tobias Lehmann, and Max-K. von Renesse, *Dean-Kawasaki dynamics: ill-posedness vs. triviality*, Electron. Commun. Probab. **24** (2019), Paper No. 8, 9. MR 3916340

[4] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proc. Natl. Acad. Sci. USA **115** (2018), no. 33, E7665–E7671. MR 3845070

[5] G. M. Rotskoff and E. Vanden-Eijnden, *Trainability and accuracy of artificial neural networks: an interacting particle system approach*, Comm. Pure Appl. Math. **75** (2022), no. 9, 1889–1935. MR 4465905

Deep Learning Theory: Statistics, Optimization and the Space Between

Sophie Langer
(joint work with Adam Krzyzak, Michael Kohler, Alina Braun)

Theory around deep learning can be roughly divided into two main directions: *Statistical Theory* and *Optimization*. From a statistical perspective, deep learning methods are often formalized as prediction problems and studied as estimators for tasks such as nonparametric regression or classification. The main goal of these analyses is to provide statistical risk bounds that depend on the number of samples used to train the estimator. Although training algorithms such as (stochastic) gradient descent could, in principle, be incorporated into the analysis, most existing statistical results, e.g., [1, 2]—focus on empirical risk minimizers and do not account for the training procedure. In this talk, we reviewed the limitations of this idealized setting and presented a first result addressing both aspects simultaneously: deriving statistical risk bounds for estimators trained via gradient descent [3]. In a simplified setting where the regression function belongs to the Barron class, that is, its Fourier transform has finite first moment, we show that a shallow neural network with sigmoidal activation, trained with gradient descent on an L_2-penalized loss, achieves the same convergence rates previously established for empirical risk minimizers over this network class. Finally, we sketched several research directions that are needed to extend these results to more complex architectures and broader statistical settings, highlighting the interplay between optimization and statistical theory in deep learning.

References

[1] Schmidt–Hieber, J. (2020). *Nonparametric regression using deep neural networks with ReLU activation function. Annals of Statistics*, 48(4), 1875–1897. doi:10.1214/19-AOS1875.

[2] Kohler, M. and Langer, S. (2021). *On the rate of convergence of fully connected deep neural network regression estimates. Annals of Statistics*, 49(4), 2231–2249. doi:10.1214/20-AOS2034.

[3] Braun, A., Kohler, M., Langer, S. and Walk, H. (2024). *Convergence rates for shallow neural networks learned by gradient descent. Bernoulli*, 30(1), 475–502.

CLTs for noisy SGD via geometric moment contraction

Johannes Schmidt-Hieber
(joint work with Jiaqi Li, Wei-Biao Wu)

Noisy (stochastic) gradient descent (S)GD encompasses a variety of important learning procedures in modern machine learning. A particularly important example is to add dropout noise to every iterate.

To analyse the behaviour of noisy SGD with fixed learning rate, a powerful alternative to the martingale CLT is to interpret SGD as a time series and to apply the machinery of geometric moment contraction and functional dependence measures. Under a fairly simple contraction condition, geometric moment contraction immediately entails the existence of a limiting stationary distribution of the iterates. This is shown by introducing a backwards process which can be proved to converge almost surely. The latter implies then the convergence to a stationary distribution of the forward process.

Regarding the shape of the stationary distribution, geometric moment contraction further implies that the limit distribution can be represented as limit of iterated functions, which provides some insights into their structure and can be used to establish a normal approximation (quenched CLT). For stronger Gaussian approximation results, one can moreover employ the concept of functional dependence measures, which measure how much the noise injection in one iterate influences subsequent iterates. Interestingly, under decay conditions on the functional dependence measure it is even possible to derive the convergence rate of the central limit theorem.

These concepts can be applied to (S)GD with dropout for a fixed learning rate. For more details we refer to [1, 2].

References

[1] J. Li, J. Schmidt-Hieber, W. B. Wu *Asymptotics of stochastic gradient descent with dropout regularization in linear models*, ArXiv:2409.07434.
[2] J. Li, Z. Lou, J. Schmidt-Hieber, W. B. Wu, *Statistical guarantees for high-dimensional stochastic gradient descent*, NeurIPS (2025), *to appear*.

Deep Learning Theory Through the Lens of Diagonal Linear Networks

Scott Pesme
(joint work with Nicolas Flammarion and Loucas Pillaud-Vivien)

Surprisingly, many optimisation phenomena which occur in complex neural networks also appear in so-called 2-layer diagonal linear networks. This rudimentary architecture, which consists of a two layer feedforward linear network with a diagonal inner weight matrix, has the advantage of revealing interesting training characteristics while keeping the theoretical analysis clean and insightful.

In this talk, I will provide an overview of various theoretical results concerning the depicted architecture, while making links with experimental observations from "real and practical" neural networks.

I will begin with a general introduction to diagonal linear networks and present key results on the optimisation trajectory of gradient flow, following the analysis of [1], which will serve as the basis for the remainder of the talk. I will then show how hyperparameters, such as the stochastic gradient descent (SGD) stepsize, influence the optimisation trajectory and thereby the generalisation ability of the obtained solution. In particular, we will see that the noise inherent to SGD drives the iterates toward solutions with smaller ℓ_1-norm than those recovered by gradient flow for the same initialisation. This part of the talk is mostly based on the results from [2].

In the second part of the talk, which is based on [3], I will focus on saddle-to-saddle dynamics. In the regression setting with vanishing initialisation, gradient flow follows a sequence of saddle points before reaching the minimum ℓ_1-norm interpolating solution. I will describe how the visited saddles and jump times can be characterised through a simple recursive procedure, akin to the Homotopy algorithm for the Lasso path. This yields an incremental activation of coordinates and provides a clear description of the trajectory under minimal assumptions.

These results open several directions for future work. A natural next step is to extend the analysis to more complex architectures that are closer to those used in practice. Promising starting points include fully connected linear networks and 2-layer ReLU networks, where many aspects of the training dynamics are still not understood.

References

[1] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, *Kernel and rich regimes in overparametrized models*, Conference on Learning Theory, PMLR, (2020), 3635–3673.

[2] S. Pesme, L. Pillaud-Vivien, and N. Flammarion, *Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity*, Advances in Neural Information Processing Systems **34** (2021), 29218–29230.

[3] S. Pesme and N. Flammarion, *Saddle-to-saddle dynamics in diagonal linear networks*, Advances in Neural Information Processing Systems **36** (2023), 7475–7505.

MAP Estimation with Denoisers: Convergence Rates and Guarantees

Scott Pesme

(joint work with Giacomo Meanti, Michael Arbel and Julien Mairal)

Denoiser models have become powerful tools for inverse problems, enabling the use of pretrained networks to approximate the score of a smoothed prior distribution. These models are often used in heuristic iterative schemes aimed at solving Maximum a Posteriori (MAP) optimisation problems, where the proximal operator of the negative log-prior plays a central role. In practice, this operator is intractable,

and practitioners plug in a pretrained denoiser as a surrogate—despite the lack of general theoretical justification for this substitution.

In this talk, based on our work [1], I will present a simple algorithm—quite close to what is already used in practice—that can be interpreted as gradient descent on a smoothed version of the proximal objective. Under a log-concavity assumption on the prior p, we show that this scheme converges to the true proximal operator. However, the hyperparameters required by the theory do not lead to satisfactory performance in practical settings. I will discuss the intuition behind this gap and outline an ongoing work, which proposes alternative hyperparameter choices that yield promising empirical results.

References

[1] S. Pesme, G. Meanti, M. Arbel, and J. Mairal, *MAP estimation with denoisers: Convergence rates and guarantees*, NeurIPS 2025 – Advances in Neural Information Processing Systems (2025).

Noisy gradient flows in machine learning: two examples

Anna Shalova

Many machine learning models are variational in nature and thus can be effectively analyzed using variational methods in general and the theory of gradient flows in particular. Gradient flows is a class of evolution equations which describe the dynamics aimed to minimize the *driving functional*. Being a bit more precise, given a functional F defined on a metric space, the gradient flow equation, formally written as

$$\partial_t x_t = -DF(x_t), \tag{1}$$

yields the trajectories which minimize F in the most efficient (from the metric perspective) way possible. The definition of DF above depends on the structure of the metric space; e.g. for Hilbert spaces DF denotes the Frechet derivative of F. Equations with a gradient flow structure allow for a large set of tools to study the solutions of the underlying models. In these two talks I gave two applications of the gradient flow formalism in machine learning.

Example 1: Noisy Gradient Descent *(based on the joint work with M. Peletier and A. Schlichting* [1]*).* The first application is related to the gradient descent method, which in the classical (noiseless) case is formulated as the iterative algorithm with the updates of the form

$$\theta_{k+1} = \theta_k - \alpha \nabla L(\theta_k).$$

Note that gradient descent as defined above is in fact the forward Euler discretization of the gradient flow dynamics (1). In [1] we consider the noisy counterpart of the gradient descent, namely the iterative algorithm with the updates of the form

$$\theta_{k+1} = \theta_k - \alpha \nabla \hat{L}(\theta_k, \eta_k), \quad \eta_k \sim \rho(\sigma), \tag{2}$$

where the noisy loss $\hat{L}$ satisfies the compatibility condition $\hat{L}(\theta, 0) = L(\theta)$ and η_k denotes the noise injected at k-th iteration. We extend the approach of [2] in order to characterize the limiting behaviour of (2) in the joint limit of small step-size $\alpha \to 0$ and small noise $\sigma \to 0$. In particular, we show that the system (2) exhibits a multiscale behaviour, where the fast dynamics is generated by the original loss function L and the slow dynamics is defined by the injected noise and the form of the noisy loss function $\hat{L}$. Our results show that the structure of the noise affects not just the form of the limiting process, but also the time scale at which the evolution takes place. We apply the theory to Dropout, label noise and classical SGD (minibatching) noise, and show that these evolve on different time scales.

Example 2: Noisy Transformers *(based on the joint work with A. Schlichting* [3]*).* As it was recently proposed in [4], the dynamics of tokens in the Transformer model in its simplified form can be described by the interacting particle system of form:

$$\dot{X}_i = -\nabla_i^{\mathbb{S}^{n-1}} \frac{1}{N} \sum_{i,j=1}^{N} W(X_i, X_j), \quad W(x, y) = \frac{1}{\beta} e^{\beta \langle x, y \rangle}.$$

In [3] we consider the noisy counterpart of the above dynamics, namely the interacting particle system described by:

$$dX_i = -\nabla_i^{\mathbb{S}^{n-1}} \frac{1}{N} \sum_{i,j=1}^{N} W(X_i, X_j) dt + \sqrt{2\gamma^{-1}} dB_t^i,$$

where B_t^i are independent Brownian motions on a sphere and $\gamma \in \mathbb{R}_+$ is the inverse temperature parameter. The mean-field limit of the above system equation is the McKean-Vlasov equation, namely the partial differential equation of form:

$$\partial_t \rho = \gamma^{-1} \Delta \rho + \operatorname{div}(\rho \nabla_x W(x, \cdot) * \rho), \tag{3}$$

which also admits a gradient flow formulation in the space of the probability measures equipped with the Wasserstein-2 distance.

In [3] we give an extensive characterization of the stationary measures of noisy transformers. In particular, we study stationary solutions of the McKean-Vlasov equation (3) on a high-dimensional sphere and other compact Riemannian manifolds. We extend the equivalence of the energetic problem formulation to the manifold setting and characterize critical points of the corresponding free energy functional. On a sphere, we employ the properties of spherical convolution to study the bifurcation branches around the uniform state. We also give a sufficient condition for an existence of a discontinuous transition point in terms of the interaction kernel and compare it to the Euclidean setting.

References

[1] A. Shalova, A. Schlichting and M. Peletier, *Singular-limit analysis of gradient descent with noise injection*, Preprint. arXiv:2404.12293 (2024)

[2] Z. Li, T. Wang and S. Arora, *What Happens after SGD Reaches Zero Loss?–A Mathematical Framework*, International Conference on Learning Representations. (2022)

[3] A. Shalova and A. Schlichting, *Solutions of stationary McKean-Vlasov equation on a high-dimensional sphere and other Riemannian manifolds*, Preprint. arXiv:2412.14813 (2024)
[4] 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)

The probabilistic effects of depth in deep learning

Mariia Seleznova

Depth plays a central role in modern deep learning, yet its probabilistic effects are subtle and not fully captured by classical theories that focus on the infinite-width limit. This talk explores how jointly scaling depth and width shapes the signal-propagation statistics of wide neural networks, highlighting a contrast between feedforward architectures with independent weights and recurrent networks with shared weights. For clarity, we outline the main intuition in a simple linear forward-chain setting, though the presented results extend to broader settings [1, 2].

Feedforward case. Consider a linear network with Glorot initialization scaling:

$$\mathbf{h}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)}, \qquad \mathbf{W}^{(\ell)}_{ij} \sim \mathcal{N}(0, 1/n).$$

If the matrices $\mathbf{W}^{(\ell)}$ are independent across layers, then $\mathbf{h}^{(\ell-1)}$ and $\mathbf{W}^{(\ell)}$ are independent, and

$$\mathbb{E}\|\mathbf{h}^{(\ell)}\|^2 = \mathrm{Tr}\big(\mathbb{E}[\mathbf{h}^{(\ell-1)}\mathbf{h}^{(\ell-1)\top}]\,\mathbb{E}[(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}]\big) = \mathbb{E}\|\mathbf{h}^{(\ell-1)}\|^2.$$

Hence signal variance is exactly preserved. The infinite-width approximations that assume such independence holds asymptotically, therefore, provide an accurate description for linear feedforward networks.

Recurrent (shared-weight) case. If all layers share the same matrix $\mathbf{W}$, so that $\mathbf{h}^{(\ell)} = \mathbf{W}\mathbf{h}^{(\ell-1)} = \mathbf{W}^\ell\mathbf{h}^{(0)}$, this independence no longer holds. In the infinite-width approximation (e.g., Tensor Programs [4]) one still predicts asymptotic variance preservation,

$$\mathbb{E}\|\mathbf{h}^{(\ell)}\|^2 = \mathbb{E}\|\mathbf{h}^{(\ell-1)}\|^2 + o(1), \qquad n \to \infty \text{ at fixed } \ell,$$

suggesting stability when width grows at fixed depth. However, finite-width effects are amplified when depth and width are scaled jointly. Under Glorot initialization, the spectral radius of $\mathbf{W}$ typically exceeds 1 by $O(n^{-1/2})$ [3], leading to exponential norm growth at depth ℓ scaled as $\ell/\sqrt{n} \to \alpha > 0$ [1]:

$$\mathbb{E}\|\mathbf{W}^t\mathbf{h}^{(0)}\|^2 \;\geq\; c\,\exp\Big(\frac{\ell^2}{2n}\Big), \qquad c > 0.$$

This illustrates a qualitative failure of the infinite-width prediction.

References

[1] N. Bar*, M. Seleznova*, Y. Alexander, G. Kutyniok, R. Giryes, *Revisiting Glorot Initialization for Long-Range Linear Recurrences*, Advances in Neural Information Processing Systems (2025).

[2] M. Seleznova, G. Kutyniok, *Neural Tangent Kernel Beyond the Infinite-Width Limit: Effects of Depth and Initialization*, International Conference on Machine Learning (2022).

[3] B. Rider, C. Sinclair, *Extremal laws for the real Ginibre ensemble*, The Annals of Applied Probability (2014).

[4] G. Yang, *Wide feedforward or recurrent neural networks of any architecture are gaussian processes*, Advances in Neural Information Processing Systems (2019).

Adaptive denoising diffusion modelling via random time reversal

Lukas Trottner

(joint work with Sören Christensen, Jan Kallsen, Claudia Strauch)

In the past ten years or so, generative modelling has become one of the most important research areas in machine learning. The impressive results of the most successful models are enabled by increasingly sophisticated probabilistic designs of algorithms – all based on the idea to learn a (random) transformation that maps easy-to-sample from noise into expressive samples – together with significant advances in the design and optimisation of neural networks that are used as approximators for the noise transformation.

In this talk I discuss a novel class of *generative diffusion models*. In their basic form, such models were introduced in [3] and provide an iterative generative procedure for sampling from a taget distribution α that builds on time-reversal of diffusion processes: given appropriate assumptions on its coefficients, the time-reversal of a diffusion process at a *fixed time* T is again a diffusion process, but now with a different time-dependent drift $\overline{b}(t,x)$ that involves the *score* $\nabla \log p_t(x)$ of the marginal distribution $p_t(x) = \int p_t(y,x)\,\alpha(dy)$ of the forward process initialised in α. Since the crucial assumption underlying generative modelling is that α is unknown but represented by a data sample, the idea of denoising diffusion models is now to first learn the score by running the forward noising process on the training data, and then to simulate the estimated backward process initialised according to an approximation of the forward terminal distribution p_T.

As a downside of time-reversal at a deterministic time T, the backward process will always be time-*inhomogeneous*. This makes the generative procedure non-adaptive to the noise level along the generated path and therefore implies that the specification of the simulation time is unclear if the intialisation does not correspond to the pure terminal noise p_T – preventing direct applications of the algorithm to different tasks such as image reconstruction.

The idea of our paper [2] is to condition a homogeneous base diffusion process using Doob's h-transform instead, which terminates the process at a suitable sampling distribution at a *random* time and thereby allows us to accomplish a denoising procedure that is again characterised an $\overleftarrow{h}$-transform of the base diffusion, whence preserving the time-homogeneous nature of the noising process. The need

to estimate the time-dependent score functions $\nabla \log p_t(x)$ is replaced with the tasks of estimating $\nabla \log \overleftarrow{h}$ (via denoising score matching) and learning the random stopping criterion of the time-reversal. Our model is particularly well suited for generating data with lower intrinsic dimensions, as the termination criterion simplifies to a first hitting rule. A key feature of the model is its adaptability to the target data, enabling a variety of downstream tasks using a pre-trained unconditional generative model. We highlight this point in [1] by demonstrating how our generative model may be used as an unsupervised learning algorithm: in high dimensions the model outputs with high probability the metric projection of a noisy observation y of some latent signal x onto the lower-dimensional support of the data, which – in the spirit of generative modelling – we don't assume to be analytically accessible, but to be only represented by an unlabeled training data set.

References

[1] S. Christensen, J. Kallsen, C. Strauch, and L. Trottner, *Model-free filtering in high dimensions via projection and score-based diffusions*, arXiv preprint arXiv:2510.23197, 2025.
[2] S. Christensen, C. Strauch, and L. Trottner, *Beyond Fixed Horizons: A Theoretical Framework for Adaptive Denoising Diffusions*, arXiv preprint arXiv:2501.19373, 2025.
[3] 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.

Low-Rank Bias in Transformers: Saddle-to-Saddle Dynamics, Symmetries, and a Surrogate Model

Katerina Papagianoulli
(joint work with Hana Tseran, Federico Pasqualotto)

Transformers have recently been observed to exhibit incremental learning dynamics, where the difference between trained and initial weights progressively increases in rank over the course of training. This talk presents a theoretical framework that explains this behavior through a gradient-flow analysis of low-rank parameterizations in attention layers. Using a simple surrogate model, we uncover saddle-to-saddle training dynamics: optimization proceeds through stages, extended near-stationary low-rank phases, followed by fast rank-increase events driven by negative curvature directions. We further show that a gradual rank increase emerges naturally in this model when the incremental dynamics are appropriately scaled to enable efficient escape from saddle points. Finally, we demonstrate that these dynamics obey conservation properties analogous to angular momentum.

Towards principled and systematic approaches to the analysis and design of first-order optimization algorithms

Adrien Taylor

An important role of the optimization community is to develop algorithms with convergence and complexity guarantees. On one hand, this allows users to trust their algorithms; on the other, it enables experts to potentially improve those algorithms. Unfortunately, such guarantees are generally obtained through abstract combinations of (potentially large numbers of) inequalities that offer little to no global insight, are perceived as the domain of experts, and may hide subtle and hard-to-detect errors. First-order algorithm particularly suffer from those issues, as many such algorithms appeared in the context of machine learning.

In this presentation, we will present a high-level overview of systematic approaches for the analysis and design of optimization algorithms. The presentation will be illustrated through concrete examples, as the core principles underlying these methods are already embedded in fundamental optimization schemes.

More precisely, we discussed the following themes

- the *primal* performance estimation problem (coined in [1], and presented using the formalism from [2, 3]), that allows to constructively obtain examples of worst-case behaviors of known optimization algorithms. The methodology presented is implemented in the PEPit package, which enables users to apply the framework without requiring direct semidefinite programing modeling steps.
- The *dual* performance estimation problem, whose feasible points correspond to *convergence proofs*. This convex problem enables to efficiently search (numerically) through the space of proofs (see, e.g., [4, 2]) and even to frame the proof verification process using symbolic computations (see, e.g., [6]).
- How to leverage the framework for designing *optimal* optimization algorithms, via minimax problem (minimize worst-case), along with examples of algorithms that were developed through this system [1, 7, 6].

This talk is based on joint works with great collaborators, who were acknowledged during the presentation.

References

[1] Y. Drori, M. Teboulle (2014). *Performance of first-order methods for smooth convex minimization: a novel approach*. Mathematical Programming 145(1)

[2] A. Taylor, J. Hendrickx, F. Glineur (2017). *Smooth strongly convex interpolation and exact worst-case performance of first-order methods*, Mathematical Programming 161(1).

[3] A. Taylor, J. Hendrickx, F. Glineur (2017). *Exact worst-case performance of first-order methods for composite convex optimization*, SIAM Journal on Optimization 27(3).

[4] B. Goujaud, A. Dieuleveut, A. Taylor (2023). *On fundamental proof structures in first-order optimization*, Conference on Decision and Control (CDC).

[5] B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, A. Dieuleveut (2024). *PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python*, Mathematical Programming Computation 16(3).

[6] A. Taylor, Y. Drori (2023). *An optimal gradient method for smooth strongly convex minimization* , Mathematical Programming 199(1).
[7] D. Kim, J. Fessler (2016). *Optimized first-order methods for smooth convex minimization*, Mathematical Programming 159(1).

On global polynomial-time computable estimators in non-linear inverse problems

Sven Wang

Statistical non-linear inverse problems are notoriously computationally difficult. For instance, high-dimensional Markov Chain Monte Carlo (MCMC) schemes used for Bayesian inference may suffer from the non-log-concavity and potential multimodality of the high-dimensional posterior distribution. Likewise, regularisation-based estimators may suffer from multimodality of loss landscapes. A key class of non-linear inverse problems is given by parameter identification problems in partial differential equations (PDEs) and stochastic differential equations (SDEs).

We give a survey of recent computational-statistical theory for such models. We begin with discussing some literature devising mixing times for MCMC under 'warm-start' conditions, based on [1]. However, even in prototypical examples, it is unclear whether polynomial-time Bayesian inference is *globally* algorithmically feasible. We then develop a theory–which is current ongoing work–for polynomial-time computation via so-called 'generalized M-estimators', which gives rise to globally polynomial-time computable estimators. This answers the algorithmic feasibility question for a certain class of inverse problems where the defining differential operators is linear in the parameter. Our theory covers several important non-linear inverse problems, such as the widely studied Darcy flow problem.

References

[1] R. Nickl and S. Wang, "On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms," *Journal of the European Mathematical Society*, vol. 26, no. 3, pp. 1031–1112, 2022.

Convergence analysis of stochastic gradient methods under gradient domination

Simon Weissmann
(joint work with Waïss Azizian, Leif Döring, Sara Klein)

Stochastic gradient descent (SGD) is a foundational optimization method in modern machine learning, yet classical convergence theory typically relies on strong convexity, an assumption that fails for most practical non-convex learning problems. In this talk, we present a convergence theory for minimizing $f : \mathbb{R}^d \to \mathbb{R}$ with SGD under a substantially weaker and widely applicable structural condition, namely β-*gradient domination*:

$$\|\nabla f(x)\| \geq c\,(f(x) - f^*)^\beta\,, \quad f^* := \inf_{x\in\mathbb{R}^d} f(x)\,, \quad \forall x \in \mathbb{R}^d. \tag{1}$$

This assumption can be viewed as a generalized, and often much weaker, form of the Polyak–Łojasiewicz condition. Importantly, its localized form has been verified in a variety of modern machine learning settings, including policy-gradient methods in reinforcement learning and the training of deep neural networks with analytic activation functions. Our main contribution in [4] is a set of *almost sure convergence rates for the last iterate of SGD under gradient domination*, achieved through a refined super-martingale analysis.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space. The SGD scheme is defined as $(\mathcal{F}_n)$-adapted stochasic process

$$X_{n+1} = X_n - \gamma_n \nabla f(X_n) + \gamma_n D_{n+1}\,,$$

where $(D_n)_{n\in\mathbb{N}}$ is an adapted sequence of martingale differences, that is $\mathbb{E}[D_{n+1} \mid \mathcal{F}_n] = 0$ for all $n \in \mathbb{N}$. Using the L-smoothness of f, the descent lemma yields

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] \le f(X_n) - \gamma_n \|\nabla f(X_n)\|^2 + \frac{L\gamma_n^2}{2}\mathbb{E}[\|D_{n+1}(X_n)\|^2 \mid \mathcal{F}_n]\,.$$

Assuming the ABC-condition

$$\mathbb{E}\big[\|D_{n+1}\|^2 \mid \mathcal{F}_n\big] \le A(f(X_n) - f^*) + B\|\nabla f(X_n)\|^2 + C, \quad n \in \mathbb{N}\,,$$

one obtains a fundamental recursion for the optimality gap $Y_n := f(X_n) - f^*$:

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \;\le\; (1 + c_1\gamma_n^2)Y_n - c_2\gamma_n Y_n^{2\beta} + c_3\gamma_n^2\,.$$

To extract almost sure convergence rates from this recursion, we apply a super-martingale convergence theorem. It asserts that if the step sizes satisfy $\gamma_n = \Theta(n^{-\theta})$ with $\theta \in (\frac{1}{2}, 1)$ then

$$Y_n \in o\Big(n^{-(1-\eta)}\Big) \qquad \text{almost surely}\,,$$

for all $\eta \in \big(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\big)$. Our argument is based on the classical Robbins–Siegmund theorem [3] and extends the recent super-martingale analysis of Liu and Yuan [2], which covers the strongly dominated case $\beta = \frac{1}{2}$. Optimizing over θ yields almost sure last-iterate convergence rates for SGD of the form

$$f(X_n) - f^* \in o\Big(n^{-\frac{1}{4\beta-1}+\varepsilon}\Big), \qquad \forall\, \varepsilon > 0,$$

which are arbitrarily close to the best known rates in expectation [1], and constitutes the first almost sure convergence rate for SGD under weak global gradient domination. In our paper [4], we further extend the convergence analysis to the setting of local gradient domination.

References

[1] I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash. *Sharp analysis of stochastic optimization under global Kurdyka-Lojasiewicz inequality*, In Advances in Neural Information Processing Systems, volume 35 (2022), 15836–15848.

[2] J. Liu and Y. Yuan. *On almost sure convergence rates of stochastic gradient methods*, In Proceedings of Thirty Fifth Conference on Learning Theory, volume 178 of Proceedings of Machine Learning Research (2022), 2963–2983.

[3] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, In Optimizing Methods in Statistics Academic Press, Cambridge (1971), 233–257.

[4] S. Weissmann, S. Klein, W. Azizian, and L. Döring, *Almost sure convergence of stochastic gradient methods under gradient domination*, Transactions on Machine Learning Research (2025).

Reporter: Sophie Langer

Participants

Julia Ackermann
Fachgruppe Mathematik und Informatik
Fakultät für Mathematik und Naturwissenschaften
Bergische Universität Wuppertal
Gaußstr. 20
42119 Wuppertal
GERMANY

Prof. Dr. Steffen Dereich
Institut für Mathematische Stochastik
Universität Münster
Einsteinstraße 62
48149 Münster
GERMANY

Aymeric Dieuleveut
Centre de Mathématiques
École Polytechnique
Plateau de Palaiseau
91128 Palaiseau Cedex
FRANCE

Prof. Dr. Benjamin Gess
Institut für Mathematik
Technische Universität Berlin
Str. des 17. Juni 136
10587 Berlin
GERMANY

Sebastian Kassing
Institut für Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Vitalii Konarovskyi
Department of Mathematics,
Faculty of Mathematics, Informatics and Natural Sciences,
University of Hamburg
Bundesstr. 55
20146 Hamburg
GERMANY

Prof. Dr. Sophie Langer
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum
GERMANY

Katerina Papagiannouli
University of Pisa &
Max-Planck-Institute for MiS
Inselstraße 22
04103 Leipzig
GERMANY

Scott Pesme
INRIA Rhone-Alpes
655 avenue de l'Europe
38334 Montbonnot, St. Ismier, Cedex
FRANCE

Prof. Dr. Johannes Schmidt-Hieber
Department of Applied Mathematics
University of Twente
Drienerlolaan 5
7522 NB Enschede
NETHERLANDS

Dr. Mariia Seleznova
Mathematisches Institut
Ludwig-Maximilians-Universität
München
Akademiestr. 7
80799 München
GERMANY

Dr. Anna Shalova
Korteweg-de Vries Institute for
Mathematics
P.O. Box 94248
1090 GE Amsterdam
NETHERLANDS

Dr. Adrien Taylor
National Institute for Research in
Computer Science and Control
INRIA – CS 42112
47 Rue Barrault
75013 Paris Cedex
FRANCE

Dr. Lukas Trottner
Universität Stuttgart
Fachbereich Mathematik
Pfaffenwaldring 57
70569 Stuttgart
GERMANY

Prof. Dr. Sven Wang
Institute of Mathematics
Swiss Federal Technology Institute of
Lausanne (EPFL)
1015 Lausanne
SWITZERLAND

Prof. Dr. Simon Weissmann
Institut für Mathematik
Universität Mannheim
68159 Mannheim
GERMANY