Theoretical Investigation of Variational Inference in High-Dimensions

PhD Thesis Proposal

Supervisors:
Bruno Loureiro (CNRS, DI-ENS & PR[AI]RIE Fellow) & Marylou Gabrié (LPENS)

1. Context

Funding — This thesis will take place within the PR[AI]RIE-PSAI project of which Bruno Loureiro is a fellow.

Scientific context — **Variational Inference (VI)** has emerged as a powerful framework for approximating high-dimensional, intractable probability distributions by optimizing over a family of *tractable* distributions [4], alternatively to sampling. It has widespread applications **in statistical physics, notably for approximating Boltzmann measures** and in Bayesian inference—where one seeks to approximate posterior distributions. Recent breakthroughs in **generative modeling** (e.g., diffusion models and normalizing flows) have shown that highly expressive parametric distributions can serve as flexible *variational families* [25, 31]. However, VI crucially differs from standard generative modeling in that the algorithm leverages knowledge of an *unnormalized* target density rather than relying on an abundance of training samples.

To this day, little theoretical work have analyzed VI. Like any learning framework, VI suffers from two main bottlenecks:

- **Approximation error**, i.e., the target distribution may lie outside the chosen variational family, introducing unavoidable bias.
- **Optimization error**, arising from the non-convexity of the parameterized KL minimization and leading to suboptimal solutions or *mode collapse* (where multimodal targets are poorly captured by focusing on fewer modes).

While Gaussian variational families have been extensively analyzed [16, 18, 9], there is comparatively little rigorous theory for more expressive families such as mixtures of Gaussians, normalizing flows, or diffusion-based models. In particular, **mode collapse** is recognized in practice but lacks a clear theoretical understanding [5, 28].

This thesis will develop theoretical tools to analyze VI leveraging **statistical physics**. This field has played a fundamental role in the development of high-dimensional probability, both from mathematical and computational perspectives. Originally motivated by the study of macroscopic properties in many-particle systems, the theoretical framework established by physicists in the late 19th century has since found fertile applications across various disciplines, particularly in computer science. The interplay between these fields is long-standing, with several influential algorithms in computer science drawing direct inspiration from statistical physics, including simulated annealing [17] and diffusion models [27].

Beyond serving as a conceptual source of inspiration, this connection has also provided powerful analytical tools for the study of high-dimensional problems motivated by machine learning, starting with the pioneering work of [12]. Indeed, over the past decade these tools have been successfully

employed to investigate multiple aspects of neural networks, such as the the interplay between over-parametrization and generalization [29, 11, 14, 26], the benefits of feature learning to generalization [8, 7], the analysis of descent-based training algorithms [33, 3, 2, 1, 22, 21, 20] and the analysis of the geometry of loss landscapes [13, 19].

While statistical physics tools have been extensively applied to supervised learning, their use in generative modeling remains largely uncharted. Developing this connection lies at the core of this thesis project.

2. Objectives and Scientific roadmap

The **central aim of this thesis is to develop theoretical tools to understand the approximation power and optimization properties** of VI in high dimensions, focusing on state-of-the-art models (mixtures, flows, and diffusion). The project will build upon methods from **statistical physics** and **high-dimensional probability**, which have proven instrumental in studying supervised learning [29, 26, 3] but remain less explored for generative modeling. In particular, the project will leverage tools such as the replica and cavity methods, dynamical mean-field theory, as well as the Kac-Rice formula — originally developed for studying the geometry of high-dimensional energy landscapes — to establish a link between the complexity of VI's loss landscapes and the computational challenges of efficiently sampling from the corresponding high-dimensional probability measures. By extending techniques that have proven successful in supervised learning, these tools will provide insights into how various aspects of the problem — such as the choice of variational family, training algorithm, and hyperparameter tuning — affect the difficulty of approximating and sampling high-dimensional measures under finite sample complexity and expressivity constraints.

In large-scale systems, such approaches often reveal phase transitions, concentration phenomena, and scaling laws crucial to understanding the success and pitfalls of algorithms, and often leading to a principled understanding of how to improve them.

The thesis will be organized on three research axis, which will provide a roadmap for addressing the challenges highlighted above with concrete and deliverable goals.

First axis: Approximation Error for Flexible Variational Families

(i) Gaussian Mixture Families. Going beyond basic Gaussian families, *finite* or *infinite* mixtures can capture complex multimodal targets [15]. We aim to derive quantitative error bounds under the reverse KL criterion, focusing initially on simplified settings (e.g., isotropic covariance) and gradually increasing complexity (e.g., full covariance, adaptive number of components).

(ii) Neural Network Parametrizations. Modern VI frequently employs normalizing flows or diffusion-based distributions. We will investigate:

- **Expressiveness**: How do architectural factors (depth, width, activation functions) constrain the variational family?
- **Continuous-time vs Discrete-time Diffusions**: Realistic implementations use discretized SDEs; understanding the gap between continuous and discrete dynamics is crucial for quantifying approximation errors.

Second axis: Optimization Dynamics and Mode Collapse

Non-convex objectives in high dimensions often lead to spurious local minima or to solutions capturing fewer modes of the target distribution. Our goal is to characterize such phenomena more rigorously:

- **Gradient Flow Analyses.** Extending [10, 18], we will derive idealized gradient-flow equations for mixtures and neural-network-based VI, comparing them to finite-step updates in practice.
- **Over-parameterization.** Mirroring results in supervised learning [23, 6], we will ask whether *increasing* the capacity (e.g., number of mixture components, network width) systematically avoids poor local minima in VI.
- **Mode Collapse Mechanisms.** Focusing on Gaussian mixtures or synthetic multi-modal targets, we will relate the loss geometry to the typical timescale or iteration count at which collapse occurs, as well as propose heuristics to mitigate it [30].

Third axis: Path-Guided and Annealing-Based Variational Methods

Annealing is a classical strategy to traverse multimodal landscapes by smoothly deforming an easy-to-sample distribution (e.g., a base Gaussian) into the target. Beyond parallel tempering and sequential Monte Carlo, the latest methods include *diffusion VI* [24, 32]. We will:

- **Analyze Annealing Schedules**: Quantify speed-accuracy trade-offs for gradually transitioning from the base to the target.
- **Optimal Path Design**: In diffusion or flow-matching frameworks, identify how to schedule the formation of modes to ensure coverage and avoid collapses.

3. Timeline and Candidate's profile

Timeline of the thesis — In Year 1, the focus will be on deepening knowledge of existing Variational Inference (VI) theory and relevant statistical physics tools, conducting initial theoretical studies on simple Gaussian mixtures, and validating predictions through small-scale experiments. In Year 2, the work will extend to more flexible families like normalizing flows, begin a rigorous analysis of training dynamics, and investigate mode collapse through numerical experiments. Finally, in Year 3, the project will formulate and test annealing strategies for VI, study the role of over-parameterization, and consolidate results into a unifying framework.

Project requirements — This thesis involves both significant computational and analytical components, demanding a candidate with a strong mathematical background and solid programming skills. Furthermore, the interdisciplinary nature of the proposed research calls for a candidate with robust training in physics, a willingness to assimilate concepts from machine learning, and the ability to communicate effectively with diverse research communities.

Non-discrimination, ouverture et transparence — L'ensemble des partenaires de PR[AI]RIE-PSAI s'engagent à soutenir et promouvoir l'égalité, la diversité et l'inclusion au sein de ses communautés. Nous encourageons les candidatures issues de profils variés, que nous veillerons à sélectionner via un processus de recrutement ouvert et transparent.

References

[1] L. Arnaboldi, Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan. Online learning and information exponents: On the importance of batch size, and time/complexity tradeoffs. *arXiv:2406.02157*, 2024.

[2] L. Arnaboldi, F. Krzakala, B. Loureiro, and L. Stephan. Escaping mediocrity: how two-layer networks learn hard single-index models with sgd. *CoRR*, 2023.

[3] L. Arnaboldi, L. Stephan, F. Krzakala, and B. Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *COLT*, pages 1199–1227. PMLR, 2023.

[4] D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[5] D. Blessing, X. Jia, J. Esslinger, F. Vargas, and G. Neumann. Beyond ELBOs: A Large-Scale Evaluation of Variational Methods for Sampling, June 2024. arXiv:2406.07423 [cs, stat].

[6] L. Chizat and F. Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[7] H. Cui, L. Pesce, Y. Dandi, F. Krzakala, Y. Lu, L. Zdeborova, and B. Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pages 9662–9695. PMLR, 21–27 Jul 2024.

[8] Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024.

[9] M. Diao, K. Balasubramanian, S. Chewi, and A. Salim. Forward-Backward Gaussian Variational Inference via JKO in the Bures-Wasserstein Space. In *ICML*, pages 7960–7991. PMLR, July 2023. ISSN: 2640-3498.

[10] J. Domke, R. Gower, and G. Garrigos. Provable convergence guarantees for black-box variational inference. *NeuRIPS*, 36:66289–66327, December 2023.

[11] S. d'Ascoli, Maria Refinetti, Giulio Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *ICML*, pages 2280–2290. PMLR, 2020.

[12] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[13] M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

[14] F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh, editors, *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR, 13–18 Jul 2020.

[15] T. Huix, A. Korba, A. Durmus, and E. Moulines. Theoretical guarantees for variational inference with fixed-variance mixture of gaussians. In *ICML*. PMLR, 2024.

[16] A. Katsevich and P. Rigollet. On the Approximation Accuracy of Gaussian Variational Inference, January 2024. arXiv:2301.02168 [math].

[17] S. Kirkpatrick, C. Gelatt Jr, and M. Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[18] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. Variational inference via Wasserstein gradient flows. *NeuRIPS*, 35:14434–14447, 2022.

[19] A. Maillard, G. Ben Arous, and G. Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning*, pages 287–327. PMLR, 2020.

[20] S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

[21] S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, and L. Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. *NeuRIPS*, 32, 2019.

[22] S. Mannelli, F. Krzakala, P. Urbani, and L. Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *ICML*, pages 4333–4342. PMLR, 2019.

[23] S. Mei, A. Montanari, and P.M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, August 2018. arXiv: 1804.06561.

[24] M. Noble, L. Grenioux, M. Gabrie, and A. Durmus. Learned Reference-based Diffusion Sampler for multi-modal distributions. In *ICLR*, October 2025.

[25] D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *ICML*, pages 1530–1538. PMLR, June 2015. ISSN: 1938-7228.

[26] D. Schröder, H. Cui, D. Dmitriev, and B. Loureiro. Deterministic equivalent and error universality of deep random features learning. In *ICML*, pages 30285–30320. PMLR, 2023.

[27] J. Sohl-dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Jmlr*, 37(ICML), 2015. arXiv: 1503.03585 ISBN: 1503.03585.

[28] R. Soletskyi, M. Gabrie, and B. Loureiro. A theoretical perspective on mode collapse in variational inference, October 2024. arXiv:2410.13300 [stat].

[29] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

[30] F. Sáez-Maldonado, J. Maroñas, and D. Hernández-Lobato. Mode Collapse in Variational Deep Gaussian Processes. In *NeurIPS 2024 Workshop BDU*, October 2024.

[31] F. Vargas, W. Grathwohl, and A. Doucet. Denoising Diffusion Samplers. September 2022.

[32] F. Vargas, S. Padhy, D. Blessing, and N. Nüsken. Transport meets Variational Inference: Controlled Monte Carlo Diffusions. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *NeuRIPS*, 35:23244–23255, 2022.