Geometry-aware concentration of empirical risk minimizers for neural networks

Advisors	Franck Iutzeler (Prof. Univ. Toulouse) & Paul Escande (CR CNRS)
Host	Institut de Mathématiques de Toulouse (IMT), Statistics and Optimization team
Location	Campus of the Université de Toulouse, Toulouse
Duration	4 to 6 months starting in early 2027
Ph.D.	Funded Ph.D. position (October 2027) available for a successful candidate
Profile	Strong background in mathematics, with an interest in machine learning and optimization
Contact	franck.iutzeler@math.univ-toulouse.fr, paul.escande@math.univ-toulouse.fr
Keywords	Machine learning, statistical learning, empirical risk minimization, optimization, neural networks

Project

This internship is part of the SO-LAR project (Stat/Optim – Landscape Analysis of Risks), funded by the French National Research Agency (ANR). SO-LAR studies the theoretical foundations behind the success of neural networks, with a focus on the interplay between the geometry of optimization landscapes and statistical generalization. Deep networks have highly non-convex losses and are massively overparameterized, yet simple stochastic methods train them efficiently and the resulting models generalize well. Explaining this paradox is the central goal of the project.

Scientific context

Training a machine learning model usually amounts to minimizing the empirical risk, that is, the average error of the model on a finite training set. Since the data distribution is unknown in most situations of interest, the empirical risk serves as a proxy for the population (or true) risk, the average error over the whole distribution. A central question in statistical learning is then to understand what guarantees obtained for the empirical problem tell us about the population one, that is, to relate

$$\min_{\theta\in\Theta} \widehat{\mathscr{F}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(f_\theta(X_i), Y_i\big) \quad \text{and} \quad \min_{\theta\in\Theta} \mathscr{F}(\theta) = \mathbb{E}_{(X,Y)\sim\mu}\big[\ell\big(f_\theta(X), Y\big)\big],$$

where f_θ is the model (e.g. a neural network) with parameters θ optimized over a set $\Theta \subseteq \mathbb{R}^p$, ℓ is the loss function, and $(X_1, Y_1), \dots, (X_n, Y_n)$ are training samples drawn from the distribution μ. These questions date back to the 1970s [12] and have led to a rich and mature theory. Most guarantees in the literature bound the excess risk or the generalization error [5, 13]. More recent works instead establish concentration guarantees on the distance between the sets of minimizers themselves [6, 7, 10]. They suggest that the concentration of empirical minimizers is closely tied to the local geometry of the empirical risk around its minimizers.

Topic

Most of the concentration results available in the literature, and in particular the most recent ones [6, 7], rely on generic local assumptions on the risks, such as a Łojasiewicz inequality around the minimizers, and do not exploit the specific structure of neural networks. However, for neural networks, the parametrization $\theta \mapsto f_\theta$ has a rich geometry. The function f_θ is invariant under symmetries of the parameters, such as neuron permutations and positive rescalings, which can be factored out through the path-lifting of the network [8, 11]. As a result, the minimizers of the risks are not isolated but belong to *invariant valleys* $\{\theta' \mid f_{\theta'} = f_\theta\}$, whose local dimension is governed by the rank of the Jacobian of $\theta \mapsto f_\theta$ [3, 4]. The goal of this internship is to incorporate these geometric insights into the analysis of the concentration of minimizers, for instance to obtain rates that depend on the local dimension of the model rather than on the number of parameters.

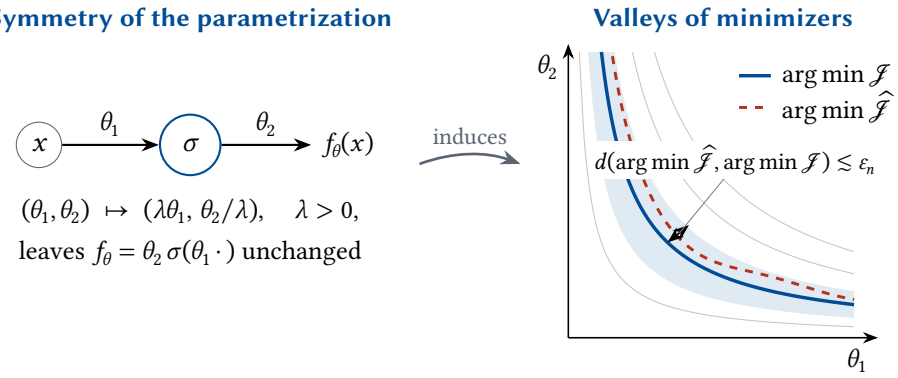


Figure 1: Left: a one-neuron network with ReLU activation $\sigma(t) = \max(t, 0)$; since $\sigma(\lambda t) = \lambda\sigma(t)$ for $\lambda > 0$, rescaling the weights leaves the computed function unchanged. Right: as a consequence, the minimizers of the population risk $\mathscr{F}$ are not isolated but form a valley (solid), and those of the empirical risk $\widehat{\mathscr{F}}$ (dashed) lie in a neighborhood of it. The width ε_n of this neighborhood, which depends on the number n of samples, is the quantity to be bounded: the internship aims to control it using the geometry of the network.

A second question concerns optimality. Current upper bounds on the concentration of minimizers do not reflect the good behavior of neural networks observed in practice [14], and no lower bounds matching this empirical behavior are known for these models: the available ones are worst-case and scale with the number of parameters [2], while uniform convergence arguments may even be provably vacuous in the overparameterized regime [9]. It is therefore hard to tell which existing results are sharp and which are an artifact of the proof techniques. Depending on the progress and the interests of the candidate, the internship will also investigate lower bounds in simple settings, such as linear or one-hidden-layer ReLU networks, and compare them with numerical experiments.

References

[1] W. Azizian, F. Iutzeler, and J. Malick. "Exact generalization guarantees for (regularized) Wasserstein distributionally robust models". In: *Advances in Neural Information Processing Systems* (2023).

[2] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks". In: *Journal of Machine Learning Research* 20.63 (2019), pp. 1–17.

[3] J. Bona-Pellissier, F. Malgouyres, and F. Bachoc. "Geometry-induced regularization in deep ReLU neural networks". In: *arXiv preprint arXiv:2402.08269* (2024).

[4] J. Bona-Pellissier, F. Malgouyres, and F. Bachoc. "Local identifiability of deep ReLU neural networks: the theory". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27549–27562.

[5] O. Bousquet and A. Elisseeff. "Stability and generalization". In: *Journal of Machine Learning Research* (2002).

[6] P. Escande. "On the concentration of the minimizers of empirical risks". In: *Journal of Machine Learning Research* (2024).

[7] P. Escande, F. Iutzeler, and I. Zekhnini. "Finite-sample concentration of the minimizers of nonconvex empirical risks". In preparation. 2026.

[8] A. Gonon, N. Brisebarre, E. Riccietti, and R. Gribonval. "A path-norm toolkit for modern networks: consequences, promises and challenges". In: *International Conference on Learning Representations*. 2024.

[9] V. Nagarajan and J. Z. Kolter. "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[10] C. Schötz. "Convergence rates for the generalized Fréchet mean via the quadruple inequality". In: *Electronic Journal of Statistics* (2019).

[11] P. Stock and R. Gribonval. "An embedding of ReLU networks and an analysis of their identifiability". In: *Constructive Approximation* (2022).

[12] V. Vapnik and A. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Theory of Probability and its Applications* (1971).

[13] M. Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge Univ. Press, 2019.

[14] C. Zhang et al. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017.