
Renormalization in the neural network-quantum field theory correspondence

Harold Erbin

Université Paris-Saclay, CEA-LIST
F-91120 Palaiseau, France
Center for Theoretical Physics, Massachusetts Institute of Technology
Cambridge, MA 02139, USA
NSF AI Institute for Artificial Intelligence and Fundamental Interactions
erbin@mit.edu

Vincent Lahoche

Université Paris-Saclay, CEA-LIST
F-91120 Palaiseau, France
vincent.lahoche@cea.fr

Dine Ousmane Samary

Université Paris-Saclay, CEA-LIST
F-91120 Palaiseau, France
Faculté des Sciences et Techniques (ICMPA-UNESCO Chair)
Université d'Abomey-Calavi, 072 BP 50, Bénin
dine.ousmanesamary@cipma.uac.bj

Abstract

A statistical ensemble of neural networks can be described in terms of a quantum field theory (NN-QFT correspondence). The infinite-width limit is mapped to a free field theory, while finite N corrections are mapped to interactions. After reviewing the correspondence, we will describe how to implement renormalization in this context and discuss preliminary numerical results for translation-invariant kernels. A major outcome is that changing the standard deviation of the neural network weight distribution corresponds to a renormalization flow in the space of networks.

1 Introduction

While neural networks (NN) perform extremely well on several tasks, they generally behave as black boxes which are hard to interpret [1, 2]. This is a problem for applications where safety can be put in jeopardy [3], but also if concrete explanations are needed, as in sciences [4–6]. Training is another concern because it is computationally expensive and has possible convergence issues. Indeed, the loss function is typically non-convex such that it can be hard to find the global minimum [7, 8]. There is also no systematic hyperparameter tuning procedure and one has to rely on random scans, possibly improved with Bayesian and bandit methods [9–12] which results in very high financial [13] and environmental costs [14–16]. Finally, the question of knowing which functions can be expressed by a given NN remains open [17, 18]: while universal approximation theorems guarantee existence [19–24], finding the appropriate architecture for a new task often boils down to trials and errors. Improving our theoretical understanding of NN is primordial for addressing these issues.

Physics provides a natural starting point for designing a theory of NN [25–27]. First, thanks to its effective descriptions, it is not necessary to know the fundamental theory. Second, efficient representations of statistical models have been developed (path integrals, Feynman diagrams, statistical mechanics. . .). Third, it allows characterizing the collective dynamics of degrees of freedom and organizing a phenomenon by scales. Applications of physics to machine learning include statistical physics [7, 28–34], renormalization [35–38], and QFT [39–45].

In this paper, we will review the neural network-quantum field theory (NN-QFT) correspondence developed in [41, 44] since it provides concrete and testable tools to improve our analytical understanding of neural network building and training. This correspondence states that, for a very general class of architectures, it is possible to associate a quantum field theory (QFT) with a statistical ensemble of NN. We focus on a fully connected NN with a single hidden layer and setup non-perturbative renormalization group equations (valid for any finite width). The main result is that varying the standard deviation of the weight distribution induces a renormalization group (RG) flow in the space of NN. Code is available at: <https://github.com/melsophos/nnqft>.

2 NN-QFT correspondence

Take a fully connected neural network $f_{\theta,N} : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^{d_{\text{out}}}$ with one hidden layer of width N :

$$f_{\theta,N}(x) = W_1 \left(g(W_0 x + b_0) \right) + b_1, \quad (1)$$

where g is the non-linear activation function, and the parameters $\theta = (W_0, b_0, W_1, b_1)$ (weights and biases) have Gaussian distributions: $W_0 \sim \mathcal{N}(0, \sigma_W^2/d_{\text{in}})$, $W_1 \sim \mathcal{N}(0, \sigma_W^2/N)$, $b_0, b_1 \sim \mathcal{N}(0, \sigma_b^2)$. Consider next a statistical ensemble of neural networks, such that a given neural network is sampled from the distribution in parameter space: $f_{\theta,N} \sim P[\theta]$. Then, there is a dual description in terms of another distribution in function space, which is induced by the parameter distribution plus the architecture: $f_{\theta,N} \sim p[f]$ [41]. Changing the parameter distribution by training corresponds to flowing in the function space.

In the large N limit (infinite width), the function distribution becomes a Gaussian process with kernel K (as a consequence of the central limit theorem) [46]: $f \sim \mathcal{N}(0, K)$. This statement generalizes to most architecture and training [47]. We denote as $S_0[f]$ the log-probability and as $G_0^{(n)}(x_1, \dots, x_n) := \mathbb{E}_0[f(x_1) \cdots f(x_n)]$ the *Gaussian expectation value* (GEV) for a product of n fields $f(x_i)$. In physics, this setting corresponds to a free QFT. At finite N , the distribution is not a Gaussian process, and we denote as

$$\Delta G^{(n)} := G^{(n)} - G_0^{(n)} \quad (2)$$

the difference between the *full expectation value* (FEV) $G^{(n)}(x_1, \dots, x_n) := \mathbb{E}[f(x_1) \cdots f(x_n)]$ and the GEV. The main message of the NN-QFT correspondence is that even at finite N , the log-probability $S[f]$ can be designed with non-Gaussian contributions to reproduce the FEVs with arbitrary precision up to the error bars of the numerical simulations. We denote as $S_{\text{int}}[f]$ the non-Gaussian contributions in $S[f]$. Furthermore, the 2 fields FEV $G^{(2)}(x, y) \equiv K(x, y)$ is N -independent and fixed by the NN itself (see Section 3).

This formulation is promising because correlation functions between outputs give a measure of learning; e.g., the 1-point function $\langle f(x) \rangle \equiv \mathbb{E}[f(x)]$ corresponds to the average prediction for input x (which is related to the idea of symmetry breaking in QFT [42]). Hence, having a QFT may allow performing (semi-)analytic predictions in advance of the outcome of the learning process.

Kernels in data-space are typically bi-local [41] such that one can expect non-local interactions. Moreover, it is not clear what are the symmetries of the inputs and outputs (in the QFT sense) for general data. With these observations, we follow an approach which can be called NN phenomenology: 1) make assumptions dictated by numerical evidence, 2) write a QFT model to match observations, 3) use the model to check theoretical facts.

3 Constructing the QFT

The expectation values $G^{(n)}$ can be computed analytically using QFT tools (“theory”) or computed from a statistical ensemble of neural networks (“measurements”). Hence, we can make an ansatz

for $S_{\text{int}}[f]$ and match the parameters by computing enough correlation functions. The choice of this ansatz especially regarding symmetries and the way the fields are coupled depend on the Gaussian kernel K . In this paper, we set $d_{\text{in}} = d_{\text{out}} = 1$ and focus on a translation-invariant activation function:

$$g(W_0x + b_0) = \frac{\exp(W_0x + b_0)}{\sqrt{\exp\left[2\left(\sigma_b^2 + \frac{\sigma_W^2}{d_{\text{in}}}x^2\right)\right]}} \quad (3)$$

such that the Gaussian kernel is [41]:

$$K(x, y) := \sigma_b^2 + K_W(x, y), \quad K_W(x, y) = \sigma_W^2 e^{-\frac{\sigma_W^2}{2d_{\text{in}}}|x-y|^2}. \quad (4)$$

In order to compute the ‘‘experimental’’ Green functions for a given N , we create n_{bags} distinct statistical ensembles of n_{nets} networks each [41, 44], and compute $\bar{G}_{\text{exp}}^{(n)}$ as the average of the (empirical) FEV:

$$G_{\text{exp}}^{(n)} = \frac{1}{n_{\text{nets}}} \sum_{\alpha=1}^{n_{\text{nets}}} f_{\alpha}(x_1) \cdots f_{\alpha}(x_n), \quad (5)$$

computed in a given bag. We furthermore define

$$\Delta G_{\text{exp}}^{(n)} := \bar{G}_{\text{exp}}^{(n)} - G_0^{(n)}, \quad (6)$$

and the normalized deviation $m_n := \Delta G_{\text{exp}}^{(n)}/G_0^{(n)}$. For the numerical investigations, we consider the points $x^{(1)}, \dots, x^{(6)} \in \{-0.01, -0.006, -0.002, 0.002, 0.006, 0.01\}$ and evaluate the Green functions for all inequivalent combinations. Moreover, all numerical tests are performed with $\sigma_b = 1, \sigma_W = 1, n_{\text{bags}} = 20, n_{\text{nets}} = 30000$, and $N \in \{2, 3, 4, 5, 10, 20, 50, 100, 500, 1000\}$. Computations ran during one week on the internal cluster of one of our institute. Empirically, we find that $m_2 \approx 0$ (the second momentum is almost independent of N) and $m_{2n} = O\left(\frac{1}{N}\right)$ for $n > 1$, the last result meaning that the empirical $2n$ -cumulant of the distribution $G_{c,\text{exp}}^{(2n)}$ must be of order $1/N^{n-1}$. The histogram of values for m_2 and m_4 are given in Figure 1.

The translation invariance of the Gaussian kernel is reminiscent of standard QFTs, where S_{int} can be expanded in powers of f coupled at the same point, namely:

$$S_{\text{int}}[f] := \sum_{n=2}^{n_0} \frac{\bar{u}_n}{(2n)!} \int d^{d_{\text{in}}}x f(x)^{2n}, \quad (7)$$

for some $n_0 \in \mathbb{N}$. We can check the validity of this ansatz experimentally. Indeed, u_4 is nothing but the magnitude of the lowest order deviation from the GEV in the perturbation theory. At higher order, this deviation receives many contributions which can be formally resummed, and we denote as $u_4(x_1, x_2, x_3, x_4)$ the full (normalized) deviation concerning the GEV:

$$u_4(x_1, x_2, x_3, x_4) = -\frac{\Delta G_{\text{exp}}^{(4)}(x_1, x_2, x_3, x_4)}{\int d^{d_{\text{in}}}x K_W(x, x_1)K_W(x, x_2)K_W(x, x_3)K_W(x, x_4)}. \quad (8)$$

In the QFT literature, $u_4(x_1, x_2, x_3, x_4)$ is called an *effective coupling constant* because it includes all quantum fluctuations. Empirically, focusing on the truncation $n_0 = 3$, we find that u_4 is negative but almost constant and u_6 remains small but positive as required for stability. Results for different σ_W and N are given below in Figure 2.

4 Renormalization group

In the previous section, we considered an *effective field theory* able to reproduce FEV corresponding to a NN ensemble. The RG is a set of techniques allowing to understand the dependency of the effective theory on a typical observation scale. The machine precision provides an example of such an observation scale, and we could consider the dependency of the parameters defining the QFT regarding the machine precision. In this paper, we consider another kind of scaling, induced by the

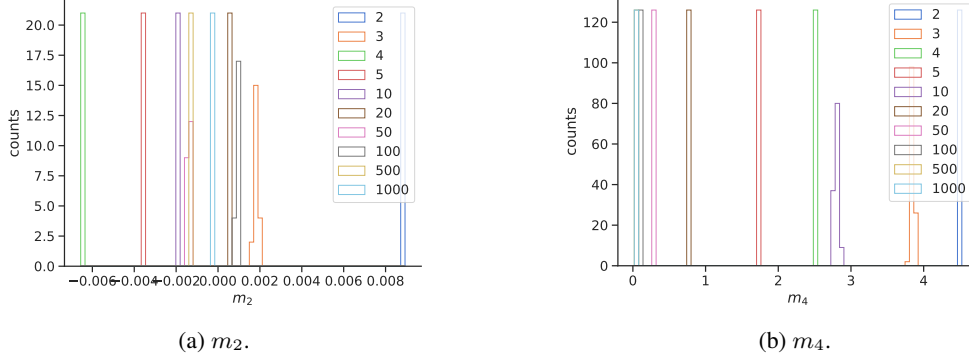


Figure 1: Normalized deviations with respect to the free theory. For m_2 , values centered around 0 and independent of N . For m_4 , the values decrease as N increases.

NN itself and called *active RG*. The motivation stems from the observation that the propagator (4) in momentum space looks like the usual Gaussian kernel in QFT at low momentum:

$$K_W(p) = (\sigma_W^2)^{1-\frac{d_{\text{in}}}{2}} \left(\frac{d_{\text{in}}}{2\pi}\right)^{\frac{d_{\text{in}}}{2}} \exp\left[-\frac{d_{\text{in}}}{2\sigma_W^2} p^2\right] \approx \frac{Z_0^{-1}}{\Lambda^2 + p^2 + O(p^2)}. \quad (9)$$

In the QFT terminology, Λ defines the *mass scale*, and the large momenta $p^2 \gg \Lambda^2$ are exponentially suppressed, blinding the physics beyond scale $p^2 \sim \Lambda^2$. Hence, in the active RG, Λ (or equivalently the standard deviation σ_W) define the typical momentum scale.

Having defined the notion of scale, we are aiming to construct a smooth interpolation between a large cut-off regime (called ultraviolet regime) and a small cut-off regime (called infrared regime). In the large cut-off regime, fluctuations are essentially frozen and the behavior of the network is mainly fixed by the saddle point of the log-probability $S[f]$. On the contrary, in the infrared regime, fluctuations are integrated out and look as a novel effective physics. The *Wetterich equation* describes how the effective description changes with the observation scale and leads to:

$$\Lambda \frac{d}{d\Lambda} \Gamma_\Lambda^{(2)}(p, -p) = -\frac{1}{2} \int \frac{d^{d_{\text{in}}} q}{(2\pi)^{d_{\text{in}}}} \Lambda \frac{dr_\Lambda}{d\Lambda}(q^2) \Gamma_\Lambda^{(4)}(p, -p, q, -q) G_\Lambda^2(q^2), \quad (10)$$

where $\Gamma_\Lambda^{(n)}$ is the n -th derivative of Γ_Λ with respect to f_{cl} , which is defined such that:

$$\Gamma_\Lambda[f_{\text{cl}}] := j \cdot f_{\text{cl}} - W_\Lambda[j] - \frac{1}{2} f_{\text{cl}} \cdot r_\Lambda \cdot f_{\text{cl}}, \quad f_{\text{cl}}(x) := \frac{\delta W_\Lambda}{\delta j}, \quad (11)$$

where $W_\Lambda[j] := \mathbb{E}[e^{-\frac{1}{2} f \cdot r_\Lambda \cdot f + j \cdot f}]$, the dot denoting the inner-product defined by integrating over the data space. The *regulator* r_Λ depends on p^2 and is designed such that $\Gamma_{\Lambda \rightarrow \infty} \rightarrow S$ (large cut-off regime) and $\Gamma_{\Lambda \rightarrow 0} \equiv \Gamma$ (vanishing cut-off regime), Γ being the full effective action, i.e. the Legendre transform of the characteristic function $\mathbb{E}[e^{j \cdot f}]$. The expectation value $K_W(p)$ being fixed by the NN, although both $\Gamma_\Lambda^{(2)}(p_1, p_2)$ and $r_\Lambda(p^2) \delta^{(d_{\text{in}})}(p_1 + p_2)$ can be arbitrary functions of the momentum p^2 , their sum is constrained to be $\Lambda^2 \exp\left(\frac{p_1^2}{\Lambda^2}\right) \delta^{(d_{\text{in}})}(p_1 + p_2)$ for any Λ . Because $\Lambda \frac{dr_\Lambda}{d\Lambda}(q^2)$ has to select only a short window of momenta in the vicinity of the scale Λ , the smooth function $\Gamma_\Lambda^{(4)}(p, -p, q, -q)$ can be expanded in power of q for Λ small enough. At zero order and using the Litim's regulator:

$$r_\Lambda(p^2) := \alpha (\Lambda^2 - p^2) \theta(\Lambda^2 - p^2), \quad (12)$$

we predict a purely scaling behavior with respect to the standard deviation σ_W for the zero momenta function $\Gamma_\Lambda^{(4)}(0, 0, 0, 0) =: u_4(\Lambda) \delta(0)$:

$$\sigma_W \frac{du_4}{d\sigma_W} = (4 - d_{\text{in}}) u_4 \implies \log u_4 = (4 - d_{\text{in}}) \log \sigma_W + \text{cst}. \quad (13)$$

This equation can be verified numerically (Figure 2). A similar equation can be derived for u_6 : $\log u_6 = (6 - 2d_{\text{in}}) \log \sigma_W + \text{cst}$.

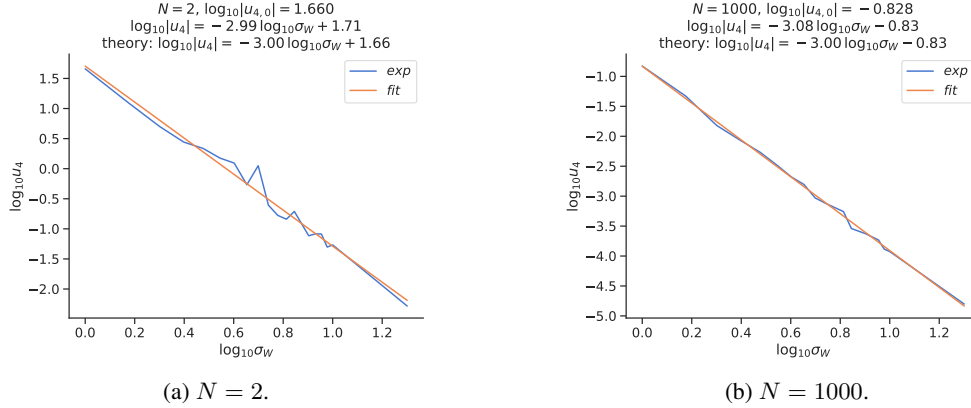


Figure 2: Dependence of u_4 in terms of σ_W , computed numerically and with the flow equation (13). Parameters: $\sigma_b = 0, \sigma_W \in \{1.0, 1.5, \dots, 10, 20\}, n_{\text{bags}} = 30, n_{\text{nets}} = 30000$.

5 Conclusion

In this paper, we have reviewed the NN-QFT correspondence and described several checks. In particular, we have derived exact renormalization equations and proved that the standard deviation σ_W is a RG flow parameter, before verifying this claim numerically.

Future directions include: increasing $d_{\text{in}}, d_{\text{out}}$ and N expansion, studying the large d_{in} limit (large data), increasing number of hidden layers and extending to non-translation invariant kernels (ReLU...) using the 2PI formalism [48], and finally studying the evolution of the QFT under training.

Broader impacts

The goal of our work is to provide analytical tools for improving the performances of neural networks and reducing the amount of training and hyperparameter tuning required. As explained in the introduction, these are important goals given the ongoing climate crisis and the need for fair access to AI techniques.

Acknowledgments and Disclosure of Funding

This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 891169. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, <http://iaifi.org/>).

References

- [1] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. October 2018.
- [2] Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021. doi: 10.1109/TETCI.2021.3100641.
- [3] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Technical report, European Commission, April 2019. URL <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf>.
- [4] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. doi: 10.1109/ACCESS.2020.2976199.
- [5] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. November 2020.

- [6] Maithra Raghu and Eric Schmidt. A survey of deep learning for scientific discovery. March 2020.
- [7] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. January 2015.
- [8] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. November 2018.
- [9] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281–305, February 2012. doi: 10.5555/2188385.2188395.
- [10] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’12, pages 2951–2959. Curran Associates Inc., December 2012. doi: 10.5555/2999325.2999464.
- [11] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18 (185):1–52, 2018. URL <http://jmlr.org/papers/v18/li16-558.html>.
- [12] Stefan Falkner, Aaron Klein, and Frank Hutter. Bobb: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*, pages 1437–1446. PMLR, July 2018. URL <http://proceedings.mlr.press/v80/falkner18a.html>.
- [13] Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training nlp models: A concise overview, April 2020.
- [14] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, June 2019.
- [15] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. In *arXiv:1910.09700 [Cs]*, November 2019.
- [16] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. January 2020.
- [17] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6231–6239. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width.pdf>.
- [18] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. June 2017.
- [19] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. doi: 10.1007/BF02551274.
- [20] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989. doi: 10.1016/0893-6080(89)90020-8.
- [21] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, January 1991. doi: 10.1016/0893-6080(91)90009-T.
- [22] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993. doi: 10.1016/S0893-6080(05)80131-5.
- [23] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, January 1999. doi: 10.1017/S0962492900002919.
- [24] Balázs Csanád Csáji. *Approximation with Artificial Neural Networks*. PhD thesis, Eötvös Loránd University, 2001.
- [25] Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017. doi: 10.1007/s10955-017-1836-5.
- [26] Lenka Zdeborová. Understanding deep learning is also a job for physicists. *Nature Physics*, pages 1–3, May 2020. doi: 10.1038/s41567-020-0929-2.

- [27] Elena Agliari, Adriano Barra, Peter Sollich, and Lenka Zdeborová. Machine learning and statistical physics: Preface. *Journal of Physics A: Mathematical and Theoretical*, 53(50):500401, November 2020. doi: 10.1088/1751-8121/abca75.
- [28] Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, January 1987. doi: 10.1016/0003-4916(87)90092-3.
- [29] E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, January 1988. doi: 10.1088/0305-4470/21/1/030.
- [30] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271–284, January 1988. doi: 10.1088/0305-4470/21/1/031.
- [31] Werner Krauth, Marc Mézard, and Jean-Pierre Nadal. Basins of attraction in a perceptron-like neural network. *Complex Syst.*, 2(4):387–408, August 1988. doi: 10.5555/56123.56124.
- [32] M. Mézard and Jean-P. Nadal. Learning in feedforward layered networks: The tiling algorithm. *Journal of Physics A: Mathematical and General*, 22(12):2191, 1989. doi: 10.1088/0305-4470/22/12/019.
- [33] Lorenza Saitta, Attilio Giordana, and Antoine Cornuéjols. *Phase Transitions in Machine Learning*. Cambridge University Press, June 2011.
- [34] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1):501–528, March 2020. doi: 10.1146/annurev-conmatphys-031119-050745.
- [35] Cédric Bény. Deep learning and the renormalization group. January 2013.
- [36] Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning. October 2014.
- [37] Cédric Bény. Inferring relevant features: From qft to pca. *International Journal of Quantum Information*, 16(08):1840012, December 2018. doi: 10.1142/S0219749918400129.
- [38] Ellen de Mello Koch, Robert de Mello Koch, and Ling Cheng. Is deep learning a renormalization group flow? *IEEE Access*, 8:106487–106505, 2020. doi: 10.1109/ACCESS.2020.3000901.
- [39] Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. A correspondence between random neural networks and statistical field theory. October 2017.
- [40] Moritz Helias and David Dahmen. Statistical field theory for neural networks. January 2019.
- [41] James Halverson, Anindita Maiti, and Keegan Stoner. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, March 2021. doi: 10.1088/2632-2153/abeca3.
- [42] Anindita Maiti, Keegan Stoner, and James Halverson. Symmetry-via-duality: Invariant neural network densities from parameter-space correlators. June 2021.
- [43] James Halverson. Building quantum field theories out of neurons. December 2021.
- [44] Harold Erbin, Vincent Lahoche, and Dine Ousmane Samary. Nonperturbative renormalization for the neural network-qft correspondence. *Machine Learning: Science and Technology*, 3(1):015027, March 2022. doi: 10.1088/2632-2153/ac4f69.
- [45] Kevin T. Grosvenor and Ro Jefferson. The edge of chaos: Quantum field theory and deep neural networks. January 2022.
- [46] Radford M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, 1996. doi: 10.1007/978-1-4612-0745-0.
- [47] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. May 2021.
- [48] Jean-Paul Blaizot, Jan M. Pawłowski, and Urko Reinosa. Functional renormalization group and 2π effective action formalism. February 2021.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
 - (b) Did you describe the limitations of your work? [Yes] See Section 5.
 - (c) Did you discuss any potential negative societal impacts of your work? [N/A]
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
 - (b) Did you include complete proofs of all theoretical results? [No] Full derivation is extremely lengthy and technical, but we gave an outline of the proof in Section 4.
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Because of computational cost.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [N/A]
 - (b) Did you mention the license of the assets? [N/A] The data we used was not licensed by the authors.
 - (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]