Renormalization in the neural network-quantum field theory correspondence

Harold Erbin, Vincent Lahoche, Dine Ousmane Samary

Renormalization in the neural network-quantum field theory correspondence

Harold Erbin*[1,2,3], Vincent Lahoche†[1], and Dine Ousmane Samary‡[1,4]

[1]Université Paris-Saclay, CEA-LIST, Gif-sur-Yvette, F-91191, France
[2]Center for Theoretical Physics, Massachusetts Institute of Technology
Cambridge, MA 02139, USA
[3]NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[4]Faculté des Sciences et Techniques (ICMPA-UNESCO Chair)
Université d'Abomey-Calavi, 072 BP 50, Bénin

December 23, 2022

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.

*erbin@mit.edu
†vincent.lahoche@cea.fr
‡dine.ousmanesamary@cea.fr

Contents

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 $f_{\theta,N} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ with one hidden layer of width N:

$$f_{\theta,N}(x) = W_1\Big(g(W_0 x + b_0)\Big) + b_1, \tag{2.1}$$

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:

$$W_0 \sim \mathcal{N}(0, \sigma_W^2/d_{\text{in}}), \quad W_1 \sim \mathcal{N}(0, \sigma_W^2/N), \quad b_0, b_1 \sim \mathcal{N}(0, \sigma_b^2)\,. \tag{2.2}$$

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]:

$$f \sim \mathcal{N}(0, K)\,. \tag{2.3}$$

This statement generalizes to most architecture and training [47]. We denote as $S_0[f]$ the (Gaussian) log-probability:

$$S_0[f] := \frac{1}{2}\int \mathrm{d}^{d_{\text{in}}}x\mathrm{d}^{d_{\text{in}}}x'\, f(x)\Xi(x,x')f(x'), \qquad \Xi := K^{-1}\,, \tag{2.4}$$

and as

$$G_0^{(n)}(x_1,\ldots,x_n) := \mathbb{E}_0[f(x_1)\cdots f(x_n)] \equiv \int \mathrm{d}f\, \mathrm{e}^{-S_0[f]}\, f(x_1)\cdots f(x_n) \tag{2.5}$$

the *Gaussian expectation value* (GEV) for a product of n fields $f(x_i)$. The measure $\mathrm{d}f$ is suitably normalized such that $\int \mathrm{d}f\, \exp(-S_0[f]) = 1$. In physics, this setting corresponds to a free QFT, K to the free propagator and $G_0^{(n)}$ to the free n-point correlation functions (also called Green functions). At finite N, the distribution is not a Gaussian process, and we denote as

$$\Delta G^{(n)} := G^{(n)} - G_0^{(n)} \tag{2.6}$$

the difference between the *full expectation value* (FEV)

$$G^{(n)}(x_1,\ldots,x_n) := \mathbb{E}[f(x_1)\cdots f(x_n)] \equiv \int \mathrm{d}f\, \mathrm{e}^{-S[f]}\, f(x_1)\cdots f(x_n) \tag{2.7}$$

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 numerical uncertainties in the simulations. We denote as $S_{\text{int}}[f]$ the non-Gaussian contributions in $S[f]$:

$$S[f] = S_0'[f] + S_{\text{int}}[f], \tag{2.8}$$

where $S_0'[f] \neq S_0[f]$ is some new Gaussian action. Indeed, the 2-point FEV $G^{(2)}(x,y)$ is N-independent and fixed by the NN, such that the Gaussian part must be different and such that:

$$G^{(2)}(x,y) = G_0^{(2)}(x,y) \equiv K(x,y)\,. \tag{2.9}$$

A complete dictionary between NN and QFT is given by Table 1. This formulation is promising because correlation functions between outputs give a measure of learning; e.g., the 1-point function $\mathbb{E}[f(x)]$ 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.

	QFT	NN / GP
x	spacetime points	data-space inputs
p	momentum space	dual data-space
f	field	neural network
$K(x,y)$	propagator	Gaussian kernel
S	action	negative log-probability
S_0	free action	Gaussian log probability
S_{int}	interactions	non-Gaussian corrections
$\langle\cdot\rangle \equiv \mathbb{E}[\cdot]$	expectation value, Green function	correlation function

Table 1: NN-QFT dictionary.

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:

$$g(W_0 x + b_0) = \frac{\exp(W_0 x + b_0)}{\sqrt{\exp\left[2\left(\sigma_b^2 + \frac{\sigma_W^2}{d_{\text{in}}}\, x^2\right)\right]}} \tag{3.1}$$

such that the Gaussian kernel is [41]:

$$K(x,y) := \sigma_b^2 + K_W(x,y), \quad K_W(x,y) = \sigma_W^2\, \mathrm{e}^{-\frac{\sigma_W^2}{2d_{\text{in}}}\,|x-y|^2}. \tag{3.2}$$

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 $\bar{G}^{(n)}_{\text{exp}}$ as the average of the (empirical) FEV:

$$G^{(n)}_{\text{exp}} = \frac{1}{n_{\text{nets}}} \sum_{\alpha=1}^{n_{\text{nets}}} f_\alpha(x_1)\cdots f_\alpha(x_n)\,, \tag{3.3}$$

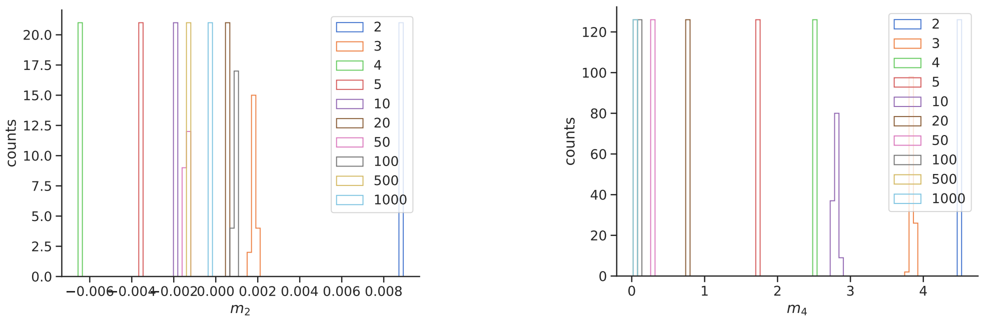
computed in a given bag. We furthermore define

$$\Delta G^{(n)}_{\text{exp}} := \bar{G}^{(n)}_{\text{exp}} - G^{(n)}_0\,, \tag{3.4}$$

and the normalized deviation $m_n := \Delta G^{(n)}_{\text{exp}} / G^{(n)}_0$. 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 $G^{(2n)}_{c,\text{exp}}$ 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:

$$S_{\text{int}}[f] := \sum_{n=2}^{n_0} \frac{\bar{u}_n}{(2n)!} \int \mathrm{d}^{d_{\text{in}}}x\, f(x)^{2n}, \tag{3.5}$$



(a) m_2.

(b) m_4.

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.

for some $n_0 \in \mathbb{N}$. We can check the validity of this ansatz experimentally. Indeed, in that expression, $\bar{u}_n$ is nothing but the magnitude of the lowest order deviation from the GEV, and is called *bare coupling*. It is different from the effective coupling u_n which is measured by the simulations and which includes quantum corrections; in perturbation theory (assuming $\bar{u}_n$ small enough), and $u_n = \bar{u}_n + \mathcal{O}(\bar{u}_n^2)$ (schematically). At higher order, this deviation receives many contributions which can be formally resumed. For the lowest order, the full (normalized) deviation from the GEV $u_4(x_1, x_2, x_3, x_4)$ reads:

$$u_4(x_1, x_2, x_3, x_4) = -\frac{\Delta G^{(4)}_{\text{exp}}(x_1, x_2, x_3, x_4)}{\int \mathrm{d}^{d_{\text{in}}} x \, K_W(x, x_1) K_W(x, x_2) K_W(x, x_3) K_W(x, x_4)}. \tag{3.6}$$

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 NN itself and called *active RG*. The motivation stems from the observation that the propagator (3.2) in momentum space looks like the usual Gaussian kernel in QFT at low momentum:

$$K_W(p) = (\sigma_W^2)^{1-\frac{d_{\text{in}}}{2}} \left(\frac{d_{\text{in}}}{2\pi}\right)^{\frac{d_{\text{in}}}{2}} \exp\left[-\frac{d_{\text{in}}}{2\sigma_W^2}\, p^2\right] \approx \frac{Z_0^{-1}}{\Lambda^2 + p^2 + O(p^2)}. \tag{4.1}$$

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:

$$\Lambda \frac{\mathrm{d}}{\mathrm{d}\Lambda} \Gamma_\Lambda^{(2)}(p,-p) = -\frac{1}{2} \int \frac{\mathrm{d}^{d_\mathrm{in}} q}{(2\pi)^{d_\mathrm{in}}} \, \Lambda \frac{\mathrm{d} r_\Lambda}{\mathrm{d}\Lambda}(q^2) \, \Gamma_\Lambda^{(4)}(p,-p,q,-q) \, G_\Lambda^2(q^2), \tag{4.2}$$

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

$$\Gamma_\Lambda[f_\mathrm{cl}] := j \cdot f_\mathrm{cl} - W_\Lambda[j] - \frac{1}{2} \, f_\mathrm{cl} \cdot r_\Lambda \cdot f_\mathrm{cl}, \qquad f_\mathrm{cl}(x) := \frac{\delta W_\Lambda}{\delta j}, \tag{4.3}$$

where

$$W_\Lambda[j] := \mathbb{E}[e^{-\frac{1}{2} f \cdot r_\Lambda \cdot f + j \cdot f}] \,, \tag{4.4}$$

the dot denoting the inner-product defined by integrating over the data space. Once again, let us note that in the power field expansion of Γ_Λ the weights are effective rather than bare couplings. The *regulator* r_Λ depends on p^2 and is designed such that $\Gamma_{\Lambda\to\infty} \to S$ (large cut-off regime) and $\Gamma_{\Lambda\to 0} \equiv \Gamma$ (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 $r_\Lambda(p^2)\, \delta^{(d_\mathrm{in})}(p_1+p_2)$ can be arbitrary functions of the momentum p^2, their sum is constrained to be $\Lambda^2 \exp\left(\frac{p_1^2}{\Lambda^2}\right) \delta^{(d_\mathrm{in})}(p_1+p_2)$ for any Λ. Because $\Lambda \frac{\mathrm{d} r_\Lambda}{\mathrm{d}\Lambda}(q^2)$ 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:

$$r_\Lambda(p^2) := \alpha \, (\Lambda^2 - p^2) \theta(\Lambda^2 - p^2) \,, \tag{4.5}$$

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

$$\sigma_W \, \frac{\mathrm{d}u_4}{\mathrm{d}\sigma_W} = (4 - d_\mathrm{in}) u_4 \quad \Longrightarrow \quad \log u_4 = (4 - d_\mathrm{in}) \log \sigma_W + \mathrm{cst}. \tag{4.6}$$

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

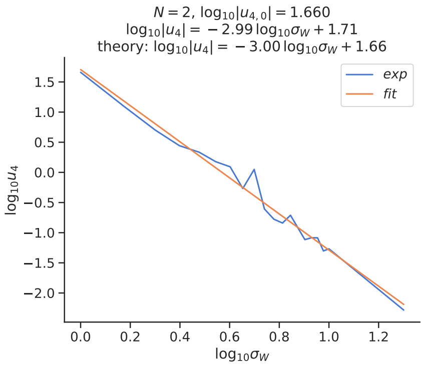
5 Conclusion

In this paper, we have reviewed the NN-QFT correspondence and described several checks. Our main result is about the derivation of exact renormalization equations where the standard deviation σ_W looks like a RG flow parameter, and the nice agreement between theoretical predictions and numerical experiments.

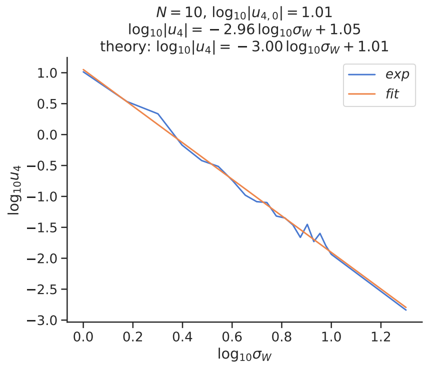
Future directions include: increasing d_in, d_out and N expansion, studying the large d_in limit (large data), increasing number of hidden layers and extending to non-translation invariant kernels (ReLU. . .) using the 2PI formalism [48], and finally studying the evolution of the QFT under training.

Acknowledgments

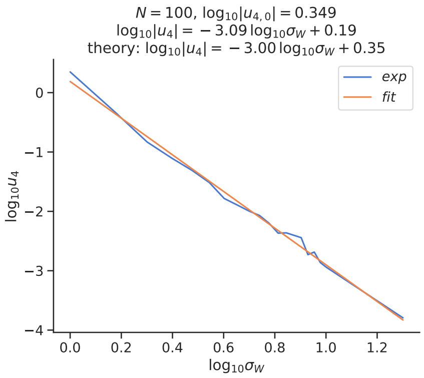
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 891169. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence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



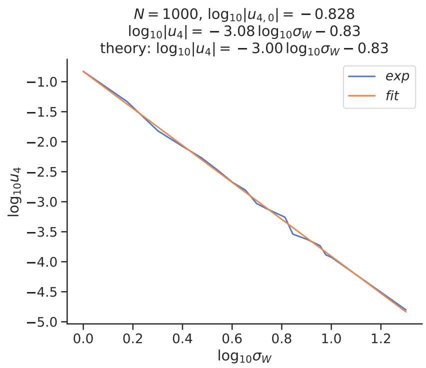
(a) $N = 2$.



(b) $N = 10$.



(c) $N = 100$.



(d) $N = 1000$.

Figure 2: Dependence of u_4 in terms of σ_W, computed numerically and with the flow equation (4.6). Parameters: $\sigma_b = 0, \sigma_W \in \{1.0, 1.5, \ldots, 10, 20\}, n_{\text{bags}} = 30, n_{\text{nets}} = 30000$.

References

[1] D. S. Weld and G. Bansal. “The Challenge of Crafting Intelligible Intelligence” (Oct. 2018). arXiv: 1803.04263.

[2] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. “A Survey on Neural Network Interpretability.” *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (Oct. 2021), pp. 726–742. DOI: 10.1109/TETCI.2021.3100641. arXiv: 2012.14261.

[3] High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI.* Tech. rep. European Commission, Apr. 2019. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[4] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. “Explainable Machine Learning for Scientific Insights and Discoveries.” *IEEE Access* 8 (2020), pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

[5] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho. “Discovering Symbolic Models from Deep Learning with Inductive Biases” (Nov. 2020). arXiv: 2006.11287.

[6] M. Raghu and E. Schmidt. “A Survey of Deep Learning for Scientific Discovery” (Mar. 2020). arXiv: 2003.11755.

[7] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. “The Loss Surfaces of Multilayer Networks” (Jan. 2015). arXiv: 1412.0233.

[8] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. “Visualizing the Loss Landscape of Neural Nets” (Nov. 2018). arXiv: 1712.09913.

[9] J. Bergstra and Y. Bengio. “Random Search for Hyper-Parameter Optimization.” *The Journal of Machine Learning Research* 13 (Feb. 2012), pp. 281–305. DOI: 10.5555/2188385.2188395.

[10] J. Snoek, H. Larochelle, and R. 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. Curran Associates Inc., Dec. 2012, pp. 2951–2959. DOI: 10.5555/2999325.2999464. arXiv: 1206.2944.

[11] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization.” *Journal of Machine Learning Research* 18.185 (2018), pp. 1–52. URL: http://jmlr.org/papers/v18/16-558.html.

[12] S. Falkner, A. Klein, and F. Hutter. “BOHB: Robust and Efficient Hyperparameter Optimization at Scale.” In: *International Conference on Machine Learning*. PMLR, July 2018, pp. 1437–1446. URL: http://proceedings.mlr.press/v80/falkner18a.html.

[13] O. Sharir, B. Peleg, and Y. Shoham. *The Cost of Training NLP Models: A Concise Overview*. Apr. 2020. DOI: 10.48550/arXiv.2004.08900. arXiv: 2004.08900.

[14] E. Strubell, A. Ganesh, and A. McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. June 2019. DOI: 10.48550/arXiv.1906.02243. arXiv: 1906.02243.

[15] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. “Quantifying the Carbon Emissions of Machine Learning.” In: *arXiv:1910.09700 [Cs]*. Nov. 2019. arXiv: 1910.09700.

[16] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. “Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning” (Jan. 2020). arXiv: 2002.05651.

[17] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. "The Expressive Power of Neural Networks: A View from the Width." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 6231–6239. arXiv: 1709.02540. URL: http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width.pdf.

[18] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. "On the Expressive Power of Deep Neural Networks" (June 2017). arXiv: 1606.05336.

[19] G. Cybenko. "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/BF02551274.

[20] K. Hornik, M. Stinchcombe, and H. White. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[21] K. Hornik. "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks* 4.2 (Jan. 1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-T.

[22] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function." *Neural Networks* 6.6 (Jan. 1993), pp. 861–867. DOI: 10.1016/S0893-6080(05)80131-5.

[23] A. Pinkus. "Approximation Theory of the MLP Model in Neural Networks." *Acta Numerica* 8 (Jan. 1999), pp. 143–195. DOI: 10.1017/S0962492900002919.

[24] B. C. Csáji. "Approximation with Artificial Neural Networks." PhD thesis. Eötvös Loránd University, 2001.

[25] H. W. Lin, M. Tegmark, and D. Rolnick. "Why Does Deep and Cheap Learning Work so Well?" *Journal of Statistical Physics* 168.6 (2017), pp. 1223–1247. DOI: 10.1007/s10955-017-1836-5. arXiv: 1608.08225.

[26] L. Zdeborová. "Understanding Deep Learning Is Also a Job for Physicists." *Nature Physics* (May 2020), pp. 1–3. DOI: 10.1038/s41567-020-0929-2.

[27] E. Agliari, A. Barra, P. Sollich, and L. Zdeborová. "Machine Learning and Statistical Physics: Preface." *Journal of Physics A: Mathematical and Theoretical* 53.50 (Nov. 2020), p. 500401. DOI: 10.1088/1751-8121/abca75.

[28] D. J. Amit, H. Gutfreund, and H Sompolinsky. "Statistical Mechanics of Neural Networks near Saturation." *Annals of Physics* 173.1 (Jan. 1987), pp. 30–67. 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 (Jan. 1988), pp. 257–270. 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 (Jan. 1988), pp. 271–284. DOI: 10.1088/0305-4470/21/1/031.

[31] W. Krauth, M. Mézard, and J.-P. Nadal. "Basins of Attraction in a Perceptron-like Neural Network." *Complex Syst.* 2.4 (Aug. 1988), pp. 387–408. DOI: 10.5555/56123.56124.

[32] M. Mézard and J.-P. Nadal. "Learning in Feedforward Layered Networks: The Tiling Algorithm." *Journal of Physics A: Mathematical and General* 22.12 (1989), p. 2191. DOI: 10.1088/0305-4470/22/12/019.

[33] L. Saitta, A. Giordana, and A. Cornuéjols. *Phase Transitions in Machine Learning*. Cambridge University Press, June 2011.

[34] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. "Statistical Mechanics of Deep Learning." *Annual Review of Condensed Matter Physics* 11.1 (Mar. 2020), pp. 501–528. DOI: 10.1146/annurev-conmatphys-031119-050745.

[35] C. Bény. "Deep Learning and the Renormalization Group" (Jan. 2013). arXiv: 1301.3124.

[36] P. Mehta and D. J. Schwab. "An Exact Mapping between the Variational Renormalization Group and Deep Learning" (Oct. 2014). arXiv: 1410.3831.

[37] C. Bény. "Inferring Relevant Features: From QFT to PCA." *International Journal of Quantum Information* 16.08 (Dec. 2018), p. 1840012. DOI: 10.1142/S0219749918400129. arXiv: 1802.05756.

[38] E. de Mello Koch, R. de Mello Koch, and L. Cheng. "Is Deep Learning a Renormalization Group Flow?" *IEEE Access* 8 (2020), pp. 106487–106505. DOI: 10.1109/ACCESS.2020.3000901. arXiv: 1906.05212.

[39] S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "A Correspondence Between Random Neural Networks and Statistical Field Theory" (Oct. 2017). arXiv: 1710.06570.

[40] M. Helias and D. Dahmen. "Statistical Field Theory for Neural Networks" (Jan. 2019). arXiv: 1901.10416.

[41] J. Halverson, A. Maiti, and K. Stoner. "Neural Networks and Quantum Field Theory." *Machine Learning: Science and Technology* (Mar. 2021). DOI: 10.1088/2632-2153/abeca3. arXiv: 2008.08601.

[42] A. Maiti, K. Stoner, and J. Halverson. "Symmetry-via-Duality: Invariant Neural Network Densities from Parameter-Space Correlators" (June 2021). arXiv: 2106.00694.

[43] J. Halverson. "Building Quantum Field Theories Out of Neurons" (Dec. 2021). arXiv: 2112.04527.

[44] H. Erbin, V. Lahoche, and D. O. Samary. "Nonperturbative Renormalization for the Neural Network-QFT Correspondence." *Machine Learning: Science and Technology* 3.1 (Mar. 2022), p. 015027. DOI: 10.1088/2632-2153/ac4f69. arXiv: 2108.01403.

[45] K. T. Grosvenor and R. Jefferson. "The Edge of Chaos: Quantum Field Theory and Deep Neural Networks" (Jan. 2022). arXiv: 2109.13247.

[46] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, 1996. DOI: 10.1007/978-1-4612-0745-0.

[47] G. Yang. "Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture Are Gaussian Processes" (May 2021). arXiv: 1910.12478.

[48] J.-P. Blaizot, J. M. Pawlowski, and U. Reinosa. "Functional Renormalization Group and 2PI Effective Action Formalism" (Feb. 2021). arXiv: 2102.13628.