On Bridging the Gap between Mean Field and Finite Width Deep Random Multilayer Perceptron with Batch Normalization

Amir Joudaki, Hadi Daneshmand and Francis Bach

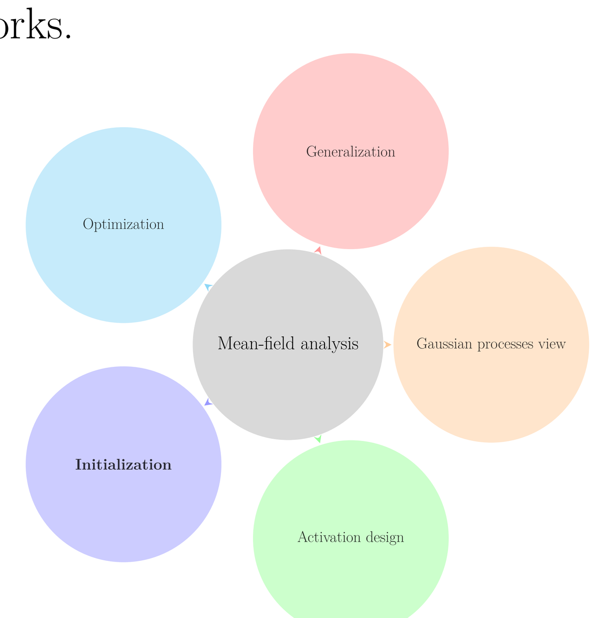
ETH Zurich, MIT-FODSI-BU, INRIA-ENS-PSL Paris

Abstract

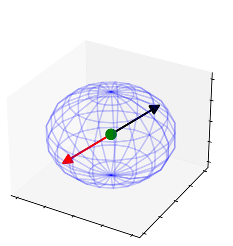
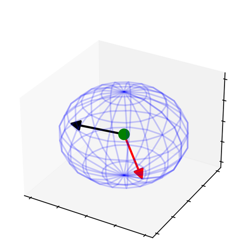
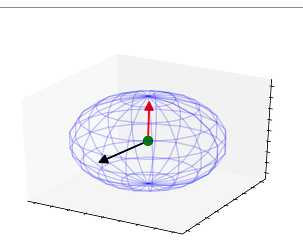
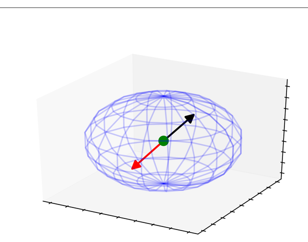
Mean-field theory is widely used in theoretical studies of neural networks. In this paper, we analyze the role of depth in the concentration of mean-field predictions for Gram matrices of hidden representations in deep multilayer perceptron (MLP) with batch normalization (BN) at initialization. It is postulated that the mean-field predictions suffer from layer-wise errors that amplify with depth. We demonstrate that BN avoids this error amplification with depth. When the chain of hidden representations is rapidly mixing, we establish a concentration bound for a mean-field model of Gram matrices. To our knowledge, this is the first concentration bound that does not become vacuous with depth for standard MLPs with a finite width.

Mean field analysis for neural networks

In the context of neural networks, mean field analysis refers to studying infinitely-wide neural networks. Mean field analysis has provided insights on pretraining, training and post training of neural networks.



Initialization and the rank collapse issue

Networks	Inputs	Outputs
BN		
Vanilla		

Background

Representations in random neural networks. Let $h_\ell \in \mathbb{R}^{d\times n}$ denote the hidden representation (layer:ℓ, batchsize:n, width:d). The sequence $\{h_\ell\}$ is a Markov chain:

$$h_{\ell+1} := W_\ell \sigma \circ \phi(h_\ell), \qquad W_\ell \sim \mathcal{N}(0, 1/d)^{d\times d},$$

where ϕ is the batch normalization [Ioffe and Szegedy, 2015]:

$$\phi(x) = \frac{x - \text{mean}(x)}{\sqrt{\text{Var}(x)}}, \qquad \forall r : \text{row}_r(\phi(h)) = \phi(\text{row}_r(h)).$$

Gram matrices. The Gram matrix G_ℓ is defined as the matrix of inner products of hidden representations at layer ℓ.

$$G_\ell := \frac{1}{d}(\sigma \circ \phi(h_\ell))(\sigma \circ \phi(h_\ell))^\top.$$

The dynamics of G_ℓ is an important topic in theoretical and practical studies of deep neural networks [Yang et al., 2019a, Pennington et al., 2018, Pennington and Worah, 2017].

Mean field analysis of Gram matrices. By letting $d \to \infty$, [Yang et al., 2019a] approximates the dynamics of G_ℓ as

$$\overline{G}_{\ell+1} = \mathbb{E}_{h\sim\mathcal{N}(0,\overline{G}_\ell)}\left[\sigma\left(\frac{\sqrt{n}Mh}{\|Mh\|}\right)^{\otimes 2}\right],$$

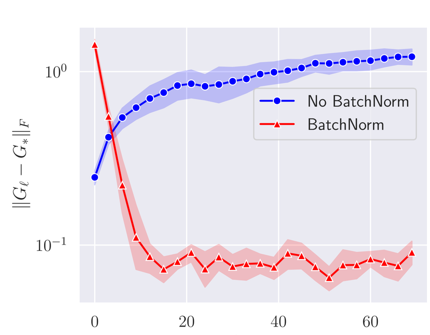
where $\overline{G}_0 = G_0$ and $M = I_n - \frac{1}{n}1_n^{\otimes 2}$. Fixed points of the above equation has the general form

$$G_* = b^*\left((1-c^*)I_n + c^* 1\!\!1_{n\times n}\right).$$

Constants b^* and c^* depend on the activation function σ.

The challenge of mean field approximation

$G_0 = \bar{G}_0 : \|\bar{G}_1 - G_1\|_F = O(1/\sqrt{d})$.
Thus, mean-field estimates suffers from $O(1/\sqrt{d})$ error per layer [Li et al., 2022].



Blessings of depth	Curses of depth
Fixed-point analysis	The accumulation of estimation error

Beyond a mean field analysis

Theorem 1 ([Daneshmand et al., 2021]) Under a spectral assumption and for neural networks with **linear activations** (i.e. $\sigma(a) := a$), it is possible to establish $O(1/\sqrt{d})$ concentration bound for mean field predictions. More precisely,

$$\|G_\ell - G_*\|_F = O\left(\frac{n}{\sqrt{d}}\right)$$

holds with a high probability for a sufficiently large ℓ.

Main result

We characterize sufficient conditions to estimate G_ℓ by G_*.

Assumption(Geometric ergodicity). *Let μ_ℓ denotes the distribution of h_ℓ. We assume the chain of hidden representations admits a unique invariant distribution. Furthermore, there is constant α ($\alpha > 0$) such that*

$$\|\mu_\ell - \mu_*\|_{tv} \leq (1-\alpha)^\ell \|\mu_0 - \mu_*\|_{tv}$$

holds almost surely for all h_0

The geometric ergodic property is established for various Markov chains, such as the Gibbs sampler, and state-space models [Eberle, 2009].

Theorem 2. *Assume the Markov chain of representations $\{h_\ell\}$ is geometric ergodic with $\alpha > 0$, and has non-degenerate fixed-point G_*. If the activation σ is uniformly bounded $|\sigma(x)| = O(|x|)$, then*

$$\|G_* - G_\ell\|_F = O\left(\kappa(G_*)(1-\alpha)^{\frac{\ell}{2}} + \frac{n}{\sqrt{d}}\alpha^{-\frac{1}{2}}\ln^{\frac{1}{2}}\left(\frac{d}{n}\right)\right)$$

holds with high probability

Remarkably, the last theorem considerably improves upon the concentration bounds for neural networks without batch normalization that become vacuous as the depth increases [Hanin and Nica, 2019, Hanin, 2022].

Theorem 2 recovers Theorem 1 as a special case when the activation is a linear function. Theorem 2 holds for a broad family of activation functions, including hyperbolic tangent and ReLU.

Validations

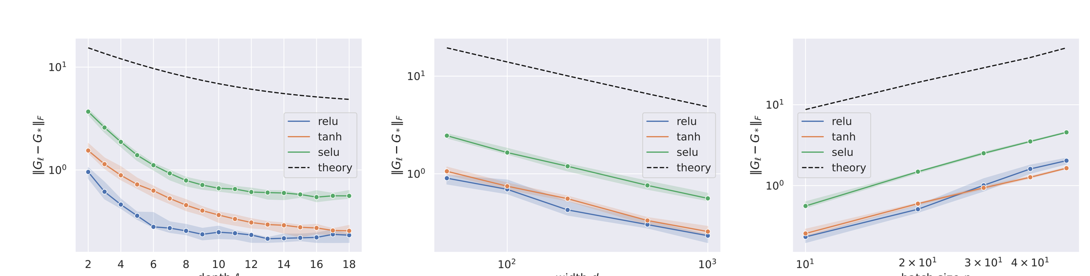


Figure 1:The dashed line shows the theoretical upper bound of Theorem 2. (a) $d = 1000$, $n = 10$ (b) $\ell = 20$, $n = 10$ (c) $d = 1000$, $\ell = 20$.

Applications

Proposition. *In the same setting as Theorem 2, for a sufficiently deep layer ℓ, $n - O(1)$ eigenvalues of G_ℓ are within $O(\sqrt{n/d})$ range of $b^*(1-c^*)$ with high probability in d.*

Combining the above result by the mean field analysis of [Yang et al., 2019b] concludes the Gram matrices are well-conditioned for deep neural networks with batch normalization. Empirical studies suggest that the conditioning of Gram matrices, G_ℓ, has a substantial impact on the training of deep neural networks [Pennington et al., 2018].

Observation

We observe empirically that the singular values of h_ℓ, which are square root of eigenvalues of G_ℓ, accurately follow the Marchenko-Pastur distribution with $\gamma = n/d$. Indeed, the spectrum obeys the Marchenko-Pastur [Pastur and Martchenko, 1967] law emerged for the eigenvalue distribution of Wishart matrices.

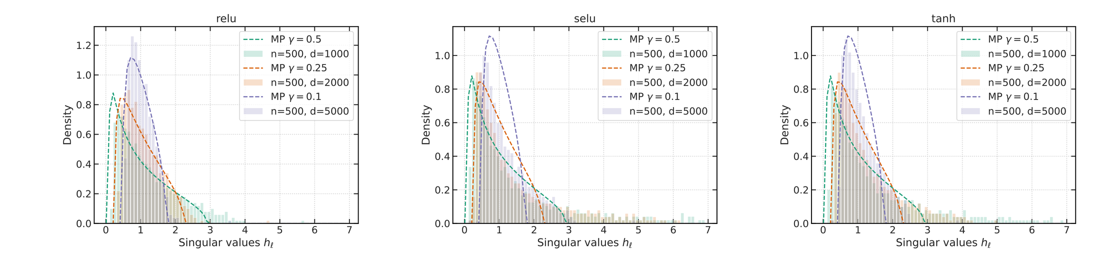


Figure 2:The density of square root of eigenvalues of the Gram matrix G_{20} when $n = 20$.

Future works

- Beyond mean field optimization for neural networks (see [Daneshmand et al., 2023]).
- A mixing analysis for representations
- Exploring other normalizations (see [Joudaki et al., 2023]).

References

[Daneshmand et al., 2021] Daneshmand, H., Joudaki, A., and Bach, F. (2021). Batch normalization orthogonalizes representations in deep random networks. *Advances in Neural Information Processing Systems.*

[Daneshmand et al., 2023] Daneshmand, H., Lee, J. D., and Jin, C. (2023). Efficient displacement convex optimization with particle gradient descent. *ICML.*

[Eberle, 2009] Eberle, A. (2009). Markov processes. *Lecture Notes at University of Bonn.*

[Hanin, 2022] Hanin, B. (2022). Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058.*

[Hanin and Nica, 2019] Hanin, B. and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989.*

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML.*

[Joudaki et al., 2023] Joudaki, A., Daneshmand, H., and Bach, F. (2023). On the impact of activation and normalization in obtaining isometric embeddings at initialization. *arXiv preprint arXiv:2305.18399.*

[Li et al., 2022] Li, M. B., Nica, M., and Roy, D. M. (2022). The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *Advances in Neural Information Processing Systems.*

[Pastur and Martchenko, 1967] Pastur, L. and Martchenko, V. (1967). The distribution of eigenvalues in certain sets of random matrices. *Math. USSR-Sbornik.*

[Pennington et al., 2018] Pennington, J., Schoenholz, S., and Ganguli, S. (2018). The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics.*

[Pennington and Worah, 2017] Pennington, J. and Worah, P. (2017). Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems.*

[Yang et al., 2019a] Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. (2019a). A mean field theory of batch normalization. *International Conference on Learning Representations.*

[Yang et al., 2019b] Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. (2019b). A mean field theory of batch normalization. In *International Conference on Learning Representations.*

contact: hdanesh@mit.edu

GitHub: