Batch Normalization Orthogonalizes Representations in DeepNets

Hadi Daneshmand, Amir Joudaki and Francis Bach

INRIA Paris, ETH Zurich, INRIA-ENS-PSL Paris

Abstract

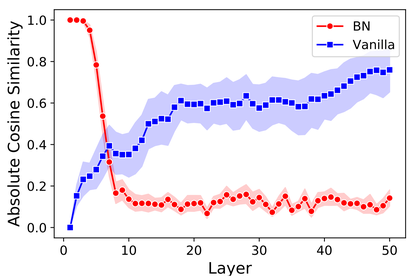
We study the interplay between modern neural components and representations in deep neural networks. In particular, we demonstrate a subtle property of batch-normalization (BN): Successive batch normalizations with random linear transformations make hidden representations increasingly orthogonal across layers of a deep neural network. We establish a non-asymptotic characterization of the interplay between depth, width, and the orthogonality of deep representations. This result has two main implications: 1) Theoretically, as the depth grows, the distribution of the representation –after the linear layers– contracts to a Wasserstein-2 ball around an isotropic Gaussian distribution. Furthermore, the radius of this Wasserstein ball shrinks with the width of the network. 2) Practically, the orthogonality of the representations directly influences the performance of stochastic gradient descent (SGD).

Batch normalization

Batch normalization [1] is the fundamental building block of modern deep neural networks. BN influences first-order optimization methods by avoiding the rank collapse in deep representation (Daneshmandet al., 2020), direction-length decoupling of optimization (Kohler et al., 2018), automatically tuning the stepsize (Arora et al., 2019b; Bjorck et al., 2018), and smoothing the optimization objective function (Santurkar et al., 2018; Karakida et al., 2019). However, the benefits of BN go beyond its critical role in optimization. For example, [2] shows that BN networks with random weights also achieve surprisingly high performance after only minor adjustments of their weights. This striking result motivates us to study the representational power ofrandom networks with BN.

BN vs. vanilla nets

We observe a stark contrast in representation of networks with BN and without BN: representations become incrementally orthogonal in BN nets, while without BN layers representations become aligned in deep layers.



Preliminaries

The chain of hidden representations:

$$H_{\ell+1} = \frac{1}{\sqrt{d}} BN(W_\ell H_\ell), \quad BN(M) = \mathrm{diag}(MM^\top)^{-1/2} M,$$

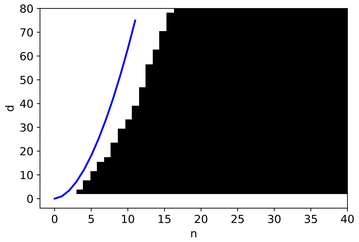
Orthogonality gap:

$$V(H) := \left\| \left(\frac{1}{\|H\|_F^2} \right) H^\top H - \left(\frac{1}{\|I_n\|_F^2} \right) I_n \right\|_F$$

Assumption $\mathcal{A}_1(\alpha, \ell)$

There exists an absolute positive constant α *such that the minimum singular value of* H_k *is greater than (or equal to)* α *for all* $k = 1, \ldots, \ell$.

We observe that $\mathcal{A}_1(\alpha_0, L)$ holds when $d = \Omega(n^2)$ for large L and an absolute constant α_0.



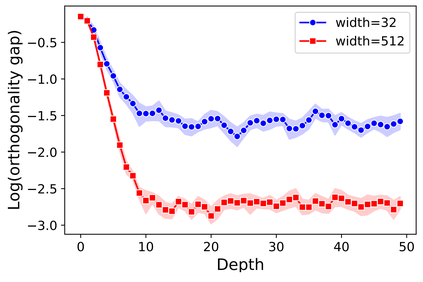
Theorem (Orthogonalization)

Under Assumption $\mathcal{A}_1(\alpha, \ell)$*, the following holds:*

$$\mathbb{E}\left[V(H_{\ell+1})\right] \le 2\left(1 - \frac{2}{3}\alpha\right)^\ell + \frac{3n}{\alpha\sqrt{d}}. \qquad (1)$$

An experimental validation

Therefore, the V (the orthogonality gap) decays at an exponential rate with depth up to a constant that is inversely proportional to the network width.



Corollary (Gaussian Approximation)

For $G \in \mathbb{R}^{d\times n}$ *with i.i.d. zero-mean* $1/d$*-variance Gaussian elements,*

$$\mathcal{W}_2\left(W_\ell H_\ell, G/\sqrt{n}\right)^2 \le 4n\left(1 - \frac{2}{3}\alpha\right)^\ell + \frac{6n^2}{\alpha\sqrt{d}} \qquad (2)$$

holds under Assumption $\mathcal{A}_1(\alpha, \ell)$.

The history of Gaussian approximation

Leveraging BN layers, we establish the first non-asymptotic Gaussian approximation for representations in deep neural networks. For vanilla networks, the Gaussianity is guaranteed only in the asymptotic regime of infinite width. Particularly, [3] links vanilla networks to Gaussian processes when their width is infinite and grows in successive layers, while our Gaussian approximation holds for networks with a finite width across layers.

∞ 1996	A single-layer MLP (Neal).
2015	Going beyond one layer (Hazan and Jaakkola).
2018	Finite-depth MLPs (Matthews et. al. and Lee et. al.) .
2021	Non-asymptotic (in)finite width and depth Linear MLPs.

Table 1:History of Gaussian approximation

Applications of Gaussian approximation

Assuming the representations are Gaussian, [4] designed novel activation functions that improve the optimization performance. Many theoretical and practical studies rely on the Gaussian output in the infinite width regime.

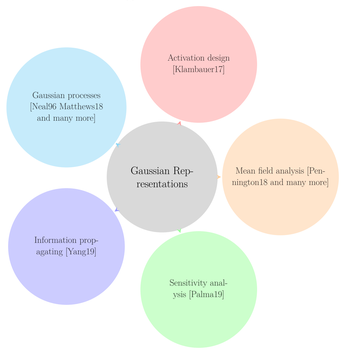
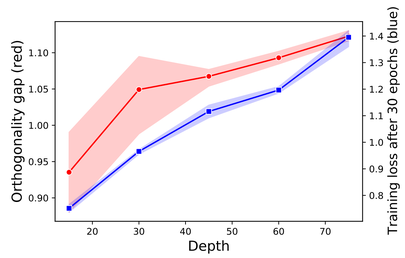


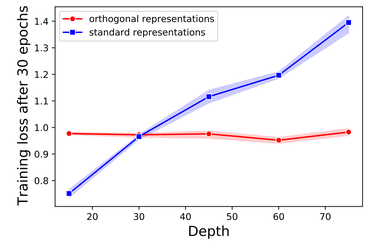
Figure 1:Applications of Gaussian approximation for deep neural network

Orthogonality and optimization

When representations are initially aligned, we observe SGD wastes many iterations to orthogonalize representations before the classification.



The influence of the orthogonality on the optimization performance inspired us to develop an initialization imposing orthogonal representations across the layers. We experimentally show that such an initialization is sufficient to accelerate SGD, with no need for BN. To enforce the orthogonality in the absence of BN, we introduce a dependency between weights of successive layers that ensures deep representations remain orthogonal by incorporating the SVD decomposition of the hidden representation of each layer into the initialization of the subsequent layer.



References

[1] Sergey Ioffe and Christian Szegedy.
Batch normalization: Accelerating deep network training by reducing internal covariate shift.
In *ICML*, 2015.

[2] Jonathan Frankle, David J Schwab, and Ari S Morcos.
Training batchnorm and only batchnorm: On the expressive power of random features in cnns.
ICLR, 2021.

[3] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani.
Gaussian process behaviour in wide deep neural networks.
ICLR, 2018.

[4] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter.
Self-normalizing neural networks.
Advances in Neural Information Processing Systems, 2017.

contact: hadi.daneshmand@gmail.com