Batch Normalization Provably Avoids Rank Collapse for Randomly Initialised Deep Networks (arXiv:2003.01652)

Presented by Hadi Daneshmand

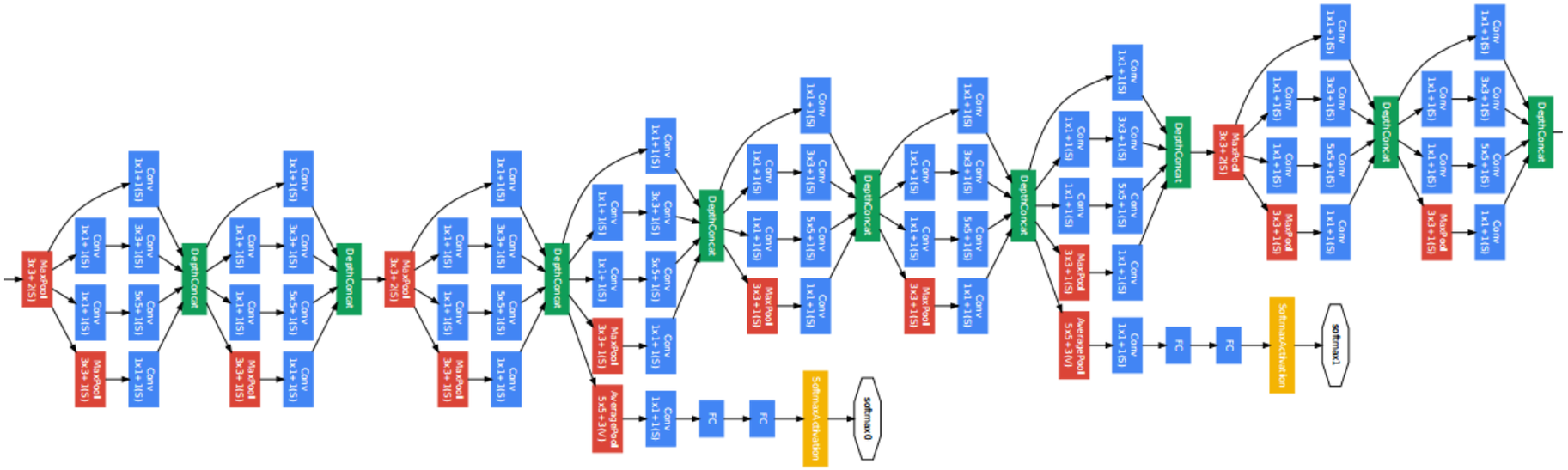
Jonas Kohler

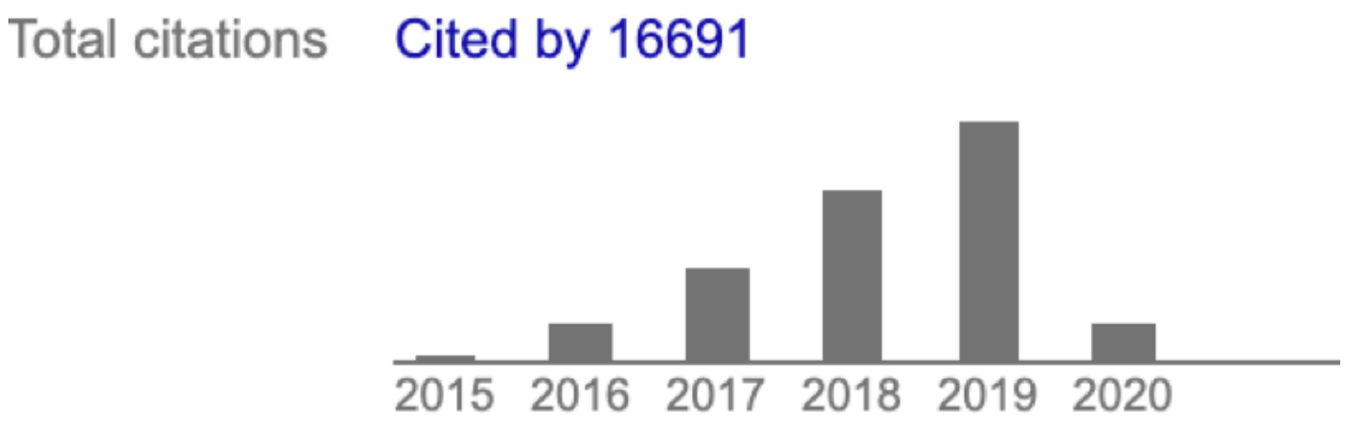
Aurelien Lucchi

Thomas Hofmann

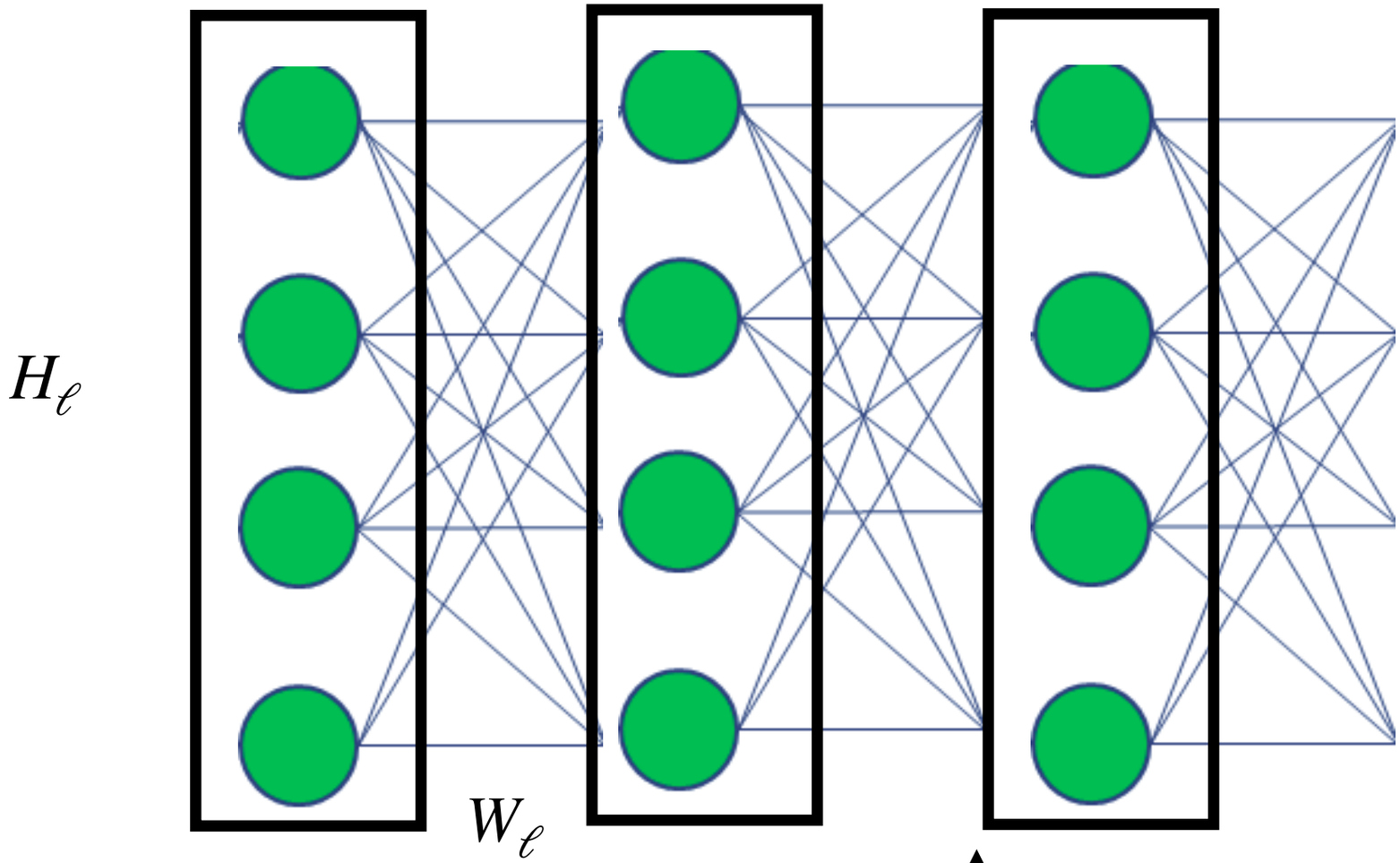
Francis Bach

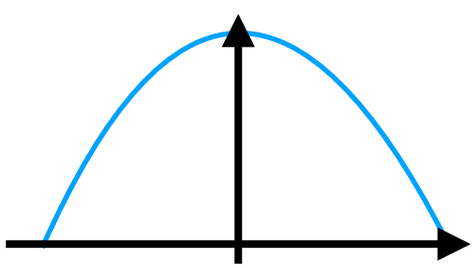
Batch Normalization (BN)

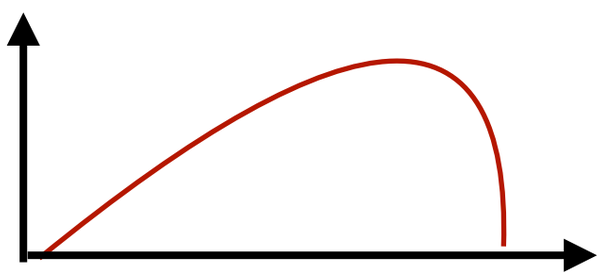




Internal covariate shift



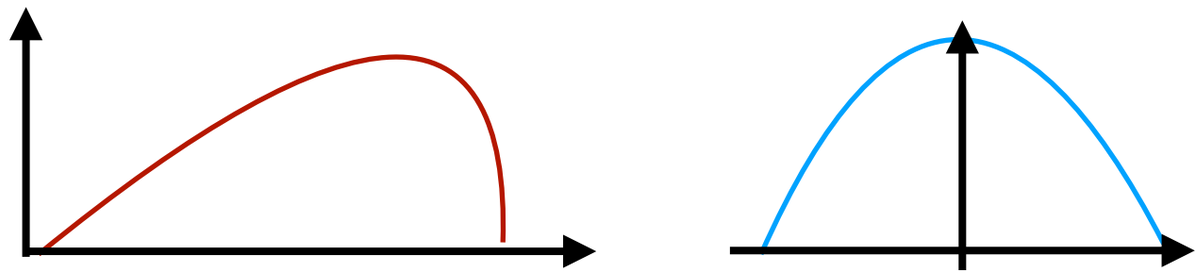




Batch normalisation (BN)

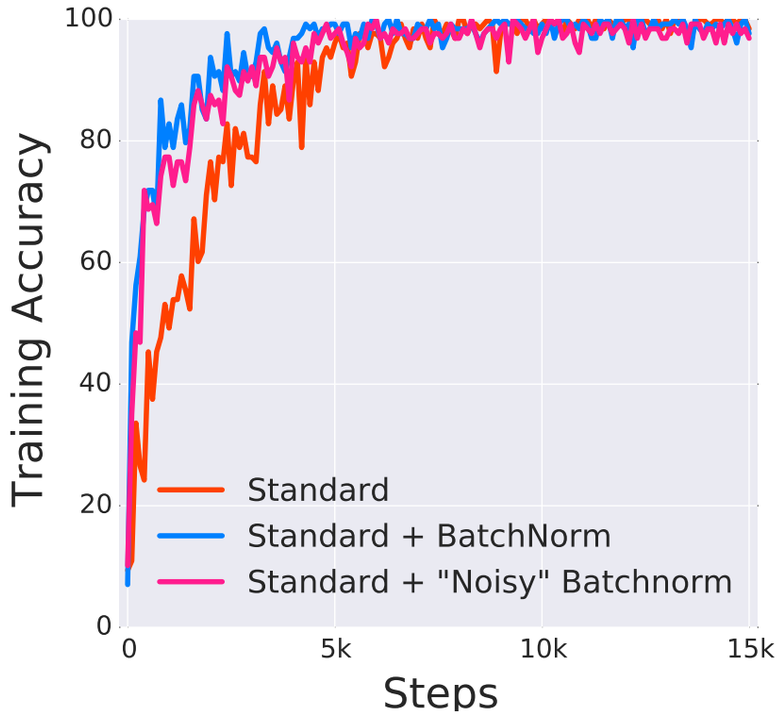
- Pre-activation matrices: $H_\ell \in \mathbb{R}^{d \times N}$
- Without BN: $H_{\ell+1} = H_\ell + \gamma W_\ell F(H_\ell)$
- With BN: $H_{\ell+1} = \text{BN}(H_\ell + \gamma W_\ell F(H_\ell))$

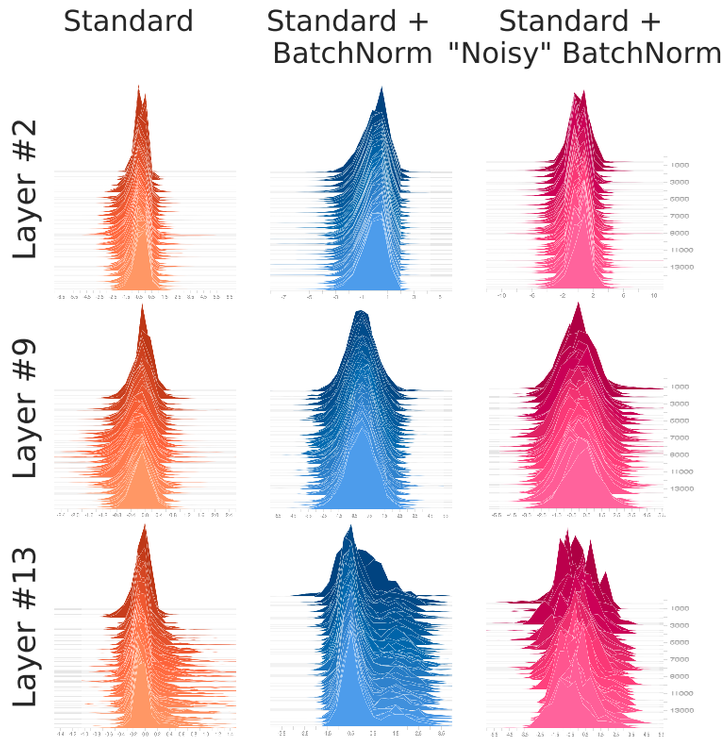
$$\textbf{BN}\left(H = \begin{bmatrix} h_{11}, \ldots, h_{1N} \\ \vdots, \vdots, \vdots \\ h_{d1}, \ldots, h_{dN} \end{bmatrix} \right) = \begin{bmatrix} h'_{11}, \ldots, h'_{1N} \\ \vdots, \vdots, \vdots \\ h'_{d1}, \ldots, h'_{dN} \end{bmatrix}$$



rows of pre-activations are **zero-mean and unit-variance**

Internal covariate shift and optimization





Santurkar, et. al (2018). How does batch normalization help optimization?… arXiv preprint arXiv:1805.11604.

BN smooths (or sharpen?) the landscape

- Claimed by Santurkar, et. al (2018). How does batch normalization help optimization?… arXiv preprint arXiv:1805.11604.

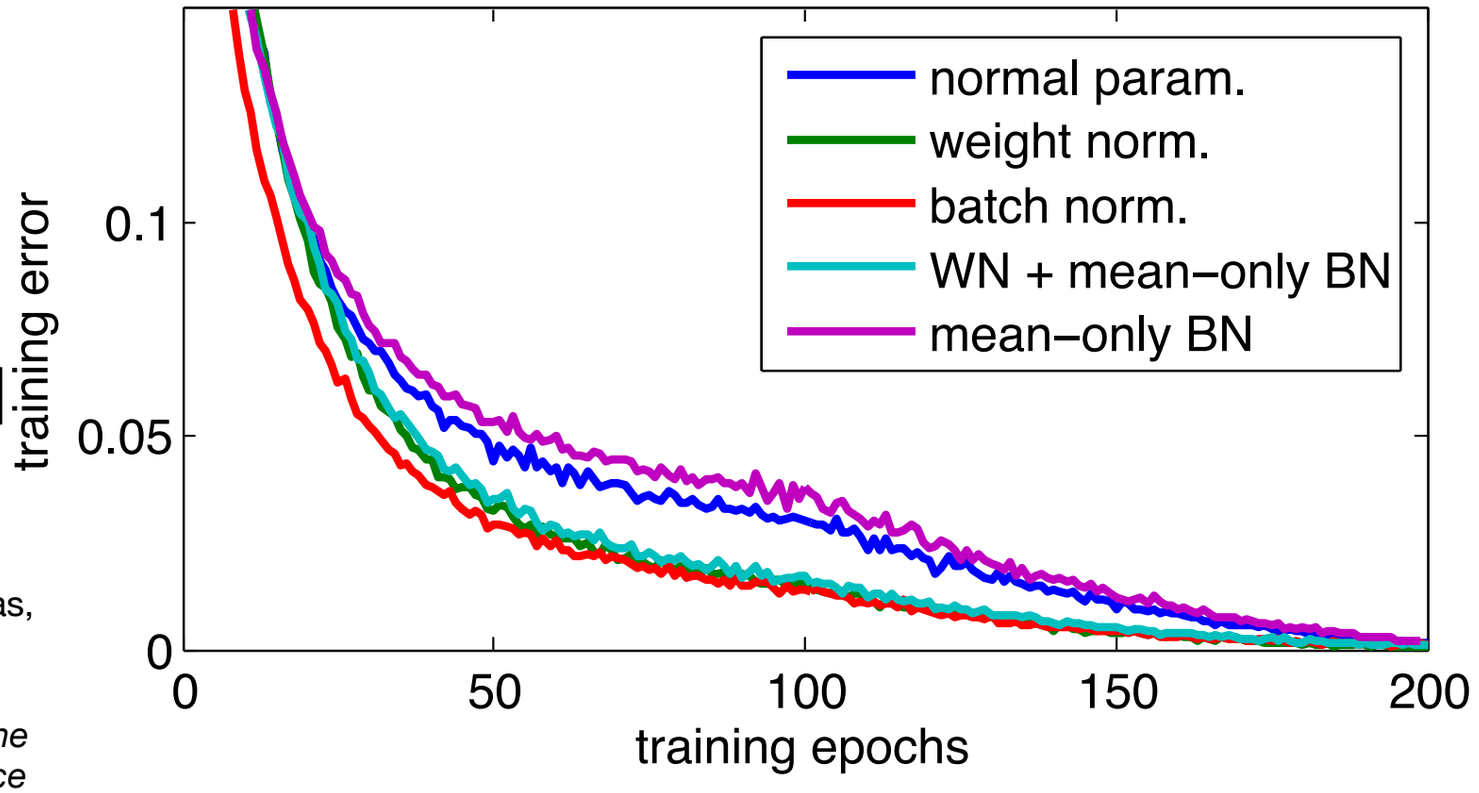
- Rejected by Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. PyHessian: Neural networks through the lens of the Hessian

.

Direction-length decoupling: weight normalisation

$$W = g\frac{V}{\|V\|}$$

A provable accelerated rate is achievable on toy examples (see Kohler, Jonas, Hadi Daneshmand, et al. "Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization." *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019.)



Salimans, Tim, and Durk P. Kingma. "Weight normalization: A simple reparameterization to accelerate training of deep neural networks." *Advances in neural information processing systems*. 2016.

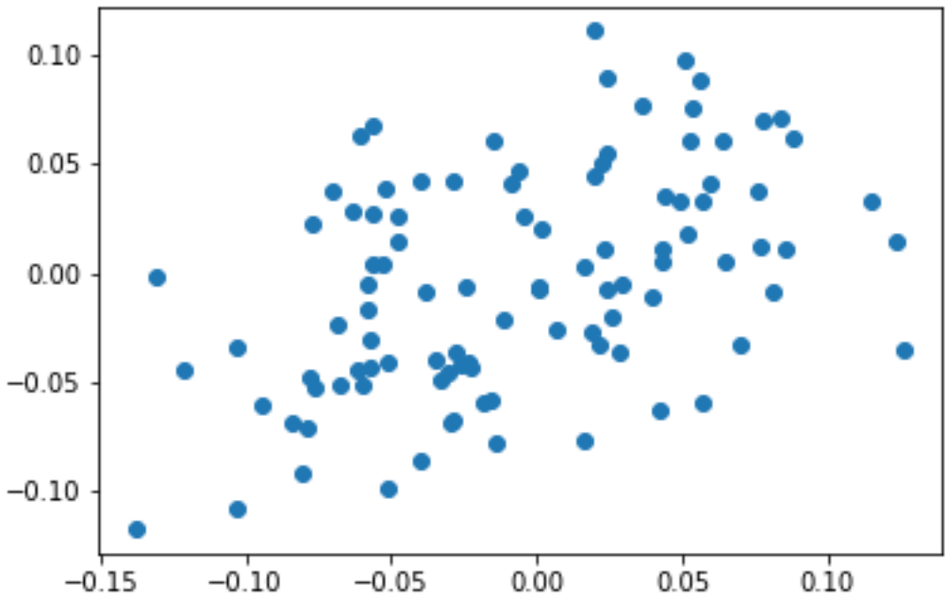
More hypotheses

- Auto-tuning of the step-size: Arora, Sanjeev, Zhiyuan Li, and Kaifeng Lyu. "Theoretical analysis of auto rate-tuning by batch normalization." *arXiv preprint arXiv: 1812.03981* (2018).

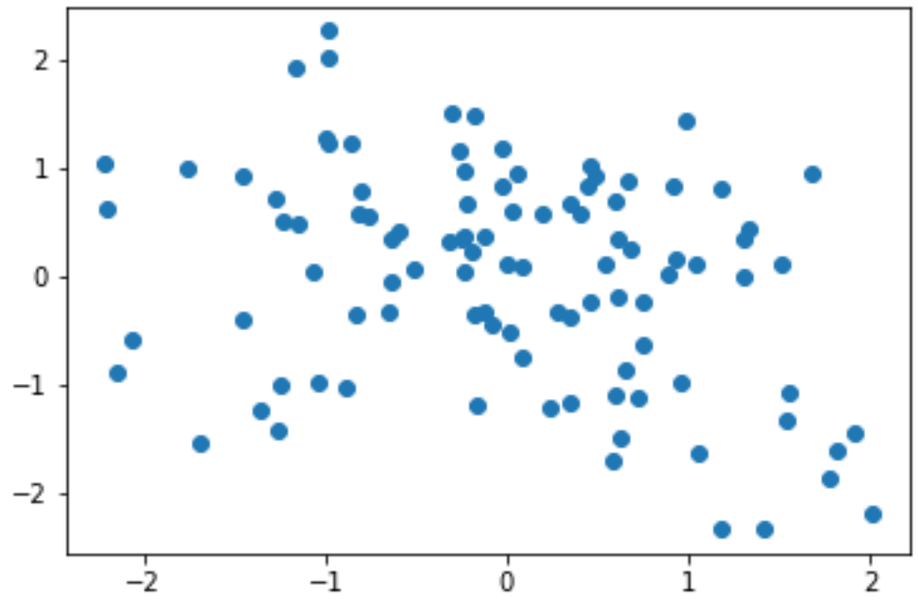
- Alleviating the sharpness of fisher information matrix: Karakida, Ryo, Shotaro Akaho, and Shun-ichi Amari. "The normalization method for alleviating pathological sharpness in wide neural networks." *Advances in Neural Information Processing Systems*. 2019.

Rank collapse issue for random nets

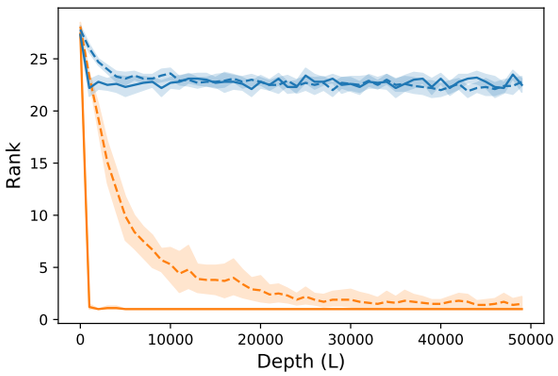
Random neural network



Random BN network



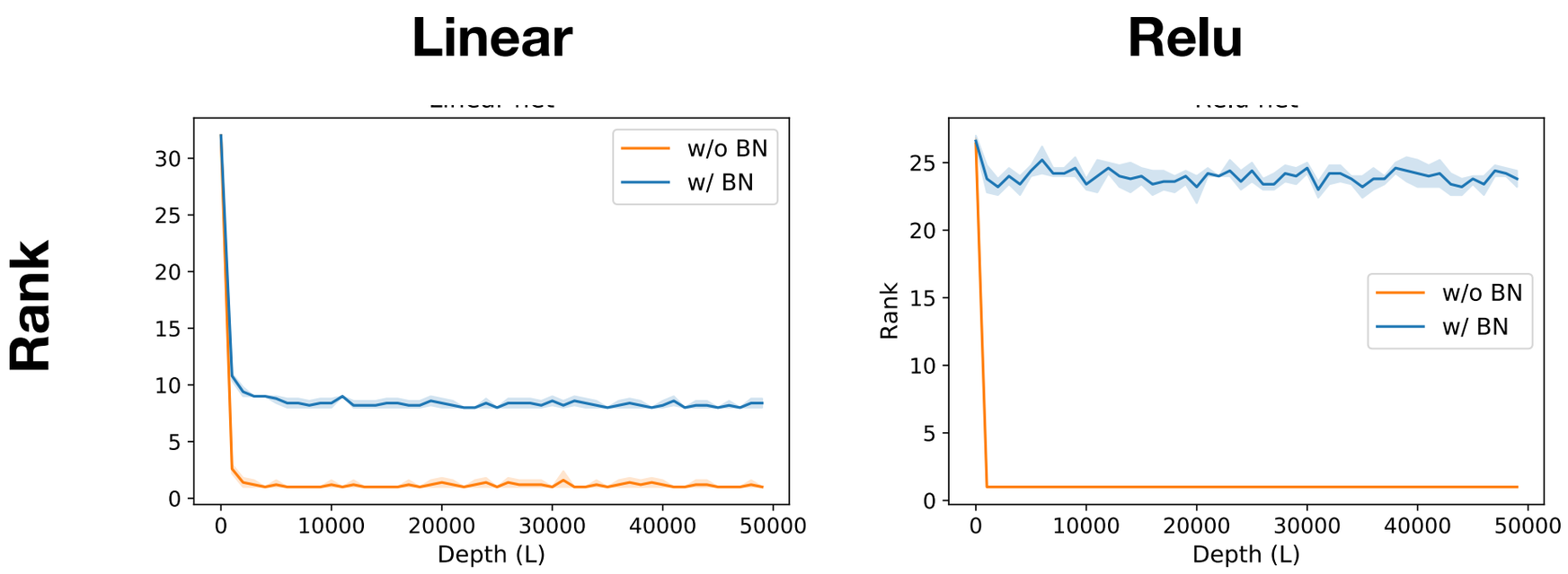
Networks without BN



Lemma

For a **vanilla network,** the rank converges to one in depth.

BN avoids the rank collapse



Number of layers

Blues: with BN **Orange: without BN**

Daneshmand, H., Kohler, J., Bach, F., Hofmann, T., and Lucchi, A. (2020). Theoretical understanding of batch-normalization: A markov chain perspective. arXiv preprint arXiv:2003.01652.

BN avoids the rank collapse

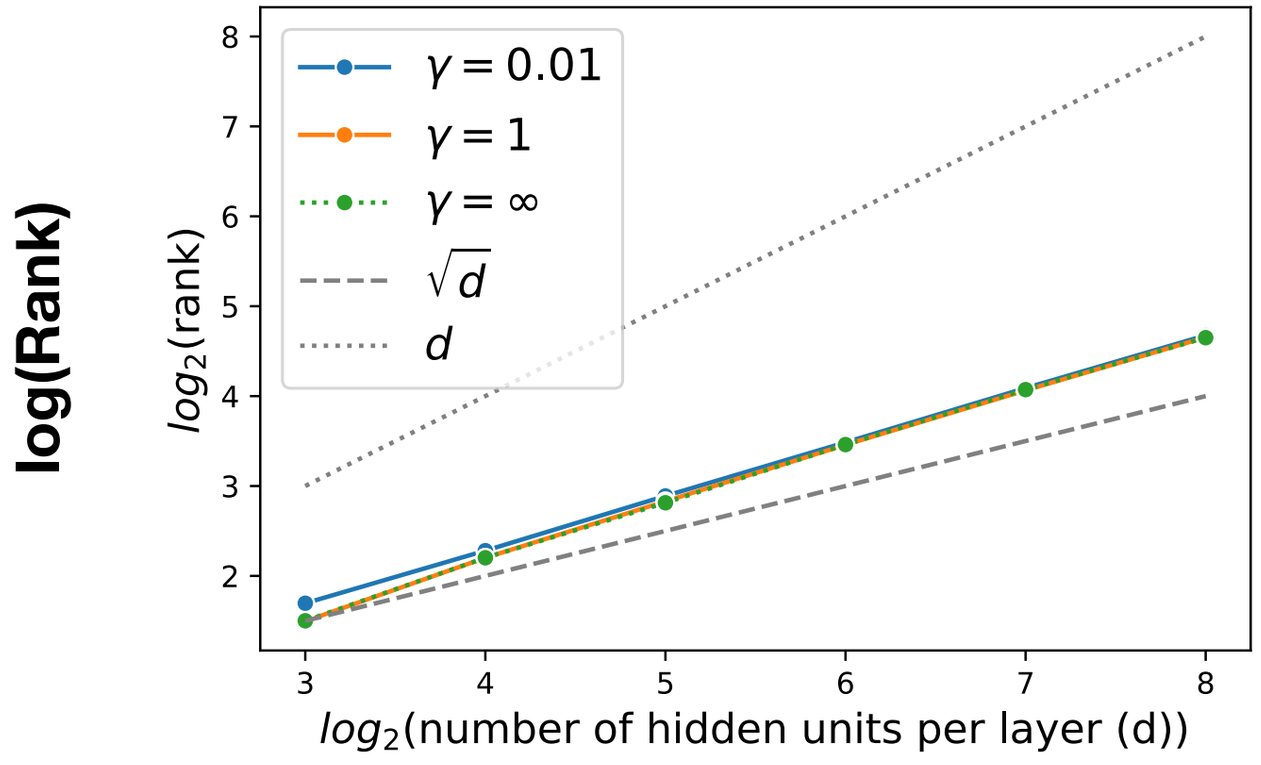
Theorem 3.3

If $F(H) = H$, $\mathrm{rank}(H_0) = d$,

$$\lim_{L\to\infty} \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{rank}_\tau(H_\ell) \geq (1-\tau)^2 \Omega(\sqrt{d})$$

holds almost surely under additional technical assumptions.

Experimental validations



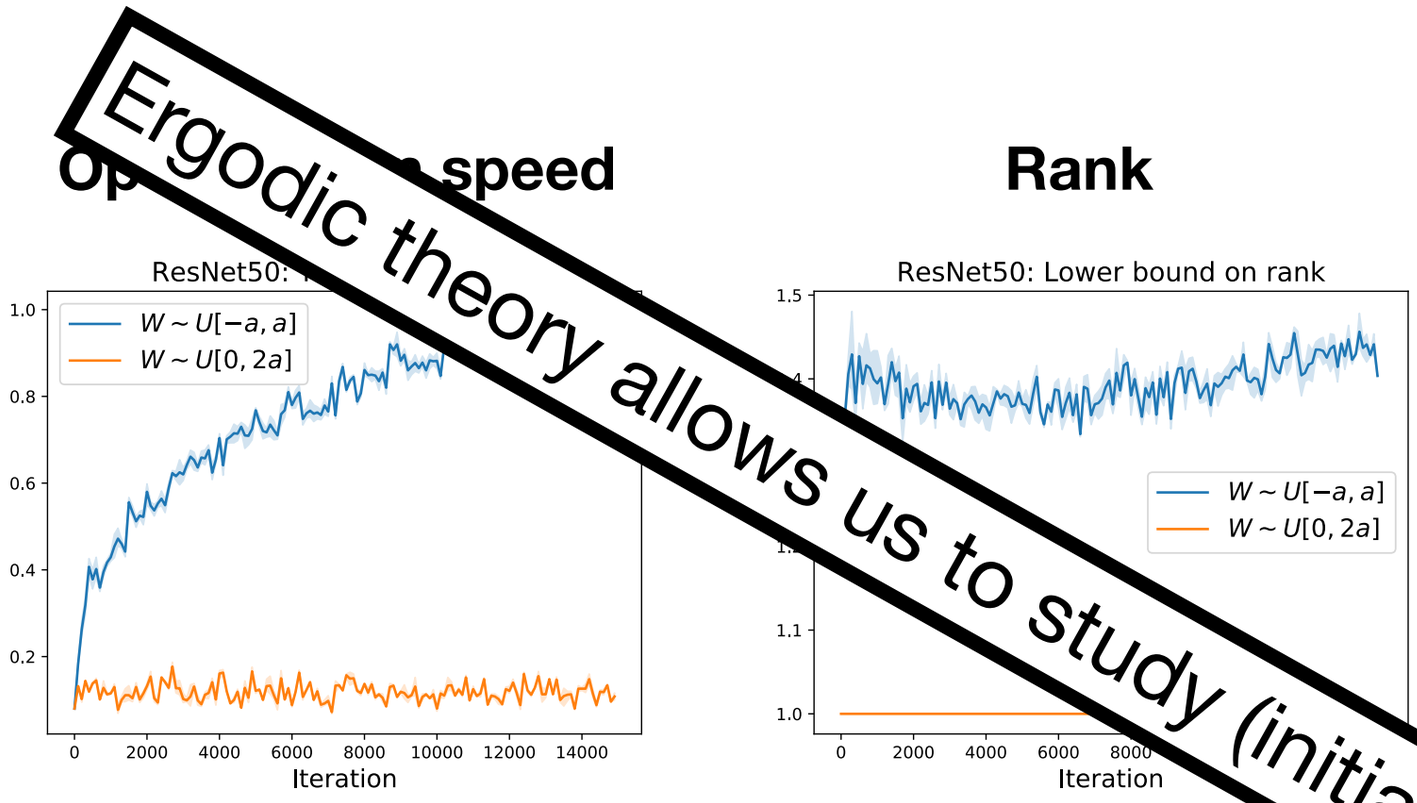
$$\frac{1}{10^6}\sum_{\ell=1}^{10^6} \textbf{rank}_\tau(H_\ell) \geq (1-\tau)^2\Omega(\sqrt{d})$$

BN: an efficient whitening

$$H_{\ell+1/2} = H_\ell + \gamma W_\ell F(H_\ell)$$

Whitening	$H_{\ell+1} = (H_{\ell+1/2} H_{\ell+1/2}^\top / N)^{-1/2} H_{\ell+1/2}$	$\textbf{rank}_\tau(H_\ell) = d$
BN	$H_{\ell+1} = (\text{diag}(H_{\ell+1/2} H_{\ell+1/2}^\top) / N)^{-1/2} H_{\ell+1/2}$	$\textbf{rank}_\tau(H_\ell) = \Omega(\sqrt{d})$

BN + initialisation = performance



Ergodic theory allows us to study (initialisation + BN)

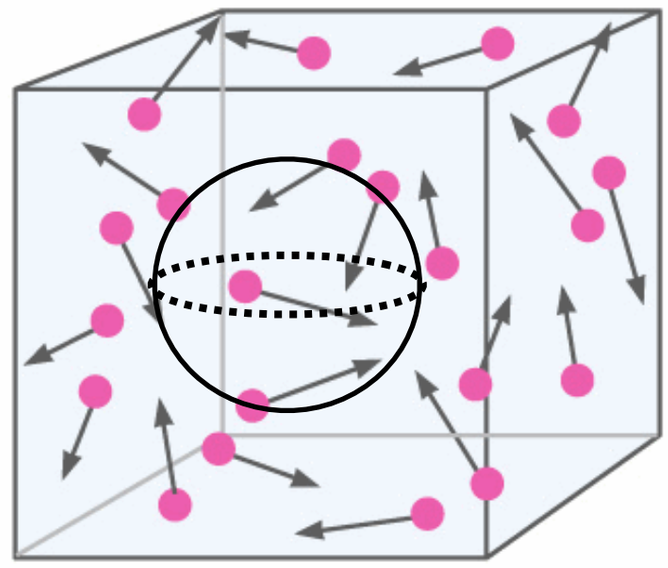
Initialization B
non-zero mean

Initialization A
zero mean

Figure 3.7: **Breaking Batchnorm:** CIFAR-10 on a VGG19 (top) (bottom) architecture with standard PyTorch initialization as well as a uniform initiali variance in $\mathbb{R}^+$. Average and 95% confidence interval of 5 independent runs.

Background: Ergodicity

- Hamiltonian dynamics: $\dot{p} = -\frac{\partial H}{\partial q}, \dot{q} = -\frac{\partial H}{\partial p}$



- Volume measure: $\rho(p, q, t)dpdq$

- Gipps' dynamics (Liouville 1838): $\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_{i=1}^{n} \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right)$

Measure-preserving systems

- System $(X, \sigma(X), \nu, T)$ is measure-preserving

$$T : X \to X, \quad \forall A \in \sigma(X) : \nu(T^{-1}(A)) = \nu(A)$$

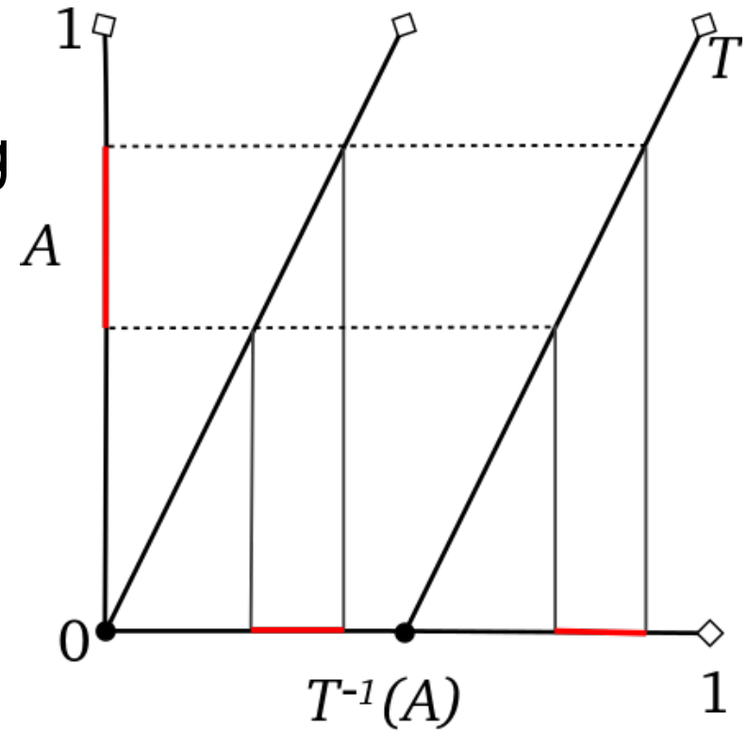


image credit: wikipedia

- Example 1:

$$T : [0,1) \to [0,1),\ T(x) = (2x \ \mathbf{mod}\ 1)$$

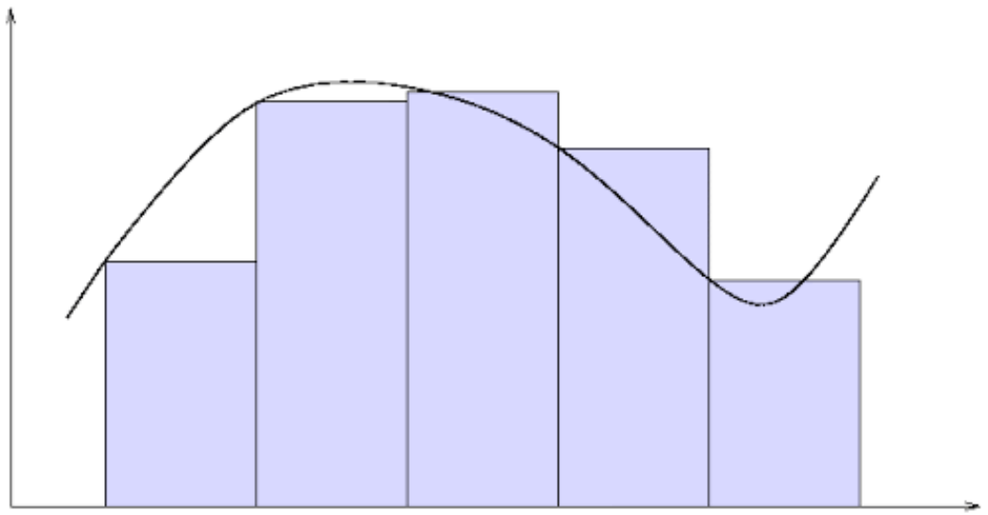
- Example 2:

Liouville's theorem (Hamiltonian): the flow of a Hamiltonian vector field on the tangent bundle of a closed connected smooth manifold is measure-preserving

Birkhoff's ergodic theorem

A measure-preserving $(X, \sigma(X), \nu, T)$

T is Ergodic if $T^{-1}(A) = A \implies \nu(A) \in \{0,1\}$



Time average **State average**

$$\frac{1}{n}\sum_{k=1}^{n} f(T^k(x)) \to \int f(x)\nu(dx)$$

Applications: Numerical integration

Birkhoff's Ergodicity for Markov chains

- Consider the following Markov chain (MC)

$$X_{n+1} = T(X_n, W_n)$$

- Invariant distribution associated with the MC

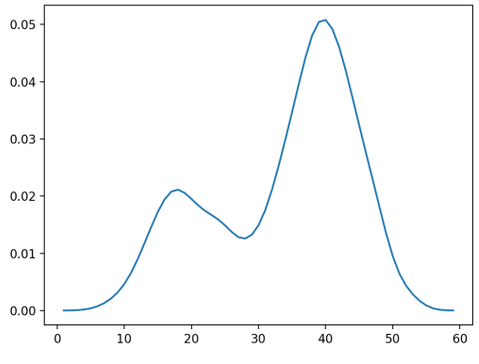
$$\int f(T(X,W))\nu(dX)\mu(dW) = \int f(X)\nu(dX)$$

- If the chain admit a unique invariant distribution, then

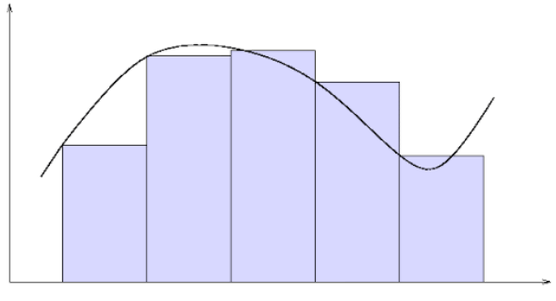
$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \mathbb{E}_\nu\left[f(X)\right]$$

$(\text{numerical integration})^{-1}$

(NI):



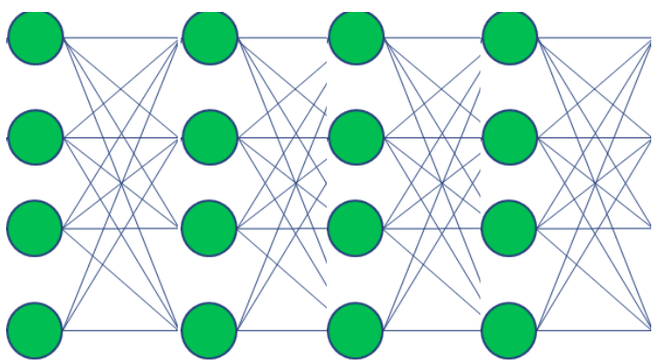
$\Longrightarrow$



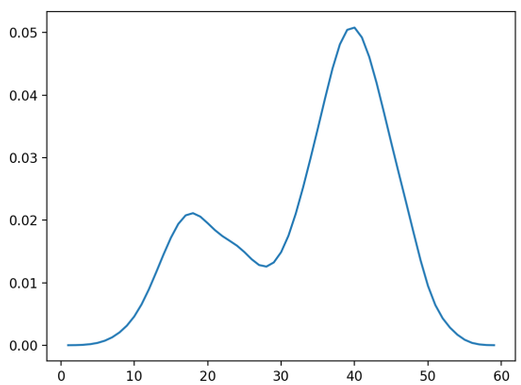
Given ν and f: $\int f(x)\nu(dx)$

$$\frac{1}{n}\sum_{k=1}^{n} f(T^k(x))$$

$(\text{NI})^{-1}$:



$\Longrightarrow$



$$\frac{1}{L}\sum_{\ell=1}^{L} f(\underbrace{T^{\ell}(H_0)}_{H_\ell})$$

Studying the invariant measure ν

Proof sketch of Thm. 3.3

1. Establishing a lower-bound on the soft rank

   - $r(H) := d^2/\|M(H)\|_F^2, \quad M(H) := HH^\top/N$

   - $\text{rank}_\tau(H) \geq (1-\tau)^2 r(H)$

2. For the invariant measure ν:

   - $\mathbb{E}_{H\sim\nu}\left[\|M(H)\|_F^2\right] \leq O(d\sqrt{d}) \implies \mathbb{E}_{H\sim\nu}\left[r(H)\right] \geq \Omega(\sqrt{d})$

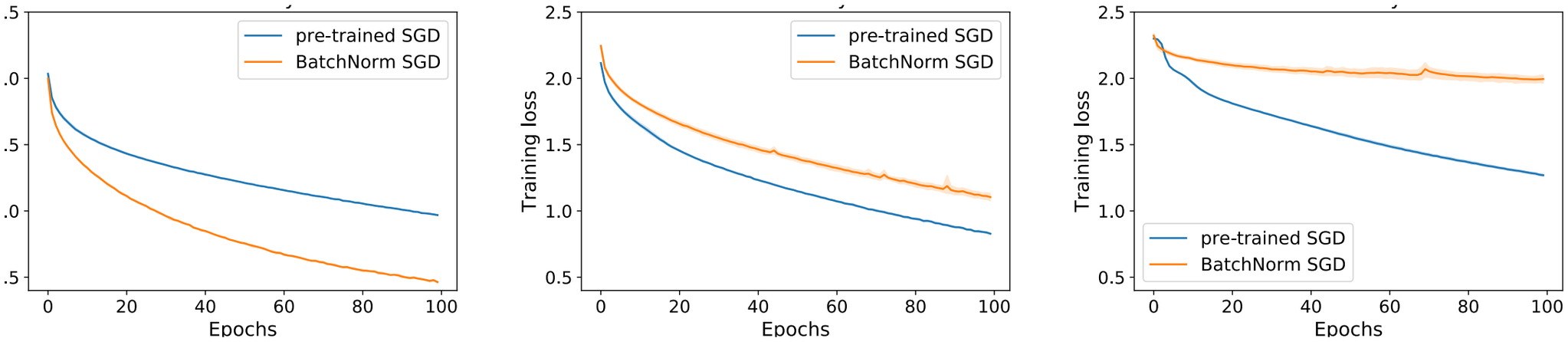
3. Ergodicity

   - $\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} \text{rank}_\tau(H_\ell) \geq \lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} (1-\tau)^2 r(H_\ell) \geq (1-\tau)^2 \mathbb{E}_{H\sim\nu}\left[r(H)\right]$

A practical implication

A pre-training step: $\max \left(r(H_L) := d^2 / \|M(H_L)\|_F^2 \right)$

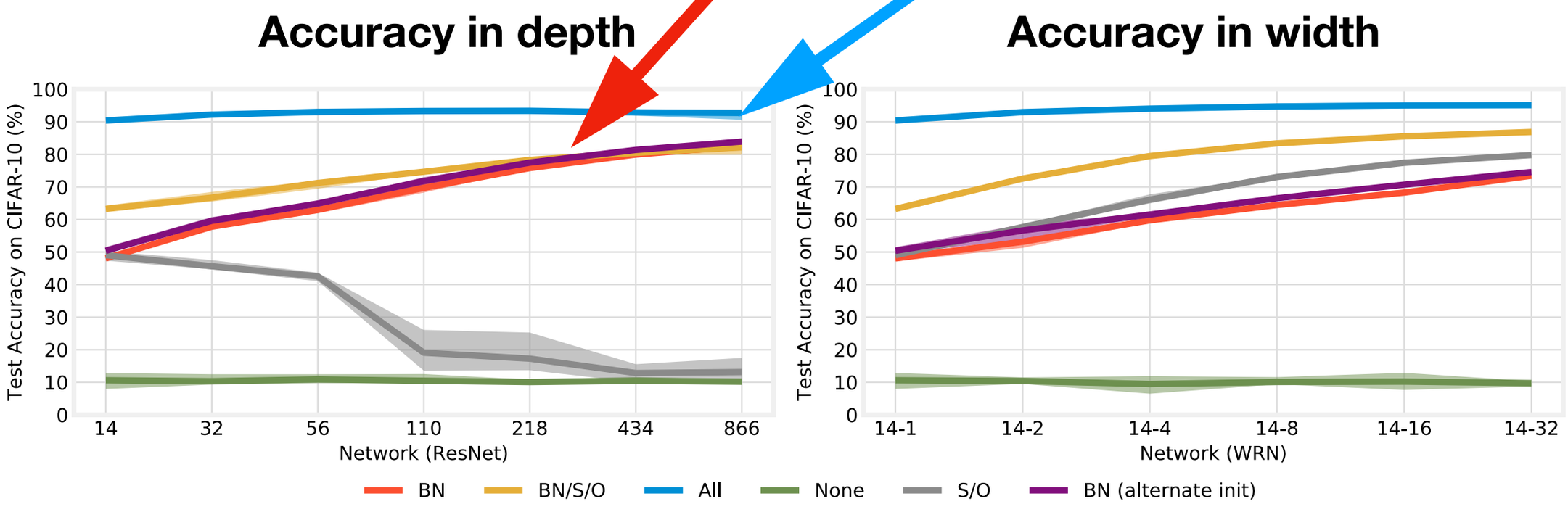
10 layers **30 layers** **50 layers**



Blues: Pre-training **Orange: BN**

Forget opt. of weights!

$$H_\ell = BN_{\alpha_\ell,\beta_\ell}(H_\ell + \gamma W_\ell F(H_\ell))$$



Jonathan Frankle and David J. Schwab and Ari S. Morcos, Training BatchNorm and Only BatchNorm: On the Expressive Power of Random Features in CNNs; arXiv 2003.00152 29 Feb 2020.

Why BN for random nets?

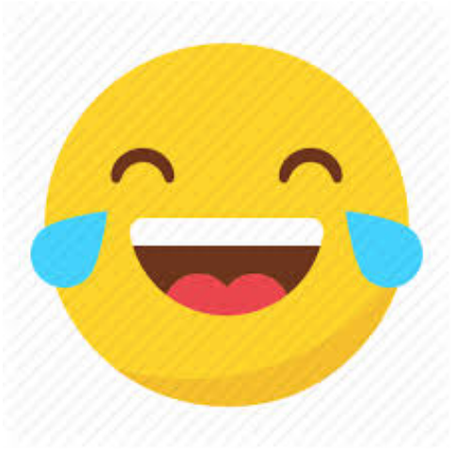
Our analysis implies that:

BN provides disguising features

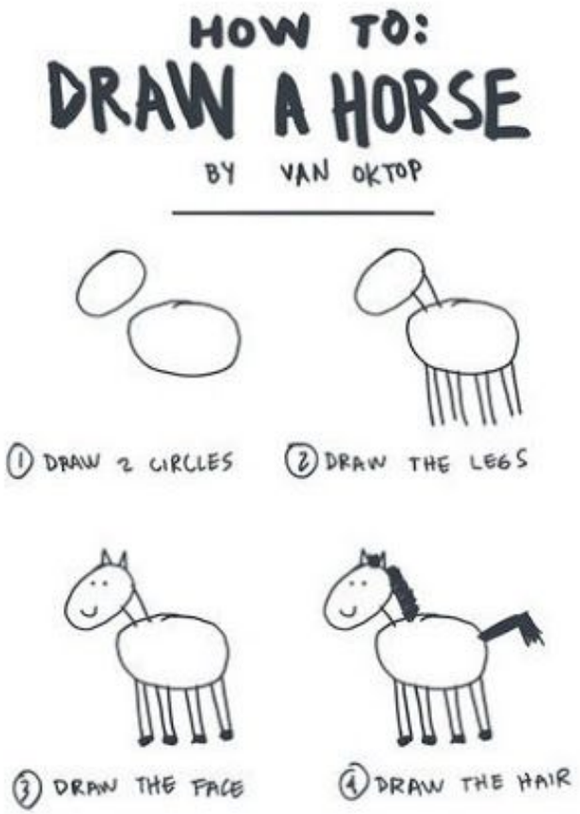
Future works

- Theoretical results for non-linear networks
- Construction of the invariant distribution
- Extension to conv-nets
- Designing better initialisation scheme for weights.

Relaxing the assumptions



Result with technical assumptions



Result presented here
without technical assumptions

Soft rank

$$\text{rank}_\tau(H) = \sum_{i=1}^{d} \mathbf{I}(\sigma_i(H)^2/N \geq \tau)$$

$$\text{rank}(H) \geq \text{rank}_\tau(H)$$