Dynamic Isometry As a Consequence of Weight Orthogonality for Faster and Better Convergence

EsterHlav

EECS 6699 Mathematics of Deep Learning
Columbia University

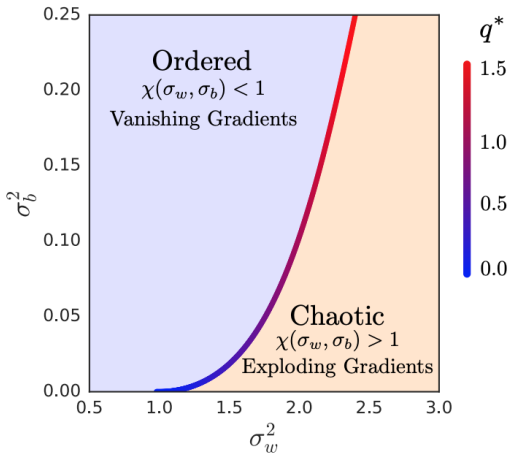
May 15, 2019

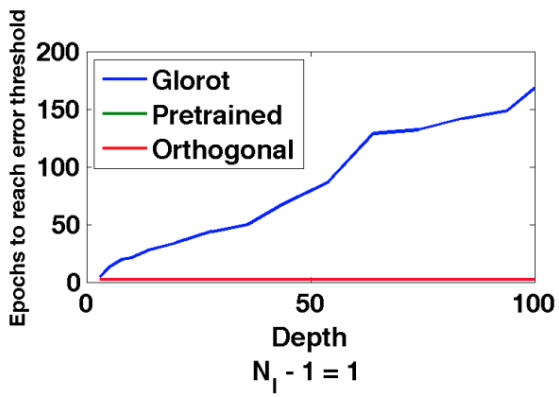
How does Orthogonal Initialization Help Neural Networks?

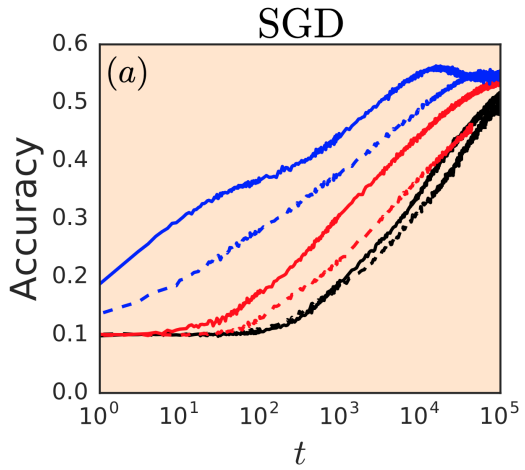
(A) Orthogonality limits the impact of **vanishing and exploding gradient** in neural networks by enforcing dynamical isometry

(B) With orthogonal weight initialization, the **difference in speed of convergence** between deep and shallow linear neural networks can become negligible

(C) **Non-linear neural networks with isometry** consistently outperform those without isometry in accuracy







What if We Maintain Orthogonality during Training?

Problem: Orthogonal initialization of weight matrices does not guarantee dynamical isometry after initialization
$\rightarrow$ How can we enforce orthogonality throughout training?

Idea: Orthogonal regularization with *hard/soft* constraint leads matrices *within/close* to the Stiefel Manifold.

Intuition and Interrogations:

(D) Can **excessive orthogonality** constraint (*i.e. hard* regularization) hurt performance?

(E) Can **specific conditions** (*e.g.* depth) enforce orthogonality in a more beneficial way than others?

$\rightarrow$ Empirical results remain inconclusive.

Outline

I. Dynamical Isometry in Neural Networks

II. Orthogonal Weight Initialization in Linear and Non-linear Networks

III. Orthogonal Regularization with Hard and Soft Constraint

IV. Experiments - Orthogonality with Gain in RNNs

I. Dynamical Isometry: Setup

Input-Output Jacobian **J**

The input-output Jacobian **J** for a Multi-Layer Perceptron with L hidden-layers and N hidden-nodes is defined as:

$$\mathbf{J} = \frac{\partial x_L}{\partial h_0} = \prod_{i=1}^{L} D_i W_i, \text{ with singular values } s_i, \tag{1}$$

with D_i a diagonal matrix consisting of $diag\{\phi'(h_{i,j})\}$.

Due to chain rule's mechanics, **J** is related to the entire network's back-propagation dynamics:

- if **J ill-conditioned**, i.e. large/low $\frac{s_{max}^2}{s_{min}^2}$, then J is a norm-exploding/vanishing mapping.
- if **J well-conditioned**, i.e. eigenvalues equal to one, *and* **J is a norm-preserving mapping**, i.e. $\|J\vec{v}\| = \|\vec{v}\|$, then dynamical isometry is reached!

Derivation of Spectral Distribution of J

Signal propagation. From [2] Pennington *et al.*

As $N \to \infty$, the empirical distribution of pre-activation $h_i = W_i x_{i-1} + b_i$ converges to $N(0, q_i)$. We obtain a recursive equation for the $\text{Var}(h_i)$:

$$q_i = \sigma_w^2 \mathbb{E}_h[\phi(\sqrt{q_{i-1}}h)^2] + \sigma_b^2 \text{ with } h \sim N(0,1), \tag{2}$$

where σ_w^2 equals to the variance of weights, σ_b^2 the variance of bias, and the initial condition $q_0 = \frac{1}{N}\sum_{i=1}^{N} h_{i,j}^2$. The recursion has a fixed point q^*:

$$q^* = \sigma_w^2 \mathbb{E}_h[\phi(\sqrt{q^*}h)^2] + \sigma_b^2 \tag{3}$$

Now we can derive a relationship between χ, **the mean of the spectral distribution of one-layer Jacobian** $J_i = D_i W_i$, and the fixed point $\mathbf{q}^*$.

Hence, if pre-activations h_i have fixed point distribution with variance q^*:

$$\chi = \sigma_w^2 \mathbb{E}_h[\phi'(\sqrt{q^*}h)^2] \tag{4}$$

(A) Characterization of Ill-/Well-Conditioning of J

It can be further shown: the **mean of spectral density of J** becomes χ^L.

So: $\chi = 1 \implies \chi^L = 1 \implies$ J well-conditioned $\implies$ dynamic isometry.
And $\chi > 1$ $(\chi < 1) \implies$ chaotic (ordered) phase of exploding (vanishing) gradients

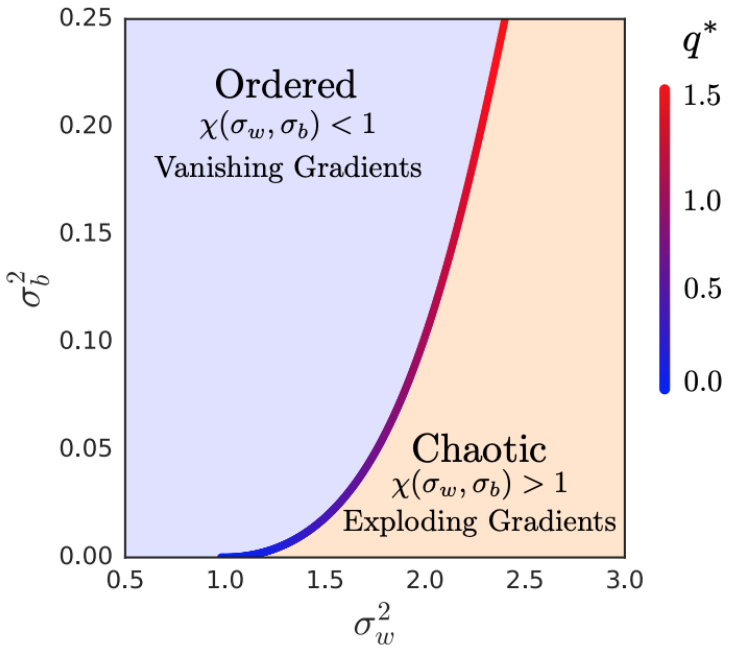


Figure 2: Order-chaos transition $\phi = tanh$. From [2] Pennington *et al.*

II. Orthogonal Weight Initialization in Linear Networks

Linear Network simplification. From [1] Saxe *et al.*

The intput-output Jacobian becomes very simple for $\phi(x) = x$:

$$\mathbf{J} = \prod_{i=1}^{L} W_i \tag{5}$$

And the **fixed point equation** becomes $q^* = \sigma_w^2 q^* + \sigma_b^2$.
Thus the critical point q^* happens only for $(\sigma_w, \sigma_b) = (1, 0)$.

- **W_i initialized Orthogonal**: **J** is orthogonal as a product of L orthogonal matrices. Thus, perfect isometry.
- **W_i initialized Gaussian**: distribution of eigenvalues of $\mathbf{J}\mathbf{J}^T$ is found explicitly as $\lambda_{max} = \frac{(L+1)^{L+1}}{L^L} \sim L$, which scales linearly with depth, and $\sigma^2_{JJ^T} = L$. Jacobian breaks dynamical isometry.

(B) Rate of Convergence in Linear Networks

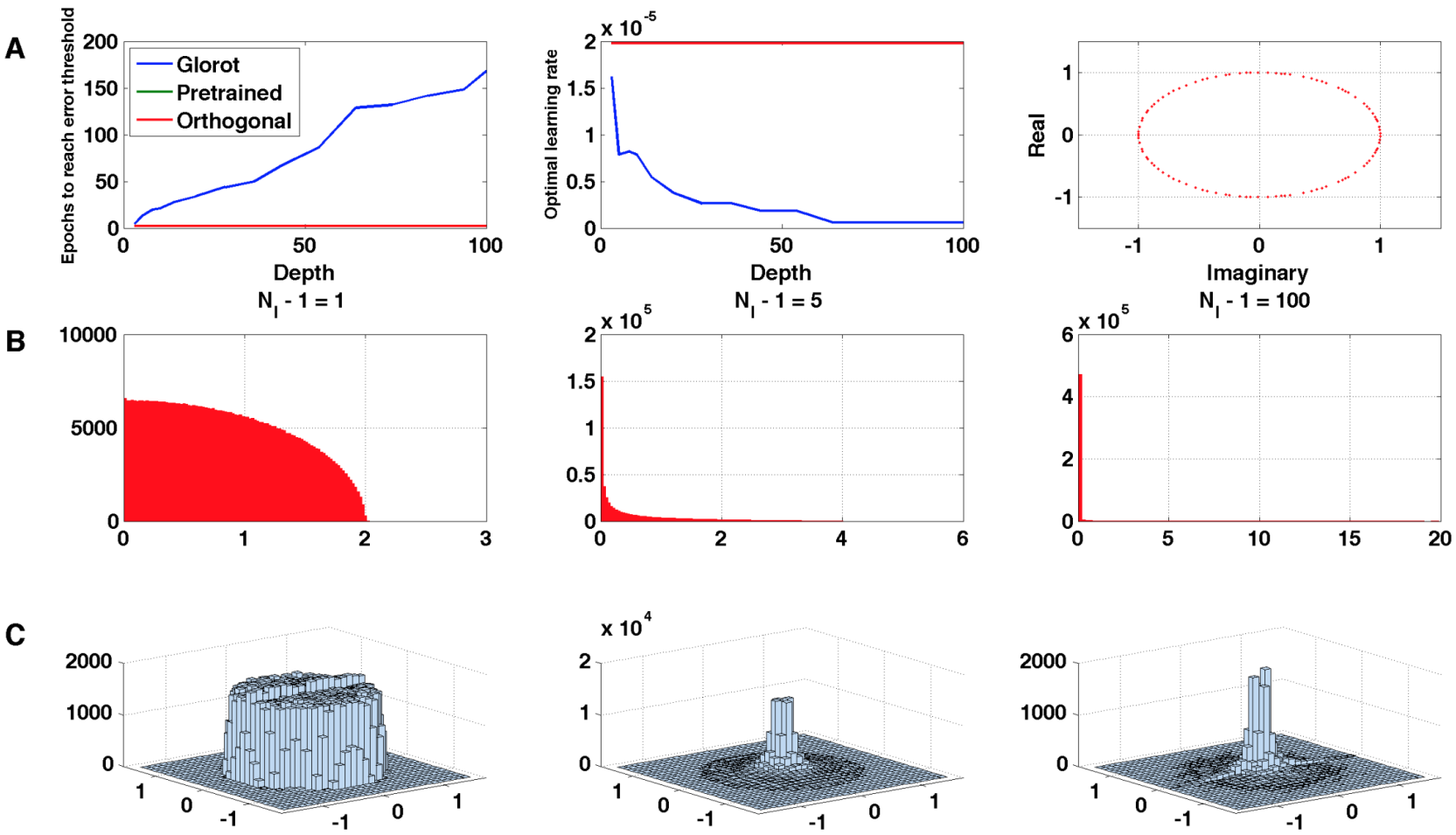


Figure 3: **(A)** Time of learning and optimal learning for MNIST, $\sigma(A^{\perp}_{100\times 100})$, **(B, C)** histograms of $\sigma(\prod^{L} W_i^G)$ for different depths. From [1] Saxe *et al.*

II. Orthogonal Weight Initialization in Non-Linear Networks

$\chi = \sigma_w^2 p(q^*)$ with $p(q^*)$ the probability that a neuron is in linear regime.

- **ReLU**: $p(q^*) = \frac{1}{2} \implies \chi = \frac{\sigma_w^2}{2}$ and for $\chi = 1$ the only critical point is $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$.
- **Hard-tanh**: $p(q^*) = erf(\frac{1}{\sqrt{2q^*}})$ and $\chi = 1$ yields a curve in (σ_w, σ_b).

In the **Gaussian** case, from [2] Pennington *et al.*:

$$\lambda_{max} \propto \sigma_w^{2L} p(q^*)^{L-1} L \text{ and } \sigma_{JJ^T}^2 = \frac{L}{p(q^*)} \geqslant L \text{ as } p(q^*) \leqslant 1 \qquad (6)$$

In the **Orthogonal** case, from [2] Pennington *et al.*:

$$\lambda_{max} \propto \sigma_w^{2L} p(q^*)^{L-1} (1 - p(q^*)) \frac{L^L}{(L-1)^{L-1}} \text{ and } \sigma_{JJ^T}^2 = \frac{1 - p(q^*)}{p(q^*)} L \qquad (7)$$

Hence, *dynamic isometry* can be achieved by **hard-tanh with orthogonal** initialization, but **not by ReLU networks**, even with orthogonality.

(C) Non-linear Neural Networks with Isometry

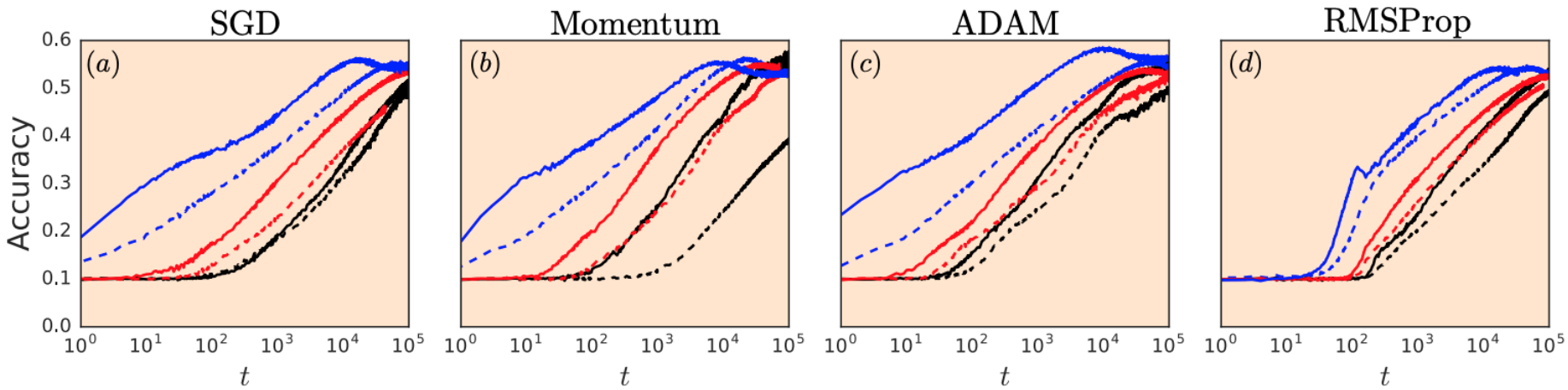


Figure 4: Learning dynamics, measured by *generalization* performance on a test set, for networks of depth 200 and width 400 trained on CIFAR-10 with different optimizers. Blue is $tanh$ with $\sigma_w^2 = 1.05$, red is $tanh$ with $\sigma_w^2 = 2$, and black is ReLU with $\sigma_w^2 = 2$. Solid lines are orthogonal and dashed lines are Gaussian initialization. The relative ordering of curves robustly persists across optimizers, and is strongly correlated with the degree to which dynamical isometry is present at initialization, as measured by $s_{\max}$ in Fig. 3. Networks with $s_{\max}$ closer to 1 learn faster, even though all networks are initialized critically with $\chi = 1$. The most isometric orthogonal $tanh$ with small σ_w^2 trains several orders of magnitude faster than the least isometric ReLU network.

Figure 4: From [2] Pennington *et al.*

III. Orthogonal Regularization: *Hard* constraint

Idea: Enforce after each batch update that all weight matrices W are within the **Stiefel manifold** $\mathcal{V}_k(\mathbb{R}^n) = \{W \in \mathbb{R}^{n \times k} | WW^T = \mathrm{I}_{\mathbb{R}^{k \times k}}\}$.

Many implementations exist such as:

- Change gradient update using *Cayley transform* to project update on Stiefel Manifold
- QR decomposition to project back the updated $\mathbb{W}$ onto the Stiefel Manifold
- Orthogonal Linear Module that optimizes a proxy matrix V such that $W = \phi(V) \in \mathcal{V}_k(\mathbb{R}^n)$

Each method yields different paths in Stiefel Manifold with different computational cost. E.g. matrix inversion and SVDs have cost $\mathcal{O}(n^3)$.

III. Orthogonal Regularization: *Soft* constraint

Idea: Penalize the deviation from Stiefel manifold in the loss function for all W by adding the term (from Single-Soft Orthogonal Regularization):

$$\lambda \left\| WW^T - \mathrm{I} \right\|_F^2, \tag{8}$$

which has an explicit gradient $4\lambda W(WW^T - \mathrm{I})$.

Recall that the *Frobenius* norm is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)} \tag{9}$$

More soft constraints can be defined such as Double Soft Orthogonality Regularization, Mutual Coherence Regularization, Spectral Restricted Isometry Property Regularization...

(D) Hard orthogonal regularization can hurt performance

Hard orthogonal regularization **can hurt performance**, potentially because of *too hard* restrictions on the network capacity.

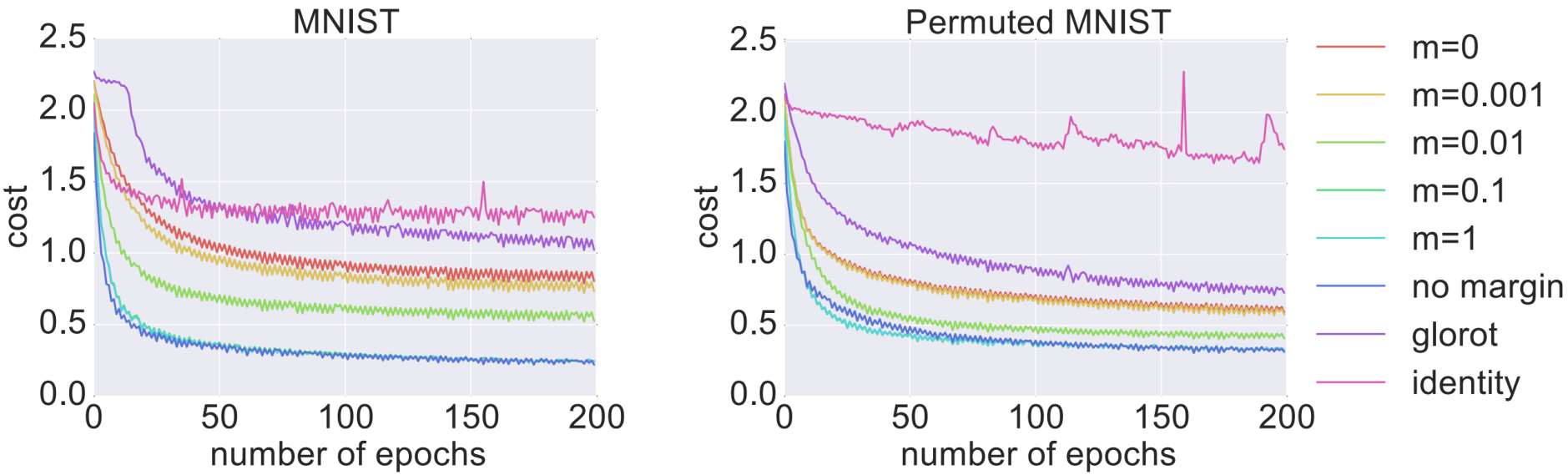


Figure 5: Sequential MNIST RNN with hard-constraint using *Cayley* transform to maintain singular values within a margin around 1: $\sigma(W) \in [1-m, 1+m]$. Best results obtained **without margin**! From [3] Vorontsov *et al.*

(E) Soft orthogonal regularization can help deep networks

Soft orthogonal regularization that is carefully selected **can be beneficial** for large and deep networks.

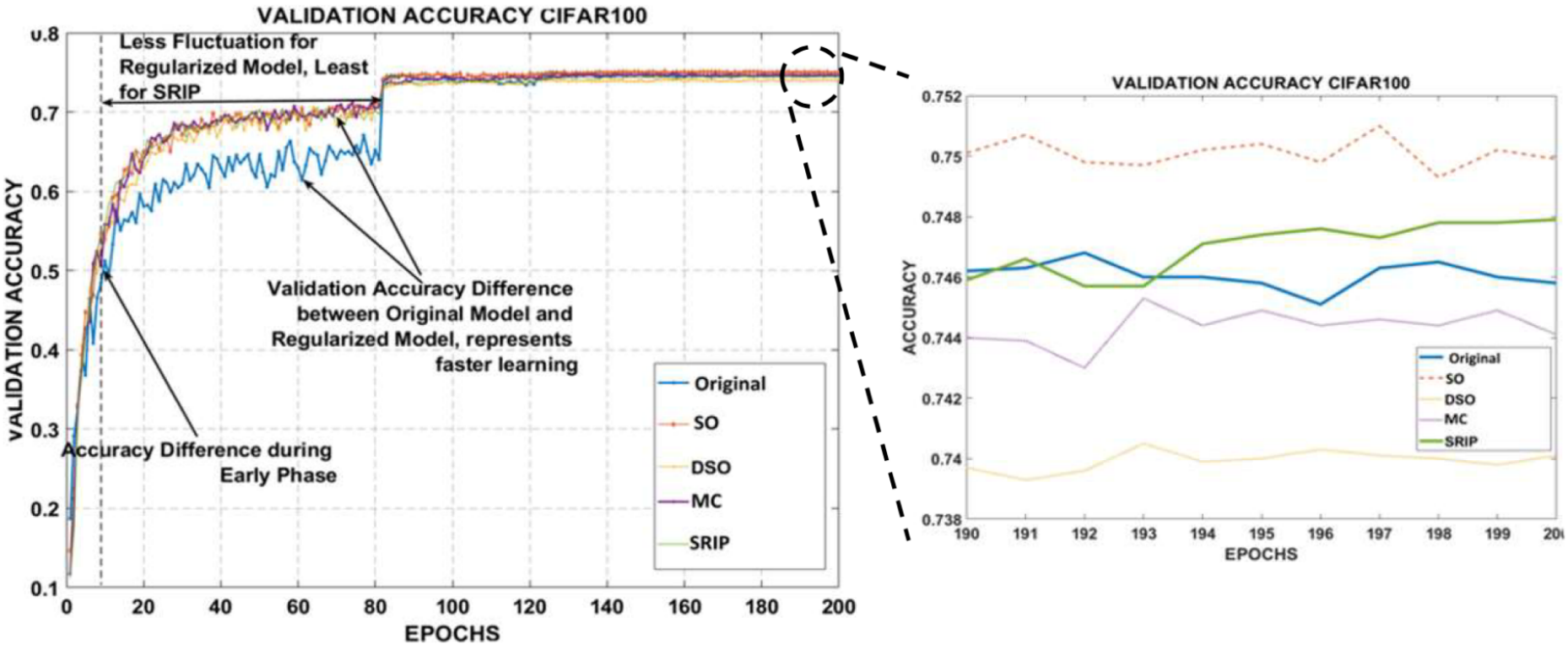


Figure 6: ResNet-110 architectures with decaying orthogonal regularization on CIFAR-100 achieve faster convergence rate. From [4] Bansal *et al.*

IV. Gain-adjusted Orthogonal Regularization with RNNs

In an **RNN of sequence length** t, the input-output Jacobian is simply $\mathbf{J} = W^t \prod_{i=1}^{t} D_i$, with W the recurrent kernel matrix. This makes RNN very sensitive to vanishing gradients

$\rightarrow$ importance of a small gain σ_w slightly higher than 1.0, to be above the "*edge of chaos*" which would fight against *contractive* nature of D_i

Gain-adjusted Orthogonal Regularization

We argue that if W is initialized as $\sigma_w V$ with $VV^T = \mathrm{I}$, then the gain will be lost during training with a standard orthogonal regularization. Hence, we propose a "**Gain-adjusted Orthogonal Regularization**":

$$\lambda \left\| \frac{1}{\sigma_w^2} WW^T - \mathrm{I} \right\|_F^2 \tag{10}$$

This way, we can maintain that $VV^T = \frac{1}{\sigma_w^2} WW^T \approx \mathrm{I}$, *i.e.* V orthogonal and W with a norm of σ_w^2.

IV. Gain-adjusted Orthogonal Regularization with RNNs

Experiment: Apply *gain-adjusted* vs *single-soft* regularizer on sequential MNIST (by row). **Result**: gain-adjusted outperforms single-soft.

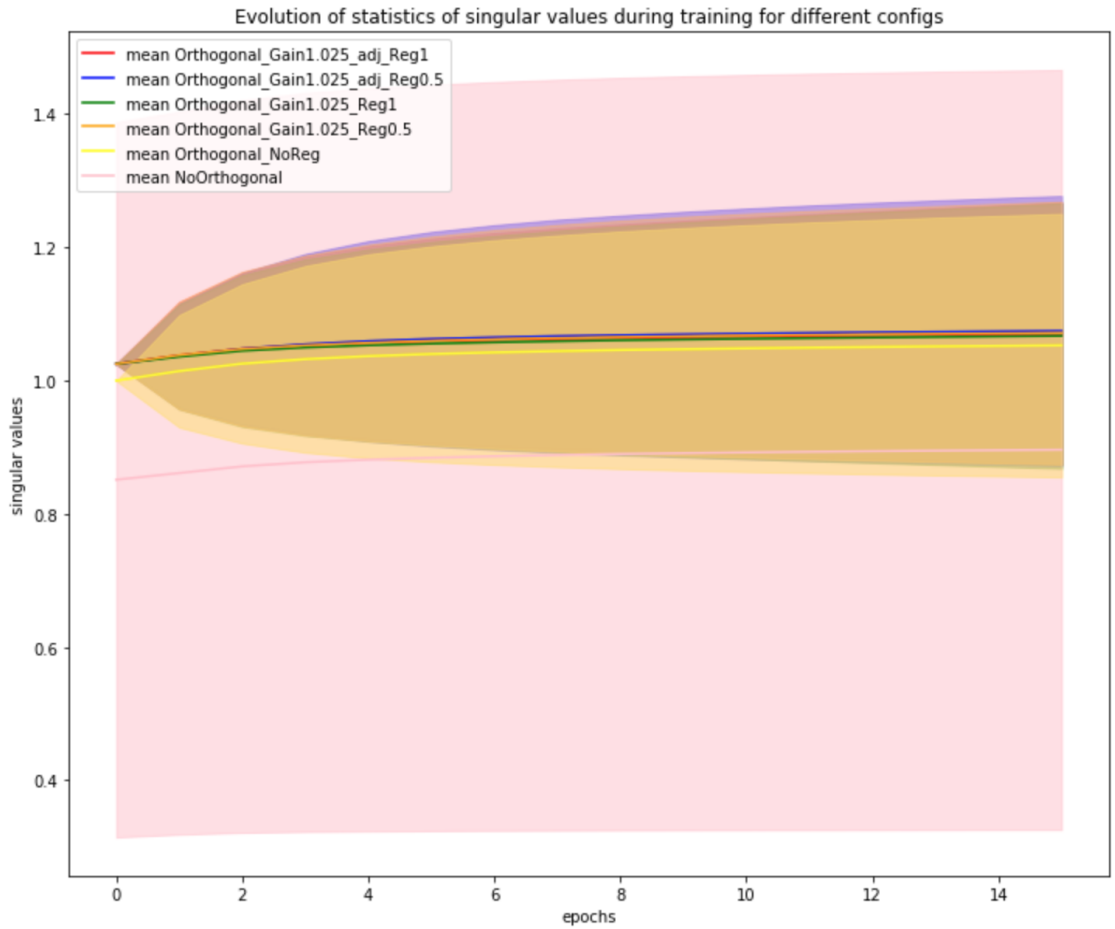


Figure 7: Eigenvalue spectrum of W during training (area is mean $\pm$ std)

IV. Gain-adjusted Orthogonal Regularization with RNNs

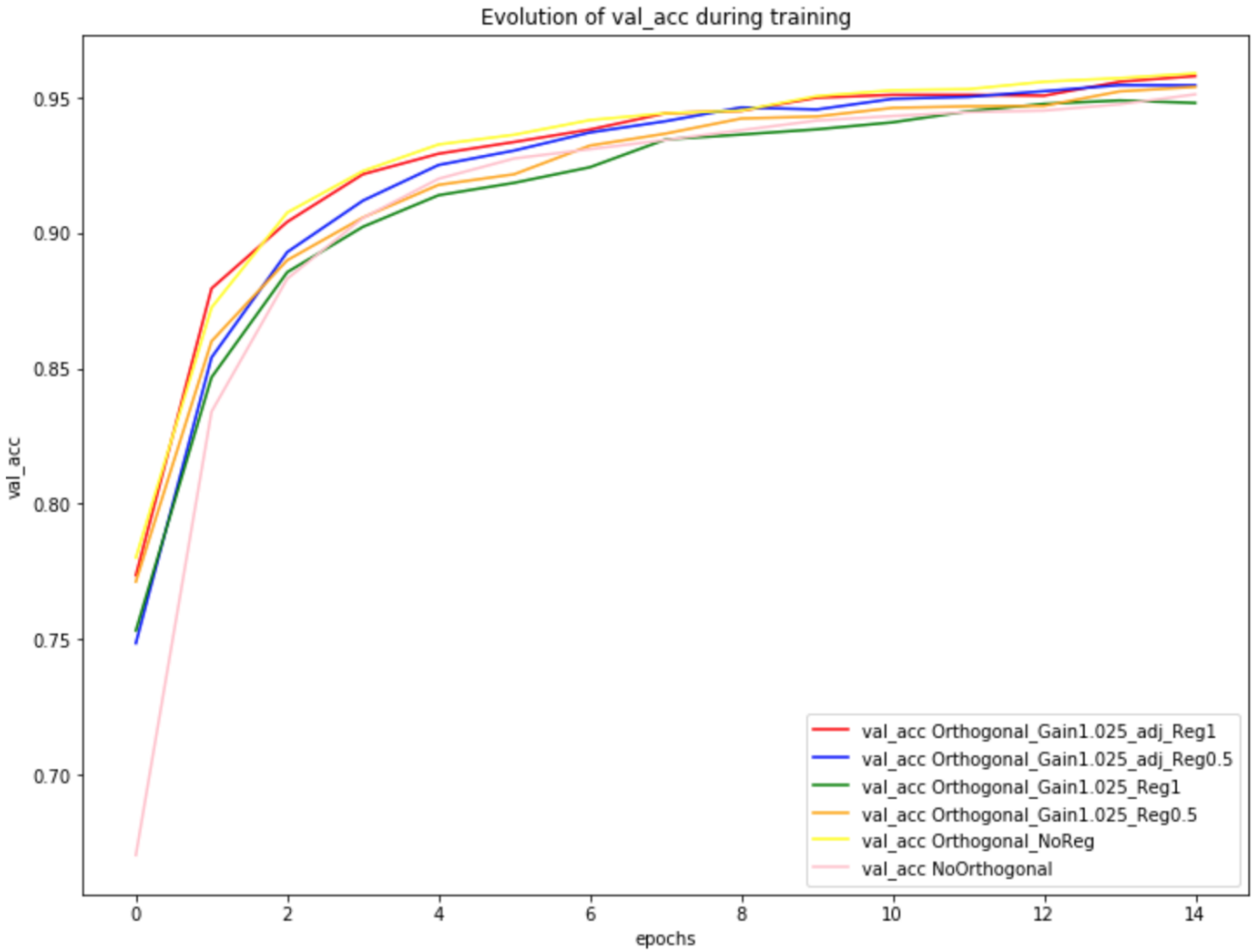


Figure 8: Evolution of validation accuracy for different configurations

References

[1] https://arxiv.org/pdf/1312.6120.pdf A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2013.

[2] Pennington, J, Schoenholz, S, and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In Advances in neural information processing systems, 2017.

[3] Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In Proc. ICML, 2017.

[4] N. Bansal, X. Chen, and Z. Wang. Can We Gain More from Orthogonality Regularizations in Training Deep CNNs? arXiv preprint arXiv:1810.09102, 2018.

References

[5] Helfrich, K., Willmott, D., and Ye, Q. Orthogonal recurrent neural networks with scaled Cayley Transform. ArXiv e-prints, 2018.

[6] Lei Huang, Xianglong Liu, Bo Lang, Admas Wei Yu, and Bo Li. Orthogonal Weight Normalization: Solution to Optimization over Multiple Dependent Stiefel Manifolds in Deep Neural Networks. CoRR, abs/1709.06079, 2018.