Fine-grained Optimization of Deep Neural Networks

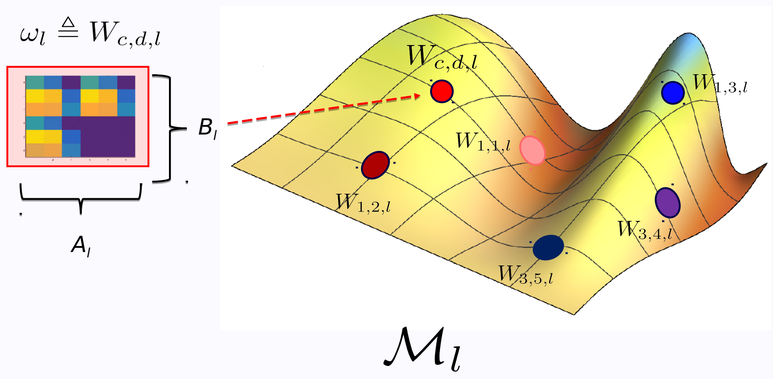
Mete Ozay
meteozay@gmail.com

Problems

We conjecture that if we can impose multiple constraints on weights of DNNs to set upper bounds of the norms of the weight matrices, and train the DNNs with these weights, then the DNNs can achieve empirical generalization errors closer to the proposed theoretical bounds. We pose two problems in order to achieve this goal;

1. Renormalization of weights to upper bound norms of their matrices.
2. Training DNNs with renormalized weights with assurance to convergence to minima.

Weight Manifolds in DNNs



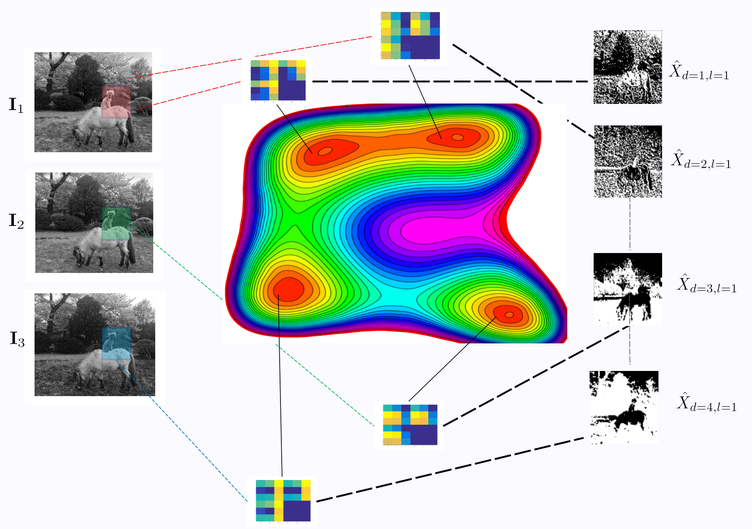
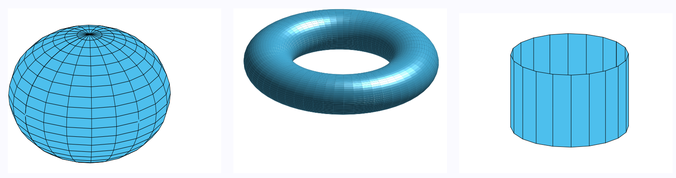


Table 1: Weights $\omega^i_{g,l} \in \mathbb{R}^{A_l \times B_l}$ belonging to the g^{th} group of size $|\mathfrak{g}|, g = 1, 2, \ldots, G_l, \forall l$ have the same size $A_l \times B_l$ for simplicity, and $\sigma(\omega^i_{g,l})$ denotes the top singular value of $\omega^i_{g,l}$. $\|\omega^i_{g,l}\|_F$, $\|\omega^i_{g,l}\|_2$, and $\|\omega^i_{g,l}\|_{2\to1}$, denotes respectively the Frobenius, spectral and $\ell_{2\to1}$ norms of the weight $\omega^i_{g,l}$.

Norms	Sphere	Stiefel	Oblique
$\|\omega^i_{g,l}\|_2$	$\sigma(\omega^i_{g,l})$	1.0	$\sigma(\omega^i_{g,l})$
$\|\omega^i_{g,l}\|_F$	1.0	$(B_l)^{1/2}$	$(B_l)^{1/2}$
$\|\omega^i_{g,l}\|_{2\to1}$	1.0	$(B_l)^{1/4}$	$(B_l)^{1/4}$

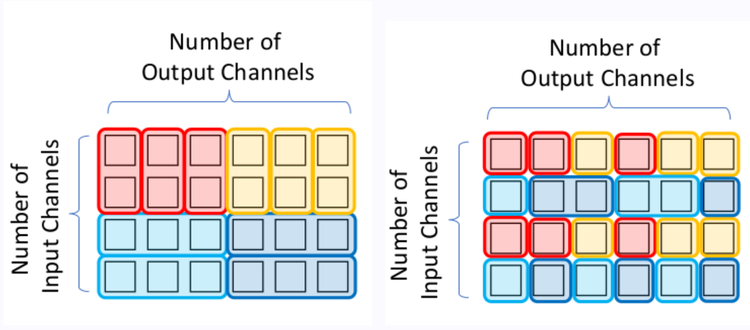


(a) An orthonormalized weight $\omega \in \mathbb{R}^{3\times1}$ ($\omega \in \mathbb{R}^{A\times B}$) resides on a two-sphere $\mathbb{S}^2$ ($\mathbb{S}^{AB-1}$) which has constant positive sectional curvature, 1. (b) A weight $\omega = (\omega_1, \omega_2)$, where each $\omega_\iota \in \mathbb{R}^{2\times1}, \iota = 1, 2$, belongs to a circle $\mathbb{S}^1$, resides on a two-torus $\mathbb{T}^2$ with varying curvature. (c) If $\omega_1 \in \mathbb{S}^1$ ($\omega_1 \in \mathbb{S}^p$) and $\omega_2 \in \mathbb{R}$ ($\omega_2 \in \mathbb{R}^{q-p}$), then ω resides on a cylinder $\mathbb{S}^1 \times \mathbb{R}$ with varying curvature.

Products of Weight Manifolds

Definition 1 *$\mathcal{G}_l = \{\mathcal{M}_{\iota,l} : \iota \in \mathcal{I}_{\mathcal{G}_l}\}$ is a set of **weight manifolds** $\mathcal{M}_{\iota,l}$ of dimension $n_{\iota,l}$, which is identified by a set of indices $\mathcal{I}_{\mathcal{G}_l}, \forall l = 1, 2, \ldots, L$.*
$\mathcal{I}_{\mathcal{G}_l}$ contains indices each of which represents an identity number (ι) of a weight that resides on a manifold $\mathcal{M}_{\iota,l}$ at the l^{th} layer.
*A subset $\mathcal{I}^g_l \subseteq \mathcal{I}_{\mathcal{G}_l}, g = 1, 2, \ldots, G_l$, is used to determine a subset $\mathcal{G}^g_l \subseteq \mathcal{G}_l$ of weight manifolds which will be aggregated to construct a **product of weight manifolds (POM)**.*
*Each $\mathcal{M}_{\iota,l} \in \mathcal{G}^g_l$ is called a **component manifold** of a product of weight manifolds which is denoted by $\mathbb{M}_{g,l}$. A weight $\omega_{g,l} \in \mathbb{M}_{g,l}$ is obtained by concatenating weights belonging to $\mathcal{M}_{\iota,l}$, $\forall \iota \in \mathcal{I}^g_l$, using $\omega_{g,l} = (\omega_1, \omega_2, \cdots, \omega_{|\mathcal{I}^g_l|})$, where $|\mathcal{I}^g_l|$ is the cardinality of $\mathcal{I}^g_l$. A $\mathcal{G}_l$ is called a **collection of POMs**.*

Example 1 *We have a weight tensor of size $3 \times 3 \times 4 \times 6$ where the number of input and output channels is 4 and 6. In total, we have $4 * 6 = 24$ weight matrices of size 3×3.*
***POMs for input and output channels (PIO):** We split the set of 24 weights into 10 subsets. For 6 output channels, we split the set of weights corresponding to 4 input channels into 3 subsets.*
We choose the sphere (Sp) for 2 subsets each containing 3 weights (depicted by light blue rectangles), and 3 subsets each containing 2 weights (depicted by red rectangles).
We choose the Stiefel manifold (St) similarly for the remaining subsets. Then, our ensemble contains 5 POMs of St and 5 POMs of Sp.



***Deterministic and random selection of channels:** In the experiments, indices of the sets are selected randomly using a hypergeometric distribution without replacement at the initialization of a training step, and fixed in the rest of the training.*

Geometry of POMs

Computation of metrics: A metric defined on a product weight manifold $\mathbb{M}_{g,l}$ can be computed by superposition (i.e. linear combination) of Riemannian metrics of its component manifolds.
Lower bounds of sectional curvatures: Sectional curvature of a product weight manifold $\mathbb{M}_{g,l}$ is lower bounded by 0.

Optimization using FG-SGD

input : T (number of iterations),
S (training set),
Θ (set of hyperparameters),
$\mathcal{L}$ (a loss function),
$\mathcal{I}^l_g \subseteq \mathcal{I}_{\mathcal{G}_l}, \forall g, l$.
output: A set of weights $\{\omega^T_{g,l}\}^L_{l=1}, \forall g$.

Initialization: Construct a collection of products of weight manifolds $\mathcal{G}_l$, initialize re-scaling parameters $\mathcal{R}^t_l$ and initialize weights $\omega^t_{g,l} \in \mathbb{M}_{g,l}$ with $\mathcal{I}^l_g \subseteq \mathcal{I}_{\mathcal{G}_l}, \forall m, l$.
for *each iteration* $t = 1, 2, \ldots, T$ **do**
  for *each layer* $l = 1, 2, \ldots, L$ **do**
    $\mathcal{L}(\omega^t_{g,l}) := \Pi_{\omega^t_{g,l}}\Big(\underset{E}{\mathcal{L}}(\omega^t_{g,l}), \Theta, \mathcal{R}^t_l\Big), \forall \mathcal{G}_l$.
    $v_t := h(\mathcal{L}(\omega^t_{g,l}), r(t, \Theta)), \forall \mathcal{G}_l$.
    $\omega^{t+1}_{g,l} := \phi_{\omega^t_{g,l}}(v_t, \mathcal{R}^t_l), \forall \omega^t_{g,l}, \forall \mathcal{G}_l$.
  end for
end for

Generalization Bounds

Norms of concatenated weights $\omega_{g,l}, \forall g$, are lower bounded by products of norms of component weights $\omega^i_{g,l}, \forall i$. Weights are rescaled at each t^{th} epoch using $\Re^t_{i,l} = \frac{\gamma_{i,l}}{\lambda^t_{i,l}}$, where $\gamma_{i,l}$ is a geometric scaling parameter and $\lambda^t_{i,l}$ is the standard deviation of features input to the i^{th} weight in the g^{th} group $\omega^i_{g,l}, \forall i, g$.

Table 2: **Comparison of generalization bounds.** $\delta_{l,F}$, $\delta_{l,2}$, and $\delta_{l,2\to1}$ denotes upper bounds of the Frobenius norm $\|\omega_l\|_F \leq \delta_{l,F}$, spectral norm $\|\omega_l\|_2 \leq \delta_{l,2}$ and the sum of the Euclidean norms for all rows $\|\omega_l\|_{2\to1} \leq \delta_{l,2\to1}$ ($\ell_{2\to1}$) of weights ω_l at the l^{th} layer of an L layer DNN using N samples.

DNNs (dynamic group scaling)
$\mathcal{O}\Big(\frac{2^L \prod_{l=1}^{L}\prod_{g=1}^{G_l}\delta_{g,l,F}}{\sqrt{N}}\Big)$
$\tilde{\mathcal{O}}\Big(\frac{\prod_{l=1}^{L}\prod_{g=1}^{G_l}\delta_{g,l,2}}{\sqrt{N}}\Big(\sum_{l=1}^{L}\prod_{g=1}^{G_l}\big(\frac{\delta_{g,l,2\to1}}{\delta_{g,l,2}}\big)^{\frac{2}{3}}\Big)^{\frac{3}{2}}\Big)$
$\tilde{\mathcal{O}}\Big(\frac{\prod_{l=1}^{L}\prod_{g=1}^{G_l}\delta_{g,l,2}}{\sqrt{N}}\sqrt{L^2\varpi\sum_{l=1}^{L}\prod_{g=1}^{G_l}\frac{\delta^2_{g,l,F}}{\delta^2_{g,l,2}}}\Big)$

Convergence of FG-SGD

Convergenge to local minima: The loss function of a non-linear DNN converges to a local minimum, and the corresponding gradient converges to zero almost surely (a.s.).
Convergenge to global minima: Loss functions of particular nonlinear DNNs (e.g. pyramidal), converge to a global minimum a.s.

Experimental Analysis

	Imagenet(Resnet-101)	**Imagenet(SENet-Resnet-101)**
Euc.	23.15 ± 0.09	22.38 ± 0.30
PIO (Sp+Ob+St)	22.83 ± 0.06	21.93 ± 0.12
PIO (Sp+Ob+St+Euc.)	22.75 ± 0.02	21.76 ± 0.09