On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications

uOttawa

Ziqiao Wang   Yongyi Mao

ICLR

Motivation and Contribution

Motivations:

- Algorithm & Distribution-dependent bound.
- Does the flatness have impact on the generalization?

Key Contributions:

- We present new information-theoretic generalization bounds for models (e.g., linear and two-layer ReLU neural networks) trained with SGD.
- Experimental study provides some insights on the SGD training of neural networks (e.g., a double descent phenomenon of gradient dispersion).
- We also design a simple regularization scheme, *Gaussian model perturbation* (GMP), which is comparably to the current SOTA.

Problem Formulation

Expected Generalization Error:

- $S = \{Z_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mu$; $\mathcal{W} \subseteq \mathbb{R}^d$; $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- Learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
- $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$; $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$
- $\text{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

SGD updates: $W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t)$ where $g(w, B_t) \triangleq 1/b \sum_{z\in B_t} \nabla_w \ell(w, z)$

Auxiliary Weight Process (only in the analysis):

- $\widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t$, for t>0, $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathrm{I}_d)$
- Let $\widetilde{W}_0 \triangleq W_0$ and $\Delta_t = \sum_{\tau=1}^t N_\tau \Longrightarrow \widetilde{W}_t = W_t + \Delta_t$.

Theoretical Results

Decomposition of the expected generalization error (Neu et al. (2021)):

$$\left|\text{gen}(\mu, P_{W_T|S})\right| = \left|\text{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathbb{E}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathbb{E}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]\right|.$$

Theorem 1 *The generalization error of SGD is upper bounded by*

$$\sqrt{\frac{R^2 d}{n}\sum_{t=1}^T \mathbb{E}\left[\log\left(\frac{\lambda_t^2}{d\sigma_t^2}\mathbb{E}\left[||g(W_{t-1}, B_t) - \mathbb{E}\left[\nabla \ell(W_{t-1}, Z)\right]||^2\right] + 1\right)\right]} + \left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|.$$

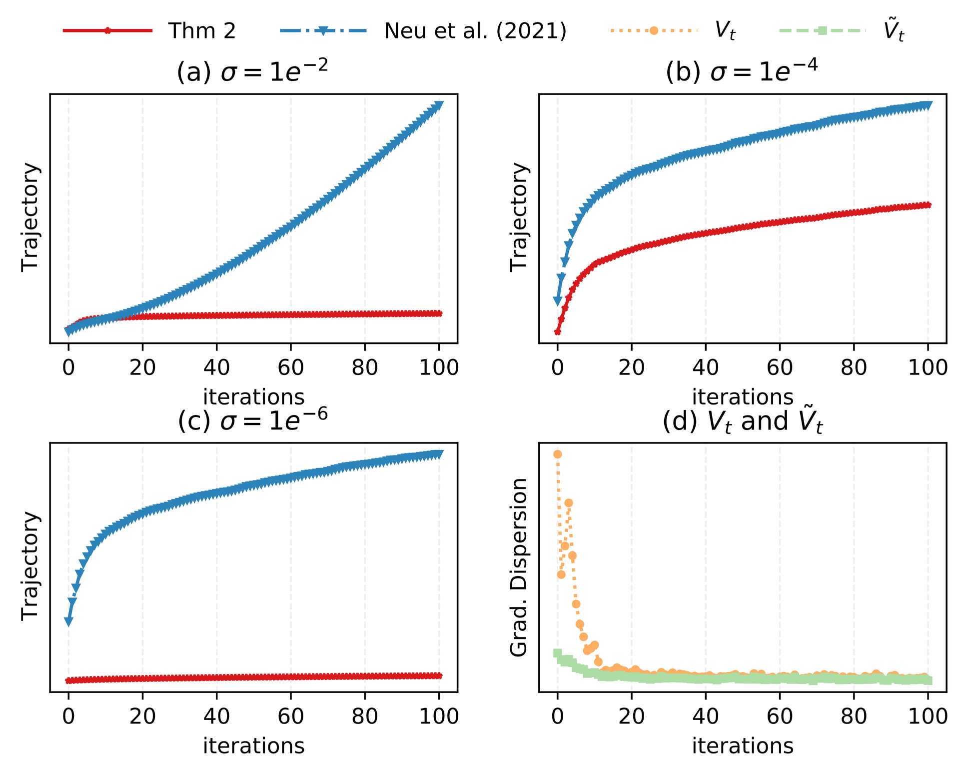
Theorem 2 *Let **gradient dispersion** $\mathbb{V}_t(w) \triangleq \mathbb{E}_S||g(w, B_t) - \mathbb{E}_{W,Z}\nabla_w \ell(W, Z)||_2^2$. Then*

$$\left|\text{gen}(\mu, P_{W_T|S})\right| \le \underbrace{\sqrt{\frac{R^2 d}{n}\sum_{t=1}^T \log\left(\frac{\lambda_t^2}{d\sigma_t^2}\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right] + 1\right)}}_{\textit{trajectory term}} + \underbrace{\left|\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]\right|}_{\textit{flatness term}}. \quad (1)$$

Assume $L_\mu(w_T) \le \mathbb{E}_{\Delta_T}\left[L_\mu(w_T + \Delta_T)\right]$ and σ_t^2 is independent of t. Then the optimal bound:

$$\text{gen}(\mu, P_{W_T|S}) \le \frac{3}{2}\left(\sum_{t=1}^T \frac{R^2\lambda_t^2 T}{n}\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right]\mathbb{E}\left[\text{Tr}\left(\text{H}_{W_T}(Z)\right)\right]\right)^{\frac{1}{3}} \quad (2)$$

Compared with the bound in Neu et al. (2021):



Application: Linear and Two-Layer ReLU Networks

- $Z = (X, Y)$; $\ell(W, Z) = \frac{1}{2}(Y - f(W, X))^2$

Theorem 3 (Linear Networks) *Upper bound:*

$$3\left(\sum_{t=1}^T \frac{R^2\lambda_t^2 T}{4n}\mathbb{E}\left[\ell(W_{t-1}, Z)\right]\right)^{\frac{1}{3}}.$$

Theorem 4 (Two-Layer ReLU Networks) *Upper bound:*

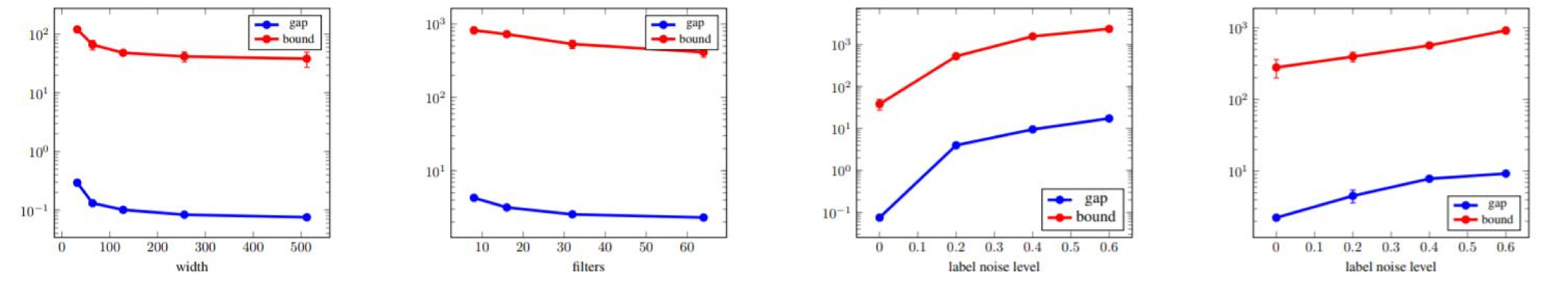
$$3\left(\sum_{r=1}^m \mathbb{E}\left[\frac{\mathbb{I}_{r,i,T}}{m}\right]\sum_{t=1}^T \frac{R^2\lambda_t^2 T}{4n}\mathbb{E}\left[\sum_{r=1}^m \frac{\mathbb{I}_{r,i,t}}{m}\ell(W_{t-1}, Z)\right]\right)^{\frac{1}{3}},$$

where $\mathbb{I}_{r,i,t} = \mathbb{I}\{W_{t-1,r}^T X_i \ge 0\}$.

Sparsely activated ReLU networks are expected to generalize better.

Experimental Results

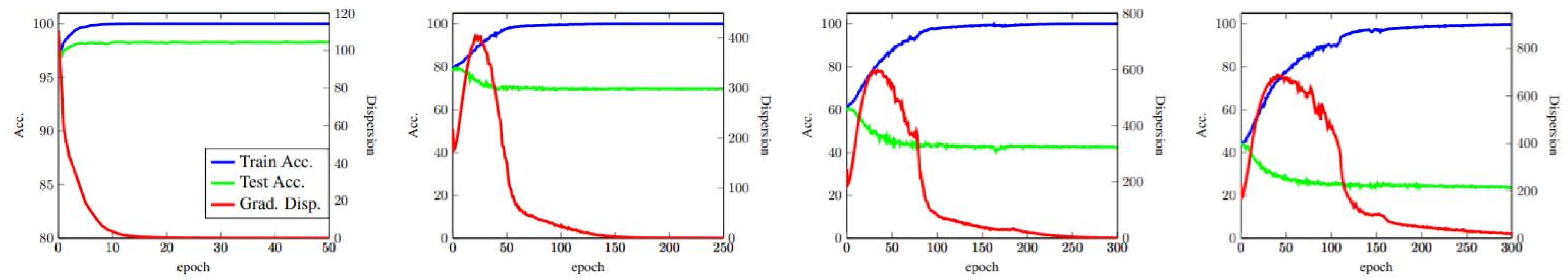
Bound Verification of Thm 2: Estimated bound and empirical generalization gap ("gap") as functions of network width ((a) and (b)) and label noise level ((c) and (d)).



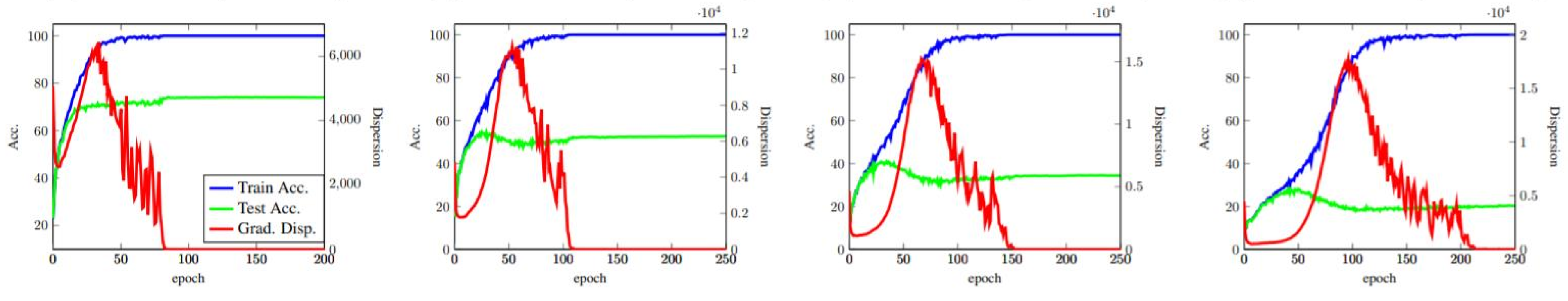
(a) MLP on MNIST (b) AlexNet on CIFAR10 (c) MLP on MNIST (d) AlexNet on CIFAR10

Epoch-wise Double Descent of Gradient Dispersion:

- $\mathbb{V}$ rapidly descends; Both training acc. and test acc. increase; $\Longrightarrow$ "Generalization"
- $\mathbb{V}$ starts increasing until it reaches a peak value; Train acc. and Test acc. diverge; $\Longrightarrow$ "Memorization"
- $\mathbb{V}$ descends again; Training and testing curves reach their respective maximum and minimum.



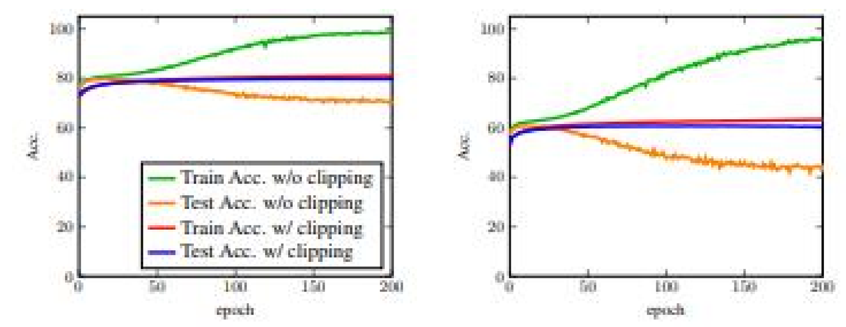
(a) noise=0 (MNIST) (b) noise=0.2 (MNIST) (c) noise=0.4 (MNIST) (d) noise=0.6 (MNIST)



(e) noise=0 (CIFAR10) (f) noise=0.2 (CIFAR10) (g) noise=0.4 (CIFAR10) (h) noise=0.6 (CIFAR10)

Implication: Dynamic Gradient Clipping

- Check if $||g(W_t, B_t)||_2 > ||g(W_{t-K}, B_{t-K})||_2$ (i.e., the model is expected to have entered the "memorization" regime)
- If so, reduce the norm of the current gradient $g(W_t, B_t)$ to α fraction of $||g(W_{t-K}, B_{t-K})||_2$
- Effectiveness is best demonstrated when labels contain noise.



(a) noise=0.2 (MNIST) (b) noise=0.4 (MNIST)

Implication: GMP

We hope the empirical risk surface at w^* is flat,

$$\min_w \frac{1}{b}\sum_{z\in B}\left((1-\rho)\ell(w, z) + \rho\frac{1}{k}\sum_{i=1}^k\left(\ell(w + \delta_i, z)\right)\right).$$

Algorithm 2 Gaussian Model Perturbation Training

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Number of noise k, Standard deviation of Gaussian distribution σ, Lagrange multiplier ρ

while $\boldsymbol{w}_t$ not converged **do**

  Update iteration: $t \leftarrow t + 1$

  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^b$ from training set S

  Sample $\Delta_j \sim \mathcal{N}(0, \sigma^2)$ for $j \in [k]$

  Compute gradient:

  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^b\left(\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}_t, z_i) + \rho\sum_{j=1}^k\left(\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}_t + \Delta_j, z_i) - \nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}_t, z_i)\right)/k\right)/b$

  Update parameter: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \lambda \cdot g_{\mathcal{B}}$

end while

Top-1 classification acc.(%) of VGG16

Method	SVHN	CIFAR-10	CIFAR-100
ERM	96.86±0.060	93.68±0.193	72.16±0.297
Dropout	97.04±0.049	93.78±0.147	72.28±0.337
L. S.	96.93±0.070	93.71±0.158	72.51±0.179
Flooding	96.85±0.085	93.74±0.145	72.07±0.271
MixUp	96.91±0.057	**94.52±0.112**	73.19±0.254
Adv. Tr.	97.06±0.091	93.51±0.130	70.88±0.145
AMP	**97.27±0.015**	94.35±0.147	74.40±0.168
GMP^3	97.18±0.057	94.33±0.094	74.45±0.256
GMP^{10}	97.09±0.068	94.45±0.158	**75.09±0.285**