On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications
Ziqiao Wang  Yongyi Mao

Motivations and Contribution

Motivations:
- Algorithm & Distribution-dependent bound.
- Does the fineness 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 regularisation scheme, Gaussian model perturbation (GMP), which is comparable to the current SOTA.

Theoretical Results

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

\[ \mathbb{E} [\mathbb{L}(W_{t+1}) - \mathbb{L}(W_t)] = \mathbb{E} \left[ \sum_{i=1}^{d} \frac{\lambda_i^2}{n} \right] + \mathbb{E} [\mathbb{L}(W_T) - \mathbb{L}(W_t)]. \]

\[ \mathbb{E} [\mathbb{L}(W_T) - \mathbb{L}(W_t)] \leq \frac{R^2}{n} \mathbb{E} \left[ \sum_{i=1}^{d} \frac{\lambda_i^2}{n} \right] \leq \frac{R^2}{n} \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i^2 \right]. \]

\[ \mathbb{E} [\mathbb{L}(W_{t+1}) - \mathbb{L}(W_t)] \leq \frac{R^2}{n} \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i^2 \right] + \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i^2 \right]. \]

Application: Linear and Two-Layer ReLU Networks

\[ Z = (X, Y); \ell(W, Z) = \frac{1}{2} \sum_{i=1}^{d} \lambda_i^2 \]

\[ \mathbb{E} [\mathbb{L}(W_{t+1}) - \mathbb{L}(W_t)] \leq \frac{R^2}{n} \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i^2 \right] + \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i^2 \right]. \]

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).

Epoch-wise Double Descent of Gradient Dispersion:
- \( V \) rapidly descends; Both training acc. and test acc. increase; \( \implies \) “Generalization”
- \( V \) starts increasing until it reaches a peak value; Train acc. and Test acc. diverge; \( \implies \) “Memorization”
- \( V \) descends again; Training and testing curves reach their respective maximum.

Implication: Dynamic Gradient Clipping

- \( \text{Check if } ||g(W_t, B_t)||_2 > ||g(W_{t-K}, B_{t-K})||_2 \) (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 \( \alpha \) fraction of \( ||g(W_t, B_t)||_2 \)
- Effectiveness is best demonstrated when labels contain noise.

Implication: GMP

We hope the empirical risk surface at \( u^* \) is flat,

\[ \min_{w} \frac{1}{n} \sum_{i=1}^{d} (1 - \rho) \ell(w, z) + \rho \sum_{i=1}^{d} (1 - \rho) \ell(w + \delta_i, z). \]

Algorithm 2: Gaussian Model Perturbation Training

Require: Training set \( S \), Batch size \( b \), Loss function \( \ell \), Initial model parameter \( w_0 \), Learning rate \( \alpha \), Number of noise \( n \), Standard deviation of Gaussian distribution \( \sigma \), Lagrange multiplier \( \rho \) while \( \alpha \), not converged do

1. Update \( \alpha \) in \( \alpha \) direction
2. Sample \( B = \{b_i\} \), \( b_i \) bernoulli random
3. Sample \( Z \sim \mathcal{N}(0, \sigma^2) \)
4. Composite \( \mathcal{X} = \mathcal{X} + \mathcal{Z} \)
5. \( g = \sum_{i=1}^{d} \ell(w + \delta_i, z) \)
6. Update parameter: \( w_{i+1} = w_i - \lambda g \)

Top-1 classification acc. (\%) of VGG16

<table>
<thead>
<tr>
<th>Method</th>
<th>SVHN</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>96.80±0.060</td>
<td>93.68±0.193</td>
<td>72.16±0.297</td>
</tr>
<tr>
<td>Dropout</td>
<td>97.04±0.049</td>
<td>93.78±0.147</td>
<td>72.28±0.337</td>
</tr>
<tr>
<td>L. S.</td>
<td>96.91±0.070</td>
<td>93.71±0.185</td>
<td>72.51±0.179</td>
</tr>
<tr>
<td>Flooding</td>
<td>96.85±0.085</td>
<td>93.74±0.145</td>
<td>72.07±0.271</td>
</tr>
<tr>
<td>MixUp</td>
<td>96.91±0.057</td>
<td>94.52±0.112</td>
<td>73.19±0.254</td>
</tr>
<tr>
<td>Adv. Tr.</td>
<td>97.08±0.091</td>
<td>93.51±0.130</td>
<td>70.88±0.145</td>
</tr>
<tr>
<td>AMP</td>
<td>97.27±0.105</td>
<td>94.35±0.147</td>
<td>74.40±0.168</td>
</tr>
<tr>
<td>GMP</td>
<td>97.18±0.057</td>
<td>94.33±0.094</td>
<td>74.45±0.256</td>
</tr>
<tr>
<td>GMP10</td>
<td>97.09±0.068</td>
<td>94.45±0.158</td>
<td>75.09±0.285</td>
</tr>
</tbody>
</table>