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

Ziqiao Wang[1] Yongyi Mao[1]

[1]University of Ottawa

Motivation

- Generalization measures (e.g., VC-dim and Rademacher complexity) in classical statistical learning theory cannot explain the success of modern deep neural networks.

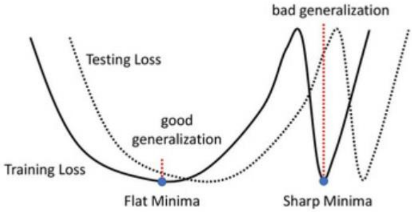
Motivation

- Generalization measures (e.g., VC-dim and Rademacher complexity) in classical statistical learning theory cannot explain the success of modern deep neural networks.
  # of parameters > # of training data & can even perfectly fit random labels
  $\Longrightarrow$ high capacity
  $\Longrightarrow$ still perform well on unseen data

- Algorithm & Distribution-dependent $\Longrightarrow$ non-vacuous generalization bound

Motivation

- ▷ Generalization measures (e.g., VC-dim and Rademacher complexity) in classical statistical learning theory cannot explain the success of modern deep neural networks.
  # of parameters > # of training data & can even perfectly fit random labels
  $\Longrightarrow$ high capacity
  $\Longrightarrow$ still perform well on unseen data
- ▷ Algorithm & Distribution-dependent $\Longrightarrow$ non-vacuous generalization bound
- ▷ Implicit bias of SGD, e.g., does the flatness have impact on the generalization?



Problem Setup

- Training dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}$, drawn i.i.d. from μ
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ by $P_{W|S}$
- Loss: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- We're interested in
  - Population risk: $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$
  - Empirical risk: $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$
  - Expected generalization error: $\mathrm{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

Lemma 1 (Thm 1., Xu&Raginsky'2017)

Assume the loss $\ell(w,Z)$ is R-subgaussian[a] for any $w \in \mathcal{W}$. The generalization error of $\mathcal{A}$ is bounded by

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2R^2}{n} I(W;S)},$$

[a] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E} \exp\left(\rho\left(X - \mathbb{E}X\right)\right) \leq \rho^2 R^2/2$.

Mutual information $I(W;S) \triangleq \mathrm{D}_{\mathrm{KL}}(P_{W,S}||P_W \otimes P_S)$.

$\Longrightarrow$ Distribution-dependent and Algorithm-dependent

Proof Technique

Lemma 2 (Donsker and Varadhan's variational formula)

For any bounded measurable function $f : \Theta \to \mathbb{R}$, *we have*

$$\mathrm{D_{KL}}(Q||P) = \sup_f \mathop{\mathbb{E}}_{\theta\sim Q}[f(\theta)] - \log \mathop{\mathbb{E}}_{\theta\sim P}[\exp f(\theta)].$$

Proof sketch of Lemma 1.

$$\begin{aligned} \mathop{\mathbb{E}}_{S,W}[L_\mu(W) - L_S(W)] &= \mathop{\mathbb{E}}_{S,W}\left[\mathop{\mathbb{E}}_{S'}[L_{S'}(W)]\right] - \mathop{\mathbb{E}}_{S,W}[L_S(W)] \\ &= \mathop{\mathbb{E}}_{P_W\otimes P_{S'}}[L_{S'}(W)] - \mathop{\mathbb{E}}_{P_{W,S}}[L_S(W)] \end{aligned}$$

Then,

$$\begin{aligned} I(W,S) &= \mathrm{D_{KL}}(P_{W,S}||P_W\otimes P_{S'}) \\ &\geq \sup_f \mathop{\mathbb{E}}_{(W,S)\sim P_{W,S}}[f(W,S)] - \log \mathop{\mathbb{E}}_{(W,S')\sim P_W\otimes P_{S'}}[\exp f(W,S')] \end{aligned}$$

Let $f(W,S) = t\cdot L_S(W)$. Recall the sub-Gaussian assumption, Lemma 1 can be obtained.

Some improved IT bounds:

- $\frac{1}{n}\sum_{i=1}^{n}\sqrt{C_1 I(W;Z_i)}$
  Bu, Y., Zou, S. and Veeravalli, V.V.. Tightening Mutual Information Based Bounds on Generalization Error. ISIT 2019.

- $\mathbb{E}\sqrt{\frac{C_2}{n-m} I^{S_J,V}(W;S_J^C)}$
  Negrea, J., Haghifam, M., Dziugaite, G.K., Khisti, A. and Roy, D.M.. Information-theoretic generalization bounds for SGLD via data-dependent estimates. NeurIPS 2019.

- $\sqrt{\frac{C_3}{n} I(W;U|\tilde{Z})}$
  Steinke, T. and Zakynthinou, L.. Reasoning about generalization via conditional mutual information. COLT 2020.

- ......

- $\sqrt{C_4 I(L;U)}$
  Haghifam, M., Moran, S., Roy, D.M. and Karolina Dziugaite, G., 2022. Understanding Generalization via Leave-One-Out Conditional Mutual Information. ISIT 2022.

Stochastic Gradient Descent (SGD)

SGD updates:

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t: learning rate
- b: batch size
- B_t denotes the batch used for the t^{th} update.

Assume SGD outputs W_T as the learned model parameter.

Stochastic Gradient Descent (SGD)

SGD updates:

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t: learning rate
- b: batch size
- B_t denotes the batch used for the t^{th} update.

Assume SGD outputs W_T as the learned model parameter.

Difficulty of using Lemma 1: $I(W_T; S) \to \infty$ in some cases
Lemma 1 is usually applied to analyze SGLD.
Pensia, A., Jog, V. and Loh, P.L.. Generalization error bounds for noisy, iterative algorithms. ISIT 2018.

Auxiliary Weight Process (only exists in the analysis)

Follow up the work of Neu et al. (2021), let $\sigma_1, \sigma_2, \ldots, \sigma_T$ be a sequence of positive real numbers.

Define

$$\widetilde{W}_0 \triangleq W_0, \quad \text{and} \;\; \widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t, \;\; \text{for } t > 0,$$

where $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_d)$ is a Gaussian noise.

$$\begin{array}{ccccccccccc}
 & & N_1 & & N_2 & & \cdots & & N_{T-1} & & N_T \\
 & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\widetilde{W}_0 & \rightarrow & \widetilde{W}_1 & \rightarrow & \widetilde{W}_2 & \rightarrow & \cdots & \rightarrow & \widetilde{W}_{T-1} & \rightarrow & \widetilde{W}_T \\
\| & \nearrow & & \nearrow & & \nearrow & & & & \nearrow & \\
W_0 & \rightarrow & W_1 & \rightarrow & W_2 & \rightarrow & \cdots & \rightarrow & W_{T-1} & \rightarrow & W_T
\end{array}$$

Let $\Delta_t = \sum_{\tau=1}^{t} N_\tau$. Notice that $\widetilde{W}_t = W_t + \Delta_t$.

Lemma 1 for noisy, iterative algorithm

- Learning algorithm $\widetilde{A}$ takes S as input and outputs $\widetilde{W}$
- Decomposition of the expected generalization gap:

$$
\begin{aligned}
&|\mathrm{gen}(\mu, P_{W_T|S})| \\
=&|\mathrm{gen}(\mu, P_{W_T|S}) + \mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) - \mathrm{gen}(\mu, P_{\widetilde{W}_T|S})| \\
=&\left|\mathbb{E}_{W,S,\Delta}[L_\mu(W_T) - L_S(W_T) + L_\mu(\widetilde{W}_T) - L_S(\widetilde{W}_T) - L_\mu(\widetilde{W}_T) + L_S(\widetilde{W}_T)]\right| \\
=&\left|\mathrm{gen}(\mu, P_{\widetilde{W}_T|S}) + \mathop{\mathbb{E}}_{W_T,\Delta_T}\left[L_\mu(W_T) - L_\mu(\widetilde{W}_T)\right] + \mathop{\mathbb{E}}_{W_T,\Delta_T,S}\left[L_S(\widetilde{W}_T) - L_S(W_T)\right]\right|.
\end{aligned}
$$

$$\Longrightarrow |\mathrm{gen}(\mu, P_{\widetilde{W}_T|S})| \leq \sqrt{\tfrac{2R^2}{n} I(\widetilde{W}_T; S)} < \infty$$

Information-theoretic bound for SGD

Lemma 3 (Thm.1, Neu et al'2021)

The generalization error of SGD is upper bounded by

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \sqrt{\frac{4R^2}{n}\sum_{t=1}^{T}\frac{\lambda_t^2}{\sigma_t^2}\mathbb{E}\left[\Psi(W_{t-1}) + \widetilde{\mathbb{V}}_t(W_{t-1})\right]} + |\mathbb{E}\left[\gamma(W_T, S) - \gamma(W_T, S')\right]|$$

- Local gradient sensitivity:
  $\Psi(w_{t-1}) \triangleq \mathbb{E}_\zeta\left[||\mathbb{E}_Z\left[\nabla_w \ell(w_{t-1}, Z)\right] - \mathbb{E}_Z\left[\nabla_w \ell(w_{t-1} + \zeta, Z)\right]||_2^2\right]$,
  $\zeta \sim \mathcal{N}(0, \sum_{i=1}^{t-1} \sigma_i^2 \mathrm{I}_d)$
- Gradient Dispersion: $\widetilde{\mathbb{V}}_t(w) \triangleq \mathbb{E}_S\left[||g(w, B_t) - \mathbb{E}_Z\left[\nabla_w \ell(w, Z)\right]||_2^2\right]$
- Local value sensitivity: $\gamma(w, s) \triangleq \mathbb{E}_{\Delta_T}\left[L_s(w + \Delta_T) - L_s(w)\right]$

Main Result: Closed Form of Optimal Bound

Let $\mathbb{V}_t(w) \triangleq \mathbb{E}_S\left[||g(w,B_t) - \mathbb{E}_{W,Z}\left[\nabla_w \ell(W,Z)\right]||_2^2\right]$.

Theorem 1

The generalization error of SGD is upper bounded by

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

Assume $L_\mu(w_T) \leq \mathbb{E}_\Delta\left[L_\mu(w_T+\Delta_T)\right]$ *and* σ_t^2 *is independent of* t. *Then*

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

Proof Sketch of Theorem 1 I

Recall that

$$|\mathrm{gen}(\mu, P_{W_T|S})| \leq \sqrt{\frac{2R^2}{n} I(\widetilde{W}_T; S)} + \left| \mathop{\mathbb{E}}_{W_T,S,S'} \left[\gamma(W_T, S) - \gamma(W_T, S')\right] \right|.$$

Notice that

$$\begin{aligned}
I(\widetilde{W}_T; S) =& I\left(\widetilde{W}_{T-1} - \lambda_T g(W_{T-1}, B_T) + N_T; S\right) \\
\leq& I\left(\widetilde{W}_{T-1}, -\lambda_T g(W_{T-1}, B_T) + N_T; S\right) && (3) \\
=& I(\widetilde{W}_{T-1}; S) + I\left(-\lambda_T g(W_{T-1}, B_T) + N_T; S|\widetilde{W}_{T-1}\right) && (4) \\
&\vdots \\
\leq& \sum_{t=1}^{T} I\left(-\lambda_t g(W_{t-1}, B_t) + N_t; S|\widetilde{W}_{t-1}\right), && (5)
\end{aligned}$$

Proof Sketch of Theorem 1 II

Lemma 4

Let X,Y and Δ be random variables which are all independent of $N \sim \mathcal{N}(0, \mathrm{I})$. Let $Z = Y + \Delta$, then for any σ and any function f, we have

$$I(f(Z,X) + \sigma N; X|Y) \leq \frac{1}{2\sigma^2}\mathbb{E}\left[||f(Z,X) - \mathbb{E}\left[f(Z,X)\right]||^2\right]$$

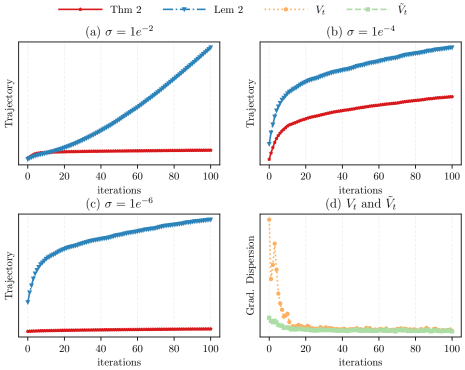
Thus,

$$\begin{aligned} I(-\lambda_t g(W_{t-1}, B_t) + \sigma_t N; S|\widetilde{W}_{t-1}) \leq & \frac{\lambda_t^2}{2\sigma_t^2}\mathbb{E}\left[||g(W_{t-1}, B_t) - \mathbb{E}\left[\nabla_w \ell(W_{t-1}, Z)\right]||^2\right] \\ = & \frac{\lambda_t^2}{2\sigma_t^2}\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right] \end{aligned}$$

Putting everything together we have the bound in Theorem 1.

In Eq. 1,

- ▷ The first term: "trajectory term"; The second term: "flatness term"
- ▷ Compared with the bound in Lemma 3:



Notice that $\Psi(W_{t-1})$ has the cumulative variance $\sum_{i=1}^{t-1} \sigma_i^2 \mathrm{I}_d$, and the gap between $\mathbb{V}_t$ and $\widetilde{\mathbb{V}}_t$ is small when W is close to local minima.

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

- Condition $L_\mu(w_T) \leq \mathbb{E}_{\Delta_T}\left[L_\mu(w_T+\Delta_T)\right] \implies$ the perturbation does not decrease the population risk.
  Also assumed in [Foret, et al.'2021] in the derivation of a PAC-Bayesian bound.

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

- Condition $L_\mu(w_T) \leq \mathbb{E}_{\Delta_T}\left[L_\mu(w_T + \Delta_T)\right] \Longrightarrow$ the perturbation does not decrease the population risk.
  Also assumed in [Foret, et al.'2021] in the derivation of a PAC-Bayesian bound.

- Eq.2 follows from Eq.1 by minimizing the bound over σ.
  Eq.2 can be computed easily and efficiently.

Application: Linear and Two-Layer ReLU Networks

▷ $Z=(X,Y)$; $X\in\mathbb{R}^{d_0}$; $||X||=1$; $f(W,\cdot):\mathbb{R}^{d_0}\to\mathbb{R}$; $\ell(W,Z)=\frac{1}{2}(Y-f(W,X))^2$

Theorem 2 (Linear Networks)

Let $f(W,X)=W^TX$. *Then,* $\mathrm{gen}(\mu,P_{W_T|S})\leq 3\left(\sum_{t=1}^{T}\frac{R^2\lambda_t^2T}{4n}\mathbb{E}\left[\ell(W_{t-1},Z)\right]\right)^{\frac{1}{3}}$.

Theorem 3 (Two-Layer ReLU Networks)

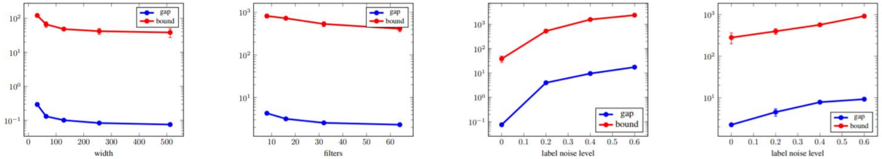
Let $f(W,X)=\frac{1}{\sqrt{m}}\sum_{r=1}^{m}A_r\mathrm{ReLU}(W_r^TX)$ *where* $A_r\sim\mathrm{unif}(\{+1,-1\})$. *We fix the second layer parameters during training. Then,*

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

where $\mathbb{I}_{r,i,t}=\mathbb{I}\{W_{t-1,r}^TX_i\geq 0\}$ *and* $\mathbb{I}$ *is the indicator function.*

$\Longrightarrow$ Sparsely activated ReLU networks are expected to generalize better.

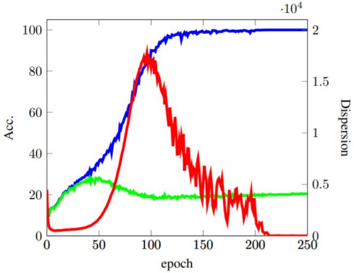
Experiment: Bound Verification of Thm 1



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

Figure 1: Estimated bound and empirical generalization gap ("gap") as functions of network width ((a) and (b)) and label noise level ((c) and (d)). Y-axis is in log scale.

Experiment: 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; Training acc. and testing acc. gradually diverge; $\Longrightarrow$ "Memorization"
- ▷ $\mathbb{V}$ descends again; Training and testing curves reach their respective maximum and minimum.

Implication: Dynamic Gradient Clipping

Algorithm 1 Dynamic Gradient Clipping

Require: Training set S, Batch size b, Loss function ℓ, Initial model parameter $\boldsymbol{w}_0$, Learning rate λ, Initial minimum gradient norm $\mathcal{G}$, Number of iterations T, Clipping parameter α, Clipping step T_c

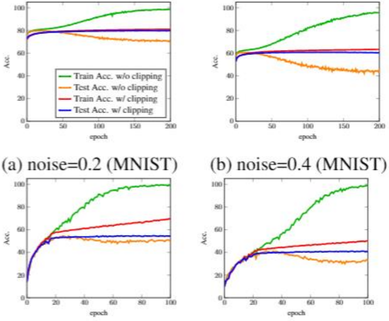
for $t \leftarrow 1$ to T **do**
  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^{b}$ from training set S
  Compute gradient:
  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i)/b$
  if $t > T_c$ **then**
    if $||g_{\mathcal{B}}||_2 > \mathcal{G}$ **then**
      $g_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot g_{\mathcal{B}}/||g_{\mathcal{B}}||_2$
    else
      $\mathcal{G} \leftarrow ||g_{\mathcal{B}}||_2$
    end if
  end if
  Update parameter: $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \lambda \cdot g_{\mathcal{B}}$
end for

Implication: Dynamic Gradient Clipping

Algorithm 1 Dynamic Gradient Clipping

Require: Training set S, Batch size b, Loss function ℓ, Initial λ, Initial minimum gradient norm $\mathcal{G}$, Number of iterations step T_c

for $t \leftarrow 1$ to T **do**
  Sample $\mathcal{B} = \{z_i\}_{i=1}^{b}$ from training set S
  Compute gradient: $g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i)/b$
  if $t > T_c$ **then**
    if $||g_{\mathcal{B}}||_2 > \mathcal{G}$ **then**
      $g_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot g_{\mathcal{B}}/||g_{\mathcal{B}}||_2$
    else
      $\mathcal{G} \leftarrow ||g_{\mathcal{B}}||_2$
    end if
  end if
  Update parameter: $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \lambda \cdot g_{\mathcal{B}}$
end for



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

(c) noise=0.2 (CIFAR10) (d) noise=0.4 (CIFAR10)

Implication: Gaussian Model Perturbation (GMP)

- We hope the empirical risk surface at w^* is flat, or insensitive to a small perturbation of w^*.

$$\min_{w} L_s(w) + \rho \mathop{\mathbb{E}}_{\Delta \sim \mathcal{N}(0,\sigma^2 \mathbf{I}_d)} [L_s(w+\Delta) - L_s(w)],$$

where ρ is a hyper-parameter.

- Replacing the expectation above with its stochastic approximation using k realizations of Δ gives rise to the following optimization problem.

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

Implication: GMP

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 w_t not converged do
    Update iteration: t ← t + 1
    Sample B = {z_i}_{i=1}^b from training set S
    Sample Δ_j ~ N(0, σ^2) for j ∈ [k]
    Compute gradient:
    g_B ← Σ_{i=1}^b ( ∇_w ℓ(w_t, z_i) + ρ Σ_{j=1}^k (∇_w ℓ(w_t + Δ_j, z_i) − ∇_w ℓ(w_t, z_i)) / k ) / b
    Update parameter: w_{t+1} ← w_t − λ · g_B
end while
```

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

- ▷ Empirical evidence shows that a small k (e.g., $k = 3$) already gives competitive performance.
- ▷ Implementing the $k + 1$ forward passes on parallel processors further reduces the computation load.

Implication: GMP on 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[1]	**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**

Table 1: Top-1 classification accuracy acc.(%) of VGG16. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy. Superscript denotes the value of k.

[1] $\min_w L_s(w) + \rho \max_\delta L_s(w+\delta) - L_s(w)$

Summary

- ▷ We derive some new information-theoretic bounds for SGD;
- ▷ Apply the bound to linear networks and two-layer ReLU networks;
- ▷ Epoch-wise double descent of gradient dispersion is observed;
- ▷ Design new regularization schemes, e.g., dynamic gradient clipping and GMP.

Thank you!

zwang286@uottawa.ca