Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels

Hao Wang, Yizhe Huang, Rui Gao, Flavio P. Calmon

hao_wang@g.harvard.edu

yizhehuang@utexas.edu

rui.gao@mccombs.utexas.edu

flavio@seas.harvard.edu

Harvard John A. Paulson School of Engineering and Applied Sciences

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Outline

- Preliminary
  - Generalization analysis
  - An information-theoretic framework
- SGLD generalization bound
  - Definition
  - Our generalization bound
  - Experiments
- Generalization amplification by iteration
  - DP-SGD algorithm
  - Our generalization bound
- Related works and open questions

Population risk minimization

Consider the following (non-convex) optimization problem:

$$\min_{w \in \mathcal{W}} \; L_\mu(w) \triangleq \mathbb{E}_{\mathsf{Z} \sim \mu} \left[\ell(w, \mathsf{Z}) \right]$$

Population risk minimization

Consider the following (non-convex) optimization problem:

$$\min_{w \in \mathcal{W}} \underline{L_\mu(w) \triangleq \mathbb{E}_{\mathsf{Z} \sim \mu}\left[\ell(w, \mathsf{Z})\right]}$$

population risk

parameter

hypothesis class

Population risk minimization

Consider the following (non-convex) optimization problem:

$$\min_{w \in \mathcal{W}} L_\mu(w) \triangleq \mathbb{E}_{\mathsf{Z} \sim \mu}\left[\ell(w, \mathsf{Z})\right]$$

loss function

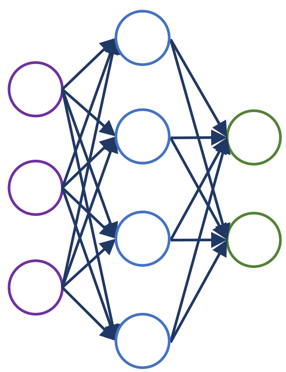
data point
following distribution μ

Population risk minimization: Neural network

Consider the following (non-convex) optimization problem:

$$\min_{w \in \mathcal{W}} L_\mu(w) \triangleq \mathbb{E}_{\mathsf{Z}\sim\mu}\left[\ell(w, \mathsf{Z})\right]$$

weight matrices



f_w

$(f_w(\mathsf{X}) - \mathsf{Y})^2$ where $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$

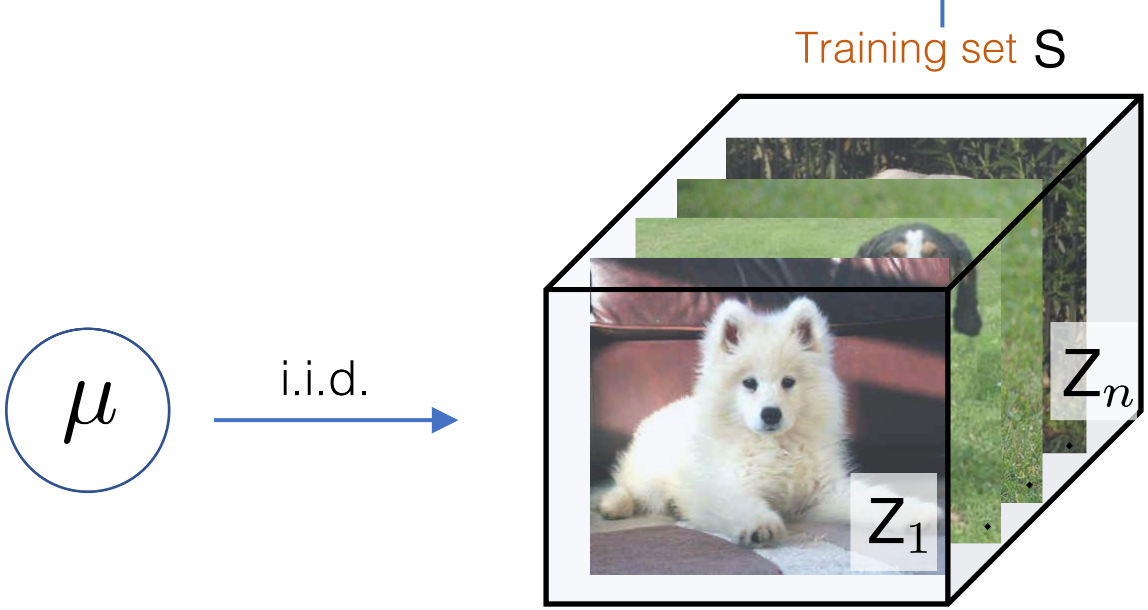
Empirical risk minimization

Consider the following (non-convex) optimization problem:

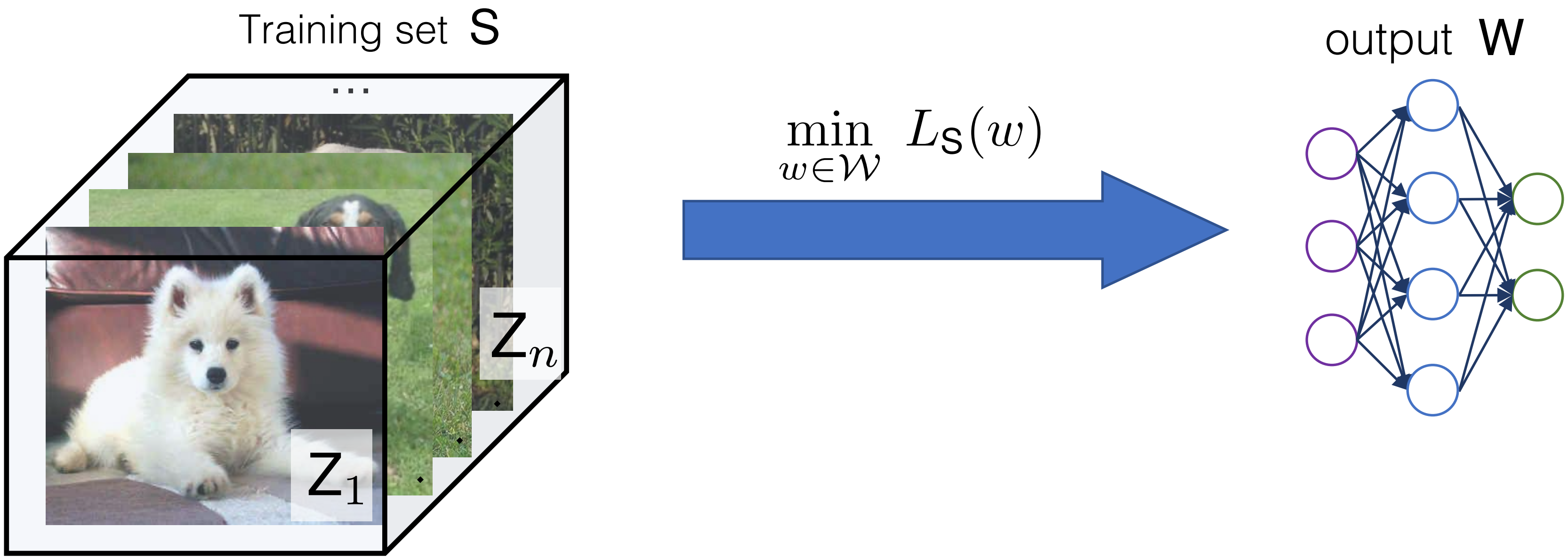
$$\min_{w \in \mathcal{W}} \; L_{\mu}(w) \triangleq \mathbb{E}_{\mathsf{Z} \sim \mu} \left[\ell(w, \mathsf{Z}) \right]$$

In practice,

$$\min_{w \in \mathcal{W}} \; L_{\mathsf{S}}(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w, \mathsf{Z}_i)$$



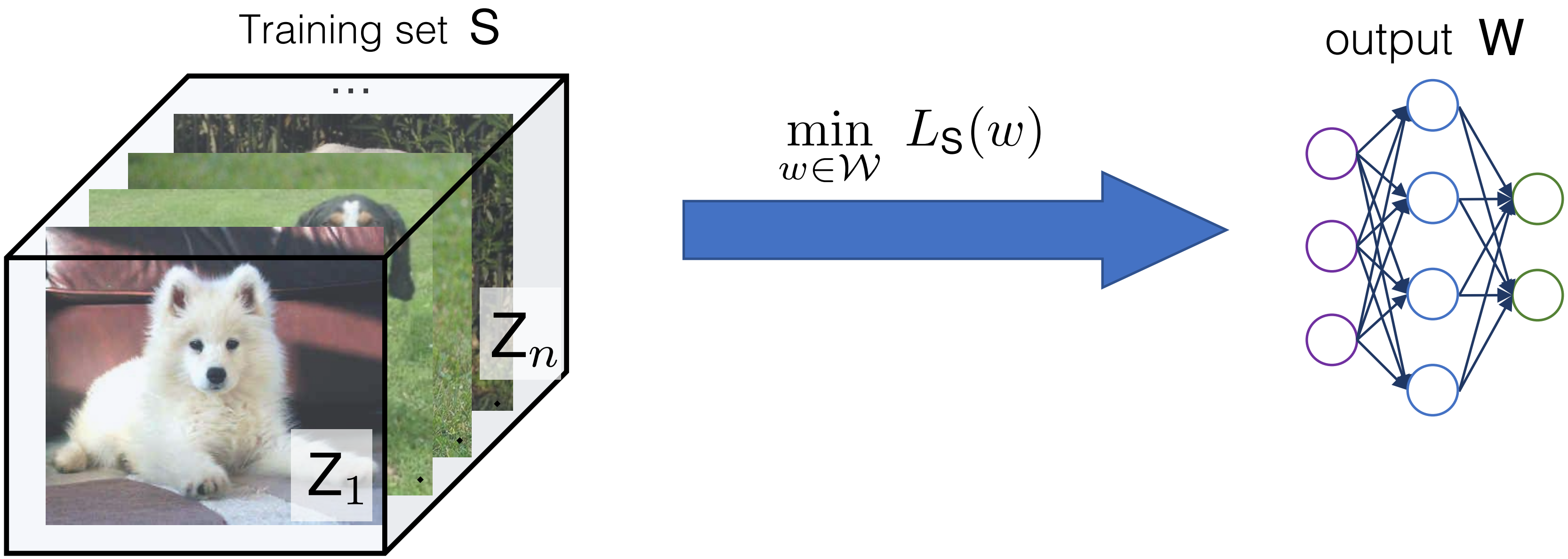
Error decomposition



$$L_\mu(\mathsf{W}) = L_{\mathsf{S}}(\mathsf{W}) + \underline{L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})}$$

generalization gap

Expected generalization gap



$$\mathbb{E}\left[L_\mu(W) - L_S(W)\right]$$

Expected generalization gap

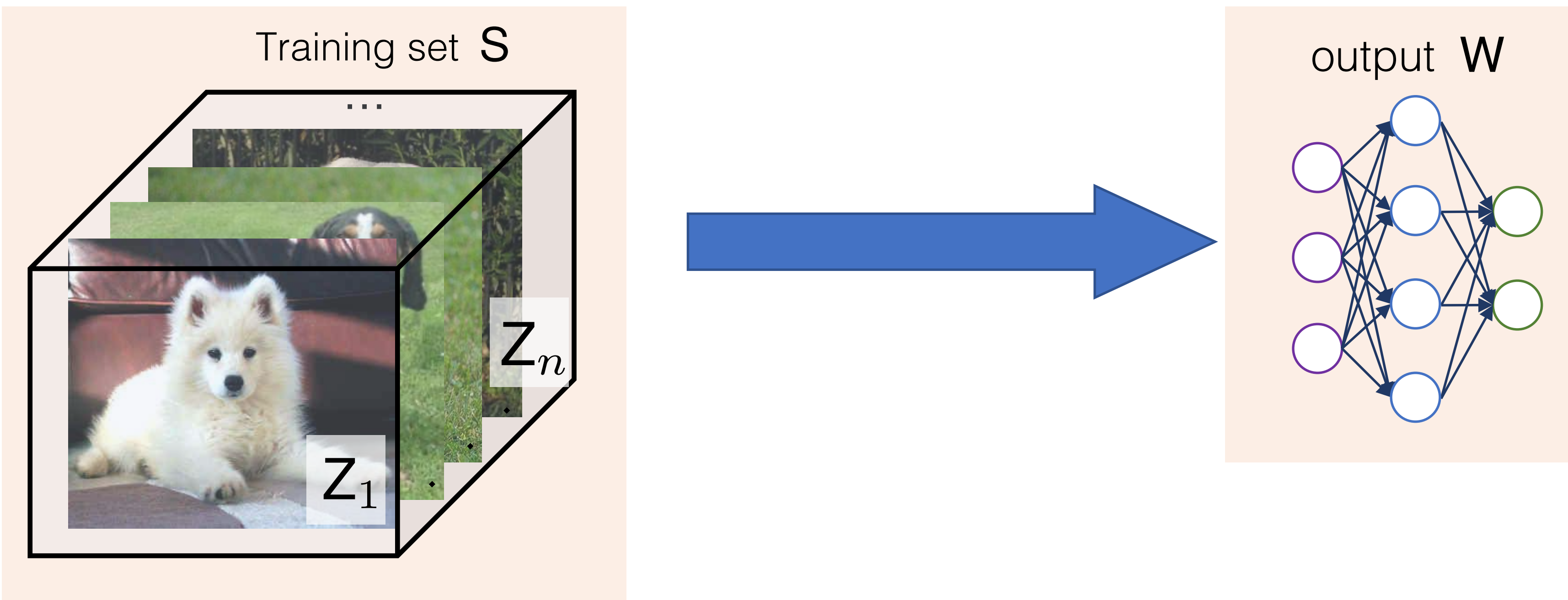
Today's focus

- VC-dimension
- Rademacher complexity
- Algorithmic stability
- PAC-Bayes
- Information theory
- …

Information-theoretic generalization bound

Theorem (Xu and Raginsky, 2017)**.** *Suppose* $\ell(w, \mathsf{Z})$ *is* σ*-sub-Gaussian under* μ *for all* w*.*

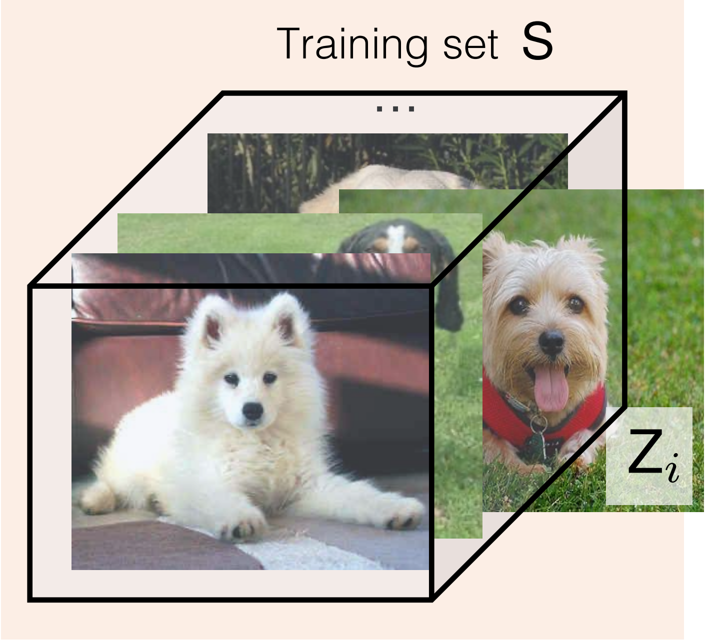
$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \sqrt{\frac{2\sigma^2}{n} I(\mathsf{W}; \mathsf{S})}.$$

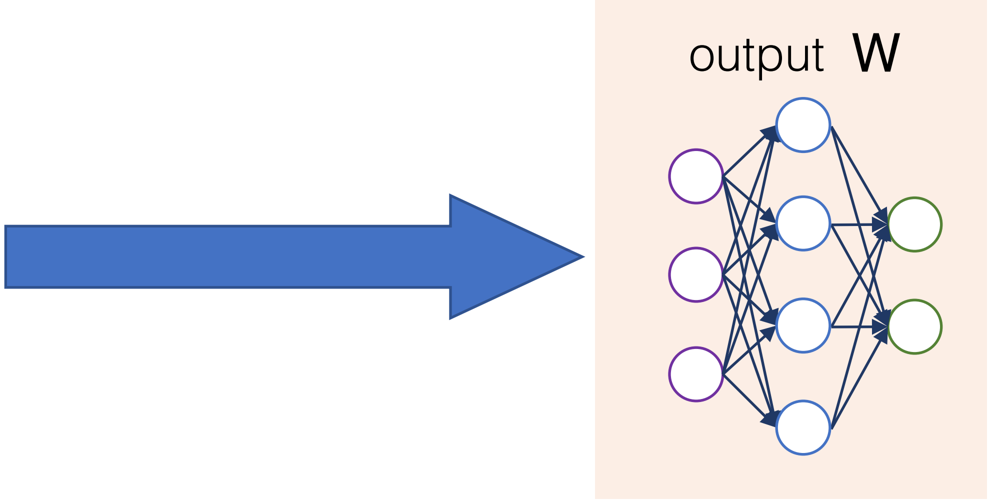


Information-theoretic generalization bound

Proposition (Bu et al., 2020). *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(\mathsf{W};\mathsf{Z}_i)}.$$





Pros and cons

Proposition (Bu et al., 2020). *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(\mathsf{W};\mathsf{Z}_i)}.$$

- Algorithm/distribution dependent
- Mild assumption
- …

Pros and cons

Proposition (Bu et al., 2020). *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(\mathsf{W};\mathsf{Z}_i)}.$$

- Algorithm/distribution dependent
- Mild assumption
- $\ldots$

A bounded loss is sufficient:
if $\ell(\cdot,\cdot) \in [0, A]$, then $\sigma = \frac{A}{2}$

Pros and cons

Proposition (Bu et al., 2020). *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(\mathsf{W};\mathsf{Z}_i)}.$$

- Mutual information is hard to compute

Outline

- Preliminary
  - Generalization analysis
  - An information-theoretic framework
- SGLD generalization bound
  - Definition
  - Our generalization bound
  - Experiments
- Generalization amplification by iteration
  - DP-SGD algorithm
  - Our generalization bound
- Related works and open questions

SGLD is popular in practice

- Privacy guarantee
- Mitigate overfitting
- Easy to analyze in theory
- …

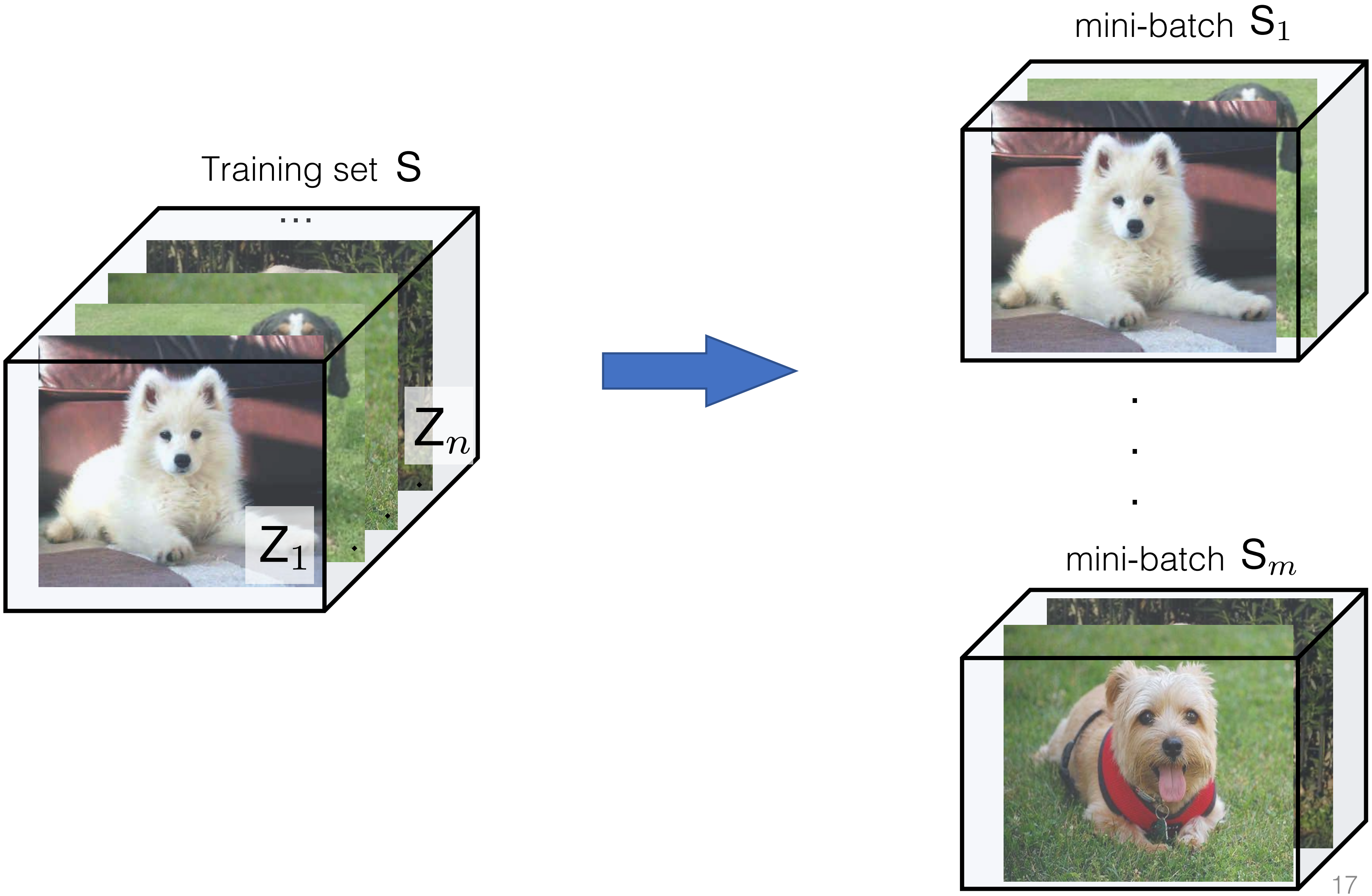
tensorflow/**privacy**

Library for training machine learning models with privacy for training data

42 Contributors · 56 Issues · 1k Stars · 321 Forks

Opacus

SGLD: mini-batches



SGLD: update rule

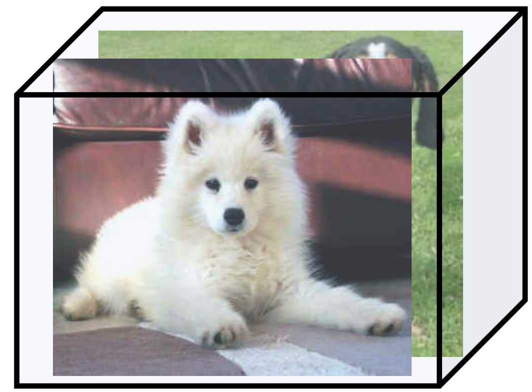
choose W_0 arbitrarily
for $t = 1, \cdots, T$

$$W_t = W_{t-1} - \eta_t \nabla_w \ell(W_{t-1}, S_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N$$

learning rate

mini-batch gradient

mini-batch S_1



.
.
.

mini-batch S_m

[Welling and Teh, 2011]

SGLD: update rule

choose W_0 arbitrarily

for $t = 1, \cdots, T$

$$\mathsf{W}_t = \mathsf{W}_{t-1} - \eta_t \nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} \mathsf{N}$$

Gaussian noise

inverse temperature

SGLD: output

choose W_0 arbitrarily

for $t = 1, \cdots, T$

$$\mathsf{W}_t = \mathsf{W}_{t-1} - \eta_t \nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} \mathsf{N}$$

output: $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$

SGLD: output

choose W_0 arbitrarily
for $t = 1, \cdots, T$

$$\mathsf{W}_t = \mathsf{W}_{t-1} - \eta_t \nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}}\mathsf{N}$$

output: $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$

Examples:

- $f(\mathsf{W}_1, \cdots, \mathsf{W}_T) = \mathsf{W}_T$
- $f(\mathsf{W}_1, \cdots, \mathsf{W}_T) = \frac{\mathsf{W}_1 + \cdots + \mathsf{W}_T}{T}$

Our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \textit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

Our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

- b: mini-batch size
- σ: sub-Gaussian constant
- n: number of samples
- β_t: inverse temperature
- η_t: learning rate

Our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$\left|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]\right| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

- b: mini-batch size
- σ: sub-Gaussian constant
- n: number of samples
- β_t: inverse temperature
- η_t: learning rate

Recall the update rule:

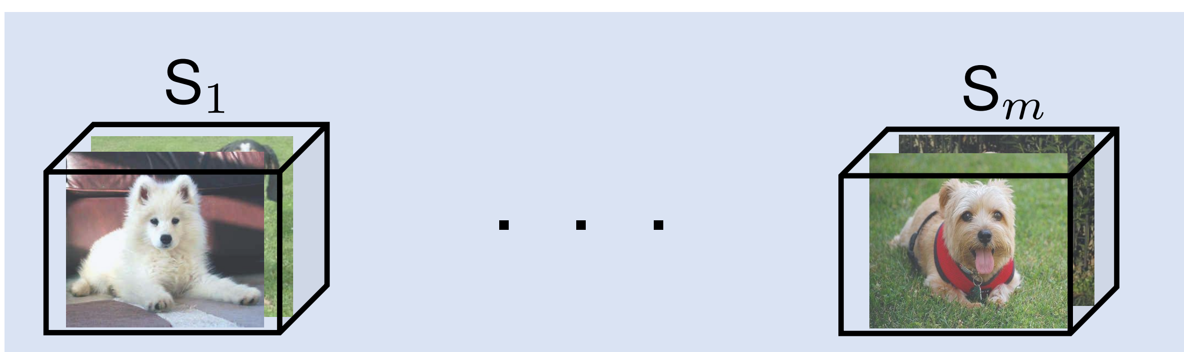
$$\mathsf{W}_t = \mathsf{W}_{t-1} - \eta_t \nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_{B_t}) + \sqrt{\frac{2\eta_t}{\beta_t}}\mathsf{N}$$

Our generalization bound

Theorem. *Suppose* $\ell(w, \mathsf{Z})$ *is* σ*-sub-Gaussian under* μ *for all* w.

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \textit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

mini-batch



Our generalization bound

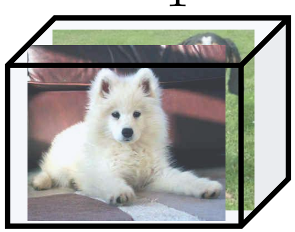
Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

$\mathcal{T}_j$ contains the indices of iterations in which S_j is used

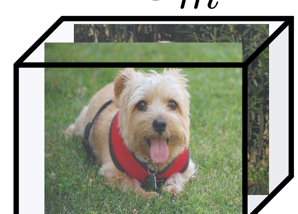
mini-batch

S_1



. . .

S_m



If the number of iterations increases,

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

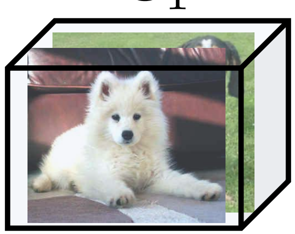
$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

more terms

invariant

S_1 S_m

mini-batch



. . .

Our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

Measure sharpness of loss landscape

Our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_\mathsf{S}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

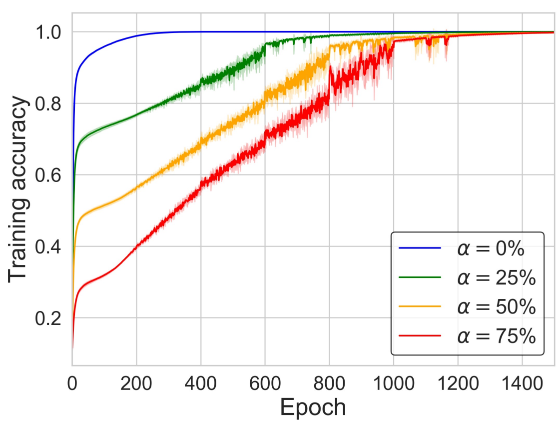
Variance of gradients

The journey matters more than the destination.

Experiment on MNIST: Label corruption

Theorem. *Suppose $\ell(w, Z)$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}[L_\mu(W) - L_S(W)]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot Var(\nabla_w \ell(W_{t-1}, S_j))}.$$

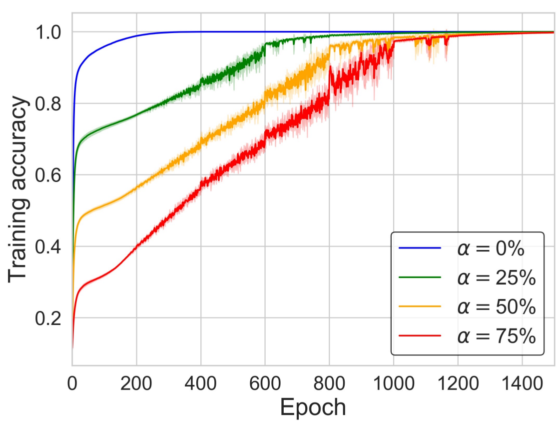


- Vary label corruption level (alpha)
- Train 3-layer neural networks on MNIST using SGLD

Experiment on MNIST: Label corruption

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$\left|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]\right| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

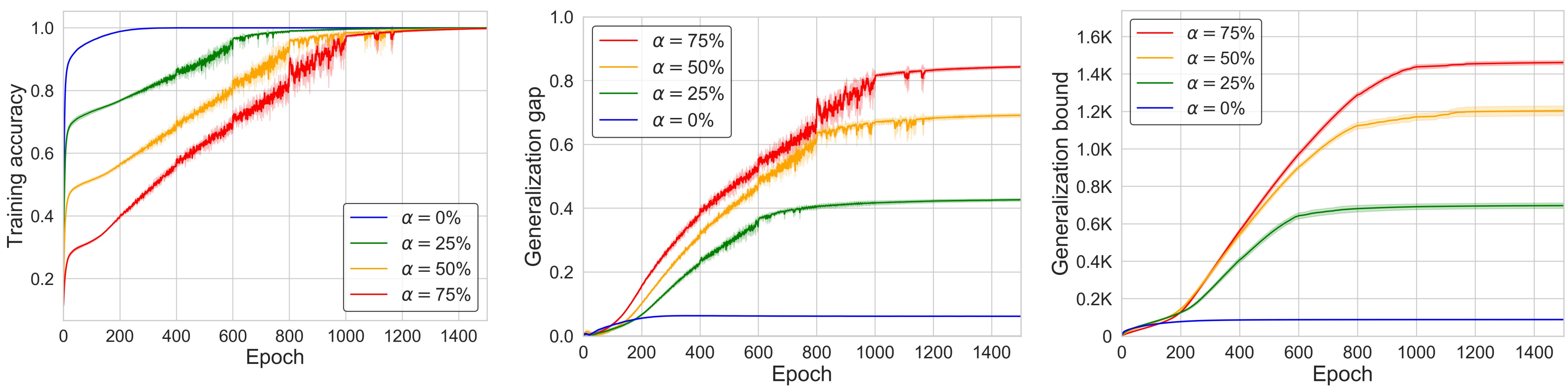


- Vary label corruption level (alpha)
- Train 3-layer neural networks on MNIST using SGLD
- Run 1500 epochs until the training accuracy = 1

Experiment on MNIST: Label corruption

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$



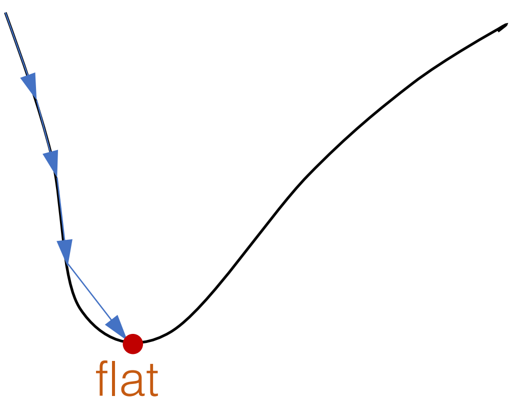
Experiment on MNIST: Label corruption

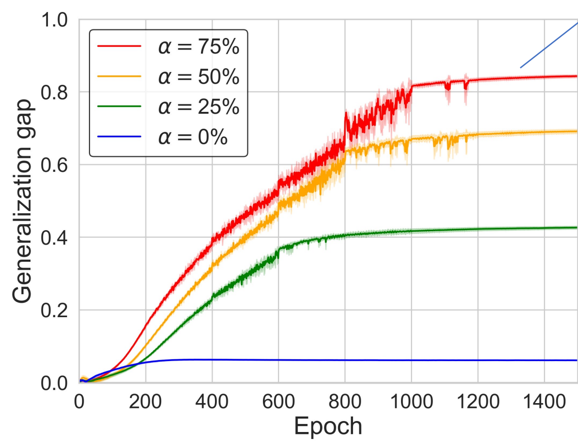
Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

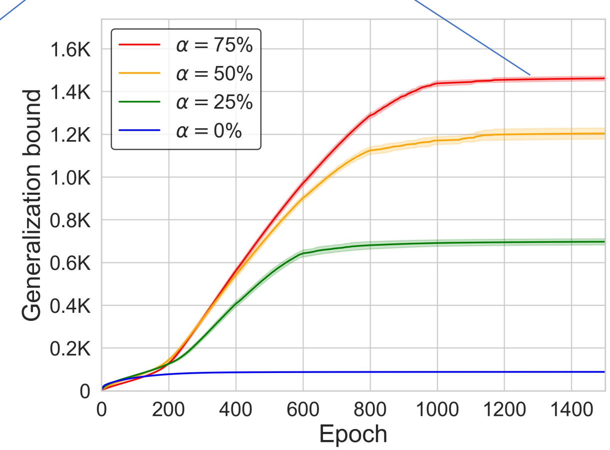
$$\left|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]\right| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

Both become stable

loss landscape





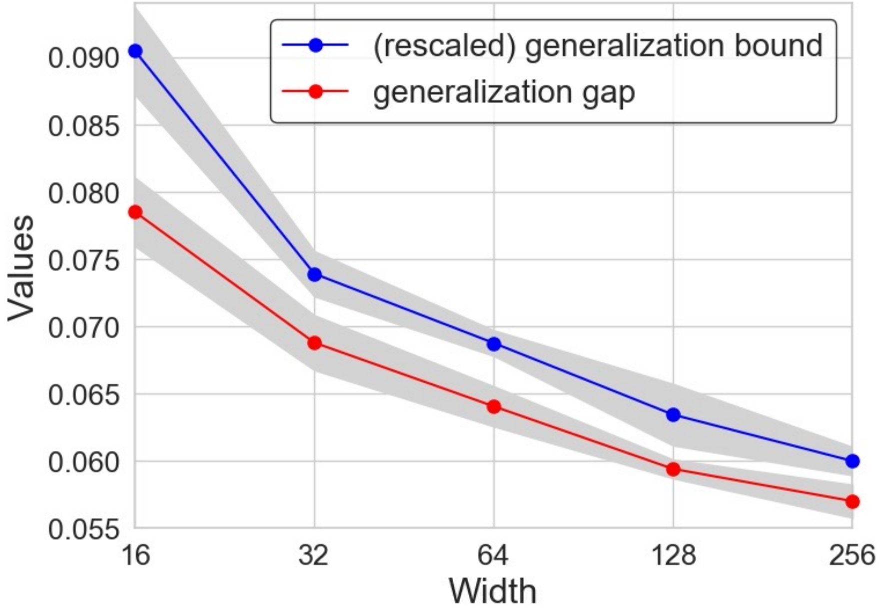


When the algorithm converges, the variance of gradients become negligible.

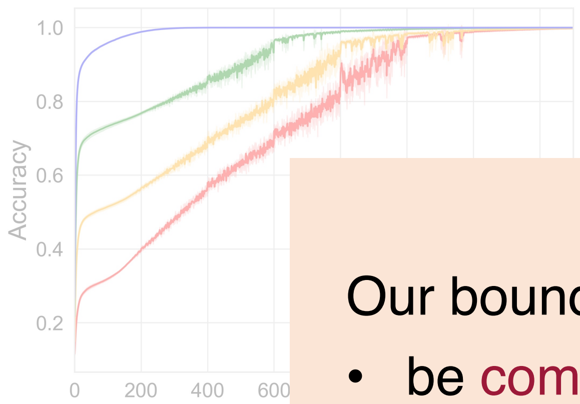
Experiment on MNIST: Network width

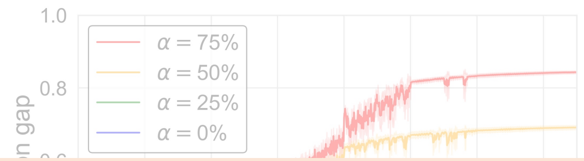
Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

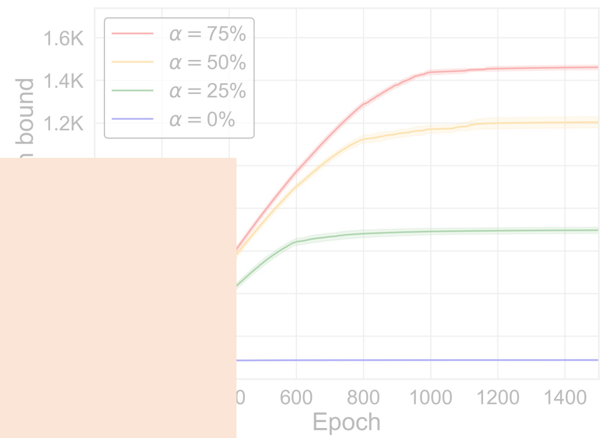
$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$



Take away

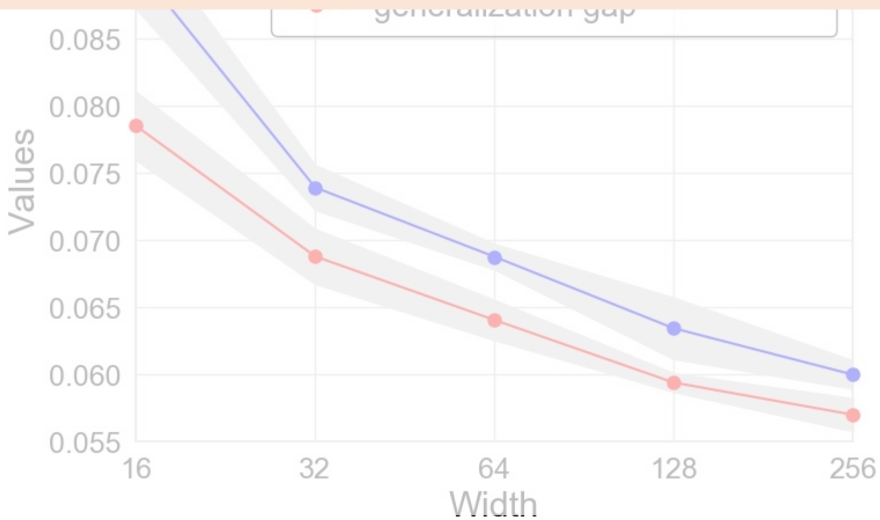






Our bound can:

- be computed from data
- explain some generalization phenomena



Outline

- Preliminary
  - Generalization analysis
  - An information-theoretic framework
- SGLD generalization bound
  - Definition
  - Our generalization bound
  - Experiments
- Generalization amplification by iteration
  - DP-SGD algorithm
  - Our generalization bound
- Related works and open questions

Recall our generalization bound

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \textit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

Here $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$ for an arbitrary function f

Upside

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \le \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}\left(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j)\right)}.$$

Here $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$ for an arbitrary function f

Can be applied to many settings

Examples:

- $f(\mathsf{W}_1, \cdots, \mathsf{W}_T) = \mathsf{W}_T$
- $f(\mathsf{W}_1, \cdots, \mathsf{W}_T) = \frac{\mathsf{W}_1 + \cdots + \mathsf{W}_T}{T}$

Downside

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$\sup_{f} |\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \textit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

Here $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$ for an arbitrary function f

A uniform bound for any function f

Question

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$|\mathbb{E}\left[L_\mu(\mathsf{W}) - L_{\mathsf{S}}(\mathsf{W})\right]| \leq \frac{\sqrt{2b}\sigma}{2n} \sum_{j=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_j} \beta_t \eta_t \cdot \mathit{Var}(\nabla_w \ell(\mathsf{W}_{t-1}, \mathsf{S}_j))}.$$

Here $\mathsf{W} = f(\mathsf{W}_1, \cdots, \mathsf{W}_T)$ for an arbitrary function f

For $\mathsf{W} = \mathsf{W}_T$, can we have a sharper bound?

Projected Differentially-Private SGD (DP-SGD)

choose W_0 arbitrarily
for $t = 1, \cdots, T$

$$\mathsf{W}_t = \mathsf{Proj}_{\mathcal{W}}\left(\mathsf{W}_{t-1} - \eta\left(g(\mathsf{W}_{t-1}, \mathsf{Z}_t) + \mathsf{N}\right)\right)$$

output: W_T

Assumptions:

- sampling without replacement
- $\|g(w,z)\|_2 \leq K$ for any w, z

Our generalization bound: time-decaying factor

Theorem. *Suppose $\ell(w, \mathsf{Z})$ is σ-sub-Gaussian under μ for all w.*

$$\left|\mathbb{E}\left[L_\mu(\mathsf{W}_T) - L_\mathsf{S}(\mathsf{W}_T)\right]\right| \leq \frac{2\sigma}{n} \sum_{t=1}^{T} \sqrt{\mathit{Var}\left(g\left(\mathsf{W}_{t-1}, \mathsf{Z}\right)\right) \cdot q^{T-t}}.$$

$$q \in (0, 1)$$

Enables the impact of early iterations to reduce with time

Proof

Theorem. *Suppose $\ell(w, Z)$ is σ-sub-Gaussian under μ for all w.*

Key proof techniques: properties of Gaussian channels

Step 1: $|\mathbb{E}[L_\mu(W_T) - L_S(W_T)]| \leq \frac{\sqrt{2\sigma^2}}{n} \sum \sqrt{I(W_T; Z_t)}$ [Bu et al., 2020]

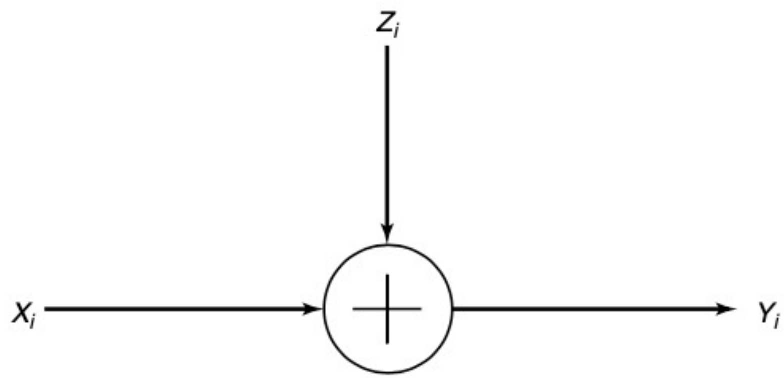


FIGURE 9.1. Gaussian channel.

Step 2:

(H

Step 3:

(How we obtain a computable bound)

Image is from [Cover and Thomas, 1999]

Outline

- Preliminary
  - Generalization analysis
  - An information-theoretic framework
- SGLD generalization bound
  - Definition
  - Our generalization bound
  - Experiments
- Generalization amplification by iteration
  - DP-SGD algorithm
  - Our generalization bound
- Related works and open questions

Related works and open questions

- SGLD generalization bounds: [Mou et al., 2018], [Li et al., 2019], [Pensia et al., 2018], [Bu et al., 2020], [Negrea et al., 2019], [Haghifam et al., 2020], [Rodriguez-Galvez et al., 2020] [Neu, 2021]

- Privacy amplification by iteration: [Feldman et al., 2018], [Asoodeh et al., 2020]

Related works and open questions

- SGLD generalization bounds: [Mou et al., 2018], [Li et al., 2019], [Pensia et al., 2018], [Bu et al., 2020], [Negrea et al., 2019], [Haghifam et al., 2020], [Rodriguez-Galvez et al., 2020] [Neu, 2021]
- Privacy amplification by iteration: [Feldman et al., 2018], [Asoodeh et al., 2020]

- Privacy can also be amplified by subsampling and shuffling. Can they improve algorithmic generalization?
- New definition of sharpness
- Tighten the decay factor.

[Chaudhuri and Mishra, 2006; Erlingsson et al., 2019]

Related works and open questions

- SGLD generalization bounds: [Mou et al., 2018], [Li et al., 2019], [Pensia et al., 2018], [Bu et al., 2020], [Negrea et al., 2019], [Haghifam et al., 2020], [Rodriguez-Galvez et al., 2020] [Neu, 2021]

- Privacy amplification by iteration: [Feldman et al., 2018], [Asoodeh et al., 2020]

- Privacy can also be amplified by subsampling and shuffling. Can they improve algorithmic generalization?

- New definition of sharpness

- Tighten the decay factor.

Related works and open questions

- SGLD generalization bounds: [Mou et al., 2018], [Li et al., 2019], [Pensia et al., 2018], [Bu et al., 2020], [Negrea et al., 2019], [Haghifam et al., 2020], [Rodriguez-Galvez et al., 2020] [Neu, 2021]

- Privacy amplification by iteration: [Feldman et al., 2018], [Asoodeh et al., 2020]

- Privacy can also be amplified by subsampling and shuffling.
  Can they improve algorithmic generalization?

- New definition of sharpness

- Tighten the decay factor.

Thanks for watching!