Stochastic Gradient Methods: Bias, Stability and Generalization

Yunwen Lei

Joint work with Shuang Zeng

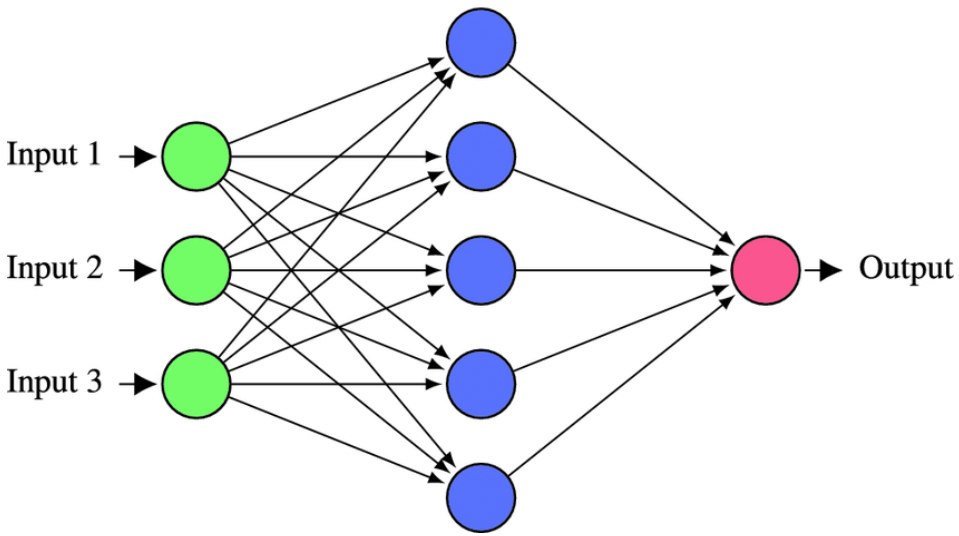
29 May, 2025
HKU IDS Interdisciplinary Workshop
Exploring the Foundations: Fundamental AI and Theoretical Machine Learning

Background

Supervised Machine Learning

- Given training examples from a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - formally $S = \{z_i = (x_i, y_i), i = 1, \ldots, n\}$, $z_i \in \mathcal{Z}$
  - Independently drawn from a probability measure ρ on $\mathcal{Z}$

- Aim to find prediction rule $g_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$, parameterized by $\mathbf{w} \in \mathcal{W}$ (model space)
  - linear models: $g_{\mathbf{w}}(x) = \langle \mathbf{w}, x \rangle$
  - neural networks: $g_{\mathbf{w}}(x) = \sigma_L(\mathbf{W}_L \sigma_{L-1}(\mathbf{W}_{L-1} \cdots \sigma_1(\mathbf{W}_1 x)))$



Population and Empirical Risk

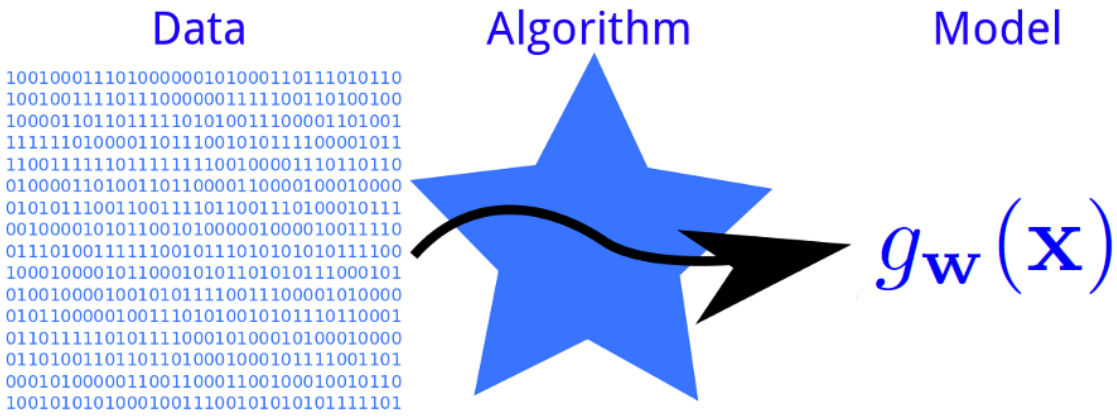
- Loss function $f(\mathbf{w}; z)$ to measure performance of $g_{\mathbf{w}}$ on an example $z = (x, y)$
  - squares loss: $f(\mathbf{w}; z) = (y - g_{\mathbf{w}}(x))^2$ for regression
  - hinge loss: $f(\mathbf{w}; z) = \max\{0, 1 - y g_{\mathbf{w}}(x)\}$ for binary classification
- Population risk (testing error) and Empirical risk (training error)

$$F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)] \quad \text{and} \quad F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}; z_i).$$

- We aim to find the best model in the hypothesis space

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}).$$

Algorithms



- A learning algorithm A with an output model $A(S) \in \mathcal{W}$
  - empirical risk minimization: $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} \text{training\_error}(\mathbf{w})$
  - regularized risk minimization:

$$A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\{\text{training\_error}(\mathbf{w}) + \text{regularizer}(\mathbf{w})\right\}$$

  - gradient descent, stochastic gradient descent, stochastic gradient descent ascent ...

Gradient Descent and Stochastic Gradient Descent

Gradient Descent (GD)

for $t = 1, 2, \ldots$ **to** T **do**
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla F_S(\mathbf{w}_t)$ for some step sizes $\eta_t > 0$
return $\mathbf{w}_{T+1}$ or an average of $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$

Stochastic Gradient Descent (SGD)

for $t = 1, 2, \ldots$ **to** T **do**
  $j_t \leftarrow$ random index from $\{1, 2, \ldots, n\}$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})$ for some step sizes $\eta_t > 0$
return $\mathbf{w}_{T+1}$ or an average of $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$

SGD uses a single example to approximate the gradient of $F_S(\mathbf{w}_t)$!

Excess Population Risk

- Algorithm A often produces models with a small training error
- This does not necessarily mean $A(S)$ has a good prediction

Generalization gap = Test Error − Training Error

Target of analysis: excess population risk

$$\mathbb{E}\big[F(A(S)) - F(\mathbf{w}^*)\big] = \mathbb{E}\Big[\underbrace{F(A(S)) - F_S(A(S))}_{\text{generalization gap}} + \underbrace{F_S(A(S)) - F_S(\mathbf{w}^*)}_{\text{optimization error}}\Big]$$

1. generalization gap: difference between testing error and training error at $A(S)$
2. optimization error: difference between $A(S)$ and $\mathbf{w}^*$ measured by training error

We will study both the generalization and optimization of learning algorithms!

SGD

SGD uses unbiased gradient estimators

$$\mathbb{E}_{j_t}[\nabla f(\mathbf{w}_t; z_{j_t})] = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{w}_t; z_i) = \nabla F_S(\mathbf{w}_t)$$

The zero-bias simplifies both the convergence and generalization analysis of SGD

- Optimization: SGD applied to convex and smooth problems achieves the convergence rate (Bottou et al., 2018)

$$\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}^*)] \lesssim \eta + \frac{1}{\eta T}$$

- Generalization: SGD applied to convex and smooth problems achieves the generalization gap (Hardt et al., 2016)

$$\mathbb{E}[F(\mathbf{w}_T) - F_S(\mathbf{w}_T)] \lesssim \frac{\eta T}{n}.$$

In practice, we often consider biased stochastic gradient methods (BSGMs)!

Biased Stochastic Gradient Method (BSGM)

At the t-th iteration, we build a possibly biased estimator $g(\mathbf{w}_t; z_{j_t})$ and update as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t g(\mathbf{w}_t; z_{j_t}).$$

- We consider a surrogate loss function $\tilde{f} : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ and build

$$\widetilde{F}_S(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \tilde{f}(\mathbf{w}; z_i). \tag{1}$$

- We define the bias

$$b_t = \mathbb{E}_{j_t}[g(\mathbf{w}_t; z_{j_t})] - \nabla \widetilde{F}_S(\mathbf{w}_t).$$

Intuition: While $g(\mathbf{w}_t; z_{j_t})$ may be a biased estimator of $\nabla F_S(\mathbf{w}_t)$, it can be an unbiased estimator of $\nabla \widetilde{F}_S(\mathbf{w}_t)$

Examples of BSGMs

Clipped-SGD

- SGD is not robust for problems with heavy-tailed noises, e.g., training of BERT
- A clipping operator (with parameter τ) is introduced to improve the robustness

$$\mathrm{clip}(\mathbf{v}, \tau) := \min\left\{1, \frac{\tau}{\|\mathbf{v}\|_2}\right\}\mathbf{v}. \tag{2}$$

- Clipped-SGD uses the estimator

$$g(\mathbf{w}_t; z_{j_t}) = \mathrm{clip}(\nabla f(\mathbf{w}_t; z_{j_t}), \tau). \tag{3}$$

- It introduces nontrivial bias

$$b_t = \mathbb{E}_{j_t}[\mathrm{clip}(\nabla f(\mathbf{w}_t; z_{j_t}), \tau)] - \nabla F_S(\mathbf{w}_t)$$

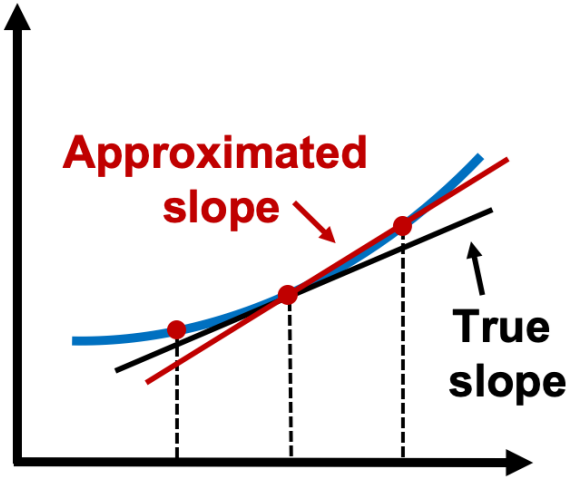
- $b_t \lesssim G^p \tau^{1-p}$ if we assume $\mathbb{E}_{j_t}[\|\nabla f(\mathbf{w}; z_{j_t})\|_2^p] \leq G^p, p \in (1, 2]$ (Zhang et al., 2020)

Zeroth-order SGD

- Gradient calculations may be infeasible in some applications
- Zeroth-order SGD approximates the gradient by finite difference

$$g(\mathbf{w}_t; z_{j_t}) = \frac{1}{K} \sum_{l=1}^{K} \underbrace{\frac{f(\mathbf{w}_t + \mu \mathbf{u}_{t,l}; z_{j_t}) - f(\mathbf{w}_t; z_{j_t})}{\mu}}_{\approx \mathbf{u}_{t,l}^\top \nabla f(\mathbf{w}_t; Z_{j_t})} \mathbf{u}_{t,l},$$

where $\mathbf{u}_{t,l} \sim \mathcal{N}(0, I_d)$ is a random direction and μ is a smoothing parameter.



- It introduces nontrivial bias (Nesterov and Spokoiny, 2017)

$$b_t = \mathbb{E}_{j_t, \mathbf{u}}[g(\mathbf{w}_t; z_{j_t})] - \nabla F_S(\mathbf{w}_t)] \quad \text{and} \quad \|b_t\|_2 \lesssim \mu d^{\frac{3}{2}}.$$

SGD with Delayed Updates

- In practical implementations, the gradients may not be immediately available, e.g., due to communication delay

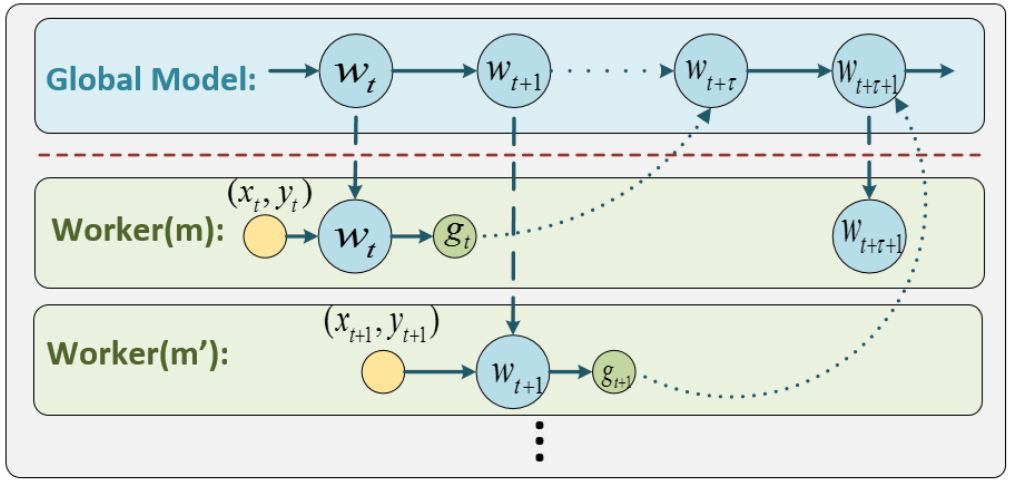


Figure in Zheng et al. (2017)

- Then, we may update a model using the outdated gradient information (τ is the delay factor)

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla f(\mathbf{w}_{t-\tau}, z_{j_{t-\tau}}), \quad t > \tau.$$

- This leads to a bias

$$b_t = \mathbb{E}_{j_{t-\tau}}[\nabla f(\mathbf{w}_{t-\tau}, z_{j_{t-\tau}})] - \nabla F_S(\mathbf{w}_t) = \nabla F_S(\mathbf{w}_{t-\tau}) - \nabla F_S(\mathbf{w}_t).$$

- The bias is of order $O(\tau)$, which affects both optimization and generalization.

Decentralized SGD

- We have m local machines with $S_k = \{z_{1,k}, z_{2,k}, \ldots, z_{n,k}\}$ in the k-th machine
- Each local machine updates its own model and communicates with its neighbors

$$\mathbf{w}_t^k = \sum_{j=1}^{m} P_{kj}\mathbf{w}_t^j - \eta_t \nabla f(\mathbf{w}_t^k; z_{j_t^k,k}), \tag{4}$$

where $P \in \mathbb{R}^{m\times m}$ is a double stochastic matrix, $j_t^k \sim \text{Unif}[n]$.
If we consider the averaged model $\bar{\mathbf{w}}_t = \frac{1}{m}\sum_{k=1}^{m} \mathbf{w}_t^k$, then

$$\bar{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_t - \frac{\eta_t}{m}\sum_{k=1}^{m} \nabla f(\mathbf{w}_t^k; z_{j_t^k,k}). \tag{5}$$

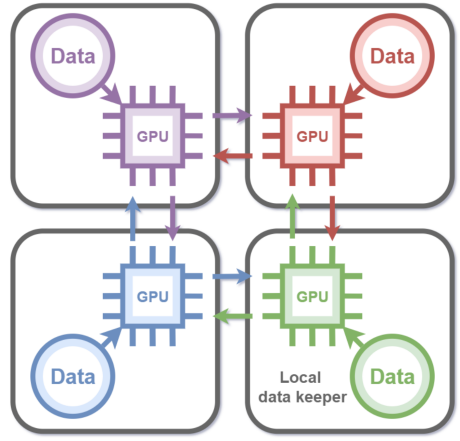


Figure in Zhu et al. (2023)

- We can formulate it as a BSGM by

$$b_t = \mathbb{E}_{j_t^k}\left[\frac{1}{m}\sum_{k=1}^{m}\nabla f(\mathbf{w}_t^k; z_{j_t^k,k})\right] - \frac{1}{m}\sum_{k=1}^{m}\nabla F_{S_k}(\bar{\mathbf{w}}_t) = \frac{1}{m}\sum_{k=1}^{m}\left(\nabla F_{S_k}(\mathbf{w}_t^k) - \nabla F_{S_k}(\bar{\mathbf{w}}_t)\right).$$

- The bias can be bounded by $\frac{1}{m}\sum_{k=1}^{m}\|\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}_t\|_2$, which is of order $\left(\sum_{j=1}^{t}\eta_j^2\right)^{\frac{1}{2}}$

Other Examples

- Top-k and Random-k sparsification: keeps k largest coordinates of the gradient vector
- Stochastic average gradient (Schmidt et al., 2017)
- and many others (Ajalloeian and Stich, 2020; Driggs et al., 2022)
- Existing theoretical analysis of BSGMs mainly focus on the convergence (Ajalloeian and Stich, 2020; Driggs et al., 2022; Hu et al., 2020; Duchi et al., 2015; Nesterov and Spokoiny, 2017).

How to study the generalization behavior?

Stability Analysis of BSGMs

Uniform Stability

A randomized algorithm A is ϵ-uniformly stable if, for any two datasets S and S' that differ by one example (neighboring dataset), we have (Bousquet and Elisseeff, 2002)

$$\sup_z \mathbb{E}_A \left[f(A(S); z) - f(A(S'); z) \right] \leq \epsilon. \tag{6}$$

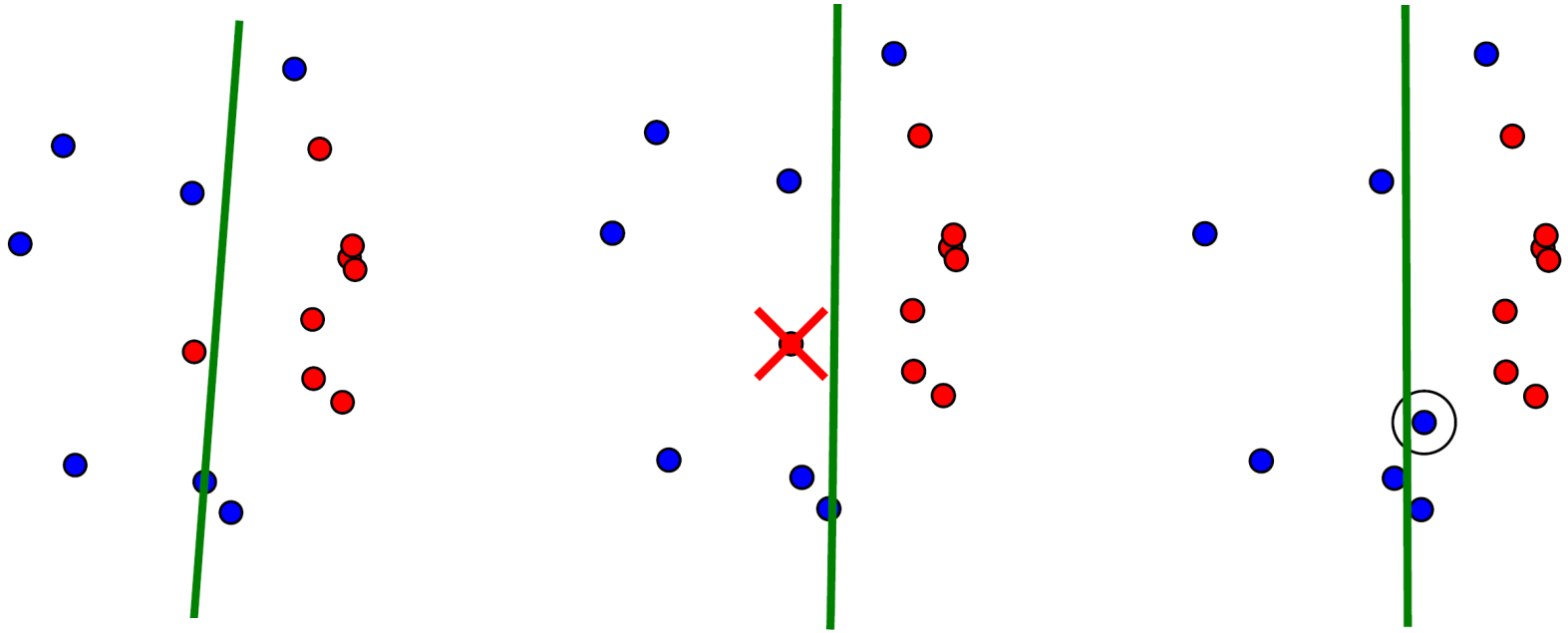


Figure Taken in Kuzborskij and Lampert (2018)

If A is uniformly stable, then it is generalizable!

Generalization by Uniform Stability

We say $\ell : \mathcal{W} \mapsto \mathbb{R}$ is G-Lipschitz if $|\ell(\mathbf{w}) - \ell(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|_2$.

If $\mathcal{A}$ is ϵ-uniform stable and f is G-Lipschitz, then (Hardt et al., 2016)

$$\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] \leq G\epsilon. \tag{7}$$

Issues with Uniform Stability

- A strong concept: consider any two neighboring datasets and any test example z

$$\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(S'); z)\big] \leq \epsilon.$$

- Requires strong assumptions to control the uniform stability: Lipschitzness, smoothness (Hardt et al., 2016)
- Cannot show the effect of optimization in improving stability and generalization.

On-Average Model Stability

- To address these issues, we introduce on-average model stability

$$S = \{z_1, z_2, \ldots, z_n\}$$
$$S' = \{z_1', z_2', \ldots, z_n'\}$$

$$\xrightarrow{\text{perturbation}}$$

$$S = \{z_1, z_2, \ldots, z_n\} \xrightarrow{A} A(S)$$
$$S^{(1)} = \{z_1', z_2, \ldots, z_n\} \xrightarrow{A} A(S^{(1)})$$
$$S^{(2)} = \{z_1, z_2', \ldots, z_n\} \xrightarrow{A} A(S^{(2)})$$
$$\vdots$$
$$S^{(n)} = \{z_1, z_2, \ldots, z_n'\} \xrightarrow{A} A(S^{(n)})$$

On-Average Model Stability

(Lei and Ying, 2020)

We say a randomized algorithm $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is on-average model ϵ-stable if

$$\mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n}\|A(S) - A(S^{(i)})\|_2^2\right] \leq \epsilon^2. \tag{8}$$

Generalization by On-average Model stability

Smoothness, Lipschitzness and Convexity

Let $\ell : \mathcal{W} \mapsto \mathbb{R}$. Let $L \geq 0$.

- We say ℓ is L-smooth if $\|\nabla \ell(\mathbf{w}) - \nabla \ell(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$.
- We say ℓ is convex if $\ell(\mathbf{w}) \geq \ell(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla \ell(\mathbf{w}') \rangle$.

Generalization by On-average Model stability (Lei and Ying, 2020)

If A is on-average model ϵ-stable, then

$$\text{generalization gap} \lesssim \epsilon^2 + \epsilon\big(\text{training error}\big)^{\frac{1}{2}}. \tag{9}$$

- If training error $= 0$, then generalization gap $\lesssim \epsilon^2$.
- This is much *faster* than generalization gap $\lesssim \epsilon$ (Hardt et al., 2016).

Stability of BSGMs

- Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}$. Construct $S^{(i)}$.
- Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced by BSGMs based on S and $S^{(i)}$.
- Recall the definition of bias:

$$b_t = \mathbb{E}_{j_t}[g(\mathbf{w}_t; S_{j_t})] - \nabla \widetilde{F}_S(\mathbf{w}_t) \quad \text{and} \quad b_t^{(i)} = \mathbb{E}_{j_t}[g(\mathbf{w}_t^{(i)}; S_{j_t}^{(i)})] - \nabla \widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)}).$$

Generalized Lipschitzness assumption

$$\mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_t}) - g(\mathbf{w}_t^{(i)}; S_{j_t}^{(i)})\|_2^2\big] \leq \mathbb{E}\big[A\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + B_{t,i}\big], \tag{10}$$

$$\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] \leq \mathbb{E}\big[\bar{A}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{B}_{t,i}\big]. \tag{11}$$

- Lipschitzness plus an additional term B
- We will verify this assumption for instantiations of BSGMs

Stability of BSGMs

Main Result

Let $\tilde{f}$ be convex. Let generalized Lipschitzness assumption hold. If $\left(A + T\bar{A}\right) \sum_{t=1}^{T} \eta_t^2 \lesssim 1$, then BSGM is on-average model ϵ-stable

$$\epsilon^2 \lesssim \underbrace{\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \eta_t^2 \mathbb{E}[B_{t,i}]}_{:=C_1} + \underbrace{\frac{\sum_{t=1}^{T} \eta_t^2}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}[\bar{B}_{t,i}]}_{:=C_2}. \tag{12}$$

- It requires a convex surrogate $\tilde{f}$, which is NOT necessary f
- To apply it, we just need to check the generalized Lipschitzness assumption!
- If $\eta_t = \eta, \mathbb{E}[B_{t,i}] \leq B_1$ and $\mathbb{E}[\bar{B}_{t,i}] \leq B_2$, then

$$C_1 \leq T\eta^2 B_1, \quad C_2 \leq T^2\eta^2 B_2 \Longrightarrow \text{bias is more important}$$

Key Idea in Generalized Lipschitzness Assumption

- As we mentioned before, the bias can be bounded for various algorithms

algorithm	Clipped-SGD	Zeroth-SGD	Delayed-SGD	D-SGD
bias^2	τ^{2-2p}	$\mu^2 d^6$	τ^2	$\sum_{t=1}^{T} \eta_t^2$

- If we directly use these bounds on bias (assume bias$^2 \leq B$), then

$$\epsilon^2 \lesssim \text{some term} + T^2\eta^2 B.$$

- For good stability, it requires $T\eta \to 0$ for which the algorithm will NOT converge!

Key Idea

Instead of considering bias, we consider the difference of bias on neighboring datasets

$$\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] \leq \mathbb{E}\big[\bar{A}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{B}_{t,i}\big]. \tag{13}$$

- This allows us to establish Eq. (13) with $\bar{A} \ll B$ and $\bar{B}_{t,i} \ll B$
- Our stability bound depends on $\bar{A}$ and $\bar{B}_{t,i}$

Applications

Stochastic Gradient Descent

- SGD uses the estimator $g(\mathbf{w}_t; z_{j_t}) = \nabla f(\mathbf{w}_t; z_{j_t})$.
- We choose $\tilde{f}(\mathbf{w}; z) = f(\mathbf{w}; z)$. Then $b_t = 0$.

Generalized Lipschitzness condition

Assume f is L-smooth. Then, the generalized Lipschitzness assumption holds with $A = L^2, B_{t,i} = \|\nabla f(\mathbf{w}_t; z_i)\|^2/n, \bar{A} = \bar{B} = 0$. That is,

$$\mathbb{E}\left[\|\nabla f(\mathbf{w}_t; S_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; S_{j_t}^{(i)})\|_2^2\right] \leq \mathbb{E}\left[L^2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \|\nabla f(\mathbf{w}_t; z_i)\|^2/n\right]. \tag{14}$$

Stability bounds

Let f be L-smooth and convex. Then SGD with $\eta_t = \eta$ is on-average model ϵ-stable:

$$\epsilon^2 \lesssim \left(\frac{\eta^2}{n} + \frac{T\eta^2}{n^2}\right) \sum_{t=1}^{T} \mathbb{E}[F_S(\mathbf{w}_t)]. \tag{15}$$

- It recovers the existing stability analysis of SGD (Hardt et al., 2016; Lei and Ying, 2020)

Zeroth-order SGD

- Zeroth-order SGD takes (μ is a smooth parameter)

$$g(\mathbf{w}_t; z_{j_t}) = \frac{1}{K} \sum_{l=1}^{K} \frac{f(\mathbf{w}_t + \mu \mathbf{u}_{t,l}; z_{j_t}) - f(\mathbf{w}_t; z_{j_t})}{\mu} \mathbf{u}_{t,l}, \quad \mathbf{u}_{t,l} \sim \mathcal{N}(0, I_d).$$

- If we choose $\tilde{f} = f$, then there is a bias which leads to a suboptimal stability bound.
- Our key idea is to consider

$$\tilde{f}(\mathbf{w}; z) = \mathbb{E}_{\mathbf{u}}\big[f(\mathbf{w} + \mu \mathbf{u}; z)\big]. \tag{16}$$

Let $\tilde{f}$ be defined in Eq. (16). (Nesterov and Spokoiny, 2017)

- If $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex and smooth, then $\mathbf{w} \mapsto \tilde{f}(\mathbf{w}; z)$ is also convex and smooth.
- We have $\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}; z_{j_t})] = \nabla \tilde{f}(\mathbf{w}; z_{j_t})$ for any $\mathbf{w} \in \mathcal{W}$ and $j \in [n]$. That is, $b_t = 0$.

Zeroth-order SGD: Stability Bounds

Generalized Lipschitzness condition

Assume f is L-smooth. Then, the generalized Lipschitzness assumption holds with $A = (1 + d/K)L^2$, $\bar{A} = \bar{B}_{t,i} = 0$,

$$B_{t,i} \lesssim \frac{d\|\nabla f(\mathbf{w}_t; z_i)\|_2^2}{Kn} + \frac{\mu^2 L^2 d^3}{K} + \frac{1}{n}\|\nabla \tilde{f}(\mathbf{w}_t; z_i)\|_2^2.$$

Stability bounds

Let f be convex and smooth. If $\sum_{t=1}^{T} \eta_t^2 \lesssim 1$, then Zeroth-order SGD is on-average model ϵ-stable

$$\epsilon^2 \lesssim \left(\frac{\eta^2}{n} + \frac{T\eta^2}{n^2}\right) \sum_{t=1}^{T} \mathbb{E}[F_S(\mathbf{w}_t)] + \mu^2 d^3 \left(\frac{1}{n} + \frac{1}{K}\right) T\eta^2 + \frac{\mu^2 d^3 T^2 \eta^2}{n^2}. \tag{17}$$

- We require $\sum_{t=1}^{T} \eta_t^2 \lesssim 1$, which is satisfied by the standard choice $\eta_t \asymp 1/\sqrt{T}$
- The existing stability analysis requires a fast-decaying $\eta_t \lesssim 1/t$ (Nikolakakis et al., 2022)

Zeroth-order SGD: Excess Risk Bounds

Convergence Analysis

Let f be L-smooth and convex, $\eta_t = \eta \lesssim 1/L$ and $\mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$, then

$$\mathbb{E}[F_S(\mathcal{A}(S)) - F_S(\mathbf{w}^*)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\left(\frac{Ld}{K} + \mu^2 L^2 d^3\right) \tag{18}$$

Excess Risk Bounds

Let f be L-smooth, convex and take $T \asymp n, K \asymp d, \eta \asymp 1/\sqrt{T}$. Then

$$\mathbb{E}[F(\mathcal{A}(S))] - F(\mathbf{w}^*) \lesssim 1/\sqrt{n}. \tag{19}$$

Clipped-SGD

Clipped-SGD uses the estimator

$$g(\mathbf{w}_t; z_{j_t}) = \mathrm{clip}(\nabla f(\mathbf{w}_t; z_{j_t}), \tau). \tag{20}$$

Moment assumption on stochastic gradient

$$\mathbb{E}_{j_t}[\|\nabla f(\mathbf{w}; z_{j_t})\|_2^p] \leq G^p, \quad \text{where } p \in (1, 2]. \tag{21}$$

Generalized Lipschitzness on bias

Let moment assumption hold and f be smooth. Then

$$\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] \leq \frac{G^{2p}}{\tau^{2p}} \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \|\nabla f(\mathbf{w}_t; z_i)\|_2^2 / n^2\big]. \tag{22}$$

- Much better than the existing bias $b_t^2 \lesssim G^{2p}\tau^{2-2p}$ in Zhang et al. (2020)

$$G^{2p}\tau^{-2p} \ll G^{2p}\tau^{2-2p} \quad \text{and} \quad G^{2p}\tau^{-2p}n^{-2} \ll G^{2p}\tau^{2-2p}$$

Clipped-SGD: Stability Bounds

Generalized Lipschitzness on gradient

Let f be L-smooth. Then

$$\mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_t}) - g(\mathbf{w}_t^{(i)}; S_{j_t}^{(i)})\|_2^2\big] \leq L^2 \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \|\nabla f(\mathbf{w}_t; z_i)\|_2^2/n\big]. \tag{23}$$

Stability bounds of Clipped-SGD

Let moment assumption hold. Let f be convex and smooth. If $\left(1 + \frac{T}{\tau^{2p}}\right) \sum_{t=1}^{T} \eta_t^2 \lesssim 1$, then Clipped-SGD is on-average model ϵ-stable

$$\epsilon^2 \lesssim \left(\frac{\eta^2}{n} + \frac{T\eta^2}{n^2}\right) \sum_{t=1}^{T} \mathbb{E}[F_S(\mathbf{w}_t)]. \tag{24}$$

- This matches the stability bounds of the standard SGD!

The first stability analysis of Clipped-SGD!

Clipped-SGD: Excess Risk Bounds

Convergence Analysis

Let moment assumption hold. Let f be convex and smooth. If $\eta_t = \eta \leq 1/(3L)$ and $G \lesssim \tau$, then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t) - F_S(\mathbf{w}_S)\big] \lesssim \frac{\mathbb{E}\big[\|\mathbf{w}_S\|_2^2\big]}{T\eta} + G^p\tau^{2-p}\eta + G^{2p}\tau^{2-2p}T\eta. \tag{25}$$

Excess Risk Bounds

Let the moment assumption hold. Let f be convex and smooth. If we take $\tau \asymp GT^{\frac{1}{p}}, \eta \asymp n^{-\frac{1}{2p-2}}, T \asymp n^{\frac{p}{2p-2}}$ and $\mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$, then

$$\mathbb{E}[F(\mathcal{A}(S))] - F(\mathbf{w}^*) \lesssim 1/\sqrt{n}. \tag{26}$$

- $\tau \asymp GT^{\frac{1}{p}}, \eta \asymp T^{-\frac{1}{p}}$ is also the choice in optimization (Nguyen et al., 2023)

Decentralized SGD

- We have m local machines with $S_k = \{z_{1,k}, z_{2,k}, \ldots, z_{n,k}\}$ in the k-th machine
- Each local machine updates its own model and communicates with its neighbors

$$\mathbf{w}_t^k = \sum_{j=1}^{m} P_{kj}\mathbf{w}_t^j - \eta_t \nabla f(\mathbf{w}_t^k; z_{j_t^k,k}), \quad P \in \mathbb{R}^{m\times m}, j_t^k \sim \text{Unif}[n].$$

Stability Bounds

Let f be convex and L-smooth. Then D-SGD with $\eta_t = \eta \lesssim 1/L$ is ϵ-model stable with

$$\epsilon^2 \lesssim \left(\frac{\eta^3}{mn(1-\lambda)^2} + \frac{T\eta^2}{m^2n^2}\right)\sum_{t=1}^{T}\mathbb{E}[F_S(\bar{\mathbf{w}}_t)], \tag{27}$$

where λ is the second largest singular value of P and $\bar{\mathbf{w}}_t = \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_t^k$.

Excess Risk Bounds

Let f be convex and L-smooth. Let $\eta \asymp 1/\sqrt{T}$ and $T \asymp mn$, then

$$\mathbb{E}[F(\mathcal{A}(S))] - F(\mathbf{w}^*) \lesssim \frac{1}{(1-\lambda)^2 mn} + \frac{1}{\sqrt{mn}}, \text{ where } \mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^{T}\bar{\mathbf{w}}_t. \tag{28}$$

Conclusion

Summary

Stability analysis of BSGMs

- We introduce generalized Lipschitzness assumption
- We develop the first general framework on the stability of BSGMs

Applications

- Zeroth-order SGD
  - We build a surrogate function to get zero bias
  - We get improved stability bounds allowing much larger step sizes
- Clipped-SGD
  - We show the bias satisfies an improved generalized Lipschitzness assumption
  - We develop the first stability analysis of Clipped-SGD
- Decentralized-SGD: we imply optimal risk bounds of order $\frac{1}{\sqrt{mn}}$ if $1-\lambda \gtrsim \frac{1}{\sqrt{mn}}$

Future directions

- Extension to nonconvex problems

Thank you!

References I

A. Ajalloeian and S. U. Stich. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Y. Hu, S. Zhang, X. Chen, and N. He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33:2759–2770, 2020.

I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

K. Nikolakakis, F. Haddadpour, D. Kalogerias, and A. Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems*, 35:31525–31541, 2022.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

S. Zheng, Q. Meng, T. Wang, W. Chen, N. Yu, Z.-M. Ma, and T.-Y. Liu. Asynchronous stochastic gradient descent with delay compensation. In *International Conference on Machine Learning*, pages 4120–4129. PMLR, 2017.

T. Zhu, F. He, K. Chen, M. Song, and D. Tao. Decentralized sgd and average-direction sam are asymptotically equivalent. In *International Conference on Machine Learning*, pages 43005–43036. PMLR, 2023.