Generalization Guarantee of SGD for Pairwise Learning

Yunwen Lei[1], Mingrui Liu[2] and Yiming Ying[3]

[1]University of Birmingham
[2]George Mason University
[3]State University of New York at Albany

y.lei@bham.ac.uk mingruil@gmu.edu yying@albany.edu

NeurIPS 2021

Pairwise Learning

- **Data**: $S = \{z_i = (x_i, y_i)\}_{i=1}^n \sim \rho$ defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- We learn a model $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$ or $h_{\mathbf{w}} : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}, \mathbf{w} \in \mathcal{W}$
- Pairwise loss: $f(\mathbf{w}; z, z')$ measures behavior of $h_{\mathbf{w}}$ over z, z'
- Population risk and Empirical risk

$$F(\mathbf{w}) = \mathbb{E}_{z,z'}\big[f(\mathbf{w}; z, z')\big], \quad F_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} f(\mathbf{w}; z_i, z_j).$$

- Risk Minimizer $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$
- Algorithm: $A : \mathcal{Z}^n \mapsto \mathcal{W}$ (output $A(S)$ when applied to S)

We are interested in studying the excess risk $F(A(S)) - F(\mathbf{w}^*)$!

Error Decomposition and SGD

Error decomposition:

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

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

Stochastic Gradient Descent (SGD)

SGD$(S, T, f, \{\eta_t\})$

for $t = 1, 2, \ldots$ **to** T **do**

  draw (i_t, j_t) uniformly over all pairs $\{(i, j) : i, j \in [n], i \neq j\}$

  $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t})$ for some step sizes $\eta_t > 0$

return $\mathbf{w}_{T+1}$ or an average of $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$

Definitions

Let $g : \mathcal{W} \mapsto \mathbb{R}$ (the following needs to hold for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$).

Smoothness We say g is L-smooth if

$$\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2.$$

Lipschitzness We say g is G-Lipschitz continuous if

$$|g(\mathbf{w}) - g(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|_2.$$

Convexity We say g is convex if

$$g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle.$$

Algorithmic Stability and Generalization

Algorithmic Stability

Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}$ be independently drawn from ρ. We denote

$$S_i = \big\{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\big\}, \quad \forall i \in [n].$$

1. We say A is ϵ-uniformly stable if for any datasets $S, \widetilde{S} \in \mathcal{Z}^n$ that differ by at most a single example we have $\sup_{z,z' \in \mathcal{Z}} \big|f(A(S); z, z') - f(A(\widetilde{S}); z, z')\big| \leq \epsilon$.
2. We say A is on-average argument ϵ-stable if $\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S_i)\|_2^2\Big] \leq \epsilon^2$.

Connection Between Stability and Generalization

1. If A is on-average argument ϵ-stable and f is smooth, then **in expectation** we have

$$\mathbb{E}[F(A(S)) - F_S(A(S))] = O(\epsilon^2 + \epsilon\sqrt{\mathbb{E}[F_S(A(S))]}).$$

2. If A is ϵ-uniformly stable and $\sigma_0^2 := \mathbb{E}_{Z,Z',S}\big[\big(f(A(S); Z, Z') - f(\mathbf{w}^*; Z, Z')\big)^2\big]$, then **with high probability** we have (Klochkov and Zhivotovskiy, 2021)

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) = \widetilde{O}\Big(\epsilon + \frac{1}{n} + \frac{\sigma_0}{\sqrt{n}}\Big).$$

SGD for Pairwise Learning: Convex and Smooth Cases

Stability Bounds

Let f be convex and L-smooth. Then SGD with T iterations is on-average argument ϵ-stable with

$$\epsilon^2 = O\Big(\frac{1}{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}[F_S(\mathbf{w}_t)]\Big).$$

Excess Generalization Bounds

Let f be convex and L-smooth. Then for SGD with $\eta_t = \eta$ and $T \asymp n$ we have

$$\mathbb{E}[F(\bar{\mathbf{w}}_T) - F_S(\mathbf{w})] = O\Big(\Big(\frac{1}{\gamma} + \gamma\eta^2\Big)\mathbb{E}[F_S(\mathbf{w})]\Big) + O\Big(\frac{1}{T\eta} + \frac{\gamma\eta}{n}\Big), \ \forall\gamma \geq 1.$$

1. We can choose $\eta \asymp 1/\sqrt{T}$ to get $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/\sqrt{n})$.
2. If $F(\mathbf{w}^*) = O(1/n)$, choosing $\eta = 2/L$ yields $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/n)$.

Our stability bounds involve training errors and get improved if training errors are small!

SGD for Pairwise Learning: Convex and Nonsmooth Cases

Stability and Excess Generalization Bounds

Let f be convex and G-Lipschitz. Then SGD with T iterations is ϵ-uniformly stable with $\epsilon = O(\sqrt{T}\eta)$. Furthermore, we can choose $\eta \asymp T^{-\frac{3}{4}}$ and $T \asymp n^2$ to get

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/\sqrt{n}).$$

- To achieve the desired bound $O(1/\sqrt{n})$, SGD requires $O(n^2)$ iterations for nonsmooth problems
- To decrease the computation cost, we develop **Iterative Localized Algorithm for Pairwise Learning** (Feldman et al., 2020)

Iterative Localized Algorithm for Pairwise Learning

Iterative Localized Algorithm for Pairwise Learning

Input: initial point $\mathbf{w}_0 = 0$, parameter $k = \lceil \frac{1}{2} \log_2 n \rceil$
for $i = 1, 2, \ldots, k$ **do**

set $T_i \asymp n_i = \lceil \frac{n}{2^i} \rceil, \gamma_i = \frac{1}{2^i \sqrt{n}}, \eta_t = \frac{\gamma_i n_i}{t+1}, \tilde{f}(\mathbf{w}; z, z') = f(\mathbf{w}; z, z') + \frac{1}{\gamma_i n_i} \|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2$
draw a sample S_i of size n_i independently from ρ
apply $\mathrm{SGD}(S_i, T_i, \tilde{f}, \{\eta_t\})$ to minimize the following problem and get $\mathbf{w}_i$

$$\widetilde{F}_{S_i}(\mathbf{w}) := \frac{1}{n_i(n_i - 1)} \sum_{z, z' \in S_i : z \neq z'} f(\mathbf{w}; z, z') + \frac{1}{\gamma_i n_i} \|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2. \tag{1}$$

Excess Generalization Bounds

Let f be convex and Lipschitz. Then with high probability $F(\mathbf{w}_k) - F(\mathbf{w}^*) = \widetilde{O}(1/\sqrt{n})$. Furthermore, it requires $O(n)$ gradient computations.

- The existing iterative localized algorithm works for pointwise learning and only leads to bounds **in expectation**.
- We derive the first $\widetilde{O}(1/\sqrt{n})$ **high-probability** bounds with $O(n)$ complexity based on algorithmic stability.

SGD for Pairwise Learning: Nonconvex and Smooth Case

Learning Rates

Let f be smooth and the variance be bounded. Consider SGD with $\eta_t = 1/\sqrt{T}$ and $T \asymp n/d$. With high probability $\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_t)\|_2^2 = O(\sqrt{d/n})$.

For nonconvex problems, we consider a different error decomposition

$$\|\nabla F(\mathbf{w}_t)\|_2^2 \leq 2\underbrace{\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2^2}_{\text{estimation error}} + 2\underbrace{\|\nabla F_S(\mathbf{w}_t)\|_2^2}_{\text{optimization error}}.$$

- We show with high probability $\|\mathbf{w}_t\|_2 \leq R_T := O(T^{\frac{1}{4}})$ if $t \leq T$.
- We use **uniform convergence of gradients** to control estimation error

$$\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2 \leq \sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq R_T} \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2 = O(R_T\sqrt{d/n}).$$

- With high probability, the optimization error satisfies

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_t)\|_2^2 = O\big(\sqrt{T}d/n + 1/\sqrt{T}\big).$$

References I

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate O(1/n). *arXiv preprint arXiv:2103.12024*, 2021.

Thank you!