Fine-Grained Analysis of Stability and Generalization for SGD

Yunwen Lei[1] and Yiming Ying[2]

[1]University of Kaiserslautern
[2]University at Albany, State University of New York (SUNY)

yunwen.lei@hotmail.com yying@albany.edu

June, 2020

Overview

Population and Empirical Risks

- Training Dataset: $S = \{z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ with each example $z_i \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Parametric model $\mathbf{w} \in \Omega \subseteq \mathbb{R}^d$ for prediction
- Loss function: $f(\mathbf{w}; z)$ measure performance of $\mathbf{w}$ on an example z
- Population risk: $F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)]$ with best model

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \Omega} F(\mathbf{w})$$

- Empirical risk: $F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{w}; z_i)$.

Excess Generalization Error

Based on the training data S, a randomized algorithm denoted by A (e.g. SGD) outputs a model $A(S) \in \Omega$ ...

- Target of analysis: excess generalization error

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

- Vast literature on optimization error: (Duchi et al., 2011; Bach and Moulines, 2011; Rakhlin et al., 2012; Shamir and Zhang, 2013; Orabona, 2014; Ying and Zhou, 2017; Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018; Bassily et al., 2018; Vaswani et al., 2019; Mücke et al., 2019) and many others

- Algorithmic stability for studying estimation error: (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Kuzborskij and Lampert, 2018; Charles and Papailiopoulos, 2018; Feldman and Vondrak, 2018) etc.

Uniform Stability Approach

Uniform Stability (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005)

A randomized algorithm A is ϵ-uniformly stable if, for any two datasets S and S' that differ by one example, we have

$$\sup_{z} \mathbb{E}_A\big[f(A(S);z) - f(A(S');z)\big] \leq \epsilon_{\text{uniform}}. \tag{1}$$

- For G-Lipschitz, strongly smooth f, SGD with step size η_t informally we have

$$\text{Generalization} \leq \text{Uniform stability} \leq \frac{1}{n}\sum_{t=1}^{T} \eta_t G^2.$$

- These assumptions are restrictive: they are not true for q-norm loss $f(\mathbf{w};z) = |y - \langle \mathbf{w}, x\rangle|^q$ ($q \in [1,2]$) and hinge loss $(1 - y\langle \mathbf{w}, x\rangle)_+$ with $\mathbf{w} \in \mathbb{R}^d$.

Can we remove these assumptions and explain the real power of SGD?

Our Results

On-Average Model Stability

To handle the general setting, we propose a new concept of stability.
Let $S = \{z_i : i = 1, \ldots, n\}$ and $\widetilde{S} = \{\tilde{z}_i : i = 1, \ldots, n\}$, and for each i, let $S^{(i)} = \{z_1, \ldots, z_{i-1}, \tilde{z}_i, z_{i+1}, \ldots, z_n\}$.

On-Average Model Stability

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

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|A(S) - A(S^{(i)})\|_2^2\Big] \le \epsilon^2. \tag{2}$$

- α-Hölder continuous gradients ($\alpha \in [0, 1]$)

$$\big\|\partial f(\mathbf{w}, z) - \partial f(\mathbf{w}', z)\big\|_2 \le \|\mathbf{w} - \mathbf{w}'\|_2^{\alpha}. \tag{3}$$

  $\alpha = 0$ means that f is Lipschitz and $\alpha = 1$ means strongly smoothness.

- If A is on-average model ϵ-stable,

$$\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] = O\Big(\epsilon^{1+\alpha} + \epsilon\big(\mathbb{E}[F_S(A(S))]\big)^{\frac{\alpha}{1+\alpha}}\Big). \tag{4}$$

- Can handle both Lipschitz functions and un-bounded gradient!

Case Study: Stochastic Gradient Descent

We study the on-average model stability ϵ_{T+1} of $\mathbf{w}_{T+1}$ from SGD ...

SGD

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

  $i_t \leftarrow$ random index from $\{1, 2, \ldots, n\}$

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

return $\mathbf{w}_{T+1}$

On-Average Model Stability for SGD

- If ∂f is α-Hölder continuous with $\alpha \in [0, 1]$, then

$$\epsilon_{T+1}^2 = O\Big(\sum_{t=1}^{T} \eta_t^{\frac{2}{1-\alpha}} + \frac{1 + T/n}{n}\big(\sum_{t=1}^{T} \eta_t^2\big)^{\frac{1-\alpha}{1+\alpha}} \Big(\sum_{t=1}^{T} \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)]\Big)^{\frac{2\alpha}{1+\alpha}}\Big)$$

- *Weighted sum of risks* (i.e. $\sum_{t=1}^{T} \eta_t^2 \mathbb{E}\big[F_S(\mathbf{w}_t)\big]$) can be estimated using tools of analyzing optimization errors

Main Results for SGD

Our Key Message (Informal)

Generalization $\leq$ On-average model stability $\leq$ Weighted sum of risks

Recall, for uniform stability with Lipschitz and smooth f, that

$$\text{Generalization} \leq \text{Uniform stability} \leq \frac{1}{n}\sum_{t=1}^{T}\eta_t G^2$$

Specifically, we have the following excess generalization bounds...

SGD with Smooth Functions

Let f be convex and strongly-smooth. Let $\bar{\mathbf{w}}_T = \sum_{t=1}^T \eta_t \mathbf{w}_t / \sum_{t=1}^T \eta_t$.

Theorem (Minimax optimal generalization bounds)

Choosing $\eta_t = 1/\sqrt{T}$ and $T \asymp n$ implies that

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

Theorem (Fast generalization bounds under low noise)

For low noise case $F(\mathbf{w}^*) = O(1/n)$, we can take $\eta_t = 1$, $T \asymp n$ and get

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

- We remove bounded gradient assumptions.
- We get the first-ever fast generalization bound $O(1/n)$ by stability analysis.

SGD with Lipschitz Functions

Let f be convex and G-Lipschitz (Not necessarily smooth! e.g. the hinge loss.)

Our on-average model stability bounds can be simplified as

$$\epsilon_{T+1}^2 = O\Big(\big(1 + T/n^2\big) \sum_{t=1}^{T} \eta_t^2\Big). \tag{5}$$

Key idea: gradient update is approximately contractive

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z) - \mathbf{w}' + \eta\partial f(\mathbf{w}'; z)\|_2^2 \leq \|\mathbf{w} - \mathbf{w}'\|_2^2 + O(\eta^2). \tag{6}$$

Theorem (Generalization bounds)

We can take $\eta_t = T^{-\frac{3}{4}}$ and $T \asymp n^2$ and get

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

We get the first generalization bound $O(1/\sqrt{n})$ for SGD with non-differentiable functions based on stability analysis.

SGD with α-Hölder continuous gradients

Let f be convex and have α-Hölder continuous gradients with $\alpha \in (0,1)$.

Key idea: gradient update is approximately contractive

$$\|\mathbf{w} - \eta \partial f(\mathbf{w}; z) - \mathbf{w}' + \eta \partial f(\mathbf{w}'; z)\|_2^2 \leq \|\mathbf{w} - \mathbf{w}'\|_2^2 + O(\eta^{\frac{2}{1-\alpha}}).$$

Theorem

- *If* $\alpha \geq 1/2$, *we take* $\eta_t = 1/\sqrt{T}$, $T \asymp n$ *and get*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

- *If* $\alpha < 1/2$, *we take* $\eta_t = T^{\frac{3\alpha-3}{2(2-\alpha)}}$, $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ *and get*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

Theorem (Fast Generalization bounds)

If $F(\mathbf{w}^*) = O(\frac{1}{n})$, *we let* $\eta_t = T^{\frac{\alpha^2+2\alpha-3}{4}}$, $T \asymp n^{\frac{2}{1+\alpha}}$ *and get* $\mathbb{E}[F(\bar{\mathbf{w}}_T)] = O(n^{-\frac{1+\alpha}{2}})$.

SGD with Relaxed Convexity

We assume f is G-Lipschitz continuous.

Non-convex f but convex F_S

- stability bound: $\epsilon^2 \leq \frac{1}{n^2}\big(\sum_{t=1}^{T}\eta_t\big)^2 + \frac{1}{n}\sum_{t=1}^{t}\eta_t^2$.
- generalization bound: if $\eta_t = 1/\sqrt{T}$ and $T \asymp n$, then

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

Non-convex f but strongly-convex F_S $(\eta_t = 1/t)$

- stability bound: $\epsilon^2 \leq \frac{1}{nT} + \frac{1}{n^2}$.
- generalization bound: if $T \asymp n$, then

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

- example: least squares regression.

References I

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

R. Bassily, M. Belkin, and S. Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

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

Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 744–753, 2018.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6 (Jan):55–79, 2005.

V. Feldman and J. Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

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.

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

J. Lin and L. Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(1): 3375–3421, 2017.

N. Mücke, G. Neu, and L. Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

References II

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224—-244, 2017.

Thank you!