# Sharper Generalization Bounds for Learning with Gradient-dominated Objective Functions

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

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

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

February 23, 2021

# Population and Empirical Risks

- Training Dataset: $S = \{z_1, \ldots, z_n\}$ with each example $z_i \in \mathcal{Z}$
- Parametric model $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ for prediction
- Loss function: $f(\mathbf{w}; z)$ measure performance of $\mathbf{w}$ on an example $z$
- Population risk: $F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)]$
- Empirical risk: $F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{w}; z_i)$.
- Algorithm $A : \mathcal{Z}^n \mapsto \mathcal{W}$ (output $A(S)$ when applied to $S$)

We are interested in Excess Generalization Error $F(A(S)) - \inf_{\mathbf{w}} F(\mathbf{w})$

# Assumptions

## Smoothness Assumption

We assume for all $z \in \mathcal{Z}$, the differentiable function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is $L$-smooth

$$\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}'; z)\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

## Polyak-Lojasiewicz (PL) Condition

We assume training errors are gradient-dominated (can be non-convex)

$$\mathbb{E}\big[F_S(\mathbf{w}) - \inf_{\mathbf{w}} F_S(\mathbf{w})\big] \leq \frac{1}{2\beta}\mathbb{E}\big[\|\nabla F_S(\mathbf{w})\|_2^2\big], \quad \forall \mathbf{w} \in \mathcal{W}. \tag{1}$$

We do not require bounded gradient assumption as $\|\nabla f(\mathbf{w}; z)\|_2 \leq G$!

# Main Results

**Theorem (Generalization bounds)**

Under PL condition and Smoothness Assumption

$$\mathbb{E}\big[F(A(S))\big] - \inf_{\mathbf{w}} F(\mathbf{w}) \leq \frac{\inf_{\mathbf{w}} F_S(\mathbf{w})}{n\beta} + \frac{F_S(A(S)) - \inf_{\mathbf{w}} F_S(\mathbf{w})}{\beta}.$$

- $F_S(A(S)) - \inf_{\mathbf{w}} F_S(\mathbf{w})$ is the optimization error
- It applies to any algorithm: SGD, SVRG, ADAM...
- Optimization helps generalization: run $A$ until optimization error $\leq 1/n$
- It significantly improves the existing results (Charles and Papailiopoulos, 2018)

$$\mathbb{E}\big[F(A(S))\big] - \inf_{\mathbf{w}} F(\mathbf{w}) \leq \frac{1}{\sqrt{n\beta}} + \sqrt{\frac{F_S(A(S)) - \inf_{\mathbf{w}} F_S(\mathbf{w})}{\beta}}.$$

- If $\inf_{\mathbf{w}} F_S(\mathbf{w}) = 0$, then it achieves bounds better than $1/(n\beta)$

# Applications to Specific Algorithms

| Algorithm | Complexity for $1/(n\beta)$ |
|---|---|
| SGD | $\frac{n}{\beta^2}$ |
| RCD | $\frac{d \log n}{\beta}$ |
| SVRG, SCSG | $\left(n + n^{\frac{2}{3}}/\beta\right) \log n$ |
| SARAH, SpiderBoost | $\left(n + 1/\beta^2\right) \log n$ |
| SNVRG | $\left(n + \sqrt{n}/\beta\right) \log^4 n$ |

Iteration complexity for different optimization algorithms to get $\mathbb{E}\big[F(A(S))\big] - \inf_{\mathbf{w}} F(\mathbf{w}) \leq 1/(n\beta)$.

- SGD: Stochastic Gradient Descent
- RCD: Randomized Coordinate Descent (Nesterov, 2012)
- SVRG: Stochastic Variance Reduction Gradient (Johnson and Zhang, 2013)
- SARAH: StochAstic Recursive grAdient algoritHm (Nguyen et al., 2017)
- SNVRG: Stochastic Nested Variance-Reduced Gradient descent (Zhou et al., 2018)

# References I


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.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22 (2):341–362, 2012.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3921–3932. 2018.