Stability and Generalization of Stochastic Gradient Methods for Minimax Problems

Yunwen Lei[*,1], Zhenhuan Yang[*,2], Tianbao Yang[3] and Yiming Ying[2]

[1]University of Birmingham
[2]University at Albany, SUNY
[3]University of Iowa

ICML 2021

Background

Minimax Problems

- A probability measure $\mathbb{P}$ on a sample space $\mathcal{Z}$ and $S = \{z_1, \ldots, z_n\} \sim \mathbb{P}$
- Minimax formulation (e.g. GAN, AUC maximization, and robust learning):

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w}, \mathbf{v}) := \mathbb{E}_{z\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{v}; z)]. \tag{1}$$

- In practice, a randomized optimization algorithm A (e.g. SGDA) is employed to solve its empirical version:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \mathbf{v}; z_i). \tag{2}$$

- The literature is vast and most of them focused on the convergence of the ouput of A, i.e. $A(S) = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$.

How is the statistical generalization of minimax optimization algorithms?

How to Define Generalization?

Three key terms: population risk, empirical risk and generalization error

Definition (Generalization Error for ERM)

- ERM problem: $\min_{\mathbf{w}\in\mathcal{W}}\{R_S(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v})\}$
- Primal population risk: $R(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w},\mathbf{v})$
- Primal generalization error: $R(\mathbf{w}) - R_S(\mathbf{w})$

How to Define Generalization?

Three key terms: population risk, empirical risk and generalization error

Definition (Generalization Error for ERM)

- ERM problem: $\min_{\mathbf{w}\in\mathcal{W}}\big\{R_S(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v})\big\}$
- Primal population risk: $R(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w},\mathbf{v})$
- Primal generalization error: $R(\mathbf{w}) - R_S(\mathbf{w})$

Definition (Weak Primal-Dual (PD) Generalization Error)

- Weak PD population risk:
  $\triangle^w(\mathbf{w},\mathbf{v}) = \sup\limits_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_A\big[F(\mathbf{w},\mathbf{v}')\big] - \inf\limits_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_A\big[F(\mathbf{w}',\mathbf{v})\big].$
- Weak PD empirical risk:
  $\triangle^w_S(\mathbf{w},\mathbf{v}) = \sup\limits_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_A\big[F_S(\mathbf{w},\mathbf{v}')\big] - \inf\limits_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_A\big[F_S(\mathbf{w}',\mathbf{v})\big].$
- Weak PD generalization error: $\triangle^w(\mathbf{w},\mathbf{v}) - \triangle^w_S(\mathbf{w},\mathbf{v})$.

How to Define Generalization?

Three key terms: population risk, empirical risk and generalization error

Definition (Generalization Error for ERM)

- ERM problem: $\min_{\mathbf{w}\in\mathcal{W}}\{R_S(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v})\}$
- Primal population risk: $R(\mathbf{w}) := \sup_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w},\mathbf{v})$
- Primal generalization error: $R(\mathbf{w}) - R_S(\mathbf{w})$

Definition (Weak Primal-Dual (PD) Generalization Error)

- Weak PD population risk:
  $\triangle^w(\mathbf{w},\mathbf{v}) = \sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_A\big[F(\mathbf{w},\mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_A\big[F(\mathbf{w}',\mathbf{v})\big]$.
- Weak PD empirical risk:
  $\triangle^w_S(\mathbf{w},\mathbf{v}) = \sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_A\big[F_S(\mathbf{w},\mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_A\big[F_S(\mathbf{w}',\mathbf{v})\big]$.
- Weak PD generalization error: $\triangle^w(\mathbf{w},\mathbf{v}) - \triangle^w_S(\mathbf{w},\mathbf{v})$.

Definition (Strong PD Generalization Error)

- Strong population risk: $\triangle^s(\mathbf{w},\mathbf{v}) = \mathbb{E}_A\Big[\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}',\mathbf{v})\Big]$.
- Strong PD empirical risk: $\triangle^s_S(\mathbf{w},\mathbf{v}) = \mathbb{E}_A\Big[\sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}',\mathbf{v})\Big]$.

Summary of Our Contributions

We have systematically studied the generalization of minimax algorithms under general settings (e.g., different generalization measures, convex-concave, nonconvex-nonconcave, and nonsmooth cases).

Alg.	Reference	Assumption	Measure
ESP	Zhang et al. (2020)	SC-SC	PD Risk
SGDA	Farnia & Ozdaglar (2020)	Smooth	Weak PD Generalization
SGDA	This work	Convex-Concave (Smooth)	Weak PD Risk
		Convex, Smooth Strongly-Concave	Primal Risk
		SC-SC	Weak PD Risk
		Weakly-Convex-Weakly-Concave	Weak PD Generalization
AGDA	This work	Strongly-Concave, PL, Smooth	Primal Risk

- SC-SC: Strongly-Convex-Strongly-Concave

Technical Tools/Concepts

The goal of generalization analysis is to analyze the population risk...

- Error decomposition

$$\triangle^s(\mathbf{w},\mathbf{v}) = \underbrace{\left(\triangle^s(\mathbf{w},\mathbf{v}) - \triangle^s_S(\mathbf{w},\mathbf{v})\right)}_{\text{generalization error}} + \underbrace{\triangle^s_S(\mathbf{w},\mathbf{v})}_{\text{optimization error}} . \tag{3}$$

- Vast literature on optimization error: (Nemirovski et al., 2009; Nedić and Ozdaglar, 2009; Balamurugan and Bach, 2016; Hsieh et al., 2019; Rafique et al., 2018; Lin et al., 2020; Luo et al., 2020; Yan et al., 2020; Yang et al., 2020; Loizou et al., 2020; Liu et al., 2020) and many others
- Generalization $\leq$ Stability: studies of stability of SGD for ERM (**????**)

  ...

Stability Concepts for Minimax Algorithms

Algorithmic Stability for Minimax Algorithms

Neighboring datasets: $S, S' \subset \mathcal{Z}$ differing by at most a single example.

Algorithmic Stability

Let A be a randomized algorithm, $\epsilon > 0$ and $\delta \in (0,1)$.

- A is ϵ-weakly-stable if for all neighboring S and S', there holds

$$\sup_{z} \Big(\sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}_A \big[f(A_{\mathbf{w}}(S), \mathbf{v}'; z) - f(A_{\mathbf{w}}(S'), \mathbf{v}'; z) \big] + \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}_A \big[f(\mathbf{w}', A_{\mathbf{v}}(S); z) - f(\mathbf{w}', A_{\mathbf{v}}(S'); z) \big] \Big) \leq \epsilon.$$

- A is ϵ-argument-stable in expectation if for all neighboring S and S'

$$\mathbb{E}_A \left[\left\| \begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S') \end{pmatrix} \right\|_2 \right] \leq \epsilon.$$

Assumptions

Lipschitz Assumption

Assume for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$, there holds

$$\left\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z)\right\|_2 \leq G \quad \text{and} \quad \left\|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z)\right\|_2 \leq G.$$

Smoothness Assumption

Assume for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$, there holds

$$\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}'; z)\end{pmatrix}\right\|_2 \leq L \left\|\begin{pmatrix}\mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}'\end{pmatrix}\right\|_2.$$

Definition (Convexity and Concavity)

1. $g : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$ is ρ-strongly-convex-strongly-concave (ρ-SC-SC) if

   $\mathbf{w} \mapsto g(\mathbf{w}, \mathbf{v})$ is ρ-strongly-convex, and $\mathbf{v} \mapsto g(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave.

2. g is convex-concave if g is 0-SC-SC. g is ρ-weakly-convex-weakly-concave (ρ-WC-WC) if $g + \frac{\rho}{2}\big(\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2\big)$ is convex-concave.

Stability Implies Generalization

Generalization error $\lesssim$ Stability for minimax algorithms ...

- Strong PD Generalization Error $\lesssim$ Argument-Stability if Strongly-Convex-Strongly-Concave

Stability Implies Generalization

Generalization error $\lesssim$ Stability for minimax algorithms ...

- Strong PD Generalization Error $\lesssim$ Argument-Stability if Strongly-Convex-Strongly-Concave
- Primal Generalization Error $\lesssim$ Argument-Stability if Strong-Concave
- Weak PD Generalization Error $\lesssim$ Weak Stability

Stability Implies Generalization

Generalization error $\lesssim$ Stability for minimax algorithms ...

- Strong PD Generalization Error $\lesssim$ Argument-Stability if Strongly-Convex-Strongly-Concave
- Primal Generalization Error $\lesssim$ Argument-Stability if Strong-Concave
- Weak PD Generalization Error $\lesssim$ Weak Stability
- High-probability bounds hold true...

Stability/Generalization of SGDA and AGDA

Stochastic Gradient Descent Ascent

Stochastic Gradient Descent Ascent (SGDA) (Nemirovski et al., 2009)

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

$$\begin{aligned} i_t &\leftarrow \text{ random index from } \{1, 2, \ldots, n\}, \\ \mathbf{w}_{t+1} &= \mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\ \mathbf{v}_{t+1} &= \mathrm{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big). \end{aligned}$$

return $(\mathbf{w}_{T+1}, \mathbf{v}_{T+1})$

Consider an average of SGDA iterates

$$\bar{\mathbf{w}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{w}_t}{\sum_{t=1}^T \eta_t} \quad \text{and} \quad \bar{\mathbf{v}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{v}_t}{\sum_{t=1}^T \eta_t}.$$

SGDA: Population Risks for Convex-Concave f

Weak PD Risks

- If f is Lipschitz, $\eta_t \asymp T^{-\frac{3}{4}}$ and $T \asymp n^2$, then $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(n^{-\frac{1}{2}})$.
- If f is Lipschitz, smooth, $\eta_t \asymp T^{-\frac{1}{2}}$, $T \asymp n$, then $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(n^{-\frac{1}{2}})$.
- If f is Lipschitz, ρ-SC-SC, $\eta_t = (\rho t)^{-1}$, $T \asymp n^2$, then $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \widetilde{O}(1/(n\rho))$.
- If f is Lipschitz, smooth, ρ-SC-SC, $\eta_t = (\rho t)^{-1}$, $T \asymp n$, then

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \widetilde{O}(1/(n\rho)).$$

Excess Primal Risk for ρ-Strongly-Concave $F(\mathbf{w}, \cdot)$

Assume f is Lipschitz, smooth, $\eta_t \asymp T^{-\frac{1}{2}}$ and $T \asymp n$.

- In expectation we have $\mathbb{E}\big[R(\bar{\mathbf{w}}_T)\big] - \inf_{\mathbf{w} \in \mathcal{W}} R(\mathbf{w}) = O((L/\rho)n^{-\frac{1}{2}})$.
- With probability $1 - \delta$, we have $R(\bar{\mathbf{w}}_T) - \inf_{\mathbf{w}} R(\mathbf{w}) = \widetilde{O}\big((L/\rho)n^{-\frac{1}{2}}\big)$.

SGDA: Nonconvex-Nonconcave Case

Weakly-Convex-Weakly-Concave

If $f(\mathbf{w}, \mathbf{v}; z)$ is ρ-WC-WC, Lipschitz and $\eta_t = c/t$, then

$$\text{weak PD generalization error} = O\Big(\Big(\Big(1 + \frac{\sqrt{T}}{n}\Big) T^{c\rho}\Big)^{\frac{2}{2c\rho+3}} \Big(\frac{1}{n}\Big)^{\frac{2c\rho+1}{2c\rho+3}}\Big).$$

- No smoothness assumption required.
- Not studied even for SGD.

Alternating Gradient Descent Ascent

Alternating Gradient Descent Ascent (AGDA)

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

$$
\begin{aligned}
i_t &\leftarrow \text{ random index from } \{1, 2, \ldots, n\}, \\
j_t &\leftarrow \text{ random index from } \{1, 2, \ldots, n\}, \\
\mathbf{w}_{t+1} &= \mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_{\mathbf{w},t} \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\
\mathbf{v}_{t+1} &= \mathrm{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_{\mathbf{v},t} \nabla_{\mathbf{v}} f(\mathbf{w}_{t+1}, \mathbf{v}_t; z_{j_t})\big).
\end{aligned}
$$

return $(\mathbf{w}_{T+1}, \mathbf{v}_{T+1})$

Alternating Gradient Descent Ascent

Two-sided PL condition (Yang et al., 2020)

Assume there exist constants $\beta_1, \beta_2 > 0$ such that

$$2\beta_1\big(F_S(\mathbf{w},\mathbf{v}) - \inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}',\mathbf{v})\big) \leq \|\nabla_{\mathbf{w}} F_S(\mathbf{w},\mathbf{v})\|_2^2,$$

$$2\beta_2\big(\sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v}') - F_S(\mathbf{w},\mathbf{v})\big) \leq \|\nabla_{\mathbf{v}} F_S(\mathbf{w},\mathbf{v})\|_2^2.$$

Excess Primal Risk Bounds

Assume f is Lipschitz, smooth, $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave, and the two-sided PL condition holds. Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ be produced by AGDA with appropriate step sizes. If $T \asymp (\frac{n}{\beta_1^2\rho^3})^{\frac{cL+1}{2cL+1}}$, then $(c \asymp 1/(\beta_1\rho^2))$

$$\mathbb{E}[R(\mathbf{w}_T)] - \inf_{\mathbf{w}} R(\mathbf{w}) = O\Big(n^{-\frac{cL+1}{2cL+1}}\beta_1^{-\frac{2cL}{2cL+1}}\rho^{-\frac{5cL+1}{2cL+1}}\Big).$$

References I

P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advance In Neural Information Processing Systems*, pages 1416–1424, 2016.

F. Farnia and A. Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. *arXiv preprint arXiv:2010.12561*, 2020.

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pages 6938–6948, 2019.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093, 2020.

M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems, 2020.

N. Loizou, H. Berard, A. Jolicoeur-Martineau, P. Vincent, S. Lacoste-Julien, and I. Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381, 2020.

L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

D. Richards and M. Rabbat. Learning with gradient descent and weakly convex losses. *arXiv preprint arXiv:2101.04968*, 2021.

Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016.

J. Zhang, M. Hong, M. Wang, and S. Zhang. Generalization bounds for stochastic saddle point problems. *arXiv preprint arXiv:2006.02067*, 2020.

Thank you!