



Fine-grained Analysis of Stability and Generalization for Stochastic Bilevel Optimization

Reporter: Xuelin Zhang

Huazhong Agricultural University

<https://zxlm1.github.io/>

This report is jointed with Professor Hong Chen.

August 7, 2024

Background

Basic Definitions & Assumptions

Quantitative Relationship between Generalization and Stability

Stability and Generalization Analysis for SSGD & TSGD

Empirical Validation

Learning Target: Stochastic Bilevel Optimization

The stochastic bilevel optimization (SBO) scheme includes several learning areas, e.g., Meta Learning, Hyperparameter Optimization, Reinforcement Learning.

We aim to analyze the generalization behavior of the following bilevel problems:

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_1}} R(x) = F(x, y^*(x)) &:= \mathbb{E}_{\xi} [f(x, y^*(x); \xi)] \\ \text{s.t. } y^*(x) &= \arg \min_{y \in \mathbb{R}^{d_2}} \{G(x, y) := \mathbb{E}_{\zeta} [g(x, y; \zeta)]\}, \end{aligned} \tag{1}$$

where $d_1, d_2 \in \mathbb{N}^+$, the outer objective function f and the inner objective function g are both continuous and differentiable, ξ, ζ are samples drawn from the validation set and training set respectively.

Upper-level Risk Definition

Given distributions $\mathbb{D}_1, \mathbb{D}_2$, we get the validation set $D_{m_1} := \{\xi_i\}_{i=1}^{m_1} \sim \mathbb{D}_1^{m_1}$ and the training set $D_{m_2} := \{\zeta_i\}_{i=1}^{m_2} \sim \mathbb{D}_2^{m_2}$ by independent sampling, where m_1 and m_2 are the sample sizes.

This paper focuses on the outer-level population risk w.r.t \mathbb{D}_1 ¹

$$R(x, y) = \mathbb{E}_{\xi \sim \mathbb{D}_1} [f(x, y(x); \xi)], \quad (2)$$

and empirical risk w.r.t D_{m_1}

$$R_{D_{m_1}}(x, y) = \frac{1}{m_1} \sum_{i=1}^{m_1} [f(x, y(x); \xi_i)]. \quad (3)$$

In order to evaluate the approximated searching of hyperparameters, we define

$$\mathbb{E}_{A, D_{m_1}, D_{m_2}} \left[R(A(D_{m_1}, D_{m_2})) - R_{D_{m_1}}(A(D_{m_1}, D_{m_2})) \right] \quad (4)$$

as the generalization gap of interest.

¹Fan Bao, et al. Stability and generalization of bilevel programming in hyperparameter optimization. NeurIPS 2021.

On-average Argument Stability for SBO

Here we introduce the analysis techniques, on-average argument stability ².

Definition 1

Let $D_{m_1} = \{z_1, \dots, z_{m_1}\}$ and $\tilde{D}_{m_1} = \{\tilde{z}_1, \dots, \tilde{z}_{m_1}\}$ be two sets drawn independently from distribution $\mathbb{D}_1^{m_1}$.

For any $i = 1, \dots, m_1$, define $D^{(i)} = \{z_1, \dots, z_{i-1}, \tilde{z}_i, z_{i+1}, \dots, z_{m_1}\}$. Denote the \mathbb{E} as the expectation of $\mathbb{E}_{D_{m_1}, D_{m_2}, \tilde{D}_{m_1}, A}$.

We say a randomized algorithm A is $l_1(\beta)$ on-average argument stable if

$$\mathbb{E} \left[\frac{1}{m_1} \sum_{i=1}^{m_1} \left\| A(D_{m_1}, D_{m_2}) - A(D_{m_1}^{(i)}, D_{m_2}) \right\|_2 \right] \leq \beta,$$

and $l_2(\beta^2)$ on-average argument stable if

$$\mathbb{E} \left[\frac{1}{m_1} \sum_{i=1}^{m_1} \left\| A(D_{m_1}, D_{m_2}) - A(D_{m_1}^{(i)}, D_{m_2}) \right\|_2^2 \right] \leq \beta^2.$$

²Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. ICML 2020.

Milder Assumptions

Based on the above definitions, we introduce the requirements of f, g in our analysis, which are milder than assumptions in ³.

Assumption 1

(Outer Function Assumption). Assume that the outer objective function f satisfies
(I) f is jointly L_f -Lipschitz.
(II) f is nonnegative, continuously differentiable and ℓ_f -smooth.

Assumption 2

(Inner Function Assumption). Assume that the inner objective function g satisfies
(I) g is jointly L_g -Lipschitz.
(II) g is continuously differentiable and ℓ_g -smooth.

³Fan Bao, et al. Stability and generalization of bilevel programming in hyperparameter optimization. NeurIPS 2021.

Definition 2

(Hölder Continuity). Let $\tau > 0, \alpha \in [0, 1]$. Gradient ∇f is (α, τ) -Hölder continuous over $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, if there holds

$$\|\nabla f(x, y; \xi) - \nabla f(x', y'; \xi)\|_2 \leq \tau \left\| \begin{array}{c} x - x' \\ y - y' \end{array} \right\|_2^\alpha$$

for all $(x, y), (x', y') \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $\xi \sim \mathbb{D}_1$.

Theorem 1

(I) If algorithm A is $l_1(\beta)$ on-average argument stable in expectation and the outer-level function f is L_f -Lipschitz continuous w.r.t. $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, there holds

$$|\mathbb{E}_{A, D_{m_1}, D_{m_2}} [R(A(D_{m_1}, D_{m_2})) - R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))]| \leq L_f \beta.$$

(II) If algorithm A is $l_2(\beta^2)$ on-average argument stable in expectation and f is nonnegative and ℓ_f -smooth w.r.t. $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, then with $\gamma > 0$,

$$\begin{aligned} & \mathbb{E}_{A, D_{m_1}, D_{m_2}} [R(A(D_{m_1}, D_{m_2})) - R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))] \\ & \leq \frac{\ell_f}{\gamma} \mathbb{E}_{A, D_{m_1}, D_{m_2}} [R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))] + \frac{(\ell_f + \gamma)\beta^2}{2}. \end{aligned}$$

(III) If algorithm A is $l_2(\beta^2)$ on-average argument stable in expectation, f is nonnegative and (α, τ) -Hölder continuous w.r.t. $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with $\alpha \in [0, 1]$, then

$$\begin{aligned} & \mathbb{E}_{A, D_{m_1}, D_{m_2}} [R(A(D_{m_1}, D_{m_2})) - R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))] \\ & \leq \frac{C_{\alpha, \tau}^2}{2\gamma} \mathbb{E}_{A, D_{m_1}, D_{m_2}} [R^{\frac{2\alpha}{1+\alpha}}(A(D_{m_1}, D_{m_2}))] + \frac{\gamma}{2} \beta^2 \end{aligned}$$

where the constant $\gamma > 0$, and $c_{\alpha, \tau} = \begin{cases} (\frac{1+\alpha}{\alpha})^{\frac{\alpha}{1+\alpha}} \tau^{\frac{1}{1+\alpha}}, & \text{if } \alpha > 0 \\ \sup_z \|\partial f(0; z)\|_2 + \tau, & \text{if } \alpha = 0 \end{cases}$.

SSGD and TSGD algorithms

Algorithm 1 Computing algorithm of SSGD

Input: Validation data $D_{m_1} = \{\xi_i\}_{i=1}^{m_1}$ and training set $D_{m_2} = \{\zeta_i\}_{i=1}^{m_2}$, the total number of iterations K , step sizes η_x, η_y .

Initialization: x_0 and y_0 .


```
1: for  $k = 1$  to  $K - 1$  do  
2:   Uniformly sample  $\xi_i \in D_{m_1}$  and  $\zeta_i \in D_{m_2}$ :  
3:    $y_{k+1} = y_k - \eta_y \nabla_y g(x_k, y_k(x_k); \zeta_i)$   
4:    $x_{k+1} = x_k - \eta_x \nabla_x f(x_k, y_k(x_k); \xi_i)$   
5: end for
```

Output: x_K and y_K .

Algorithm 2 Computing algorithm of TSGD

Input: Validation data $D_{m_1} = \{\xi_i\}_{i=1}^{m_1}$ and training set $D_{m_2} = \{\zeta_i\}_{i=1}^{m_2}$, the total number of inner iterations T and outer iterations K , step sizes η_x and η_y .

Initialization: x_0 and y_0^0 .


```
1: for  $k = 0$  to  $K - 1$  do  
2:   for  $t = 0$  to  $T - 1$  do  
3:     Uniformly sample  $\zeta_i \in D_{m_2}$ :  
4:      $y_k^{t+1} = y_k^t - \eta_y \nabla_y g(x_k, y_k^t(x_k); \zeta_i)$   
5:   end for  
6:   Uniformly sample  $\xi_i \in D_{m_1}$ :  
7:    $x_{k+1} = x_k - \eta_x \nabla_x f(x_k, y_k^T(x_k); \xi_i)$   
8:    $y_{k+1}^0 = y_k^T$    
9: end for
```

Output: x_K and y_K^0 .

Algorithm 3 Computing algorithm of UD (Bao et al., 2021)

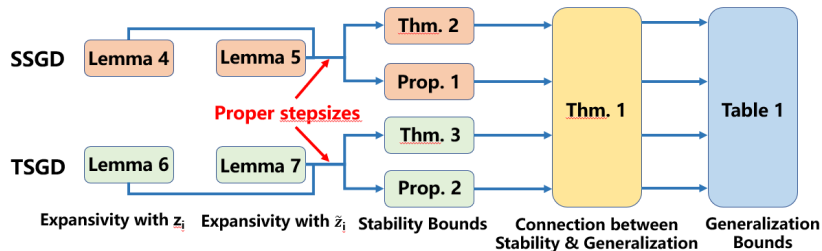
Input: Validation data $D_{m_1} = \{\xi_i\}_{i=1}^{m_1}$ and training set $D_{m_2} = \{\zeta_i\}_{i=1}^{m_2}$, the total number of inner iterations T and outer iterations K , step sizes η_x and η_y .

Initialization: x_0 and y^0 .

```
for  $k = 0$  to  $K - 1$  do  
   $y_k^0 = y^0$    
  for  $t = 0$  to  $T - 1$  do  
    Uniform sampling  $\zeta_i \in D_{m_2}$ :  
     $y_k^{t+1} = y_k^t - \eta_y \nabla_y g(x_k, y_k^t(x_k); \zeta_i)$   
  end for  
  Uniform sampling  $\xi_i \in D_{m_1}$ :  
   $x_{k+1} = x_k - \eta_x \nabla_x f(x_k, y_k^T(x_k); \xi_i)$   
   $y_{k+1}^0 = y_k^T$   
end for
```

Output: x_K and y_K^0 .

Main Proof Process



Main Theoretical Results in Summary

Algorithms	Stability	SC-SC	C-C	NC-NC
SSGD (Theorem 2)	l_1	$\mathcal{O}\left(\frac{K}{m_1}\right)$	$\mathcal{O}\left(\frac{K^{C_4} \ln(K)}{m_1}\right)$	—
	l_2	$\mathcal{O}\left(\frac{(m_1+K)K}{m_1^2}\right)$	$\mathcal{O}\left(\frac{(m_1+K)K^{2C_4-1} \ln^2(K)}{m_1^2}\right)$	—
TSGD (Theorem 3)	l_1	$\mathcal{O}\left(\frac{KT^{C_5}}{m_1}\right)$	$\mathcal{O}\left(\frac{\sqrt{2}^K T^{C_6} \ln(T)}{m_1 K}\right)$	$\mathcal{O}\left(\frac{\sqrt{2}^K K^{C_2 T^{1+C_3}} T}{m_1}\right)$
	l_2	$\mathcal{O}\left(\frac{(m_1+K)KT^{2C_5}}{m_1^2}\right)$	$\mathcal{O}\left(\frac{(m_1+K)2^K T^{2C_6} \ln^2(T)}{m_1^2 K^2}\right)$	$\mathcal{O}\left(\frac{(m_1+K)2^K K^{2C_2 T^{1+C_3}} T^2}{m_1^2}\right)$
SSGD (Proposition 1)	l_1	$\mathcal{O}\left(\frac{1}{m_1}\right)$	$\mathcal{O}\left(\frac{1}{m_1}\right)$	$\mathcal{O}\left(\frac{K^{C_1}}{m_1}\right)$
	l_2	$\mathcal{O}\left(\frac{m_1+K}{m_1^2 \sqrt{K}}\right)$	$\mathcal{O}\left(\frac{m_1+K}{m_1^2 \sqrt{K}}\right)$	$\mathcal{O}\left(\frac{(m_1+K)K^{2C_1}}{m_1^2}\right)$
TSGD (Proposition 2)	l_1	$\mathcal{O}\left(\frac{K}{m_1}\right)$	$\mathcal{O}\left(\frac{\sqrt{2}^K}{m_1 K}\right)$	$\mathcal{O}\left(\frac{\sqrt{2}^K K^{C_2 T^{C_3}} T}{m_1}\right)$
	l_2	$\mathcal{O}\left(\frac{(m_1+K)K}{m_1^2}\right)$	$\mathcal{O}\left(\frac{(m_1+K)2^K}{m_1^2 K^2}\right)$	$\mathcal{O}\left(\frac{(m_1+K)2^K K^{C_2 T^{C_3}} T^2}{m_1^2}\right)$

Summary of the generalization bounds under different settings. For briefly, l_1 (l_2) represents the l_1 (l_2) on-average argument stability and $C_1 - C_6$ are positive constants.

SC, C and NC stand for strongly convex, convex and non-convex respectively; m_1 is the number of validation samples; K and T are the total numbers of outer and inner iterations. Assume that the output model has a small empirical risk

$$\mathbb{E} \left[R_{D_{m_1}}(A(D_{m_1}, D_{m_2})) \right] = \mathcal{O}(m_1^{-1}).$$

Empirical Validation

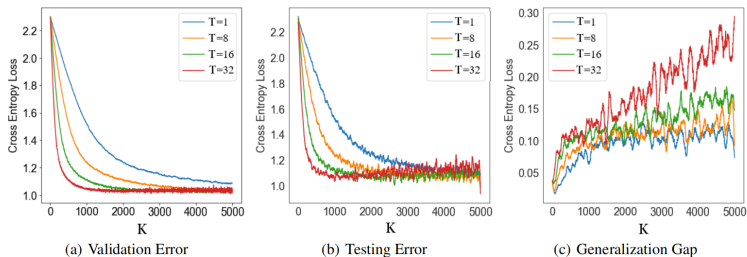
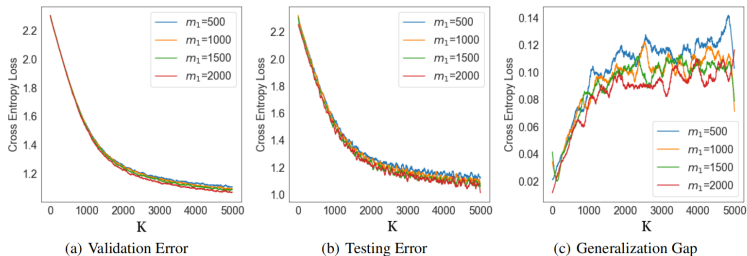


Figure 1: Results of hyperparameter optimization in data reweighting with varying T and K



My Related Published or Ongoing Works

- Meta additive model for auto weighting and sparse approximation (Under Review)
- S^2 MAM: Semi-supervised Meta Additive Model (Under Review)
- Generalized Sparse Additive Model with Unknown Link Function (Under Review)
- Fine-grained analysis of stability and generalization for stochastic bilevel optimization (*IJCAI* 2024)
- Error Density-dependent Empirical Risk Minimization (ESWA 2024)
- Neural Partially Linear Additive Model (Frontiers of Computer Science 2023)
- Robust variable structure discovery based on tilted empirical risk minimization (Applied Intelligence 2023)
- Stepdown SLOPE for controlled feature selection (AAA/ 2023)
- Robust Manifold Learning via Bilevel Cycle GAN (Ongoing Work)



Thanks!