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

Xuelin Zhang[1], **Hong Chen**[1,2,*], **Bin Gu**[3], **Tieliang Gong**[4] and **Feng Zheng**[5]
[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[3]School of Artificial Intelligence, Jilin University, Jilin 130012, China
[4]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[5]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
zhangxuelin@webmail.hzau.edu.cn, chenh@mail.hzau.edu.cn

Abstract

Stochastic bilevel optimization (SBO) has been integrated into many machine learning paradigms recently including hyperparameter optimization, meta learning, reinforcement learning, etc. Along with the wide range of applications, there have been abundant studies on concerning the computing behaviors of SBO. However, the generalization guarantees of SBO methods are far less understood from the lens of statistical learning theory. In this paper, we provide a systematical generalization analysis of the first-order gradient-based bilevel optimization methods. Firstly, we establish the quantitative connections between the on-average argument stability and the generalization gap of SBO methods. Then, we derive the upper bounds of on-average argument stability for single timescale stochastic gradient descent (SGD) and two timescale SGD, where three settings (nonconvex-nonconvex (NC-NC), convex-convex (C-C) and strongly-convex-strongly-convex (SC-SC)) are considered respectively. Experimental analysis validates our theoretical findings. Compared with the previous algorithmic stability analysis, our results do not require the re-initialization of the inner-level parameters before each iteration and are suited for more general objective functions.

1 Introduction

In this paper, we focus on establishing stability and generalization analysis for the stochastic bilevel optimization (SBO) (Bracken and McGill, 1973; Ji *et al.*, 2021; Bao *et al.*, 2021) defined as follows:

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

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.

For this bilevel optimization scheme, we often call $\min_{y\in\mathbb{R}^{d_2}} \mathbb{E}_\zeta[g(x,y;\zeta)]$ as the inner (or lower-level) problem, and name $\min_{x\in\mathbb{R}^{d_1}} \mathbb{E}_\xi\left[f\left(x, y^*(x); \xi\right)\right]$ as the outer (or upper-level) problem. The goal of (1) is to minimize the outer objective function $R(x)$ (also $F(x, y^*(x))$) with respect to (w.r.t.) the model parameter x, where parameter $y^*(x)$ is derived from the inner minimization formulation.

The SBO formulation in (1), stemming from (Bracken and McGill, 1973), has attracted increasing attention in many machine learning applications including hyper-parameter optimization (Franceschi *et al.*, 2017, 2018; Lorraine and Duvenaud, 2018; MacKay *et al.*, 2019; Okuno *et al.*, 2021; Zhang *et al.*, 2023), generative adversarial learning (Pfau and Vinyals, 2016), meta learning (Franceschi *et al.*, 2018; Bertinetto *et al.*, 2018; Zügner and Günnemann, 2019; Rajeswaran *et al.*, 2019; Ji *et al.*, 2020), and reinforcement learning (Tschiatschek *et al.*, 2019). Indeed, there are rich computing methods to implement this bilevel optimization scheme, as well as theoretical works on optimization convergence analysis (Li *et al.*, 2020; Ji *et al.*, 2021). However, the generalization analysis of SBO is still far less understood from the viewpoint of statistical learning theory (STL), e.g., algorithmic stability and generalization analysis (Hardt *et al.*, 2016; Lei and Ying, 2020; Bao *et al.*, 2021).

Stability-based generalization analysis can be traced back to the 1970s (Rogers and Wagner, 1978) and has achieved rapid developments in STL, see e.g., (Bousquet and Elisseeff, 2002; Elisseeff *et al.*, 2005; Hardt *et al.*, 2016; Liu *et al.*, 2017; Lei and Ying, 2020; Lei *et al.*, 2021a; Deng *et al.*, 2021; Kuzborskij and Lampert, 2018). To match the characterizations of various algorithms, different definitions of algorithmic stability have been formulated (including the uniform stability (Bousquet and Elisseeff, 2002), uniform argument stability (Liu *et al.*, 2017), locally elastic stability (Deng *et al.*, 2021), on-average stability (Kuzborskij and Lampert, 2018) and on-average argument stability (Lei and Ying, 2020)) to

*Corresponding author.

better investigate their generalization bounds. The on-average argument stability was proposed in (Lei and Ying, 2020) to establish the fine-grained generalization analysis of single-level pointwise stochastic gradient descent (SGD). Subsequently, Lei et al. extended the stability-based generalization assessment to the pairwise SGD (Lei *et al.*, 2021a), where systematic strategies have been provided to make a better balance between generalization error and optimization error. As far as we know, there is only one study exploring the generalization analysis of SBO (Bao *et al.*, 2021), which presents an expectation generalization bound w.r.t. the validation set via the uniform stability approach. However, the theoretical analysis of (Bao *et al.*, 2021) is limited to unrolled differentiation (UD) based algorithms with re-initialization in inner-level for hyper-parameter optimization, which may not be applicable to other commonly used optimization algorithms, e.g., single timescale SGD (SSGD) (Zhou *et al.*, 2022a; Liu *et al.*, 2022; Chen *et al.*, 2022) and two timescale SGD (TSGD) (Zhou *et al.*, 2022a; Liu *et al.*, 2022; Hong *et al.*, 2023). Therefore, it is important to further investigate the generalization guarantees for general SBO formulation to cover wider bilevel optimization algorithms.

To address the aforementioned issue, this paper establishes the fine-grained stability and generalization analysis for general first-order bilevel optimization methods. Our main contributions are summarized as follows:

- Firstly, we establish the quantitative connection between the generalization gap of bilevel optimization methods and on-average argument stability. Especially for the l_2 on-average argument stability, the derived stability-based generalization bounds involve the empirical risks, which is consistent with the previous analysis for singe-level optimization (Lei and Ying, 2020; Lei *et al.*, 2021a).
- Secondly, this paper provides several stability bounds of bilevel optimization methods associated with both SSGD and TSGD algorithms, where different conditions of objective functions (i.e., SC-SC, C-C, NC-NC) are considered. Moreover, we extend the results to the more general setting by relaxing the restriction (e.g., Lipschitz continuity and smoothness assumptions) of the optimization objective. As far as we know, this is the first systemic generalization analysis for first-order SGD-based bilevel optimization under the low-noise setting.
- Finally, we conduct experimental evaluations for bilevel optimization methods including hyperparameter optimization. Empirical results validate our theoretical findings about the relationship between the generalization gap and the size of the validation set as well as the maximum value of inner (outer) iterations.

To better evaluate our results, we compare them with the most related work on stability and generalization analysis (Bao *et al.*, 2021) from the following perspectives:

- *Optimization strategy.* The previous UD-based hyperparameter optimization (Algorithm 1 in (Bao *et al.*, 2021)) requires reinitialization in the inner-level parameters before each iteration. Different from this special case, this paper considers the SBO algorithms where the parameters in inner-level and outer-level are both updated continuously (e.g., SSGD (Zhou *et al.*, 2022a; Chen *et al.*, 2022) and TSGD (Zhou *et al.*, 2022a; Liu *et al.*, 2022; Hong *et al.*, 2023)). The iteration strategy matching our analysis has been used extensively in practice (Ji *et al.*, 2021; Liu *et al.*, 2022; Ghadimi and Wang, 2018). Especially for theoretical analysis of TSGD, it is challenging to deal with the gradient summation during the inner iterations and the previous analysis technique (Bao *et al.*, 2021) can not be extended to this case directly.
- *Analysis tool.* Different from uniform stability used in (Bao *et al.*, 2021), this paper develops the analysis technique of on-average argument stability to provide the fine-gained generalization bounds under low noise settings, where the stability bounds involve a weighted sum of empirical risks instead of the uniform Lipschitz constants.
- *Conditions of objective functions.* Similar to the previous stability analysis in (Lei *et al.*, 2020; Shen *et al.*, 2020; Zhou *et al.*, 2022b), the objective functions in (Bao *et al.*, 2021) are assumed to be bounded, third order continuously differentiable and smooth. Here, we merely need the bilevel objective functions to be non-negative, smooth and Lipschitz continuous, where the last condition for the outer-level function can be further removed by the l_2 on-average argument stability. Detailed stability results have been derived for both SSGD and TSGD algorithms under NC-NC, C-C and SC-SC settings. In addition, we also establish generalization bounds by replacing the smooth condition with the weaker Hölder continuous assumption.

2 Problem Formulation

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 $\mathbb{D}_1$ and empirical risk w.r.t D_{m_1} [1], which are defined respectively as

$$R(x, y) = \mathbb{E}_{\xi\sim\mathbb{D}_1}[f(x, y(x); \xi)]$$
$$\text{and } R_{D_{m_1}}(x, y) = \frac{1}{m_1}\sum_{i=1}^{m_1}[f(x, y(x); \xi_i)],$$

where $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is an objective function and $y(x)$ is the inner model parameter given the outer model parameter x (also see (1)).

Let $(x, y(x))$ in (1) be estimated by a stochastic algorithm A with data D_{m_1}, D_{m_2}, i.e. $A(D_{m_1}, D_{m_2})$. Similar to the

[1] Thus we consider adding corruptions to D_{m_1} to access the generalization behavior of the meta-learner (Thrun, 1998) at upper level.

previous works (Bao *et al.*, 2021; Hoffer *et al.*, 2017; Keskar *et al.*, 2017), in order to evaluate the approximated searching of hyperparameters, we define

$$\mathbb{E}_{A,D_{m_1},D_{m_2}}\left[R\left(A\left(D_{m_1},D_{m_2}\right)\right)-R_{D_{m_1}}\left(A\left(D_{m_1},D_{m_2}\right)\right)\right] \tag{2}$$

in the upper (outer) level as the generalization gap of A, which measures the difference between the population risk $R(A)$ and the empirical risk $R_{D_{m_1}}(A)$.

The following conditions have been used to characterize the theoretical properties of objective functions in (1).

Definition 1. *(Joint Lipschitz Continuity (Ji* et al.*, 2021; Liu* et al.*, 2022)). An objective function f is jointly L_f-Lipschitz over $\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, if there holds*

$$|f(x,y;\xi)-f(x',y';\xi)|\leq L_f\sqrt{\|x-x'\|_2^2+\|y-y'\|_2^2}$$

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

Definition 2. *(Joint Smoothness (Lei* et al.*, 2021b)). An objective function f is ℓ_f-smooth over $\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, if*

$$\|\nabla f(x,y;\xi)-\nabla f(x',y';\xi)\|_2\leq\ell_f\sqrt{\|x-x'\|_2^2+\|y-y'\|_2^2}$$

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

Definition 3. *(Strong Convexity). A function ψ is μ-strongly-convex over a set X, if $\forall t,t'\in X$,*

$$\psi(t')+\langle\nabla\psi(t'),t-t'\rangle+\frac{\mu}{2}\|t-t'\|_2^2\leq\psi(t).$$

Definition 4. *(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 $\xi\sim\mathbb{D}_1$.

The above conditions for objective functions have been used extensively in convergence analysis for bilevel optimization (Ji *et al.*, 2021; Ghadimi and Wang, 2018; Liu *et al.*, 2022) and stability-based generalization analysis for single-level optimization methods (Hardt *et al.*, 2016; Lei *et al.*, 2021b). Moreover, the Hölder continuity is much weaker than the Lipschitz continuity and smoothness (Lei and Ying, 2020; Nesterov, 2015). If Definition 4 holds with $\alpha=1$, then f is smooth (see Definition 2). And if Definition 4 holds with $\alpha=0$, f becomes Lipschitz continuous as in Definition 1 and can be non-differentiable (Lei and Ying, 2020). The objective functions satisfying Definition 4 include the mean absolute function, the hinge function and some of their variants (Lei and Ying, 2020; Steinwart and Christmann, 2008).

Definition 5. *(On-average Argument Stability (Lei and Ying, 2020)). Let $D_{m_1}=\{z_1,\ldots,z_{m_1}\}$ and $\widetilde{D}_{m_1}=\{\tilde{z}_1,\ldots,\tilde{z}_{m_1}\}$ be two sets drawn independently from distribution $\mathbb{D}_1^{m_1}$. For any $i=1,\ldots,m_1$, define $D^{(i)}=\{z_1,\ldots,z_{i-1},\tilde{z}_i,z_{i+1},\ldots,z_{m_1}\}$. Denote the $\mathbb{E}$ as the expectation of $\mathbb{E}_{D_{m_1},D_{m_2},\widetilde{D}_{m_1},A}$. We say a randomized algorithm A is $l_1(\beta)$ on-average argument stable if*

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

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

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

Remark 1. *The on-average argument stability measures the average sensitivity (stability) of output parameters of the learning algorithm when at most one validation sample is changed. Definition 5 is different from Definition 1 in (Bao* et al.*, 2021), where the uniform stability is evaluated by the drift of prediction error of hyperparameter optimization algorithm and the boundedness of loss function often is required.*

Based on the above definitions, we introduce the requirements of f,g in our analysis.

Assumption 1. *(Outer Function Assumption). Assume that the outer objective function f in* (1) *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 in* (1) *satisfies*

(I) g is jointly L_g-Lipschitz.

(II) g is continuously differentiable and ℓ_g-smooth.

3 Quantitative Relationship between Generalization and Stability

This section states that the generalization gap of (1) can be bounded by the on-average argument stability. Before providing the detailed conclusion of Theorem 1, we first introduce the self-bounding property definition.

Lemma 1. *(Self-bounding property). Assume that for all $z\in\mathcal{D}$, the map $w\mapsto f(w;z)$ is nonnegative, and $w\mapsto\partial f(w;z)$ is (α,τ)-Hölder continuous with $\alpha\in[0,1]$. Then we have*

$$\|\partial f(w,z)\|_2\leq c_{\alpha,\tau}f^{\frac{\alpha}{1+\alpha}}(w,z),\quad\forall w\in\mathbb{R}^d,z\in\mathcal{D},$$

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

The self-bounding property of f with (α,τ)-Hölder continuous (sub)gradient contains the specific Lipschitz continuous ($\alpha=0$) and smoothness ($\alpha=1$) conditions (Lei *et al.*, 2021b).

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. $(x,y)\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, denote $\mathbb{E}$ as $\mathbb{E}_{A,D_{m_1},D_{m_2}}$, there holds*

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

(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. $(x,y)\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, denote $\mathbb{E}$ as $\mathbb{E}_{A,D_{m_1},D_{m_2}}$, then

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

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.

for $k = 1$ to $K-1$ **do**
  Uniformly sample $\xi_i \in D_{m_1}$ and $\zeta_i \in D_{m_2}$:
  $y_{k+1} = y_k - \eta_y \nabla_y g\left(x_k, y_k\left(x_k\right); \zeta_i\right)$
  $x_{k+1} = x_k - \eta_x \nabla_x f\left(x_k, y_k\left(x_k\right); \xi_i\right)$
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.

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

Output: x_K and y_K^0.

where the constant $\gamma > 0$.

(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

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

for $D_{m_1} \sim \mathbb{D}_1^{m_1}$ and $D_{m_2} \sim \mathbb{D}_2^{m_2}$, where the constant $\gamma > 0$.

Remark 2. *Theorem 1 validates the connection between on-average argument stability and the generalization gap. Especially, the smoothness assumption is further relaxed by the Hölder continuity in Theorem 1(III).*

Remark 3. *Different from the uniform stability technique employed in (Bao* et al.*, 2021), the on-average argument stability further exploits the Lipschitz continuous (L_f) or smooth properties (ℓ_f) of the objective function as well as the stability parameter (β) to bound the algorithmic generalization gap. Especially, there is a trade-off between the empirical risk and the algorithmic stability bound.*

Remark 4. *There are several advantages of l_2 on-average argument stability in Theorem 1 (II), where Assumption 1(I) is removed and the low noise assumption can be used to obtain a fine-gained result instead of the Lipschitz constant (Lei and Ying, 2020). If algorithm A is $l_2(\beta^2)$ on-average argument stable, then we derive the upper bound of generalization gap*

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

Output: x_K and y_K^0.

with $\sqrt{2\ell_f \mathbb{E}\left[R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))\right]}\beta + \ell_f\beta^2/2$ by taking $\gamma = \sqrt{2\ell_f \mathbb{E}\left[R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))\right]}/\beta$. Moreover, if the output model achieves a small empirical risk (e.g., low noise assumption $\mathbb{E}\left[R_{D_{m_1}}(A(D_{m_1}, D_{m_2}))\right] = \mathcal{O}(m_1^{-1})$), we get that $\mathbb{E}\left[R\left(A\left(D_{m_1}, D_{m_2}\right)\right) - R_{D_{m_1}}\left(A\left(D_{m_1}, D_{m_2}\right)\right)\right] = \mathcal{O}\left(\beta^2 + \beta/\sqrt{m_1}\right)$.

4 Stability Analysis for Stochastic Bilevel Optimization

To solve bilevel optimization formulation (1), some gradient-based algorithms are designed based on the single timescale or two timescale strategies (Ji *et al.*, 2021; Chen *et al.*, 2022; Liu *et al.*, 2020, 2022; Zhou *et al.*, 2022a). In the following, we introduce the computing approaches for (1) (SSGD in Algorithm 1 and TSGD in Algorithm 2), and then establish their generalization assessments by presenting their algorithmic stability bounds.

4.1 Stability and Generalization Analysis for SSGD

Let η_x and η_y be the step sizes for updating x and y. According to Theorem 1, the on-average argument stable metrics in Definition 5 for SSGD algorithm A with K iterations $\left\|A(D_{m_1}, D_{m_2}) - A\left(D_{m_1}^{(i)}, D_{m_2}\right)\right\|_2$ can be measured by

$$\sqrt{\|x_K - x_K^{(i)}\|_2^2 + \|y_K - y_K^{(i)}\|_2^2}.$$

Now we state the upper bounds of on-average argument stability for SSGD in Algorithm 1.

Theorem 2. *Suppose that Assumptions 1, 2 hold and Algorithm A is SSGD with K iterations. Denote $\ell = \max\{\ell_f, \ell_g\}$, $\eta = \max\{\eta_x, \eta_y\}$.*

(I) Assume that the bilevel optimization problem (1) *is SC-SC with strong convexity parameters μ_f and μ_g. Let the step sizes satisfy that $\frac{2(\mu_f+\mu_g) - \sqrt{4(\mu_f+\mu_g)^2 - 2\left(\ell_f^2+\ell_g^2\right)}}{2\left(\ell_f^2+\ell_g^2\right)} \leq \eta_x =$*

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

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

$\eta_y \leq \frac{2(\mu_f+\mu_g)+\sqrt{4(\mu_f+\mu_g)^2-2\left(\ell_f^2+\ell_g^2\right)}}{2\left(\ell_f^2+\ell_g^2\right)}$. *Then, A is $l_1(\beta)$ on-average argument-stable in expectation with*

$$\beta = \frac{2C}{m_1}\sum_{k=1}^{K}\sqrt{2\ell_f E_{A,D_{m_1}}[R_{D_{m_1}}(x_k, y_k)] + L_g^2}$$

and $l_2(\beta^2)$ on-average argument-stable in expectation with

$$\beta^2 = \frac{4(m_1+K)eC^2}{m_1^2}\sum_{k=1}^{K}(2\ell_f E_{A,D_{m_1}}[R_{D_{m_1}}(x_k, y_k)] + L_g^2),$$

where $C = \frac{2(\mu_f+\mu_g)+\sqrt{4(\mu_f+\mu_g)^2-2\left(\ell_f^2+\ell_g^2\right)}}{2\left(\ell_f^2+\ell_g^2\right)}$.

(II) Assume that the bilevel optimization problem (1) *is C-C. If $\eta \leq \frac{c_1\ln(K)}{\sqrt{2}K\ell}$ for some $c_1 > 0$, then A is $l_1(\beta)$ on-average argument-stable in expectation with $\beta =$*

$$\frac{\sqrt{2}c_1\ln(K)K^{c_1-1}}{m_1\ell}\sum_{k=1}^{K}\sqrt{2\ell_f E_{A,D_{m_1}}[R_{D_{m_1}}(x_k, y_k)] + L_g^2}.$$

And A is $l_2(\beta^2)$ on-average argument-stable in expectation, where $\beta^2 =$

$$\frac{2c_1^2(m_1+K)eK^{2c_1-2}\ln^2(K)}{m_1^2\ell^2}\sum_{k=1}^{K}(2\ell_f E_{A,D_{m_1}}[R_{D_{m_1}}(x_k, y_k)] + L_g^2).$$

Remark 5. *Theorem 2 demonstrates that the algorithmic stability can be improved when the model can achieve a relatively small optimization error. In addition, the ℓ_f-smooth assumption also can be replaced by Hölder continuous condition. In order to obtain tighter bounds for the SSGD algorithm, we further derive its algorithmic stability with refined step sizes in Proposition 1 in Appendix C.*

Combining Theorems 1 and 2, the algorithmic generalization bounds of SSGD are further summarized in Table 1 under the low noise settings (small empirical risk). As shown in Table 1, the generalization bounds of some SSGD algorithms achieve the rate of $\mathcal{O}(m_1^{-1})$ under the limitations of step sizes in Theorem 2. From Table 1, one can easily find that objective functions with better (convexity) properties usually lead to better algorithmic stability and generalization performance, which is consistent with the existing stability and generalization analysis for single-level problems (Lei and Ying, 2020; Kuzborskij and Lampert, 2018; Lei *et al.*, 2021b).

4.2 Stability and Generalization Analysis for TSGD

Now we turn to establish the stability bounds of the TSGD algorithm with different inner and outer functions (i.e., NC-NC, C-C and SC-SC).

Assume that f is ℓ_f-smooth and g is ℓ_g-smooth. Let η_x and η_y be the step sizes for updating x and y, respectively. Denote $\nabla_y g(x, y)$ as the partial derivative of function g over variable y. y_K^t represents the inner parameter y in K-th outer loop and t-th inner loop. For the TSGD algorithm A with K outer iterations and T inner iterations, the argument stability $\left\|A(D_{m_1}, D_{m_2}) - A\left(D_{m_1}^{(i)}, D_{m_2}\right)\right\|_2$ is measured by $\sqrt{\|x_K - x_K^{(i)}\|_2^2 + \|y_K^0 - (y_K^0)^{(i)}\|_2^2}$, where

$$y_K^0 = y_{K-1}^T = y_{K-1}^0 - \sum_{t=0}^{T-1}\eta_y\nabla_y g(x_{K-1}, y_{K-1}^t).$$

Remark 6. *Analogous to TSGD algorithm, the UD algorithm employed in (Bao* et al.*, 2021) (see Algorithm 3) also involves two layers of nested loops but requires re-initialization in the inner level before each new outer loop. In their stability analysis, the inner-level parameter updates are not considered, but used to determine the constants of Lipschitz continuity and smoothness of the outer-level function. This paper considers the general TSGD algorithms where both inner-level and outer-level parameters are updated continuously, e.g. $y_{K-1}^T = y_K^0$. The gradient summation of the inner-level parameter is relatively complex and makes it difficult to utilize the (smooth or convex) properties as (Bao* et al.*, 2021;*

Hardt et al.*, 2016) directly, which brings challenges to the stability analysis.*

Theorem 3. *Suppose that Assumptions 1 and 2 hold and algorithm A is TSGD with T inner loops and K outer loops. Denote $\ell = \max\{\ell_f, \ell_g\}$, $\eta = \max\{\eta_x, \eta_y\}$, $E[R_{D_{m_1}}] = E_{A,D_{m_1}}[R_{D_{m_1}}(x_k, y_k)]$.*

***(I)** Assume that the bilevel optimization problem is SC-SC with strong convexity parameters μ_f and μ_g. Denote $\rho_1 = \frac{2(T\mu_g+\mu_f-T\ell)}{2(1+T^2)\ell^2}$ and $\rho_2 = \frac{\sqrt{4(T\ell-\mu_f-T\mu_g)^2-2(1+T^2)\ell^2(1-\frac{c_1\ln(T)}{K})}}{2(1+T^2)\ell^2}$ for simplicity.*

Let the step sizes satisfy that $\rho_1 - \rho_2 \le \eta \le \rho_1 + \rho_2 = C_1$ for some positive constant c_1, C_1. Then A is $l_1(\beta)$ on-average argument-stable in expectation with

$$\beta = \frac{2T^{\frac{c_1}{2}}C_1}{m_1}\sum_{k=1}^{K}\sqrt{2\ell_f E[R_{D_{m_1}}] + T^2L_g^2}$$

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

$$\beta^2 = \frac{4(m_1+K)eT^{c_1}C_1^2}{m_1^2}\sum_{k=1}^{K}(2\ell_f E[R_{D_{m_1}}] + T^2L_g^2).$$

(II) Assume that the bilevel optimization problem is C-C. When $\eta \le \frac{c_2\ln(T)}{\sqrt{1+T^2}K\ell}$ for some $c_2 > 0$, A is $l_1(\beta)$ on-average argument-stable in expectation with

$$\beta = \frac{2c_2\ln(T)T^{c_2}}{m_1\sqrt{1+T^2}K\ell}\sum_{k=1}^{K}2^{\frac{K-k}{2}}\sqrt{2\ell_f E[R_{D_{m_1}}] + T^2L_g^2},$$

and is $l_2(\beta^2)$ on-average argument-stable with $\beta^2 =$

$$\frac{4c_2^2(K+m_1)\ln^2(T)T^{2c_2}e}{m_1^2(1+T^2)K^2\ell^2}\sum_{k=1}^{K}2^{K-k}(2\ell_f E[R_{D_{m_1}}] + T^2L_g^2).$$

(III) Assume that the bilevel optimization problem is NC-NC. Denote $\eta_{y,t}$ as the inner step size in t-th inner loop, denote $\eta_k = \max\{\eta_{x,k}, \eta_{y,k}\}$ as the outer step size in k-th outer loop. Let $\eta_{y,t} \le \frac{c_3}{\ell_g(t+1)}$, $\eta_k \le \frac{c_4}{\ell k}$ and $\sum_{x=a}^{b} f(x) \le c_5\int_a^b f(x)dx$ for some positive constants c_3, c_4, c_5, then A is $l_1(\beta)$ on-average argument-stable in expectation with

$$\beta = \frac{2c_4}{m_1\ell}\sum_{k=1}^{K}2^{\frac{K-k}{2}}(\frac{K}{k})^{c_4c_5T^{c_6}\sqrt{1+T^2}}\sqrt{2\ell_f E[R_{D_{m_1}}] + T^2L_g^2},$$

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

$$\frac{4(m_1+K)c_4^2e}{m_1^2\ell^2}\sum_{k=1}^{K}(\frac{K}{k})^{2c_4c_5T^{c_6}\sqrt{1+T^2}}2^{K-k}(2\ell_f E[R_{D_{m_1}}] + T^2L_g^2).$$

Notice that $T^{c_6}\sqrt{1+T^2}$ with $\eta_{y,t} \le \frac{c_3}{\ell_g(t+1)}$ is obtained from Lemma 6 for NC-NC in Appendix D*, where the original form is $T^{c_0c_1}\sqrt{1+T^2}$ with $\eta_{y,t} \le \frac{c_0}{\ell_g(t+1)}$.*

Remark 7. *After integrating Theorems 1 and 3, we summarize the generalization bounds of Algorithm 2 in Table 1. Similar to Theorem 2, the results of Theorem 3 also demonstrate that the total numbers of the validation samples m_1 (↑), the inner iterations T (↓) and outer iterations K (↓) directly affect the generalization performance (↑) of TSGD algorithms. We also observe that the impacts of K and T on generalization are suppressed for Algorithm 2 with SC-SC (or C-C) with a small enough step size. In order to obtain tighter bounds w.r.t. Theorems 2 and 3, we further derive the corresponding results with refined step sizes in Propositions 1 and 2 in Appendix C, D. The results shown in Table 1 are comparable to Bao* et al. *(2021) with the bound of $\mathcal{O}(\frac{K^c}{m})$ where $0 < c < 1$. Relaxing the stepsize limitations, especially for SC-SC, is a meaningful direction, which is left for future work.*

5 Empirical Evaluations

This section empirically validates our theoretical findings on two real-world datasets. We consider Algorithm 2 here, since it is equal to Algorithm 1 as $T = 1$. The distributions of the testing samples and validation samples are assumed to be same, but can differ from the training data (Ren *et al.*, 2018; Bao *et al.*, 2021). Similar to (Bao *et al.*, 2021), we focus on evaluating the generalization behavior of outer-level problems based on the validation set. All experiments are implemented in Python on an Intel Core i7 with 32 GB memory. Implemented codes (including (Bao *et al.*, 2021) for hyperparameter optimization) and data sets (including the MNIST data (LeCun, 1998) and the Omnilot data (Lake *et al.*, 2015)) are from publicly available sources.

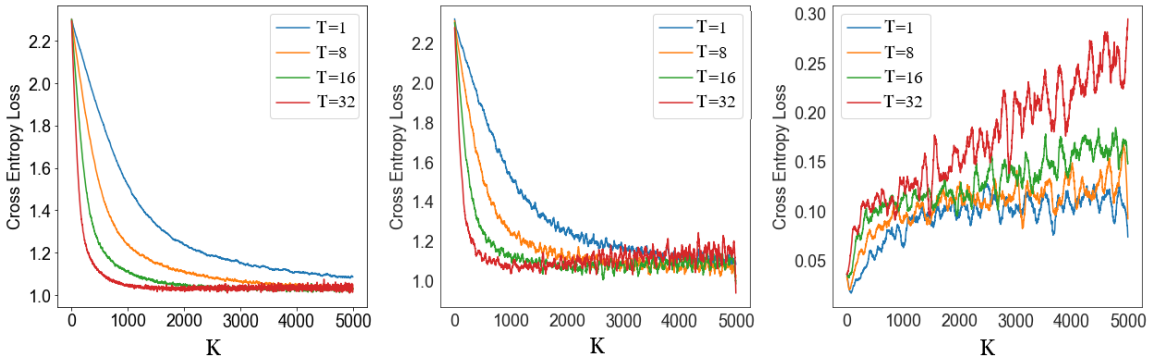
This section considers the general hyperparameter optimization formulation (Ji *et al.*, 2021). Given the training set D_{train} and the validation set D_{val}, the hyperparameter optimization scheme can be formulated as

$$\min_x \mathcal{R}_{D_{\text{val}}}(x) = \frac{1}{|D_{\text{val}}|}\sum_{\xi\in D_{val}} h(x, y^*(x); \xi)$$
$$\text{s. t. } y^* = \arg\min_y \underbrace{\frac{1}{|D_{\text{train}}|}\sum_{\zeta\in D_{\text{train}}}(h(x, y; \xi) + \Omega_{y,x})}_{\mathcal{R}_{D_{\text{train}}}(x,y)},$$

where h is the loss function, $\Omega_{y,x}$ is the regularizer and $|D_{\text{train}}|$ represents the size of training data.

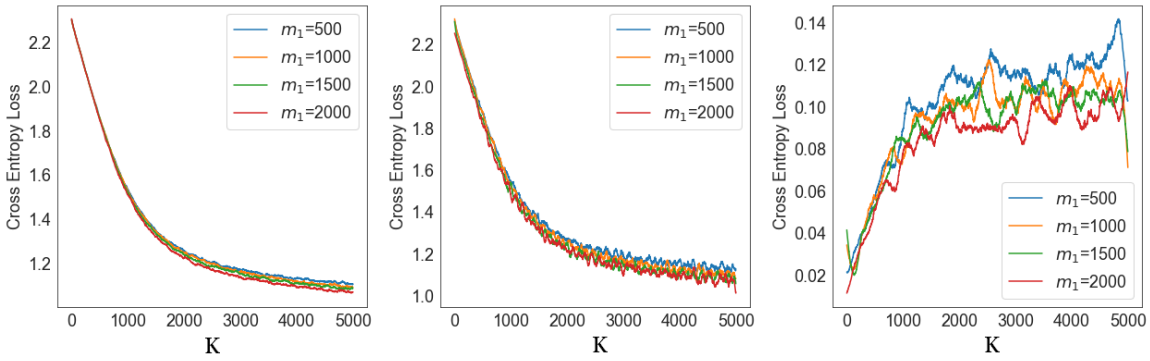
5.1 Experiment Settings

We evaluate the impact of several factors on the generalization gap (2) based on the famous MNIST data (LeCun, 1998), which totally consists of more than 6×10^5 handwritten figures with the size of 28×28. Following the same task of data reweighting in (Bao *et al.*, 2021), we corrupt the labels of training samples randomly with the probability of 50% and employ a fully connected network (with size of 784/256/10) with cross-entropy loss for classification. Initially, we randomly select 2000, 2000, and 1000 figures for training, validation and testing, respectively. Meanwhile, set the initial batch size as 8, the maximum number of inner iterations as $T = 5000$, and the number of outer iterations as $K = 5000$. The initial step sizes for inner and outer minimization problems are 0.01 and 5, respectively. For the given parameter settings, each experiment is randomly repeated five times on one GeForce GTX 1660 SUPER GPU, and the average results are reported.



(a) Validation Error (b) Testing Error (c) Generalization Gap

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



(a) Validation Error (b) Testing Error (c) Generalization Gap

Figure 2: Results of hyperparameter optimization in data reweighting with varying K and m_1

5.2 Experimental Results

The generalization gap defined in (2) is estimated by the divergence between the validation error and the testing error.

Impact of iteration numbers K and T. Now we evaluate the impact of parameters (e.g., the numbers of validation samples m_1, inner iteration T and outer iteration K) on the generalization performance. Figure 1 shows the curves of validation error, testing error and the generalization gap under different settings of maximum inner iteration T and maximum outer loop K. Figures 1(a) and 1(b) imply that the classification model might be overfitting with increasing testing errors as $K > 3000$ and $T = 32$. Besides, Figure 1(c) demonstrates that too large K and T may reduce the generalization ability of the hyperparameter optimization method due to overfitting. This empirical finding is consistent with our theoretical results and the previous related analysis (Franceschi *et al.*, 2018; Bao *et al.*, 2021).

Impact of sample size m_1 with $T = 1$. Figure 2 presents the results of SSGD (Algorithm 1) under different choices of K and m_1. From Figure 2(c), we observe that a small sample size $m_1 = 500$ leads to an increase in validation error and testing error. This indicates that the larger number of validation samples is beneficial to reduce the generalization gap. The above empirical findings match our theoretical results, see e.g., Theorem 3 and Table 1.

Based on theoretical analysis and empirical evaluations, we can get some understanding of the generalization performance of bilevel optimization. Explicitly, the generalization ability of SBO often can be improved with the increase of m_1 and proper iteration numbers K, T, where too small iterations may cause underfitting and large ones can lead to overfitting. Usually, it is beneficial for generalization through setting appropriate learning rates, especially for NC-NC. For real applications, the trade-off between m_1, T, and K is of great importance to guarantee the effectiveness of SBO methods.

6 Conclusion

This paper established the stability and generalization analysis for stochastic bilevel optimization with first-order gradient-based approximate algorithms. Our theoretical results are obtained by developing the analysis technique associated with the on-average argument stability, and can cover wider bilevel optimization algorithms under low noise settings. Compared with the state-of-the-art analysis (Bao *et al.*, 2021), our theoretical results do not require reinitializing the inner-level parameter before each iteration and suit for objective functions under milder conditions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 62376104 and 12071166), the Fundamental Research Funds for the Central Universities of China (No. 2662023LXPY005), and HZAU-AGIS Cooperation Fund (No. SZYJY2023010).

References

Fan Bao, Guoqiang Wu, Chongxuan Li, Jun Zhu, and Bo Zhang. Stability and generalization of bilevel programming in hyperparameter optimization. In *Advances in Neural Information Processing Systems*, pages 4529–4541, 2021.

Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Jerome Bracken and James T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022.

Zhun Deng, Hangfeng He, and Weijie Su. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning*, pages 2590–2600. PMLR, 2021.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892, 2021.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of SGD for pairwise learning. In *Advances in Neural Information Processing Systems*, pages 21216–21228, 2021.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

Junyi Li, Bin Gu, and Heng Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for

bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pages 6305–6315. PMLR, 2020.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.

Jonathan Lorraine and David Duvenaud. Stochastic hyperparameter optimization through hypernetworks. *arXiv preprint arXiv:1802.09419*, 2018.

Matthew MacKay, Paul Vicol, Jonathan Lorraine, David Duvenaud, and Roger B. Grosse. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *International Conference on Learning Representations*, 2019.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

Takayuki Okuno, Akiko Takeda, Akihiro Kawana, and Motokazu Watanabe. On lp-hyperparameter learning via bilevel nonsmooth optimization. *Journal of Machine Learning Research*, 22:245:1–245:47, 2021.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018.

William Rogers and T. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6:506–514, 1978.

Wei Shen, Zhenhuan Yang, Yiming Ying, and Xiaoming Yuan. Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, 18(05):887–927, 2020.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.

Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems*, pages 4147–4157, 2019.

Xuelin Zhang, Yingjie Wang, Liangxuan Zhu, Hong Chen, Han Li, and Lingjuan Wu. Robust variable structure discovery based on tilted empirical risk minimization. *Applied Intelligence*, 53(14):17865–17886, 2023.

Xiao Zhou, Renjie Pi, Weizhong Zhang, Yong Lin, Zonghao Chen, and Tong Zhang. Probabilistic bilevel coreset selection. In *International Conference on Machine Learning*, pages 27287–27302. PMLR, 2022.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of sgd in nonconvex optimization. *Machine Learning*, pages 1–31, 2022.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations*, 2019.