On the Convergence of Stochastic Compositional Gradient Descent Ascent Method

Hongchang Gao[1*], **Xiaoqian Wang**[2], **Lei Luo**[3] and **Xinghua Shi**[1]

[1]Department of Computer and Information Sciences, Temple University, PA, USA
[2]School of Electrical and Computer Engineering, Purdue University, IN, USA
[3]JD Finance America Corporation, Mountain View, CA, USA
hongchang.gao@temple.edu, joywang@purdue.edu, luoleipitt@gmail.com, mindyshi@temple.edu

Abstract

The compositional minimax problem covers plenty of machine learning models such as the distributionally robust compositional optimization problem. However, it is yet another understudied problem to optimize the compositional minimax problem. In this paper, we develop a novel efficient stochastic compositional gradient descent ascent method for optimizing the compositional minimax problem. Moreover, we establish the theoretical convergence rate of our proposed method. To the best of our knowledge, this is the first work achieving such a convergence rate for the compositional minimax problem. Finally, we conduct extensive experiments to demonstrate the effectiveness of our proposed method.

1 Introduction

In recent years, minimax optimization has attracted increasing attention in the community of machine learning. This is mainly due to the fact that a broad range of machine learning models can be formulated as the minimax optimization problem, including generative adversarial networks [Goodfellow *et al.*, 2014], adversarial training of deep neural networks [Madry *et al.*, 2017], and distributionally robust machine learning models [Chen *et al.*, 2017]. In the meanwhile, numerous machine learning models can be formulated as the compositional optimization problem, such as policy evaluation in reinforcement learning [Sutton and Barto, 2018], risk-averse portfolio optimization [Rockafellar, 2007], and sparse additive models [Wang *et al.*, 2017a]. Given the growing importance of the compositional minimax problem in machine learning, in this paper, we are interested in optimizing the compositional minimax problem as follows:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(g(x),y) \triangleq \mathbb{E}_\zeta f(\mathbb{E}_\xi[g(x;\xi)],y;\zeta)\ , \qquad (1)$$

where $g(\cdot):\mathcal{X}\to\mathbb{R}^p$, $f(\cdot,\cdot):(\mathbb{R}^p,\mathcal{Y})\to\mathbb{R}$, $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact sets. In particular, $f(g(x),y)$ is a *strongly concave* function with respect to y for all $x\in\mathcal{X}$, and $f(g(x),y)$ is a *nonconvex* compositional function regarding x for all $y\in\mathcal{Y}$. A typical example of this kind of compositional minimax problem is the distributionally robust compositional optimization problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{\sum_{i=1}^n\left(y_i f_i\Big(\frac{1}{m}\sum_{j=1}^m g_j(x)\Big)-(y_i-\frac{1}{n})^2\right)\right\}, \qquad (2)$$

where $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{Y}=\{y\in\mathbb{R}^n|\sum_{i=1}^n y_i=1, y_i\geq 0,\forall i\}$. The loss function for each sample $f_i\Big(\frac{1}{m}\sum_{j=1}^m g_j(x)\Big)$ is a compositional function, and $y\in\mathbb{R}^n$ weights each sample to handle noisy data to learn a robust model. The goal of this paper is to develop efficient algorithms to solve the compositional minimax optimization problem in Eq. (1).

A lot of efforts [Lin *et al.*, 2020; Luo *et al.*, 2020; Xu *et al.*, 2020; Huang *et al.*, 2020; Chen *et al.*, 2020; Zhang *et al.*, 2020; Tran-Dinh *et al.*, 2020; Yan *et al.*, 2020] have been recently made to develop efficient algorithms to optimize the minimax optimization problem. For instance, [Lin *et al.*, 2020] proposed a stochastic gradient descent ascent method for the nonconvex-strongly-concave minimax problem. [Luo *et al.*, 2020; Xu *et al.*, 2020] developed a variance reduced stochastic gradient descent ascent (SGDA) method to accelerate the convergence speed of SGDA. [Huang *et al.*, 2020] studied an optimization method for the minimax problem on Riemannian manifold. [Chen *et al.*, 2020] presented a projection-free approach for the convex-strongly-concave constrained minimax problem. However, all of these approaches only focus on the *non-compositional* machine learning model where a stochastic gradient is an unbiased estimation of the full gradient. Thus, these methods are not applicable to optimize the compositional minimax optimization problem in Eq. (1).

Due to the two-level stochasticity in the compositional loss function, a stochastic gradient is a *biased* estimation of the full gradient. It is much more challenging to optimize the *minimax* problem in Eq. (1). Although there exist some works studying how to deal with the two-level stochasticity in optimizing the compositional *minimization* problem, these prior work is not applicable for the minimax problem in Eq. (1). For instance, [Wang *et al.*, 2017a] proposed a stochastic compositional gradient descent (SCGD) method to handle the two-level stochasticity. After that, a string of variance reduced variants [Yuan *et al.*, 2019; Yang and Hu, 2020;

*Corresponding author

Zhang and Xiao, 2019b] have been proposed to accelerate the convergence speed. However, for these methods, it is not clear how to optimize the maximization subproblem in Eq. (1). Thus, it is necessary and important to design new optimization methods for Eq. (1).

In this study, to address the aforementioned challenges, we develop a novel stochastic compositional gradient descent ascent (SCGDA) method for optimizing Eq. (1). In particular, when optimizing the minimization subproblem, we use a new strategy to estimate the inner function value $g(x)$ and its gradient $\nabla g(x)$, which helps to control the estimation variance of the stochastic compositional gradient. Meanwhile, when optimizing the maximization problem, we employ the standard stochastic gradient ascent method to update y. Our theoretical result indicates that SCGDA can achieve the convergence rate of $O(\kappa^4/\epsilon^4)$, where κ is the condition number of the loss function. To the best of our knowledge, this is the first work achieving such a convergence rate. Additionally, our experimental results confirm the effectiveness of the proposed optimization algorithm. The contributions of this work are summarized as follows:

- We proposed a new optimization algorithm to optimize the compositional *minimax* optimization problem. This is the first work studying how to optimize this kind of minimax compositional problem.
- We theoretically proved that our proposed algorithm enjoys a convergence rate of $O(\kappa^4/\epsilon^4)$. This is the first work achieving such a convergence rate.
- We conducted extensive experiments and our experimental results confirm the effectiveness of the proposed algorithm.

2 Related Works

In this section, we briefly revisit related works to motivate our proposed approaches.

Minimax optimization problem The minimax optimization problem is an important type of models in machine learning, which motivates numerous efforts to develop efficient algorithms to solve the problem. Currently, the basic idea for minimax optimization is to alternatively optimize the minimization and maximization subproblems. For instance, based on the regular SGD method, [Lin *et al.*, 2020] developed the stochastic gradient descent ascent (SGDA) method, which uses stochastic gradient to alternatively update the two subproblems. [Luo *et al.*, 2020; Xu *et al.*, 2020] applied a recursive variance reduction technique [Fang *et al.*, 2018] for SGDA to improve the convergence rate of SGDA. [Yan *et al.*, 2020] provided an epoch-wise stochastic gradient descent ascent method for strongly-convex-strongly-concave problems. Moreover, [Huang *et al.*, 2020] proposed the Riemannian stochastic gradient descent ascent method and some variants for the Riemannian minimax optimization problem. [Qiu *et al.*, 2020] reformulated nonlinear Temporal-Difference (TD) learning as a minimax optimization problem and proposed the single-timescale stochastic gradient descent ascent method. [Chen *et al.*, 2020] developed a stochastic Frank-Wolfe method to optimize the constrained minimax problem. All of these methods fail to handle the compositional structure in the compositional minimax optimization problem shown in Eq. (1).

Compositional minimization problem The compositional optimization problem is defined as follows:

$$\min_x f(g(x)) \triangleq \mathbb{E}_\zeta f(\mathbb{E}_\xi[g(x;\xi)];\zeta)\ . \tag{3}$$

A typical challenge to optimize the compositional minimization problem is that the stochastic gradient is not an unbiased estimation of the full gradient. To address this issue, various methods have been proposed in recent years. For instance, [Wang *et al.*, 2017a] uses the stochastic gradient for the inner function and uses a momentum-like method to compute the inner function value when computing the stochastic gradient. However, its convergence rate is as slow as $O(1/\epsilon^8)$ for nonconvex problems, which is much worse than the regular SGD method. Consequently, some accelerated variants [Wang *et al.*, 2017a; Wang *et al.*, 2017b] have been proposed to improve the convergence rate. However, their convergence rates are still inferior to that of the regular SGD method. Recently, based on the variance reduction technique such as SPIDER [Fang *et al.*, 2018]and STORM [Cutkosky and Orabona, 2019], a series of methods [Zhang and Xiao, 2019b; Zhang and Xiao, 2019a; Yuan *et al.*, 2019; Yang and Hu, 2020] have been proposed to further improve the convergence rate. Nonetheless, these methods can only be applied to the compositional *minimization* problem rather than the *minimax* problem.

To sum up, the aforementioned minimax optimization algorithms and compositional minimization algorithms cannot be applied to optimize the compositional minimax problem defined in Eq. (1). Therefore, in this study, we will develop new optimization algorithms to optimize this challenging problem.

Algorithm 1 The Stochastic Compositional Gradient Descent Ascent Method (SCGDA)

Initialization: x_1, $y_1 = y^*(x_1)$, $u_1' = \nabla g(x_1)$, $u_1 = g(x_1)$, $w_1 = \nabla_y f(u_1, y_1)$, $v_1 = (u_t')^T w_1$, $\alpha > 0$, $\gamma > 0$, $\lambda > 0$, $\eta_t > 0$

1: **for** $t = 1, \cdots, T$ **do**
2: Update x:
$x_{t+1} = x_t - \gamma \eta_t v_t$
3: Update y:
$\tilde{y}_{t+1} = \mathcal{P}(y_t + \lambda w_t)$
$y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$
4: Randomly sample a subset $B_{\xi,t+1}$, compute $\nabla g(x_{t+1}; B_{\xi,t+1})$ and $g(x_{t+1}; B_{\xi,t+1})$,
$u_{t+1} = (1-\alpha\eta_t)u_t + \alpha\eta_t g(x_{t+1}; B_{\xi,t+1})$
$u'_{t+1} = (1-\alpha\eta_t)u'_t + \alpha\eta_t \nabla g(x_{t+1}; B_{\xi,t+1})$
5: Randomly sample a subset $B_{\zeta,t+1}$, compute $\nabla_g f(u_{t+1}, y_{t+1}; B_{\zeta,t+1})$ and $\nabla_y f(u_{t+1}, y_{t+1}; B_{\zeta,t+1})$,
$v_{t+1} = (u'_{t+1})^T \nabla_g f(u_{t+1}, y_{t+1}; B_{\zeta,t+1})$
$w_{t+1} = \nabla_y f(u_{t+1}, y_{t+1}; B_{\zeta,t+1})$
6: **end for**

3 Stochastic Compositional Gradient Descent Ascent Method

In this section, we present the details of our proposed algorithm and provide its theoretical convergence rate.

In Algorithm 1, we propose a stochastic compositional gradient descent ascent (SCGDA) method to optimize the compositional minimax problem defined in Eq. (1). Here, a key challenge is that the stochastic gradient regarding x is NOT an unbiased estimation of the full compositional gradient shown as follows:

$$\mathbb{E}_{\xi,\zeta}[\nabla g(x;\xi)^T \nabla_g f(g(x;\xi),y;\zeta)] \neq \nabla g(x)^T \nabla_g f(g(x),y) \ . \tag{4}$$

To alleviate this issue, when optimizing the compositional *minimization* problem, [Wang *et al.*, 2017a] employs the following strategy[1] to estimate the compositional gradient:

$$\begin{aligned} u_t &= (1-\alpha\eta_{t-1})u_{t-1} + \alpha\eta_{t-1} g(x_t;B_{\xi,t}) \ , \\ v_t &= \nabla g(x_t;B_{\xi,t})^T \nabla_g f(u_t;B_{\zeta,t}) \ , \end{aligned} \tag{5}$$

where $g(x_t;B_{\xi,t}) = \frac{1}{|B_{\xi,t}|}\sum_{j\in B_{\xi,t}} g_j(x_t)$, $\nabla g(x_t;B_{\xi,t}) = \frac{1}{|B_{\xi,t}|}\sum_{j\in B_{\xi,t}} \nabla g_j(x_t)$, $\alpha > 0$, and $\eta_t > 0$. Here, u_t estimates $g(x)$ and v_t is the estimation of $\nabla g(x)^T \nabla_g f(g(x))$. With the strategy in Eq. (5), u_t is supposed to have a smaller estimation variance than $g(x_t;B_{\xi,t})$. In fact, this strategy of controlling variance is also used in the stochastic Frank-Wolfe method [Mokhtari *et al.*, 2020] and TD learning [Qiu *et al.*, 2020]. However, theoretical results in [Wang *et al.*, 2017a] indicate that its convergence rate is as slow as $O(1/\epsilon^8)$.

In Algorithm 1, when $t > 1$, we use the following strategy to compute u_t to estimate the inner function $g(x_t)$ and u_t' for its gradient $\nabla g(x_t)$:

$$\begin{aligned} u_t &= (1-\alpha\eta_{t-1})u_{t-1} + \alpha\eta_{t-1} g(x_t;B_{\xi,t}) \ , \\ u_t' &= (1-\alpha\eta_{t-1})u_{t-1}' + \alpha\eta_{t-1} \nabla g(x_t;B_{\xi,t}) \ . \end{aligned} \tag{6}$$

In this way, both u_t and u_t' will have a small estimation variance, which will be shown in the next section. Based on u_t and u_t' , we compute the stochastic compositional gradient regarding x as follows:

$$v_t = (u_t')^T \nabla_g f(u_t,y_t;B_{\zeta,t}) \ , \tag{7}$$

where $\nabla_g f(u_t,y_t;B_{\zeta,t}) = \frac{1}{|B_{\zeta,t}|}\sum_{i\in B_{\zeta,t}} \nabla_g f_i(u_t,y_t)$. Then, we use the following strategy to update x:

$$x_{t+1} = x_t - \gamma\eta_t v_t \ , \tag{8}$$

where $\gamma > 0$. This strategy can facilitate tightly bounding the estimation variance $\mathbb{E}[\|u_t - g(x_t)\|^2]$ and $\mathbb{E}[\|u_t' - \nabla g(x_t)\|^2]$, which will be demonstrated in the next section.

To optimize the maximization subproblem regarding y, we directly compute the stochastic gradient $w_t = \nabla_y f(u_t,y_t;B_{\zeta,t}) = \frac{1}{|B_{\zeta,t}|}\sum_{i\in B_{\zeta,t}} \nabla_y f_i(u_t,y_t)$ and update y as follows:

$$\begin{aligned} \tilde{y}_{t+1} &= \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t) \ , \\ y_{t+1} &= y_t + \eta_t(\tilde{y}_{t+1} - y_t) \ , \end{aligned} \tag{9}$$

where $\lambda > 0$, $0 < \eta_t < 1$, and $\mathcal{P}_{\mathcal{Y}}(\cdot)$ is a projection operator to make sure that $\tilde{y}_{t+1}$ satisfy the constraint $\mathcal{Y}$. Since $\mathcal{Y}$ is a convex set, the combination between $\tilde{y}_{t+1}$ and y_t in Eq. (9) still lies in $\mathcal{Y}$.

In the following, we will investigate the convergence rate of Algorithm 1. We first introduce some commonly-used assumptions for compositional optimization. To make these assumptions easy to follow, we denote $\nabla f(a,b) = (\nabla_a f(a,b), \nabla_b f(a,b))$ for $(a,b) \in \mathcal{A}\times\mathcal{B}$ where $\mathcal{A} = \{g(x)|x\in\mathcal{X}\}$ and $\mathcal{B} = \mathcal{Y}$.

Assumption 1. *(Smoothness) For $\forall (a_1,b_1),(a_2,b_2) \in \mathcal{A}\times\mathcal{B}$, there exists $L > 0$, such that*

$$\|\nabla f(a_1,b_1) - \nabla f(a_2,b_2)\| \leq L\|(a_1,b_1) - (a_2,b_2)\| \ . \tag{10}$$

For $\forall x_1, x_2 \in \mathcal{X}$, there exists $L_g > 0$, such that

$$\|\nabla g(x_1) - \nabla g(x_2)\| \leq L_g\|x_1 - x_2\| \ . \tag{11}$$

Assumption 2. *(Bounded gradient) There exist $C_g > 0$ and $C_f > 0$, such that*

$$\begin{aligned} &\mathbb{E}[\|\nabla g(x;\xi)\|^2] \leq C_g^2, \ \forall x\in\mathcal{X} \ , \\ &\mathbb{E}[\|\nabla f(a,b;\zeta)\|^2] \leq C_f^2, \ \forall (a,b)\in\mathcal{A}\times\mathcal{B} \ . \end{aligned} \tag{12}$$

Assumption 3. *(Bounded variance) There exist $\sigma_f > 0$, $\sigma_g > 0$, and $\sigma_g' > 0$, such that*

$$\begin{aligned} &\mathbb{E}[\|\nabla f(a,b;\zeta) - \nabla f(a,b)\|^2] \leq \sigma_f^2, \ \forall (a,b)\in\mathcal{A}\times\mathcal{B} \ , \\ &\mathbb{E}[\|\nabla g(x;\xi) - \nabla g(x)\|^2] \leq \sigma_{g'}^2, \ \forall x\in\mathcal{X} \ , \\ &\mathbb{E}[\|g(x;\xi) - g(x)\|^2] \leq \sigma_g^2, \ \forall x\in\mathcal{X} \ . \end{aligned} \tag{13}$$

Assumption 4. *(Strong concavity) For $\forall a\in\mathcal{A}$ and $\forall b_1,b_2\in\mathcal{B}$, there exists $\mu > 0$, such that*

$$f(a,b_1) \leq f(a,b_2) + \langle \nabla_b f(a,b_2), b_1 - b_2\rangle - \frac{\mu}{2}\|b_1 - b_2\|^2 \ . \tag{14}$$

Additionally, we introduce two auxiliary functions: $\Phi(x) = \max_{y\in\mathcal{Y}} f(g(x),y)$ and $y^*(x) = \arg\max_{y\in\mathcal{Y}} f(g(x),y)$. Here, $\Phi(x)$ is L_Φ-smooth where $L_\Phi = 2C_g^2L^2/\mu + C_fL_g$. We use Φ_* to represent the minimum value of $\Phi(x)$. Based on these assumptions and definitions, we have the following convergence result.

Theorem 1. *Given Assumptions 1-4, for Algorithm 1, by setting $\alpha = 3$, $\eta_t = \eta \leq \min\{\frac{1}{24}, \frac{1}{2\gamma L_\Phi}\}$, $|B_{\zeta_t}| = |B_{\xi_t}| = B$, $\lambda < \frac{1}{6L}$, and $\gamma < 1\Big/\Big(\frac{10C_gL^2\sqrt{(1+3C_g^2)}}{\lambda\mu^2} + \frac{10C_gL^2\sqrt{(1+3C_g^2)}}{\mu} + \frac{6L_gC_fL}{\mu}\Big)$, we have*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[\|\nabla\Phi(x_t)\| + L\|y^*(x_t) - y_t\|]\Big) \\ &\leq \frac{2\sqrt{(\Phi(x_1) - \Phi_*)}}{\sqrt{\gamma\eta T}} + \frac{2\sqrt{3}C_g\sigma_f}{\sqrt{B}} + \frac{2\sqrt{3}C_f\sigma_{g'}}{\sqrt{B}} \\ &\quad + \frac{5\sqrt{2(1+2C_g^2)}\sigma_f L}{\mu\sqrt{B}} + \frac{5\sqrt{2(1+3C_g^2)}\sigma_g L^2}{\mu\sqrt{B}} \\ &\quad + \frac{6\sigma_{g'}C_fL}{\mu\sqrt{B}} \ . \end{aligned} \tag{15}$$

[1] Note that this strategy is designed for the compositional *minimization* problem, not the compositional *minimax* problem. Here, we just use it to motivate our method.

From Theorem 1, it can be seen that $\eta\gamma = O(\frac{1}{\kappa^2})$ where $\kappa = L/\mu$. Then, by setting $B = T$, we can get the convergence rate $O(\frac{\kappa}{\sqrt{T}})$. Hence, to achieve ϵ-accuracy solution, the total sample complexity is $B \times T = O(\kappa^4/\epsilon^4)$. Note that, although it is incomparable with SCGD [Wang *et al.*, 2017a] for the compositional minimization problem, we observe that our sample complexity has a much better dependence on ϵ.

4 Convergence Analysis

In this section, we present the high-level idea of the convergence analysis for Theorem 1.

As discussed in the last section, our Algorithm 1 results in a tighter bound for $\mathbb{E}[\|u_t - g(x_t)\|^2]$ and $\mathbb{E}[\|u'_t - \nabla g(x_t)\|^2]$, which makes our algorithm converge faster comparing with existing work. In particular, we have the following bound for these two variances.

Lemma 1. *Given Assumptions 1-3 for Algorithm 1, by setting $\eta_t \leq \min\{\frac{1}{8\alpha}, 1\}$ and $|B_{\zeta_t}| = |B_{\xi_t}| = B$, we can get*

$$\begin{aligned}\mathbb{E}[\|u_t - g(x_t)\|^2] &\leq (1-\alpha\eta_{t-1})\mathbb{E}[\|u_{t-1} - g(x_{t-1})\|^2] \\ &+ \frac{9\eta_{t-1}\gamma^2 C_g^2}{8\alpha}\mathbb{E}[\|v_{t-1}\|^2] + \frac{\alpha^2\eta_{t-1}^2\sigma_g^2}{B} .\end{aligned} \tag{16}$$

Lemma 2. *Given Assumptions 1-3 for Algorithm 1, by setting $\eta_t \leq \min\{\frac{1}{8\alpha}, 1\}$ and $|B_{\zeta_t}| = |B_{\xi_t}| = B$, we can get*

$$\begin{aligned}\mathbb{E}[\|u'_t - \nabla g(x_t)\|^2] &\leq (1-\alpha\eta_{t-1})\mathbb{E}[\|u'_{t-1} - \nabla g(x_{t-1})\|^2] \\ &+ \frac{9\eta_{t-1}\gamma^2 L_g^2}{8\alpha}\mathbb{E}[\|v_{t-1}\|^2] + \frac{\alpha^2\eta_{t-1}\sigma_{g'}^2}{B} .\end{aligned} \tag{17}$$

Note that, [Wang *et al.*, 2017a] employs a similar way to estimate $g(x_t)$ and obtains a similar bound for the estimation variance. However, our bound is tighter. In particular, for our bound in Lemma 1, the learning rate $\eta_t < 1$ lies in the nominator of the second term on RHS while it resides in the denominator in the bound of [Wang *et al.*, 2017a]. Thus, our bound for the estimation variance is much tighter.

Furthermore, the following two lemmas demonstrate that controlling the estimation variance of $g(x)$ and $\nabla g(x)$ can benefit controlling the estimation variance of the compositional gradient $\nabla_x f(g(x), y)$ and that of the gradient $\nabla_y f(g(x), y)$.

Lemma 3. *Given Assumptions 1-3, for Algorithm 1, by setting $|B_{\zeta_t}| = |B_{\xi_t}| = B$, we can get*

$$\begin{aligned}\mathbb{E}[\|v_t - \nabla_x f(g(x_t), y_t)\|^2] &\leq 3C_f^2\mathbb{E}[\|u'_t - \nabla g(x_t)\|^2] \\ &+ \frac{3C_g^2\sigma_f^2}{B} + 3C_g^2L^2\mathbb{E}[\|u_t - g(x_t)\|^2] .\end{aligned} \tag{18}$$

Lemma 4. *Given Assumptions 1-3, for Algorithm 1, by setting $|B_{\zeta_t}| = |B_{\xi_t}| = B$, we can get*

$$\begin{aligned}&\mathbb{E}[\|w_t - \nabla_y f(g(x_t), y_t)\|^2] \\ &= L^2\mathbb{E}[\|u_t - g(x_t)\|^2] + \frac{\sigma_f^2}{B} .\end{aligned} \tag{19}$$

Moreover, we need an additional lemma to prove Theorem 1.

Lemma 5. *Given Assumptions 1-3, for Algorithm 1, by setting $\lambda \leq \frac{1}{6L}$ and $\eta_t < 1$, we can get*

$$\begin{aligned}&\mathbb{E}[\|y_{t+1} - y^*(x_{t+1})\|^2] \\ &\leq (1 - \frac{\eta_t\mu\lambda}{4})\mathbb{E}[\|y^*(x_t) - y_t\|^2] - \frac{3\eta_t}{4}\mathbb{E}[\|\tilde{y}_{t+1} - y_t\|^2] \\ &+ \frac{25\eta_t\lambda}{6\mu}\mathbb{E}[\|\nabla_y f(g(x_t), y_t) - w_t\|^2] \\ &+ \frac{25\eta_t\gamma^2L^2C_g^2}{6\lambda\mu^3}\mathbb{E}[\|v_t\|^2] .\end{aligned} \tag{20}$$

Based on aforementioned lemmas, we will provide the main proof for Theorem 1.

Proof. Due to the smoothness of $\Phi(x)$, we can get

$$\begin{aligned}\Phi(x_{t+1}) &\leq \Phi(x_t) - \gamma\eta_t\langle\nabla\Phi(x_t), v_t\rangle + \frac{\gamma^2\eta_t^2L_\Phi}{2}\|v_t\|^2 \\ &= \Phi(x_t) - \frac{\gamma\eta_t}{2}\|\nabla\Phi(x_t)\|^2 + \Big(\frac{\gamma^2\eta_t^2L_\Phi}{2} - \frac{\gamma\eta_t}{2}\Big)\|v_t\|^2 \\ &+ \frac{\gamma\eta_t}{2}\|\nabla\Phi(x_t) - v_t\|^2 \\ &\leq \Phi(x_t) - \frac{\gamma\eta_t}{2}\|\nabla\Phi(x_t)\|^2 + \Big(\frac{\gamma^2\eta_t^2L_\Phi}{2} - \frac{\gamma\eta_t}{2}\Big)\|v_t\|^2 \\ &+ \gamma\eta_t\|\nabla\Phi(x_t) - \nabla_x f(g(x_t), y_t)\|^2 \\ &+ \gamma\eta_t\|\nabla_x f(g(x_t), y_t) - v_t\|^2 \\ &\leq \Phi(x_t) - \frac{\gamma\eta_t}{2}\|\nabla\Phi(x_t)\|^2 + \gamma\eta_tC_g^2L^2\|y^*(x_t) - y_t\|^2 \\ &+ \gamma\eta_t\|\nabla_x f(g(x_t), y_t) - v_t\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2 ,\end{aligned} \tag{21}$$

where the last inequality follows from $\eta_t \leq \frac{1}{2\gamma L_\Phi}$ and the following inequality:

$$\begin{aligned}&\|\nabla\Phi(x_t) - \nabla_x f(g(x_t), y_t)\|^2 \\ &= \|\nabla_x f(g(x_t), y^*(x_t)) - \nabla_x f(g(x_t), y_t)\|^2 \\ &= \|\nabla g(x_t)^T\nabla_g f(g(x_t), y^*(x_t)) \\ &\quad - \nabla g(x_t)^T\nabla_g f(g(x_t), y_t)\|^2 \\ &\leq C_g^2L^2\|y^*(x_t) - y_t\|^2 ,\end{aligned} \tag{22}$$

where the last inequality follows from Assumptions 1 and 2. Furthermore, according to Lemma 3 and taking expectation for both sides, we can get

$$\begin{aligned}&\mathbb{E}[\Phi(x_{t+1})] \\ &\leq \mathbb{E}[\Phi(x_t)] - \frac{\gamma\eta_t}{2}\mathbb{E}[\|\nabla\Phi(x_t)\|^2] \\ &+ \gamma\eta_tC_g^2L^2\mathbb{E}[\|y^*(x_t) - y_t\|^2] - \frac{\gamma\eta_t}{4}\mathbb{E}[\|v_t\|^2] \\ &+ 3\gamma\eta_tC_f^2\mathbb{E}[\|u'_t - \nabla g(x_t)\|^2] \\ &+ 3\gamma\eta_tC_g^2L^2\mathbb{E}[\|u_t - g(x_t)\|^2] + \frac{3\gamma\eta_tC_g^2\sigma_f^2}{B} .\end{aligned} \tag{23}$$

Furthermore, define the Lyapunov function

$$\begin{aligned} P_t &= \mathbb{E}[\Phi(x_t)] + A_1\mathbb{E}[\|y_t - y^*(x_t)\|^2] \\ &\quad + A_2\mathbb{E}[\|u_t - g(x_t)\|^2] + A_3\mathbb{E}[\|u_t' - \nabla g(x_t)\|^2] \,, \end{aligned} \tag{24}$$

where $A_1 = \frac{\gamma L^2(1+2C_g^2)}{\lambda\mu}$, $A_2 = \frac{25\gamma L^4(1+3C_g^2)}{6\alpha\mu^2}$, and $A_3 = \frac{3\gamma C_f^2 L^2}{\alpha\mu^2}$, then according to Lemmas 1, 2, 4, 5, we get

$$\begin{aligned} & P_{t+1} - P_t \\ \leq & -\frac{\gamma\eta_t}{2}\mathbb{E}[\|\nabla\Phi(x_t)\|^2] + \gamma\eta_t C_g^2 L^2\mathbb{E}[\|y^*(x_t) - y_t\|^2] \\ & + 3\gamma\eta_t C_f^2\mathbb{E}[\|u_t' - \nabla g(x_t)\|^2] + \frac{3\gamma\eta_t C_g^2\sigma_f^2}{B} \\ & + 3\gamma\eta_t C_g^2 L^2\mathbb{E}[\|u_t - g(x_t)\|^2] + \frac{3\gamma\eta_t C_f^2\sigma_{g'}^2}{B} \\ & - \frac{\eta_t\mu\lambda A_1}{4}\mathbb{E}[\|y^*(x_t) - y_t\|^2] - \frac{3\eta_t A_1}{4}\mathbb{E}[\|\tilde{y}_{t+1} - y_t\|^2] \\ & + \frac{25\eta_t\lambda A_1}{6\mu}\mathbb{E}[\|\nabla_y f(g(x_t), y_t) - w_t\|^2] - \frac{\gamma\eta_t}{4}\mathbb{E}[\|v_t\|^2] \\ & + \frac{25\eta_t\gamma^2 L^2 C_g^2 A_1}{6\lambda\mu^3}\mathbb{E}[\|v_t\|^2] - \alpha\eta_t A_2\mathbb{E}[\|u_t - g(x_t)\|^2] \\ & + \frac{9\eta_t\gamma^2 C_g^2 A_2}{8\alpha}\mathbb{E}[\|v_t\|^2] + \frac{\alpha^2\eta_t\sigma_g^2 A_2}{B} + \frac{\alpha^2\eta_t\sigma_{g'}^2 A_3}{B} \\ & - \alpha\eta_t A_3\mathbb{E}[\|u_t' - \nabla g(x_t)\|^2] + \frac{9\eta_t\gamma^2 L_g^2 A_3}{8\alpha}\mathbb{E}[\|v_t\|^2] \\ \leq & -\frac{\gamma\eta_t}{2}\mathbb{E}[\|\nabla\Phi(x_t)\|^2] - \frac{3\eta_t A_1}{4}\mathbb{E}[\|\tilde{y}_{t+1} - y_t\|^2] \\ & - (\frac{\eta_t\mu\lambda A_1}{4} - \gamma\eta_t C_g^2 L^2)\mathbb{E}[\|y^*(x_t) - y_t\|^2] \\ & + (3\gamma\eta_t C_g^2 L^2 + \frac{25\eta_t\lambda L^2 A_1}{6\mu} - \alpha\eta_t A_2)\mathbb{E}[\|u_t - g(x_t)\|^2] \\ & + (3\gamma\eta_t C_f^2 - \alpha\eta_t A_3)\mathbb{E}[\|u_t' - \nabla g(x_t)\|^2] \\ & + \Big(\frac{9\eta_t\gamma^2 C_g^2 A_2}{8\alpha} + \frac{25\eta_t\gamma^2 L^2 C_g^2 A_1}{6\lambda\mu^3} + \frac{9\eta_t\gamma^2 L_g^2 A_3}{8\alpha} \\ & - \frac{\gamma\eta_t}{4}\Big)\mathbb{E}[\|v_t\|^2] + \frac{\alpha^2\eta_t\sigma_g^2 A_2}{B} + \frac{\alpha^2\eta_t\sigma_{g'}^2 A_3}{B} \\ & + \frac{3\gamma\eta_t C_g^2\sigma_f^2}{B} + \frac{3\gamma\eta_t C_f^2\sigma_{g'}^2}{B} + \frac{25\eta_t\lambda\sigma_f^2 A_1}{6\mu B} \,. \end{aligned} \tag{25}$$

By setting $\alpha = 3$ and $\gamma < 1\Big/\Big(\frac{10C_g L^2\sqrt{(1+3C_g^2)}}{\lambda\mu^2} + \frac{10C_g L^2\sqrt{(1+3C_g^2)}}{\mu} + \frac{6L_g C_f L}{\mu}\Big)$, we can get

$$\begin{aligned} & P_{t+1} - P_t \\ \leq & -\frac{\gamma\eta_t}{2}\mathbb{E}[\|\nabla\Phi(x_t)\|^2] - \frac{\gamma\eta_t L^2}{2}\mathbb{E}[\|y^*(x_t) - y_t\|^2] \\ & + \frac{3\gamma\eta_t C_g^2\sigma_f^2}{B} + \frac{3\gamma\eta_t C_f^2\sigma_{g'}^2}{B} + \frac{25\eta_t\lambda\sigma_f^2 A_1}{6\mu B} \\ & + \frac{9\eta_t\sigma_g^2 A_2}{B} + \frac{9\eta_t\sigma_{g'}^2 A_3}{B} \,. \end{aligned} \tag{26}$$

By summing t over $1, \cdots, T$ and setting $\eta_t = \eta$, we get

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[\|\nabla\Phi(x_t)\|^2 + L^2\|y^*(x_t) - y_t\|^2]\Big) \\ \leq & \frac{2(P_1 - P_{T+1})}{\gamma\eta T} + \frac{6C_g^2\sigma_f^2}{B} + \frac{6C_f^2\sigma_{g'}^2}{B} \\ & + \frac{50\sigma_f^2 L^2(1+2C_g^2)}{6\mu^2 B} + \frac{50\sigma_g^2 L^4(1+3C_g^2)}{2\mu^2 B} \\ & + \frac{18\sigma_{g'}^2 C_f^2 L^2}{\mu^2 B} \,. \end{aligned} \tag{27}$$

From the initialization condition, it is easy to get

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[\|\nabla\Phi(x_t)\|^2 + L^2\|y^*(x_t) - y_t\|^2]\Big) \\ \leq & \frac{2(\Phi(x_1) - \Phi_*)}{\gamma\eta T} + \frac{6C_g^2\sigma_f^2}{B} + \frac{6C_f^2\sigma_{g'}^2}{B} \\ & + \frac{50\sigma_f^2 L^2(1+2C_g^2)}{6\mu^2 B} + \frac{50\sigma_g^2 L^4(1+3C_g^2)}{2\mu^2 B} \\ & + \frac{18\sigma_{g'}^2 C_f^2 L^2}{\mu^2 B} \,. \end{aligned} \tag{28}$$

Finally, we can get

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[\|\nabla\Phi(x_t)\| + L\|y^*(x_t) - y_t\|]\Big) \\ \leq & \sqrt{\frac{2}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[\|\nabla\Phi(x_t)\|^2 + L^2\|y^*(x_t) - y_t\|^2]\Big)} \\ \leq & \frac{2\sqrt{(\Phi(x_1) - \Phi_*)}}{\sqrt{\gamma\eta T}} + \frac{2\sqrt{3}C_g\sigma_f}{\sqrt{B}} + \frac{2\sqrt{3}C_f\sigma_{g'}}{\sqrt{B}} \\ & + \frac{5\sqrt{2(1+2C_g^2)}\sigma_f L}{\mu\sqrt{B}} + \frac{5\sqrt{2(1+3C_g^2)}\sigma_g L^2}{\mu\sqrt{B}} \\ & + \frac{6\sigma_{g'}C_f L}{\mu\sqrt{B}} \,, \end{aligned} \tag{29}$$

which completes the proof. □

5 Experiments

In this section, we conduct experiments to verify the convergence of our proposed SCGDA in Algorithm 1.

In our experiments, we use SCGDA to optimize the distributionally robust linear value function approximation in reinforcement learning [Zhang and Xiao, 2019a]. The value function in reinforcement learning is an important component to compute the reward. In detail, given a Markov decision process (MDP) $\{\mathcal{S}, P^\pi, R, r\}$ where $\mathcal{S} = \{1, 2, \cdots, S\}$ represents the state space, $P^\pi_{s,s'}$ denotes the transition probability from state s to state s' for a given policy π, $R_{s,s'}$ is the reward when state s goes to state s', and r is the discount factor, then the value function at state s is defined as

$V(s) = \sum_{s'=1}^{S} P^{\pi}_{s,s'}(R_{s,s'} + rV(s'))$. To estimate the value function, a typical choice is to parameterize it with a linear function: $\tilde{V}_w(s) = z_s^T w$ where $z_s \in \mathbb{R}^d$ is fixed and $w \in \mathbb{R}^d$ is the model parameter which needs to be optimized. To obtain the model parameter w, we need to optimize the following problem:

$$\min_{w} \frac{1}{S} \sum_{s=1}^{S} \Big(\tilde{V}_w(s) - \sum_{s'=1}^{S} P^{\pi}_{s,s'}(R_{s,s'} + r\tilde{V}_w(s')) \Big)^2 . \quad (30)$$

Obviously, the loss function is a compositional function [Yuan *et al.*, 2019; Zhang and Xiao, 2019c]. Here, we are interested in its distributionally robust variant, which is defined as follows:

$$\min_{w \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} \frac{1}{S} \sum_{s=1}^{S} y_s \Big(\tilde{V}_w(s) - \sum_{s'=1}^{S} P^{\pi}_{s,s'}(R_{s,s'} + r\tilde{V}_w(s')) \Big)^2 + \sum_{i=1}^{d} \frac{\beta w_i^2}{1 + w_i^2} - \|y - \frac{1}{S}\|^2 , \quad (31)$$

where $w = [w_i] \in \mathbb{R}^d$ is the model parameter, $\mathcal{Y} = \{y = [y_s] \in \mathbb{R}^S | \sum_{s=1}^{S} y_s = 1, y_s \geq 0, \forall s\}$, $\beta > 0$. Obviously, it is nonconvex regarding w and strongly concave regarding y.

Following [Yuan *et al.*, 2019], we generate an MDP which has 400 states and each state is associated with 10 actions. Regarding the transition probability, $P^{\pi}_{s,s'}$ is drawn from $[0, 1]$ uniformly. Additionally, to guarantee the ergodicity, we add 10^{-5} to $P^{\pi}_{s,s'}$. Then, we use SCGDA to optimize Eq. (31) on this dataset. Note that, our method is the first stochastic compositional gradient descent ascent method where there are no baseline methods to compare with. Thus, we will conduct experiments to show the convergence of our proposed SCGDA under different conditions.

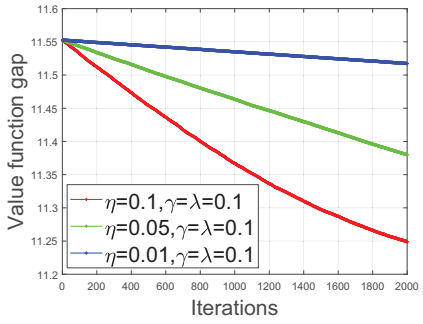

Figure 1: The convergence of SCGDA with different η.

In our experiments, we set the batch size to 20, $\alpha = 3$, $\beta = 10^{-5}$. Then, we verify the convergence performance of SCGDA with different learning rates η. Specifically, in Figure 1, we fix $\gamma = \lambda = 0.1$ and change η to show the value function gap $\frac{1}{S}\sum_{s=1}^{S} \Big(\tilde{V}_w(s) - \sum_{s'=1}^{S} P^{\pi}_{s,s'}(R_{s,s'} + r\tilde{V}_w(s')) \Big)^2$ regarding the number of iterations. It can be seen that the value function gap decreases with the training going on for all cases, which confirms that our algorithm is effective to optimize Eq. (31). Additionally, with a larger learning rate $\eta = 0.1$, our SCGDA method converges much faster than the other two cases.

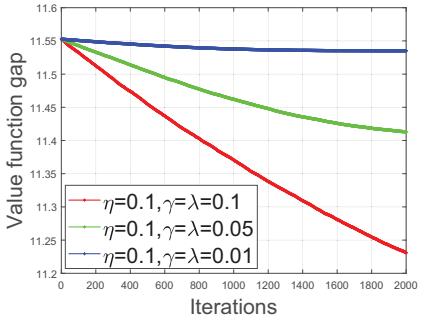

Figure 2: The convergence of SCGDA with different λ and γ.

Furthermore, in Figure 2, we fix the learning rate η and change λ, as well as γ. Here, we set $\lambda = \gamma$ to make the minimization subproblem and maximization subproblem update in the single-timescale manner. Similarly, from Figure 2, it can be seen that the value function gap decreases with the optimization going on for all cases, and larger parameters λ, γ lead to a much faster convergence rate.

Finally, we plot the learned parameter y in Figure 3. Here, we set η to 0.1 and $\gamma = \lambda = 0.1$. Then, we plot $y - \frac{1}{S}$ in Figure 3 for all sates. We observe that our method can learn different weights for different states, which is consistent with the idea of the distributionally robust optimization.

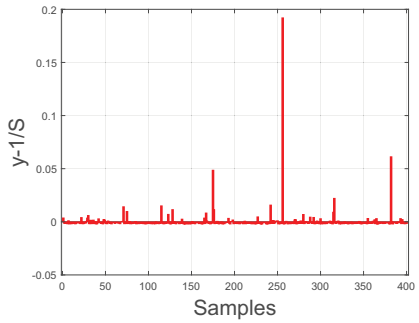

Figure 3: The learned y with $\eta = 0.1$, $\gamma = 0.1$, and $\lambda = 0.1$.

6 Conclusions

In this paper, we proposed a novel stochastic compositional gradient descent ascent method for optimizing the compositional minimax problem. We provided the theoretical convergence rate of the proposed method for the nonconvex-strongly-concave compositional minimax problem. We then conducted extensive experimental results that confirmed the effectiveness of our proposed algorithm.

Acknowledgments

This work is partially supported by NSF (#1955890 to X.W. and #1750632 to X.S.).

References

[Chen *et al.*, 2017] Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. *Advances in Neural Information Processing Systems*, 30:4705–4714, 2017.

[Chen *et al.*, 2020] Cheng Chen, Luo Luo, Weinan Zhang, and Yong Yu. Efficient projection-free algorithms for saddle point problems. *arXiv preprint arXiv:2010.11737*, 2020.

[Cutkosky and Orabona, 2019] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

[Fang *et al.*, 2018] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[Huang *et al.*, 2020] Feihu Huang, Shangqian Gao, and Heng Huang. Gradient descent ascent for min-max problems on riemannian manifold. *arXiv preprint arXiv:2010.06097*, 2020.

[Lin *et al.*, 2020] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[Luo *et al.*, 2020] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mokhtari *et al.*, 2020] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21(105):1–49, 2020.

[Qiu *et al.*, 2020] Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *arXiv preprint arXiv:2008.10103*, 2020.

[Rockafellar, 2007] R Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. Informs, 2007.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Tran-Dinh *et al.*, 2020] Quoc Tran-Dinh, Deyi Liu, and Lam M Nguyen. Hybrid variance-reduced sgd algorithms for nonconvex-concave minimax problems. *arXiv preprint arXiv:2006.15266*, 2020.

[Wang *et al.*, 2017a] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[Wang *et al.*, 2017b] Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *The Journal of Machine Learning Research*, 18(1):3721–3743, 2017.

[Xu *et al.*, 2020] Tengyu Xu, Zhe Wang, Yingbin Liang, and H Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *arXiv preprint arXiv:2006.09361*, 2020.

[Yan *et al.*, 2020] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[Yang and Hu, 2020] Jiaojiao Yang and Wenqing Hu. Stochastic recursive momentum method for non-convex compositional optimization. *arXiv preprint arXiv:2006.01688*, 2020.

[Yuan *et al.*, 2019] Huizhuo Yuan, Xiangru Lian, and Ji Liu. Stochastic recursive variance reduction for efficient smooth non-convex compositional optimization. *arXiv preprint arXiv:1912.13515*, 2019.

[Zhang and Xiao, 2019a] Junyu Zhang and Lin Xiao. A composite randomized incremental gradient method. In *International Conference on Machine Learning*, pages 7454–7462, 2019.

[Zhang and Xiao, 2019b] Junyu Zhang and Lin Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.

[Zhang and Xiao, 2019c] Junyu Zhang and Lin Xiao. A stochastic composite gradient method with incremental variance reduction. In *Advances in Neural Information Processing Systems*, pages 9078–9088, 2019.

[Zhang *et al.*, 2020] Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhi-Quan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *arXiv preprint arXiv:2010.15768*, 2020.