ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447666

A METHOD FOR BILEVEL OPTIMIZATION WITH CONVEX LOWER-LEVEL PROBLEM

Han Shen* *Santiago Paternain** *Gaowen Liu*† *Ramana Kompella*† *Tianyi Chen**

* Rensselaer Polytechnic Institute, United States
† Cisco Systems, United States

ABSTRACT

Gradient-based bilevel optimization methods have been applied to a wide range of applications including hyper-parameter optimization, meta-learning, and model pruning. However, it is known that the bilevel optimization problem is difficult to solve, and the finite-time guarantee has only been established for simpler bilevel problems with a strongly-convex lower-level problem. In this work, we propose an iterative bilevel optimization method that sequentially solves simple approximate problems of the original problem. Despite the lack of strong convexity in the lower level, we show that the proposed method converges to an ϵ-stationary-point with an iteration complexity of $\mathcal{O}(\epsilon^{-1})$. Experiments have verified the effectiveness of the method.

***Index Terms*—** Bilevel optimization, difference-of-convex, convex-concave procedure, hyperparameter tunning

1. INTRODUCTION

We consider the following bilevel optimization problem

$$\min_{x\in C, y\in\mathbb{R}^{d_y}} f(x,y) \qquad \text{s.t.} \quad u(x,y)\leq 0$$
$$y\in \arg\min_{y'\in U(x)} g(x,y') \qquad (1)$$

where C is a closed convex subset of $\mathbb{R}^{d_x}$; f is a differentiable real function on $\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$; g and u are convex functions on $\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$; $U(x)$ is $U(x) \coloneqq \{y\in\mathbb{R}^{d_y} : u(x,y)\leq 0\}$ and is assumed to be nonempty for any $x\in C$.

Bilevel optimization plays an increasingly important role in the machine learning area. Bilevel optimization method has been applied to hyper-parameter optimization [1, 2], active learning [3], federated learning [4], meta learning [5], reinforcement learning [6] and adversarial learning [7].

In the simpler case of (1) where there is no functional constraint $u(x,y)\leq 0$ and the lower-level function g is strongly-convex in y, the lower-level solution is unique and differentiable in x [8]. Later it has been shown that this class of bilevel optimization problems can be solved with gradient descent by differentiating through the lower-level minima, implying that this class of bilevel problems are no harder than single-level optimization problems; see e.g., [9, 10, 11, 12]. However, when g is not strongly-convex and there are lower-level constraints, there can be multiple lower-level minima and the minima is no longer differentiable. In general, the bilevel optimization problem in the form of (1) can be extremely challenging to solve. This paper focuses on the bilevel problem (1) where the lower-level function g and constraint function u are convex (but not strongly convex). To put this work in context, we provide a brief review of the related works in the next section.

1.1. Related works

The bilevel problem can be dated back to [13]. Recently, the gradient-based bilevel optimization methods have gained growing popularity in the machine learning area [14, 2, 15, 16, 4, 17, 11, 12]. **Implicit differentiation.** A branch of gradient-based method relies on the implicit gradient method [8], of which the finite-time rate was established in [9, 18] under strongly-convex lower-level problem. Later, the convergence rate was improved in [10, 11] and an extension to the constrained strongly-convex lower-level was studied in [19]. Despite the strong finite-time convergence guarantee, the implicit gradient-based methods only apply to the bilevel problems with a strongly convex lower level. **Unroll differentiation.** Another branch of methods rely on unrolling the lower-level solution to multiple gradient steps, which then allows explicit differentiation [20, 21, 22]. The unroll differentiation methods and the implicit differentiation methods oftentimes require higher-order derivatives which could be resource-hungry. **DC bilevel method.** Recently, a method based on the DC bilevel method in [23] has been proposed in [24]. However, the methods in [23, 24] only apply to fully convex bilevel problems with convex upper and lower level objectives. In a recent work [25], a conditional gradient approach has been developed for a special class of bilevel problem called the simple bilevel optimization problem. In this context, our method extends [24, 25] to general bilevel optimization with a non-convex upper level and provides the first convergence rate.

1.2. Our contribution

Compared to previous works, our contribution is two-folds.

C1) A method for non-convex bilevel problems with a convex lower-level problem. We propose an convex-concave procedure for non-convex bilevel optimization

This work was supported by National Science Foundation MoDL-SCALE Grant 2134168 and a Cisco Research Award.

problems with convex lower-level defined in (1). As will be introduced later in detail, we first reformulate (1) via a value function-based formulation and then at each iteration, we solve a locally-approximated problem of the reformulated problem. As the subproblem is convex, we call for efficient off-the-shelve solvers.

C2) Finite-time convergence matching projected gradient descent for nonconvex single-level problems. Further, we establish the connection between the proposed algorithm and the projected gradient method. Despite the lack of strong convexity in the lower-level, we show the finite-time convergence of the proposed method to the first-order stationary point with an iteration complexity of $\mathcal{O}(\epsilon^{-1})$. This matches the rate of the projected gradient method for non-convex problem [26].

2. PRELIMINARY

In this section, we define the notations and preliminary results that will be used later.

A differentiable convex function F defined on a convex set $C \subseteq \mathbb{R}^d$ satisfies the following inequality

$$F(x') \geq F(x) + \langle \nabla F(x),\, x' - x\rangle,\ \forall x, x' \in C$$

which implies that the function value of a convex function always upper bounds its local linearization at any $x \in C$.

Given $x \in C$, we generalize the concept of derivative to the so called sub-derivative $\xi \in \mathbb{R}^d$ at x, which satisfies that

$$F(x') \geq F(x) + \langle \xi,\, x' - x\rangle,\ \forall x' \in C. \tag{2}$$

We call the set of all sub-derivatives of F at x as the sub-differential of F at x, denoted as $\partial F(x)$. Formally, we define

$$\partial F(x) := \{\xi \in \mathbb{R}^d : F(x') \geq F(x) + \langle \xi,\, x' - x\rangle,\ \forall x' \in C\}.$$

Define $v(x) := \min_{y \in U(x)} g(x, y)$. The following proposition summarizes the known properties of $v(x)$.

Proposition 2.1 ([23, Theorem 3])**.** *Suppose g and u are convex functions, then $v(x)$ is a convex function and*

$$\partial v(x) \supseteq \{\xi \in \mathbb{R}^{d_x} : (\xi, 0) \in \partial g(x, y_x) + \gamma_x \partial u(x, y_x)\}$$

where y_x and γ_x are respectively any optimal solution and its KKT multiplier for the following convex optimization problem

$$\min_y\ g(x, y) \quad \text{s.t.} \quad u(x, y) \leq 0. \tag{3}$$

This proposition suggests that we can obtain $\xi \in \partial v(x)$ by finding any solution of the convex optimization problem (3).

3. CONVEX-CONCAVE PROCEDURE FOR BILEVEL OPTIMIZATION

In this section, we first give a value-function reformulation of (1), and then develop a convex-concave procedure for bilevel optimization method **(CCCP-BO)** to solve the problem.

3.1. Value function reformulation

Since g, u are convex, the value function $v(x)$ is convex by Proposition 2.1. To exploit the convexity of g and v, we consider the following reformulation of (1), given by

$$\min_{x \in C, y \in U(x)} f(x, y) \quad \text{s.t.} \quad g(x, y) - v(x) \leq 0. \tag{4}$$

The above problem is equivalent to (1) since given any $x \in C$, it holds that

$$\{y \in U(x) : g(x, y) - v(x) \leq 0\} \tag{5a}$$

$$= \{y \in U(x) : g(x, y) - v(x) = 0\} \tag{5b}$$

$$= \{y \in U(x) : y \in \arg\min_{y \in U(x)} g(x, y)\}. \tag{5c}$$

Since g and v are convex, the functional constraint in (4) has a *difference-of-convex* (DC) structure [27]. This enables the usage of DC techniques that will be introduced later. However, it is known that (4) violates the constraint qualifications such as the Mangasarian Fromovitz constraint qualification (MFCQ) [28]. To overcome this issue, with constant $\epsilon > 0$, we consider the following approximate value function reformulation [23]

$$\mathcal{BP}: \min_{x \in C, y \in U(x)} f(x, y) \quad \text{s.t.} \quad g(x, y) - v(x) \leq \epsilon. \tag{6}$$

The above relaxation accounts for the situation that in practice, the lower level problem usually cannot be exactly solved.

3.2. Algorithm development

Since $v(x)$ is a convex function, the lower-level problem in $\mathcal{BP}$ has a difference-of-convex structure. In order to deal with the concave part $-v(x)$, the constrained convex-concave procedure [27], which is a celebrated DC method, suggests 'convexifying' the DC structure by linearizing the concave $-v(x)$. By slight abuse of notations $z := (x, y)$ and $f(z) = f(x, y)$, this gives rise to the following approximation of $\mathcal{BP}$

$$\min_{z \in \mathcal{Z}} f(z) \quad \text{s.t.} \quad g(z) - v(x') - \langle \xi,\, x - x'\rangle \leq \epsilon \tag{7}$$

where $\mathcal{Z} := \{(x, y) : x \in C,\ y \in U(x)\}$ and x' is a point different from x and $\xi \in \partial v(x')$.

Compared to $\mathcal{BP}$, the formulation in (7) replaces $-v(x)$ with its local linearization at x' which is the upper bound of $-v(x)$ by convexity of $v(x)$ (2):

$$g(x, y) - v(x) \leq g(x, y) - v(x') - \langle \xi,\, x - x'\rangle.$$

It is then clear that if the constraint in (7) is satisfied, the original constraint in $\mathcal{BP}$ is satisfied. Moreover, the constraint in (7) is now convex thanks to the linearization of $-v(x)$. Thus we can view (7) as a simplification of $\mathcal{BP}$.

Next we linearize the upper level at some point $z' = (x', y')$ and obtain

$$\min_{z \in \mathcal{Z}} \langle \nabla f(z'),\, z - z'\rangle \quad \text{s.t.}\ g(z) - v(x') - \langle \xi,\, x - x'\rangle \leq \epsilon.$$

Algorithm 1 CCCP-BO for bilevel optimization

```
Initialization: Initialize $x_1 \in C$ and solve for $y_1 \in U(x_1)$ such that $g(x_1, y_1) - v(x_1) \leq \epsilon$. Pick $\rho > 0$ and $K \in \mathbb{Z}_+$.
for $k = 1$ to $K$ do
    Compute $\xi_k \in \partial v(x_k)$ following Proposition 2.1.
    Get $z_{k+1} := (x_{k+1}, y_{k+1})$ by solving (9).
end for
```

To ensure z staying close to z' so that the linearization is accurate, we add a proximal term and obtain

$$\begin{aligned} \min_{z \in \mathcal{Z}} \quad & \langle \nabla f(z'),\, z - z' \rangle + \frac{\rho}{2}\|z - z'\|^2 \\ \text{s.t.} \quad & g(z) - v(x') - \langle \xi,\, x - x' \rangle \leq \epsilon \end{aligned} \tag{8}$$

where $\rho > 0$ is the proximal constant. Given z', (8) is a strongly-convex problem with a convex constraint. We consider solving (8) iteratively, resulting in the following iteration.

$$\begin{aligned} z_{k+1} = & \arg\min_{z \in \mathcal{Z}} \langle \nabla f(z_k),\, z - z_k \rangle + \frac{\rho}{2}\|z - z_k\|^2 \\ & \text{s.t. } g(z) - v(x_k) - \langle \xi_k,\, x - x_k \rangle \leq \epsilon \end{aligned} \tag{9}$$

where $\xi_k \in \partial v(x_k)$. Here z_k is a natural choice of the proximal point z'. As will be shown later in Lemma 4.1, the constraint set in (9) is nonempty, closed and convex, thus the solution z_{k+1} always exists. The sub-problems (9) and (3) can be solved by, e.g., the interior point methods or the primal-dual methods. In the case where g, u are lipschitz-continuous, the sub-problems can be efficiently solved by [29] with a convergence rate of $\mathcal{O}(1/t)$ where t is the iteration number.

4. CONVERGENCE ANALYSIS THROUGH THE LENS OF PROXIMAL GRADIENT DESCENT

In this section, we will study the convergence of CCCP-BO. We first state the assumptions needed for the analysis.

Assumption 1. *Assume C is a closed convex subset of $\mathbb{R}^{d_x}$, f is a differentiable real function on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, and g, u are continuous convex functions on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. Assume $U(x) = \{y \in \mathbb{R}^{d_y} : u(x, y) \leq 0\}$ is nonempty for any $x \in C$.*

Define the time-variant constraint set in (9) as

$$D_k := \{z \in \mathcal{Z} : g(z) - v(x_k) - \langle \xi_k,\, x - x_k \rangle \leq \epsilon\}. \tag{10}$$

Note x is part of z in the above definition. Before we study the convergence of z_k, we first show some useful properties of D_k in the following lemma.

Lemma 4.1. *Consider Algorithm 1. Suppose Assumption 1 holds. Then given any k, $z_k \in D_k$ and D_k is closed convex.*

Proof. We prove the result by induction. Assume there exists $k \in \{1, ..., K\}$ such that $z_k \in D_k$. Since g is continuous and convex, the inverse-image $\mathcal{S}_k = \{z \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} : g(z) - v(x_k) - \langle \xi_k,\, x - x_k \rangle \leq \epsilon\}$ is closed and convex. Since $\mathcal{Z}$ is closed convex, $D_k = \mathcal{Z} \cap \mathcal{S}_k$ is closed and convex.

Furthermore, since D_k is nonempty, closed and convex, the sub-problem (9) admits solutions and z_{k+1} exists. Since $z_{k+1} \in D_k$ and $v(x_k)$ is convex (2), it follows that

$$g(z_{k+1}) - v(x_{k+1}) \leq g(z_{k+1}) - (v(x_k) + \langle \xi_k,\, x_{k+1} - x_k \rangle) \leq \epsilon$$

which along with $z_{k+1} \in \mathcal{Z}$ indicates $z_{k+1} \in D_{k+1}$. Finally, since $z_1 \in D_1$, this lemma holds from induction. □

To gain further insight into the behavior of the algorithm, we then introduce the following lemma.

Lemma 4.2. *Under Assumption 1, the update* (9) *is equivalent to the following projected gradient update*

$$z_{k+1} = \mathrm{Proj}_{D_k}\Big(z_k - \frac{1}{\rho}\nabla f(z_k)\Big) \tag{11}$$

where Proj_{D_k} denotes the projection operator to D_k in (10).

Proof. Define $z^* = \arg\min_z F(z)$ where

$$F(z) := \langle \nabla f(z_k),\, z - z_k \rangle + \frac{\rho}{2}\|z - z_k\|^2. \tag{12}$$

By the optimality condition, it follows $z^* = z_k - \frac{1}{\rho}\nabla f(z_k)$. For any z, it follows from the definition of F that

$$\begin{aligned} & F(z) - F(z^*) \\ & = \langle \nabla f(z_k),\, z - z_k \rangle + \frac{\rho}{2}\|z - z_k\|^2 \\ & \quad - \langle \nabla f(z_k),\, -\frac{1}{\rho}\nabla f(z_k) \rangle - \frac{1}{2\rho}\|\nabla f(z_k)\|^2 \\ & = \frac{\rho}{2}\|z - z_k\|^2 + \langle \nabla f(z_k),\, z - z_k \rangle + \frac{1}{2\rho}\|\nabla f(z_k)\|^2 \\ & = \frac{\rho}{2}\Big\|z - z_k + \frac{1}{\rho}\nabla f(z_k)\Big\|^2 = \frac{\rho}{2}\|z - z^*\|^2. \end{aligned} \tag{13}$$

Then by (9), we have

$$\begin{aligned} z_{k+1} = \arg\min_{z \in D_k} F(z) & = \arg\min_{z \in D_k} F(z) - F(z^*) \\ & = \arg\min_{z \in D_k} \|z - z^*\| \quad \text{By (13)} \\ & = \mathrm{Proj}_{D_k}\Big(z_k - \frac{1}{\rho}\nabla f(z_k)\Big). \end{aligned} \tag{14}$$

This proves the result. □

Lemma 4.2 indicates the dynamic of Algorithm 1 is equivalent to that of the projected gradient descent on a set D_k given by a simplification of the lower level problem.

A common convergence metric adopted in the analysis of non-convex projected gradient method is the so-called *projected gradient* [26]. Formally, it is defined as

$$g_k := \rho(z_k - z_{k+1}). \tag{15}$$

To establish the convergence of g_k, we need the following regularity assumption on f.

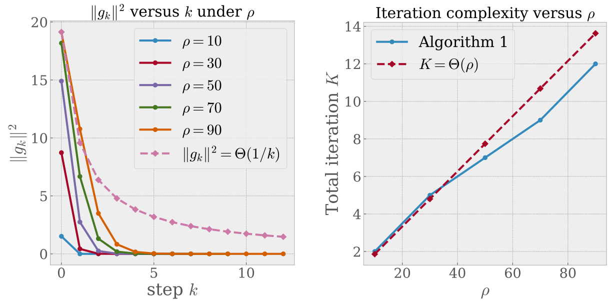

Fig. 1. Decay of $\|g_k\|^2$ under different ρ (left); iterations for Algorithm 1 to achieve $\|g_k\| \leq 10^{-3}$ versus ρ (right).

Assumption 2. *Assume $f(x,y)$ is L-lipschitz smooth w.r.t. (x,y) and is bounded from below by constant C_f.*

This assumption is standard in the analysis of gradient-descent type methods for non-convex objectives [26]. With the assumption, we characterize the convergence of Algorithm 1 in the following theorem.

Theorem 4.3. *Consider running Algorithm 1 for K steps. Under Assumption 1 and 2, if we choose $\rho \geq L$, it holds that*

$$\min_{k\in\{1,\dots,K\}} \|g_k\|^2 \leq \frac{2\rho(f(x_1,y_1)-C_f)}{K}. \tag{16}$$

Proof. By the Lipschitz smoothness of f, it holds that

$$\begin{aligned} f(z_{k+1}) - f(z_k) &\leq \langle \nabla f(z_k),\, z_{k+1}-z_k\rangle + \frac{L}{2}\|z_{k+1}-z_k\|^2 \\ &= -\frac{1}{\rho}\langle \nabla f(z_k),\, g_k\rangle + \frac{L}{2}\|z_{k+1}-z_k\|^2 \end{aligned} \tag{17}$$

where the equality follows from the definition of g_k in (15). By the first-order optimality condition of z_{k+1} for the sub-problem (9), we have

$$\langle \nabla f(z_k) + \rho(z_{k+1}-z_k),\, z_{k+1}-z\rangle \leq 0,\ \forall z \in D_k.$$

By Lemma 4.1, $z_k \in D_k$. Choosing $z = z_k$ above yields

$$-\frac{1}{\rho}\langle \nabla f(z_k),\, g_k\rangle \leq -\rho\|z_{k+1}-z_k\|^2. \tag{18}$$

Substituting the above inequality into (17) gives

$$\begin{aligned} f(z_{k+1}) - f(z_k) &\leq \Big(-\rho + \frac{L}{2}\Big)\|z_{k+1}-z_k\|^2 \\ &\overset{(15)}{=} \Big(-\frac{1}{\rho} + \frac{L}{2\rho^2}\Big)\|g_k\|^2. \end{aligned} \tag{19}$$

Choosing $\rho \geq L$, then $-\frac{1}{\rho} + \frac{L}{2\rho^2} \leq -\frac{1}{2\rho}$. This along with the above inequality implies

$$\|g_k\|^2 \leq 2\rho(f(z_k) - f(z_{k+1})). \tag{20}$$

Telescoping (20) on $k = 1, 2, \dots, K$ yields

$$\sum_{k=1}^{K} \|g_k\|^2 \leq 2\rho(f(z_1) - C_f) \tag{21}$$

where we have used $f(x_{K+1}, y_{K+1}) \geq C_f$. □

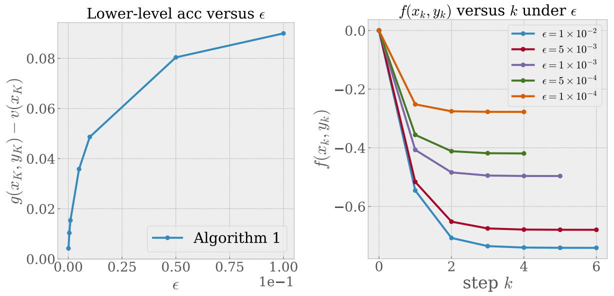

Fig. 2. Lower-level accuracy at last-iterate versus choice of ϵ in (6) (left); decay of $f(x_k, y_k)$ under different ϵ (right).

5. PRELIMINARY SIMULATIONS

In this section, we test Algorithm 1 on a synthetic problem to verify the theoretical results. Consider the following problem

$$\begin{aligned} \min_{x\in\mathbb{R}^{10},y\in\mathbb{R}^{10}}\ & \sin(c^\top x + d^\top y) + \ln(\|x+y\|^2+1) \\ \text{s.t. } & y \in \operatorname*{argmin}_{y\in\mathbb{R}^{10}} \frac{1}{2}(y+x)^\top A(y+x). \end{aligned}$$

where $d, c \in \mathbb{R}^{10}$, and $A \in \mathbb{R}^{10} \times \mathbb{R}^{10}$ is a randomly-generated non-zero matrix which is positive semi-definite but not positive-definite. It can be shown that Assumption 1 and 2 are satisfied in this problem. In all the tests, the sub-problem (9) in Algorithm 1 is solved by the primal-dual method, and (3) is solved by gradient descent.

We first test our algorithm with different ρ and report the results in Figure 1. It can be observed from Figure 1 (left) that given ρ, the decay rate of $\|g_k\|^2$ is $\mathcal{O}(1/k)$. While it can be observed from Figure 1 (right) that convergence rate is $\mathcal{O}(\rho)$. The dependence on k, ρ is consistent with Theorem 4.3. We then test the impact of ϵ in the value-function reformulation (6). Figure 2 (left) indicates with a smaller ϵ, the lower-level will be more accurate since $g(x_K, y_K) - v(x_K)$ is smaller. In the meantime, the lower-level constraint set in (6) will be smaller so that the optimal value of f will be larger, which can be observed from Figure 2 (right).

6. CONCLUSIONS

In this paper, we introduce an algorithm to solve the bilevel optimization problem with convex lower-level problem. We exploit the convexity of the lower-level by reformulating it with the value function and then utilizing the DC techniques. We then prove the resulting algorithm converges at a rate of $\mathcal{O}(1/k)$ through the lens of projected gradient method. Preliminary experiments are provided to verify our results. Future research that will be pursued include the following two dimensions: i) extending the proposed CCCP-BO algorithm to the stochastic and variance reduced variants; and ii) quantifying the overall iteration and sample complexity by considering the complexity of solving the subproblem (9).

7. REFERENCES

[1] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proc. of International Conference on Machine Learning*, 2015.

[2] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in *Proc. of International Conference on Machine Learning*, 2018.

[3] Z. Borsos, M. Tagliasacchi, and A. Krause, "Semi-supervised batch active learning via bilevel optimization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[4] S. Lu, X. Cui, M. Squillante, B. Kingsbury, and L. Horesh, "Decentralized bilevel optimization for personalized client learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[5] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. of International Conference on Machine Learning*, 2017.

[6] C. Cheng, T. Xie, N. Jiang, and A. Agarwal, "Adversarially trained actor critic for offline reinforcement learning," in *Proc. of International Conference on Machine Learning*, 2022.

[7] H. Jiang, Z. Chen, Y. Shi, B. Dai, and T. Zhao, "Learning to defend by learning to attack," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2021.

[8] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *Proc. of International Conference on Machine Learning*, 2016.

[9] S. Ghadimi and M. Wang, "Approximation methods for bilevel programming," *arXiv preprint arXiv:1802.02246*, 2018.

[10] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, "A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic," *arXiv preprint arXiv:2007.05170*, 2020.

[11] T. Chen, Y. Sun, and W. Yin, "Tighter analysis of alternating stochastic gradient method for stochastic nested problems," in *Proc. of Advances in Neural Information Processing Systems*, 2021.

[12] H. Shen and T. Chen, "A single-timescale analysis for stochastic approximation with multiple coupled sequences," in *Proc. of Advances in Neural Information Processing Systems*, 2022.

[13] H. Stackelberg, *The Theory of Market Economy*, Oxford University Press, 1952.

[14] S. Sabach and S. Shtern, "A first order method for solving convex bilevel optimization problems," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 640–660, 2017.

[15] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo, "On the iteration complexity of hypergradient computation," in *Proc. of International Conference on Machine Learning*, 2020.

[16] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang, "A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton," in *Proc. of International Conference on Machine Learning*, 2020.

[17] A. Ghosh, M. Mccann, and S. Ravishankar, "Bilevel learning of l1 regularizers with closed-form gradients (blorc)," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[18] K. Ji, J. Yang, and Y. Liang, "Provably faster algorithms for bilevel optimization and applications to meta-learning," in *Proc. of International Conference on Machine Learning*, 2021.

[19] I. Tsaknakis, P. Khanduri, and M. Hong, "An implicit gradient-type method for linearly constrained bilevel problems," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[20] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, "Forward and reverse gradient-based hyperparameter optimization," in *Proc. of International Conference on Machine Learning*, 2017.

[21] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[22] R. Liu, Y. Liu, S. Zeng, and J. Zhang, "Towards gradient based bilevel optimization with non-convex followers and beyond," in *Proc. of Advances in Neural Information Processing Systems*, 2021.

[23] J. Ye, X. Yuan, S. Zeng, and J. Zhang, "Difference of convex algorithms for bilevel programs with applications in hyperparameter selection," *arXiv preprint arXiv:2202.03397*, 2022.

[24] L. Gao, J. Ye, H. Yin, S. Zeng, and J. Zhang, "Value function based difference-of-convex algorithm for bilevel hyperparameter selection problems," in *Proc. of International Conference on Machine Learning*, 2022.

[25] R. Jiang, N. Abolfazli, A. Mokhtari, and E. Hamedani, "A conditional gradient-based method for simple bilevel optimization with convex lower-level problem," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2023, pp. 10305–10323.

[26] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, 2016.

[27] A. Smola, S. Vishwanathan, and T. Hofmann, "On first-order meta-learning algorithms," in *Proc. of International Workshop on Artificial Intelligence and Statistics*, 2005.

[28] Y. Jane and D. Zhu, "Optimality conditions for bilevel programming problems," *Optimization*, vol. 33, no. 1, pp. 9–27, 1995.

[29] H. Yu and M. Neely, "A simple parallel algorithm with an o(1/t) convergence rate for general convex programs," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 759–783, 2017.