An Improved Unconstrained Approach for Bilevel Optimization

Xiaoyin Hu, Nachuan Xiao, Xin Liu, and Kim-Chuan Toh

December 26, 2022

Abstract

In this paper, we focus on the nonconvex-strongly-convex bilevel optimization problem (BLO). In this BLO, the objective function of the upper-level problem is nonconvex and possibly nonsmooth, and the lower-level problem is smooth and strongly convex with respect to the underlying variable y. We show that the feasible region of BLO is a Riemannian manifold. Then we transform BLO to its corresponding unconstrained constraint dissolving problem (CDB), whose objective function is explicitly formulated from the objective functions in BLO. We prove that BLO is equivalent to the unconstrained optimization problem CDB. Therefore, various efficient unconstrained approaches, together with their theoretical results, can be directly applied to BLO through CDB. We propose a unified framework for developing subgradient-based methods for CDB. Remarkably, we show that several existing efficient algorithms can fit the unified framework and be interpreted as descent algorithms for CDB. These examples further demonstrate the great potential of our proposed approach.

1   Introduction

In this paper, we focus on the following nonconvex-strongly-convex bilevel optimization problem

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^p} \quad f(x,y) \qquad \text{(upper-level problem)}$$
$$\text{s. t.} \quad y = \arg\min_{y\in\mathbb{R}^p} g(x,y), \quad \text{(lower-level problem)} \qquad \text{(BLO)}$$

where the functions f and g satisfy the following blanket assumptions,

Assumption 1.1. Blanket assumptions

1. *f is possibly nonsmooth and M_f-Lipschitz continuous over $\mathbb{R}^n \times \mathbb{R}^p$.*

2. *The function $g(x,y)$ is twice differentiable and μ-strongly convex with respect to y for any fixed x, i.e. $\nabla^2_{yy} g(x,y) \succeq \mu I_p$ holds for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$.*

3. *The gradient $\nabla g(x,y)$ is L_g-Lipschitz continuous.*

4. *The Hessian matrices $\nabla^2_{yy} g(x,y)$ and $\nabla^2_{xy} g(x,y)$ are Q_g-Lipschitz continuous.*

5. *$\nabla^2_{yy} g(x,y)$ is continuously differentiable over $\mathbb{R}^n \times \mathbb{R}^p$.*

Problem BLO has attracted a lot of attention in the current era of big data and artificial intelligence due to its close connection with various real-world applications, including reinforcement learning [36], hyperparameter optimization [49, 32, 21, 41], and meta learning [20, 48]. Interested reader can refer to several survey papers [14, 40] and the references therein for details.

1

The blanket assumption 1.1 is commonly assumed in a great number of existing works. In particular, Assumptions 1.1 is satisfied in the applications discussed in [18, 43, 47, 22, 24, 38, 25, 33]. It should be noted that although we assume the Lipschitz smoothness of $\nabla^2_{yy} g(x, y)$, it is only necessary in the theoretical analysis, and we do not involve the computation of any third-order derivatives in our proposed methods throughout this paper.

1.1 Existing works

Recently, nonconvex-strongly-convex bilevel optimization problems with Lipschitz smooth objective functions have been extensively studied. For any given $x \in \mathbb{R}^n$, we denote $y^\star(x)$ as the unique minimizer of the lower-level problem, i.e., $y^\star(x) := \arg\min_{y \in \mathbb{R}^p} g(x, y)$. Since the lower-level problem of BLO is assumed to be strongly convex with respect to y, $y^\star(x)$ is differentiable with respect to x by the implicit function theorem. Therefore, BLO is equivalent to the following unconstrained optimization problem that only involves the x-variable,

$$\min_{x \in \mathbb{R}^n} \quad \Phi(x) := f(x, y^\star(x)). \tag{1.1}$$

Various existing efficient approaches are developed based on solving the unconstrained optimization problem (1.1). However, $\Phi(x)$ is implicitly formulated since the solution to the lower-level problem usually does not have a closed-form expression [27]. Therefore, it is usually intractable to compute the exact function value and derivatives of $\Phi(x)$.

Some of the existing approaches [18, 47, 24, 25, 33], referred to as *double-loop* approaches, are developed by introducing inner loops in each iteration to obtain an approximated estimation for $y^\star(x)$. Then these approaches inexactly evaluate $\nabla \Phi(x)$ through the approximated solution for the lower-level problem and chain rule. Although their theoretical properties are simple to analyze, these algorithms may suffer from poor performance as one has to take multiple steps in the inner loop to solve the lower-level problem to a desired accuracy [35]. It is usually challenging to balance the computational cost of the inner loops and the overall performance of these algorithms.

Furthermore, several *single-loop* approaches [11, 27, 35] are proposed to minimize $\Phi(x)$ by updating the x- and y-variables simultaneously, hence avoiding inner loops for an approximated solution of the lower-level problem. In each iteration, these single-loop approaches update the x-variable by taking an approximated gradient descent step to $\Phi(x)$, while the y-variable is updated to track $y^\star(x)$ by taking a descent step for the lower-level problem [27, 35] or other specifically designed schemes [11]. Although prior arts [27, 35, 11, 54] use $\Phi(x)$ as the merit function in their theoretical analysis, these existing single-loop approaches cannot be simply interpreted as approximated gradient descent methods to minimize $\Phi(x)$. Therefore, establishing the related theoretical analysis for these approaches becomes more complicated and challenging in these existing works.

Though solving BLO with smooth objective functions has been intensively studied, how to solve BLO with a nonsmooth upper-level objective function is relatively less explored. Due to the implicit formulation of $\Phi(x)$, existing single-loop and double-loop approaches have to approximately solve the lower-level problems and evaluate $\nabla \Phi(x)$ inexactly. Therefore, without the assumption on the Lipschitz smoothness of f, the above-mentioned approaches have no theoretical guarantee. On the other hand, computing the exact subdifferential of Φ requires the exact solution to the lower-level problem, which is usually expensive to achieve in practice. As a result, it is challenging to develop algorithms for BLO based on $\Phi(x)$.

Apart from those existing approaches developed for minimizing $\Phi(x)$ over $\mathbb{R}^n$, several other existing approaches [26, 44] reshape BLO as the following single-level optimization problem with equality constraints [55]

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^p} \quad f(x, y)$$
$$\text{s. t.} \quad \nabla_y g(x, y) = 0. \tag{1.2}$$

Then (1.2) can be solved by employing existing approaches for constrained optimization, including polynomial optimization approach (when the functions involved are polynomials) [46], sequential quadratic programming methods [15, 53], etc. However, these approaches treat BLO as a constrained

optimization problem with p equality constraints, hence they are usually not as efficient as those aforementioned single-loop and double-loop approaches in practice [27].

1.2 Motivation

Our motivation in this paper comes from the *constraint dissolving* approaches [52] for Riemannian optimization. Let $\mathcal{M}$ be the feasible region of (1.2), i.e.

$$\mathcal{M} := \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^p : \nabla_y g(x,y) = 0\}. \tag{1.3}$$

As $g(x,y)$ is twice-order differentiable and strongly convex with respect to y, the constraints $\nabla_y g(x,y) = 0$ satisfy linear independent constraint qualification (LICQ) for any $(x,y) \in \mathcal{M}$. Therefore, the implicit function theorem ensures that $\mathcal{M}$ is a Riemannian manifold embedded in $\mathbb{R}^n \times \mathbb{R}^p$ [34]. Although various Riemannian optimization approaches are developed in recent years [1, 6, 28], the required geometrical materials of the manifold $\mathcal{M}$ are usually expensive to compute. For example, computing the retraction of $\mathcal{M}$ can be regraded as computing a projection from the tangent space to $\mathcal{M}$, which is as expensive as solving the lower-level subproblem exactly. To our best knowledge, there is no efficient Riemannian optimization approach developed for solving BLO.

When f is assumed to be Lipschitz smooth over $\mathbb{R}^n \times \mathbb{R}^p$, [52] proposes a general framework for developing the constraint dissolving function for BLO, which takes the form as

$$f(\tilde{\mathcal{A}}(x,y)) + \frac{\beta}{2} \left\| \nabla_y g(x,y) \right\|^2. \tag{1.4}$$

Here, the mapping $\tilde{\mathcal{A}} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n \times \mathbb{R}^p$ satisfying the following assumptions is called the constraint dissolving mapping.

Assumption 1.2. • $\tilde{\mathcal{A}}$ *is locally Lipschitz continuous over* $\mathbb{R}^n \times \mathbb{R}^p$.

- $\tilde{\mathcal{A}}(x,y) = (x,y)$ *for any* $(x,y) \in \mathcal{M}$.

- *The Jacobian of* $(\nabla_y g) \circ \tilde{\mathcal{A}}$ *equals to* 0 *for any* $(x,y) \in \mathcal{M}$.

As illustrated in [52, Lemma 3.3], any constraint dissolving mapping $\tilde{\mathcal{A}}$ will drive any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ closer to the feasible region $\mathcal{M}$ with the feasibility violation locally quadratically converges to zero. This property plays a crucial role in establishing the equivalence between the original bilevel optimization problem (BLO) and minimizing the constraint dissolving function (1.4). The detailed proof can be found at [52, Section 3].

Moreover, [52] provides some practical schemes for constructing the constraint dissolving mapping, see [52, Section 4.1] for instances. However, [52] focuses on smooth optimization over the Riemannian manifold. Existing constraint dissolving approaches for nonsmooth optimization are only developed for special manifolds [29]. Furthermore, the equivalence established in [52] only holds in a neighborhood of the feasible region $\mathcal{M}$. For general nonsmooth cases, how to choose an appropriate constraint dissolving operator $\tilde{\mathcal{A}}$ for (1.4) and establish the equivalence between (BLO) and (1.4) over $\mathbb{R}^n \times \mathbb{R}^p$ rather than a neighborhood of $\mathcal{M}$ remain to be studied.

1.3 Contributions

In this paper, we consider the mapping $(x,y) \mapsto (x, \mathcal{A}(x,y))$, where $\mathcal{A}$ is defined by

$$\mathcal{A}(x,y) := y - \left(\nabla_{yy}^2 g(x,y) \right)^{-1} \nabla_y g(x,y) \tag{1.5}$$

as a special choice of $\tilde{\mathcal{A}}$. Substituting this $\tilde{\mathcal{A}}$ into (1.4), we obtain a constraint dissolving function for bilevel optimization (CDB)

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^p} \quad h(x,y) := f\left(x, \mathcal{A}(x,y)\right) + \frac{\beta}{2} \left\| \nabla_y g(x,y) \right\|^2. \tag{CDB}$$

We prove that such an $\tilde{\mathcal{A}}$ satisfies Assumption 1.2 and hence is a constraint dissolving mapping [52]. Clearly, h can be explicitly formulated from f and the derivatives of g. Under mild conditions, we prove that BLO and CDB have the same stationary points over $\mathbb{R}^n \times \mathbb{R}^p$ from the perspective of both the Clarke subdifferential and the conservative field [4]. As a result, the bilevel optimization problem BLO is equivalent to the unconstrained optimization problem CDB, and various optimization approaches for unconstrained nonsmooth optimization can be directly implemented to solve BLO through CDB.

We propose a unified framework for developing subgradient-based methods to solve CDB and prove their global convergence. We provide several illustrative examples on how to develop single-loop subgradient-based methods and how to establish their convergence properties from the proposed framework. Moreover, we can interpret the updating schemes in the deterministic versions of several existing single-loop algorithms [27, 11, 35] as approximated gradient-descent steps for CDB. Therefore, we provide a clear explanation for the updating schemes in these existing algorithms, extend these algorithms to nonsmooth cases and prove their convergence properties based on our proposed framework. These examples further highlight the significant advantages and great potentials of CDB.

2 Preliminaries

2.1 Basic notations

Let $\langle \cdot, \cdot \rangle$ be the standard inner product and $\|\cdot\|$ be the ℓ_2-norm of a vector or an operator. $\mathbb{B}_\delta(x, y) := \{(\tilde{x}, \tilde{y}) \in \mathbb{R}^n \times \mathbb{R}^p : \|\tilde{x} - x\|^2 + \|\tilde{y} - y\|^2 \leq \delta^2\}$ refers to the ball centered at (x, y) with radius δ. Moreover, for a given set $\mathcal{X}$, $\text{dist}(x, \mathcal{X})$ denotes the distance between x and a set $\mathcal{X}$, i.e. $\text{dist}(x, \mathcal{X}) := \arg\min_{y \in \mathcal{X}} \|x - y\|$, $\text{cl}\,\mathcal{X}$ denotes the closure of $\mathcal{X}$ and $\text{conv}\,\mathcal{X}$ denotes the convex hull of $\mathcal{X}$. For any differentiable function $g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, let $\nabla_x g$ and $\nabla_y g$ be the partial derivatives of g with respect to x and y, respectively. Moreover, $\nabla^2_{xy} g(x, y)$ and $\nabla^2_{yy} g(x, y)$ denotes the partial Jacobian of $\nabla_y g(x, y)$ with respect to variable x and y, respectively. More precisely,

$$\nabla^2_{xy} g(x,y) := \begin{bmatrix} \frac{\partial^2 g(x,y)}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 g(x,y)}{\partial x_1 \partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g(x,y)}{\partial x_n \partial y_1} & \cdots & \frac{\partial^2 g(x,y)}{\partial x_n \partial y_p} \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad \nabla^2_{yy} g(x,y) := \begin{bmatrix} \frac{\partial^2 g(x,y)}{\partial y_1 \partial y_1} & \cdots & \frac{\partial^2 g(x,y)}{\partial y_1 \partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g(x,y)}{\partial y_p \partial y_1} & \cdots & \frac{\partial^2 g(x,y)}{\partial y_p \partial y_p} \end{bmatrix} \in \mathbb{R}^{p \times p},$$

and $\nabla^2_{yx} g(x, y)$ is the transpose of $\nabla^2_{xy} g(x, y)$. Furthermore, $\nabla^3_{xyy} g(x, y)$ is the partial derivative of $\nabla^2_{xy} g(x, y)$ with respect to variable y, which is expressed as the linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^{n \times p}$ by $\nabla^3_{xyy} g(x, y)[d] := \lim_{t \to 0} \frac{1}{t} \left(\nabla^2_{xy} g(x, y + td) - \nabla^2_{xy} g(x, y) \right)$. Similarly, $\nabla^3_{yyy} g(x, y)$ is the partial derivative of $\nabla^2_{yy} g(x, y)$ with respect to variable y, which is expressed as the linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^{p \times p}$ by $\nabla^3_{yyy} g(x, y)[d] := \lim_{t \to 0} \frac{1}{t} \left(\nabla^2_{yy} g(x, y + td) - \nabla^2_{yy} g(x, y) \right)$. Under Assumption 1.1, it is easy to verify that the inequalities $\left\| \nabla^3_{yyx} g(x, y)[d_x] \right\| \leq Q_g \|d_x\|$ and $\left\| \nabla^3_{yyy} g(x, y)[d_y] \right\| \leq Q_g \|d_y\|$ hold for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ and $(d_x, d_y) \in \mathbb{R}^n \times \mathbb{R}^p$.

2.2 Clarke subdifferential

Definition 2.1. *For any given locally Lipschitz continuous function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, the generalized directional derivative of f at (x, y) in the direction $(d_x, d_y) \in \mathbb{R}^n \times \mathbb{R}^p$, denoted by $f^\circ(x, y; d_x, d_y)$, is defined as*

$$f^\circ(x, y; d_x, d_y) := \limsup_{(\tilde{x}, \tilde{y}) \to (x, y),\, t \downarrow 0} \frac{f(\tilde{x} + t d_x, \tilde{y} + t d_y) - f(\tilde{x}, \tilde{y})}{t}.$$

Then the generalized gradient or the Clarke subdifferential of f at (x, y), denoted by $\partial f(x, y)$, is defined as

$$\partial f(x, y) := \big\{ (w_x, w_y) \in \mathbb{R}^n \times \mathbb{R}^p :$$
$$\langle w_x, d_x \rangle + \langle w_y, d_y \rangle \leq f^\circ(x, y; d_x, d_y), \text{ for all } (d_x, d_y) \in \mathbb{R}^n \times \mathbb{R}^p \big\} .$$

Remark 2.2. *For any locally Lipschitz continuous function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, its Clarke subdifferential is compact and convex for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$. Moreover, the mapping $(x, y) \mapsto \partial f(x, y)$ is outer-semicontinuous over $\mathbb{R}^n \times \mathbb{R}^p$ [13].*

Definition 2.3. *We say that f is (Clarke) regular at $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ if for every direction $(d_x, d_y) \in \mathbb{R}^n \times \mathbb{R}^p$, the one-sided directional derivative*

$$f^\star(x, y; d_x, d_y) := \lim_{t \downarrow 0} \frac{f(x + td_x, y + td_y) - f(x, y)}{t}$$

exists and $f^\star(x, y; d_x, d_y) = f^\circ(x, y; d_x, d_y)$.

Definition 2.4. *For any given locally Lipschitz continuous function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, the δ-Goldstein subdifferential of f at (x, y) is defined as*

$$\partial_\delta f(x, y) = \mathrm{cl\,conv} \left(\cup_{(\tilde{x}, \tilde{y}) \in \mathbb{B}_\delta(x, y)} \partial f(\tilde{x}, \tilde{y}) \right).$$

The following proposition present some basic properties of δ-Goldstein subdifferential, which are mainly from the upper-semicontinuity of ∂f, as illustrated in [7, Theorem 3.1] and [57, Lemma 7].

Proposition 2.5. *For any given locally Lipschitz continuous function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, it holds that*

$$\lim_{\delta \to 0} \partial_\delta f(x, y) = \partial f(x, y).$$

2.3 Conservative field

In this subsection, we introduce the concept of conservative field, which generalizes Clarke sub-differential for a broad class of nonsmooth functions. For simplicity, we provide a self-contained description and highlight some essential ingredients for our theoretical analysis. Interested readers can refer to several recent papers [4, 10] for more details.

Definition 2.6. *A set-valued mapping $\mathcal{D} : \mathbb{R}^m \rightrightarrows \mathbb{R}^s$ is a mapping from $\mathbb{R}^m$ to a collection of subsets of $\mathbb{R}^s$. $\mathcal{D}$ is said to have closed graph if the graph of $\mathcal{D}$, defined by*

$$\mathrm{graph}(\mathcal{D}) := \{ (w, z) \in \mathbb{R}^m \times \mathbb{R}^s : w \in \mathbb{R}^m, z \in \mathcal{D}(w) \},$$

is a closed set.

Definition 2.7. *An absolutely continuous curve is a continuous mapping $\gamma : \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}^p$ whose derivative γ' exists almost everywhere in $\mathbb{R}$ and $\gamma(t) - \gamma(0)$ equals to the Lebesgue integral of γ' between 0 and t for all $t \in \mathbb{R}_+$, i.e.,*

$$\gamma(t) = \gamma(0) + \int_0^t \gamma'(\tau) d\tau, \qquad \text{for all } t \in \mathbb{R}_+.$$

With the concept of absolutely continuous curve, we can present the definition of a conservative set-valued field.

Definition 2.8. *Let $\mathcal{D}$ be a set-valued mapping from $\mathbb{R}^n \times \mathbb{R}^p$ to subsets of $\mathbb{R}^n \times \mathbb{R}^p$. Then we call $\mathcal{D}$ as a conservative field whenever it has closed graph, nonempty compact values, and for any absolutely continuous curve $\gamma : [0, 1] \to \mathbb{R}^n \times \mathbb{R}^p$ satisfying $\gamma(0) = \gamma(1)$, we have*

$$\int_0^1 \max_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle dt = 0, \tag{2.1}$$

where the integral is understood in the Lebesgue sense.

Remark 2.9. *When the set-valued mapping $\mathcal{D}$ has compact values and closed graph, then the mapping $t \mapsto \max_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle$ is Lebesgue measurable [4, Lemma 1]. Therefore, the path integral in (2.1) is well-defined. Furthermore, the equation (2.1) can be replaced by $\int_0^1 \min_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \, dt = 0$.*

Definition 2.10. *Let $\mathcal{D}$ be a conservative field in $\mathbb{R}^n \times \mathbb{R}^p$. Then with any given $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^p$, we can define a function through*

$$f(x,y) = f(x_0, y_0) + \int_0^1 \max_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \, dt = f(x_0, y_0) + \int_0^1 \min_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \, dt \tag{2.2}$$

for any absolutely continuous curve γ that satisfies $\gamma(0) = (x_0, y_0)$ and $\gamma(1) = (x, y)$. Then f is called a potential function for $\mathcal{D}$, and we also say $\mathcal{D}$ admits f as its potential function, or that $\mathcal{D}$ is a conservative field for f.

It is worth mentioning that any conservative field defines a unique potential function up to a constant, since the value of the integral does not depend on the selection of the path in (2.2). Moreover, for any f that is a potential function for some conservative field $\mathcal{D}$, ∂f is a conservative field that admits f as its potential function, and $\partial f(x,y) \subseteq \text{conv}(\mathcal{D}(x,y))$ holds for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ [4, Corollary 1].

As a result, for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ that is a first-order stationary point of f, then it holds that $0 \in \partial f(x,y) \subseteq \text{conv}(\mathcal{D}(x,y))$. Thus the stationarity of the potential function f can be characterized by its corresponding conservative field $\mathcal{D}$ as illustrated in the following definition.

Definition 2.11. *Given a fixed conservative field $\mathcal{D} : \mathbb{R}^n \times \mathbb{R}^p \rightrightarrows \mathbb{R}^n \times \mathbb{R}^p$ that admits f as a potential function, then we say (x,y) is a $\mathcal{D}$-stationary point for f if $0 \in \mathcal{D}(x,y)$.*

Similar to the definition on conservative field, we present the definition on conservative mapping as follows.

Definition 2.12. *Let $F : \mathbb{R}^d \to \mathbb{R}^m$ be a locally Lipschitz function. $J_F : \mathbb{R}^d \rightrightarrows \mathbb{R}^{m \times d}$ is called a conservative mapping for F, if for any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^d$, the function $t \mapsto F(\gamma(t))$ satisfies*

$$\frac{d(F \circ \gamma)}{dt}(t) = V\gamma'(t), \quad \text{for all } V \in J_F(\gamma(t)) \text{ and a.e. } t \in [0,1].$$

When we choose $m = 1$ in Definition 2.12, the definition on conservative mapping is equivalent to the definition on conservative field in Definition 2.8, as illustrated in [4, Remark 7]. The following propositions illustrate that the chain rule and sum rule hold for conservative fields.

Proposition 2.13 (Lemma 7 in [4]). *Let $F_1 : \mathbb{R}^d \to \mathbb{R}^m$ and $F_2 : \mathbb{R}^m \to \mathbb{R}^s$ be locally Lipschitz continuous mappings, $J_{F_1} : \mathbb{R}^d \to \mathbb{R}^{m \times d}$ and $J_{F_2} : \mathbb{R}^d \to \mathbb{R}^{s \times m}$ be their associated conservative mappings. Then the mapping $x \mapsto J_{F_2}(F_1(x)) J_{F_1}(x)$ is a conservative mapping for $F_2 \circ F_1$.*

Proposition 2.14 (Corollary 4 in [4]). *Let $f_1, ..., f_n$ be locally Lipschitz continuous functions for the conservative fields $\mathcal{D}_{f_1}, ..., \mathcal{D}_{f_n}$, respectively. Then $f = \sum_{i=1}^n f_i$ is a potential function for $\mathcal{D}_f = \sum_{i=1}^n \mathcal{D}_{f_i}$.*

2.4 Additional assumption and stationarity

In this subsection, we present the basic assumptions on BLO as well as the definition of its stationarity. In the rest of this paper, we assume the objective function f to be a potential function for a certain conservative field.

Assumption 2.15. *f is a potential function of a conservative set-valued field $\mathcal{D}_f : \mathbb{R}^n \times \mathbb{R}^p \rightrightarrows \mathbb{R}^n \times \mathbb{R}^p$, which has convex values, and satisfies*

$$\sup_{x \in \mathbb{R}^n, \, y \in \mathbb{R}^p, \, \xi \in \mathcal{D}_f(x,y)} \|\xi\| \leq M_f,$$

for some constant $M_f > 0$.

6

The following remark illustrates that Assumption 2.15 is general enough to cover most applications of BLO.

Remark 2.16. *It is worth mentioning that any Clarke regular function is a potential function for some conservative fields [16]. However, the Clarke regularity is too restrictive in practice, which excludes some important applications of BLO, in particular, training the neural network built from nonsmooth activation functions.*

To this end, [16] reviews the concept of Whitney stratifiable functions, and prove that any locally Lipschitz function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ that is Whitney C^1-stratifiable is a potential function for ∂f in [16, Theorem 5.8]. Whitney stratifiable functions are general enough to cover several important classes of functions, including semi-algebraic functions, and semi-analytic functions.

Additionally, several recent works [4, 10, 5] focus on the optimization of definable functions (i.e., functions that are definable in an o-minimal structure [50, 16]), which are all Whitney C^s-stratifiable functions for any $s \geq 1$ [50]. The finite summation and composition of definable functions are also definable, hence various nonsmooth functions can be easily recognized as definable functions. As shown in [51, 16, 4], the finite composition among semi-algebraic functions, $\exp$ and $\log$ is definable. Therefore, most common activation functions and loss functions, including sigmoid, hyperbolic tangent, softplus, ReLU [2], Leaky-ReLU [42], piecewise polynomial activations, ℓ_1-loss, MSE loss, hinge loss, logistic loss and cross-entropy loss are all definable in some o-minimal structures. Furthermore, for any nonsmooth deep neural network built from definable loss functions and activation functions, its objective function is also definable, hence is a potential function for a certain conservative field (e.g., its Clarke subdifferential).

Remark 2.17. *Assumption 2.15 implies that f is M_f-Lipschitz continuous over $\mathbb{R}^n \times \mathbb{R}^p$, as illustrated in [5, Remark 3(d)].*

Based on Assumption 1.1 and Assumption 2.15, we make the following definitions on the stationarity of BLO and CDB.

Definition 2.18 ([13])**.** *For any given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, we say (x, y) is a first-order stationary point of BLO if there exists $(d_x, d_y) \in \partial f(x, y)$ such that*

$$\begin{cases} 0 = d_x - \nabla^2_{xy} g(x,y) \nabla^2_{yy} g(x,y)^{-1} d_y, \\ 0 = \nabla_y g(x,y). \end{cases}$$

Similarly, the stationarity of CDB can be stated in the following definition.

Definition 2.19. *For any given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, we say that (x, y) is a first-order stationary point of CDB if $0 \in \partial h(x, y)$.*

On the other hand, we can characterize the stationarity of BLO from the perspective of conservative field.

Definition 2.20. *Suppose f is a potential function admitted by a convex-valued conservative field $\mathcal{D}_f$. Then for any given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, we say that (x, y) is a $\mathcal{D}_f$-stationary point of BLO if there exists $(d_x, d_y) \in \mathcal{D}_f(x, y)$ such that*

$$\begin{cases} 0 = d_x - \nabla^2_{xy} g(x,y) \nabla^2_{yy} g(x,y)^{-1} d_y, \\ 0 = \nabla_y g(x,y). \end{cases}$$

It directly follows from Definition 2.18 and Definition 2.20 that all the ∂f-stationary points of BLO are its first-order stationary points.

Definition 2.21. *Suppose h in CDB is a potential function admitted by a convex-valued conservative field $\mathcal{D}_h$, we say that $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ is a $\mathcal{D}_h$-stationary point of CDB if $0 \in \mathcal{D}_h(x, y)$.*

Remark 2.22. *$\mathcal{D}_f$ is a generalization of the Clarke subdifferential of f, whose expression depends on how to achieve the "subdifferential" of $f(x, y)$ [4, 5]. As illustrated in [5], the conservative field $\mathcal{D}_f$ is not unique and may differs from ∂f in a dense set, hence may lead to infinitely many spurious stationary points for BLO.*

However, we should keep in mind that the most important case for us is $\mathcal{D}_f = \partial f$. In the following remark, we discuss how to approximately evaluate ∂f in practice.

Although in some cases, directly computing one element from $\partial f(x,y)$ may be intractable [5], there are already several randomized approaches [7, 8, 19, 56, 45, 9, 39] developed for approximately evaluating one element for $\partial f(x,y)$ in practice.

Some existing approaches approximate $\partial f(x,y)$ by random sampling of gradients [7, 8, 9]. In these approaches, with a given radius δ, we randomly sample $\{(x_1, y_1), ..., (x_s, y_s)\} \subset \mathbb{B}_\delta(x,y)$. Since f is differentiable at $\{(x_1, y_1), ..., (x_s, y_s)\}$ almost surely, let $C_\delta = \text{conv}\{\nabla f(x_i, y_i) : 1 \leq i \leq s\} \subset \partial_\delta f(x,y)$, and it holds from Proposition 2.5 and [7, Theorem 3.1] that $\lim_{\delta \to 0} \text{dist}(\partial f(x,y), C_\delta) = 0$. By choosing one element from C_δ, we get an approximated evaluation for an element in $\partial f(x,y)$.

Furthermore, some recent works approximate $\partial f(x,y)$ by the randomized smoothing approaches [19, 56, 45, 39]. In these approaches, we first uniformly sample $(\zeta_x, \zeta_y) \in \mathbb{B}_\delta(0)$, and approximate $\partial f(x,y)$ by

$$\tilde{\partial}_\delta f(x,y; \zeta_x, \zeta_y) := \frac{n+p}{2\delta}(f(x + \delta\zeta_x, y + \delta\zeta_y) - f(x - \delta\zeta_x, y - \delta\zeta_y)) \cdot \begin{bmatrix} \zeta_x \\ \zeta_y \end{bmatrix}.$$

From [39, Theorem 3.1, Lemma D.1], it holds that

$$\mathbb{E}_{\zeta_x, \zeta_y}[\tilde{\partial}_\delta f(x,y; \zeta_x, \zeta_y)] \in \partial_\delta(x,y), \text{ and } \mathbb{E}_{\zeta_x, \zeta_y}[\|\tilde{\partial}_\delta f(x,y; \zeta_x, \zeta_y)\|^2] \leq 16\sqrt{2\pi}(n+p)M_f^2.$$

Then Proposition 2.5 illustrates that $\lim_{\delta \to 0} \text{dist}\left(\partial f(x,y), \mathbb{E}_{\zeta_x, \zeta_y}[\tilde{\partial}_\delta f(x,y; \zeta_x, \zeta_y)]\right) = 0$. Therefore, $\tilde{\partial}_\delta f(x,y; \zeta_x, \zeta_y)$ approximates $\partial f(x,y)$ with noises.

3 Theoretical properties

3.1 Equivalence: Clarke subdifferential

In this subsection, we study the equivalence between BLO and CDB based on the Clarke subdifferential. We first define

$$J_{\mathcal{A},x}(x,y) := -\nabla_{xy}^2 g(x,y)\nabla_{yy}^2 g(x,y)^{-1} + \nabla_{xyy}^3 g(x,y)[\nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)]\nabla_{yy}^2 g(x,y)^{-1}, \quad (3.1)$$

$$J_{\mathcal{A},y}(x,y) := \nabla_{yyy}^3 g(x,y)[\nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)]\nabla_{yy}^2 g(x,y)^{-1}. \quad (3.2)$$

Then the following proposition characterizes the expression of $\partial h(x,y)$ for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$.

Proposition 3.1. *For any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, it holds that*

$$\partial h(x,y) \subseteq \left\{ \begin{bmatrix} d_x + J_{\mathcal{A},x}(x,y)d_y + \beta\nabla_{xy}^2 g(x,y)\nabla_y g(x,y) \\ J_{\mathcal{A},y}(x,y)d_y + \beta\nabla_{yy}^2 g(x,y)\nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \partial f(x, \mathcal{A}(x,y)) \right\}.$$

Here the equality holds when f is Clarke regular.

Proposition 3.1 can be verified through direct calculation, hence we omit its proof for simplicity.

Proposition 3.2. *For any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, suppose $(x,y) \in \mathcal{M}$ is a first-order stationary point of CDB, then (x,y) is a first-order stationary point of BLO.*

Furthermore, when f is Clarke regular, then for any given $(x,y) \in \mathcal{M}$, (x,y) is a first-order stationary point of BLO if and only if it is a first-order stationary point of CDB.

Proof. Since $(x,y) \in \mathcal{M}$ is a first-order stationary point of CDB, it follows from the optimality conditions of CDB that $0 \in \partial h(x,y)$. Together with the fact that $0 = \nabla_y g(x,y)$ and Proposition 3.1, there exists $(d_x, d_y) \in \partial f(x,y)$ such that $0 = d_x - \nabla_{xy}^2 g(x,y)\nabla_{yy}^2 g(x,y)^{-1}d_y$, which coincides with the optimality conditions of BLO. Therefore, we obtain that (x,y) is a first-order stationary point of BLO.

Furthermore, when f is assumed to be Clarke regular, and $(x,y) \in \mathcal{M}$ is a first-order stationary point of BLO, Proposition 3.1 illustrates that there exists $(d_x, d_y) \in \partial f(x,y)$ such that $0 = d_x - \nabla_{xy}^2 g(x,y)\nabla_{yy}^2 g(x,y)^{-1}d_y \in \partial h(x,y)$. Therefore, (x,y) is a first-order stationary point of CDB. This completes the proof. □

Proposition 3.2 illustrates that any first-order stationary point of CDB on $\mathcal{M}$ is also a first-order stationary point of BLO. In the rest of this subsection, we aim to show that with a sufficiently large penalty parameter β, any first-order stationary point of CDB lies on $\mathcal{M}$.

Lemma 3.3. *The Lipschitz constant for $\nabla^2_{yy}g(x,y)^{-1}$ is no greater than $\frac{Q_g}{\mu^2}$.*

Proof. Firstly, notice that $\left\|(A+tE)^{-1} - \left(A^{-1} - tA^{-1}EA^{-1}\right)\right\| = \mathcal{O}(t^2)$ holds for any symmetric non-singular matrix A and any square symmetric matrix E. Therefore, the following inequality holds for any $d_x \in \mathbb{R}^n$

$$\left\|\nabla^2_{yy}g(x+td_x,y)^{-1} - \nabla^2_{yy}g(x,y)^{-1}\right\|$$

$$= t\left\|\nabla^2_{yy}g(x,y)^{-1}\nabla^3_{yyx}g(x,y)[d_x]\nabla^2_{yy}g(x,y)^{-1}\right\| + \mathcal{O}(t^2) \leq \frac{Q_g}{\mu^2}t\|d_x\| + \mathcal{O}(t^2).$$

Similarly, for any $d_y \in \mathbb{R}^p$, it holds that

$$\left\|\nabla^2_{yy}g(x,y+td_y)^{-1} - \nabla^2_{yy}g(x,y)^{-1}\right\|$$

$$= t\left\|\nabla^2_{yy}g(x,y)^{-1}\nabla^3_{yyy}g(x,y)[d_y]\nabla^2_{yy}g(x,y)^{-1}\right\| + \mathcal{O}(t^2) \leq \frac{Q_g}{\mu^2}t\|d_y\| + \mathcal{O}(t^2).$$

Therefore, we can conclude that the Lipschitz constant for $\nabla^2_{yy}g(x,y)^{-1}$ is no greater than $\frac{Q_g}{\mu^2}$. $\qquad\square$

Lemma 3.4. *For any given $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, it holds that*

$$\|y^\star(x) - \mathcal{A}(x,y)\| \leq \frac{Q_g}{2\mu^3}\|\nabla_y g(x,y)\|^2.$$

Proof. For any $v \in \mathbb{R}^p$, it follows from the mean-value theorem that there exists $\xi \in \mathbb{R}^p$ such that

$$v^\top \nabla_y g(x,\mathcal{A}(x,y)) = v^\top \nabla_y g(x,y - \nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y))$$

$$= v^\top \nabla_y g(x,y) - v^\top \nabla^2_{yy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)$$

$$+ \frac{1}{2}v^\top \nabla^3_{yyy}g(x,\xi)[\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)]\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)$$

$$\leq \frac{Q_g}{2}\|v\|\left\|\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)\right\|^2 \leq \frac{Q_g}{2\mu^2}\|v\|\|\nabla_y g(x,y)\|^2.$$

As a result, it holds that $\left\|\nabla_y g(x,\mathcal{A}(x,y))\right\| \leq \frac{Q_g}{2\mu^2}\|\nabla_y g(x,y)\|^2$. Then from the fact that $g(x,y)$ is μ-strongly convex with respect to y, we obtain that

$$\|\mathcal{A}(x,y) - y^\star(x)\| \leq \frac{1}{\mu}\left\|\nabla_y g(x,\mathcal{A}(x,y)) - \nabla_y g(x,y^\star(x))\right\| \leq \frac{Q_g}{2\mu^3}\|\nabla_y g(x,y))\|^2.$$

This completes the proof. $\qquad\square$

Remark 3.5. *Lemma 3.4 illustrates that for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, it holds that $\left\|\nabla_y g(x,\mathcal{A}(x,y))\right\| = \mathcal{O}(\|\nabla_y g(x,y)\|^2)$. As a result, the mapping $(x,y) \mapsto (x,\mathcal{A}(x,y))$ satisfies the Assumption 1.2, and hence it is a constraint dissolving mapping for BLO.*

Proposition 3.6. *Suppose $\beta \geq \frac{M_f Q_g}{\mu^3}$ and $\Phi(x)$ is bounded below in $\mathbb{R}^n$. Then $h(x,y)$ is bounded below.*

Proof. We conclude from Lemma 3.4 that

$$h(x,y) - f(x,y^\star(x)) = f(x,\mathcal{A}(x,y)) + \frac{\beta}{2}\|\nabla_y g(x,y)\|^2 - f(x,y^\star(x))$$

$$\geq -M_f\|\mathcal{A}(x,y) - y^\star(x)\| + \frac{\beta}{2}\|\nabla_y g(x,y)\|^2 \geq 0,$$

9

which implies that

$$\inf_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^p} h(x,y) \geq \inf_{x\in\mathcal{M}} f(x,y^\star(x)) = \inf_{x\in\mathbb{R}^n} \Phi(x) > -\infty,$$

hence completes the proof. □

Lemma 3.7. *For any given* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, *and any* $d \in \mathbb{R}^p$, *it holds that*

$$\limsup_{t\to 0}\left|\frac{f(x,\mathcal{A}(x,y+td)) - f(x,\mathcal{A}(x,y))}{t}\right| \leq \frac{M_f Q_g}{\mu^2}\|\nabla_y g(x,y)\|\,\|d\|.$$

Proof. Let $z_t := y + td$. Then it follows from the expression of $\mathcal{A}$ that

$$\begin{aligned}
&|f(x,\mathcal{A}(x,z_t)) - f(x,\mathcal{A}(x,y))|\\
=\ &\left|f(x,z_t - \nabla_{yy}^2 g(x,z_t)^{-1}\nabla_y g(x,z_t)) - f(x,y - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y))\right|\\
\leq\ &\left|f(x,z_t - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t)) - f(x,y - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y))\right|\\
&+ \left|f(x,z_t - \nabla_{yy}^2 g(x,z_t)^{-1}\nabla_y g(x,z_t)) - f(x,z_t - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t))\right|.
\end{aligned}$$

Notice that

$$\left\|td - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t) + \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)\right\| \leq t^2\frac{Q_g}{\mu}\|d\|^2,$$

hence we achieve the following inequality,

$$\begin{aligned}
&\left|f(x,z_t - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t)) - f(x,y - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y))\right|\\
\leq\ &M_f\left\|\left(z_t - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t)\right) - \left(y - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)\right)\right\|\\
\leq\ &t^2\frac{M_f Q_g}{\mu}\|d\|^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\left|f(x,z_t - \nabla_{yy}^2 g(x,z_t)^{-1}\nabla_y g(x,z_t)) - f(x,z_t - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t))\right|\\
\leq\ &M_f\left\|\nabla_{yy}^2 g(x,z_t)^{-1}\nabla_y g(x,z_t) - \nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,z_t)\right\|\\
\leq\ &M_f\left\|\nabla_{yy}^2 g(x,z_t)^{-1} - \nabla_{yy}^2 g(x,y)^{-1}\right\|\,\|\nabla_y g(x,z_t)\|\\
\leq\ &t\frac{Q_g M_f}{\mu^2}\|\nabla_y g(x,z_t)\|\,\|d\| \leq t\frac{Q_g M_f}{\mu^2}\|\nabla_y g(x,y)\|\,\|d\| + t^2\frac{Q_g M_f L_g}{\mu^2}\|d\|^2.
\end{aligned}$$

Therefore, we obtain that

$$\limsup_{t\to 0}\left|\frac{f(x,\mathcal{A}(x,y+td)) - f(x,\mathcal{A}(x,y))}{t}\right| \leq \frac{Q_g M_f}{\mu^2}\|\nabla_y g(x,y)\|\,\|d\|,$$

and the proof is completed. □

Theorem 3.8. *Suppose* $\beta \geq \frac{2Q_g M_f}{\mu^3}$. *If* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ *is a first-order stationary point of CDB, then* $(x,y) \in \mathcal{M}$ *and hence is a first-order stationary point of BLO.*

Proof. Suppose (x, y) is a stationary point of CDB, we have $0 \in \partial h(x, y)$. Therefore, it follows from Definition 2.1 that $0 \leq h^\circ(x, y; 0, -\nabla_y g(x, y))$.

Notice that $\nabla_y g(x, y)$ is differentiable, then it holds that

$$\lim_{t \to 0} \frac{\left\| \nabla_y g(x, y - t \nabla_y g(x, y)) \right\|^2 - \left\| \nabla_y g(x, y) \right\|^2}{t}$$
$$= -2 \nabla_y g(x, y)^\top \nabla_{yy}^2 g(x, y) \nabla_y g(x, y) \leq -2\mu \left\| \nabla_y g(x, y) \right\|^2.$$

Therefore, it holds from Lemma 3.7 that

$$0 \leq h^\circ(x, y; 0, -\nabla_y g(x, y)) = \limsup_{(\tilde{x}, \tilde{y}) \to (x, y),\, t \downarrow 0} \frac{h(\tilde{x}, \tilde{y} - t \nabla_y g(x, y)) - h(\tilde{x}, \tilde{y})}{t}$$

$$= \limsup_{(\tilde{x}, \tilde{y}) \to (x, y),\, t \downarrow 0} \frac{f(\tilde{x}, \mathcal{A}(\tilde{x}, \tilde{y} - t \nabla_y g(x, y))) - f(\tilde{x}, \mathcal{A}(\tilde{x}, \tilde{y}))}{t}$$

$$+ \frac{\beta}{2} \lim_{t \to 0} \frac{\left\| \nabla_y g(x, y - t \nabla_y g(x, y)) \right\|^2 - \left\| \nabla_y g(x, y) \right\|^2}{t}$$

$$\leq -\mu\beta \left\| \nabla_y g(x, y) \right\|^2 + \limsup_{(\tilde{x}, \tilde{y}) \to (x, y)} \frac{M_f Q_g}{\mu^2} \left\| \nabla_y g(\tilde{x}, \tilde{y}) \right\| \left\| \nabla_y g(x, y) \right\|$$

$$\leq -\frac{\mu\beta}{2} \left\| \nabla_y g(x, y) \right\|^2 \leq 0.$$

Therefore, we conclude that $\nabla_y g(x, y) = 0$ and $(x, y) \in \mathcal{M}$. Thus (x, y) is a first-order stationary point of BLO by Proposition 3.2. □

Corollary 3.9. *Suppose f is Clarke regular and $\beta \geq \frac{2 Q_g M_f}{\mu^3}$. Then BLO and CDB have the same first-order stationary points over $\mathbb{R}^n \times \mathbb{R}^p$.*

The proof straightforwardly follows from Theorem 3.8 and Proposition 3.2. Hence we omit its details for simplicity.

3.2 Equivalence: conservative field

In this subsection, we study the equivalence between BLO and CDB based on the concept of conservative field. With the set-valued mapping $\mathcal{D}_h(x, y)$ defined by

$$\mathcal{D}_h(x, y) := \left\{ \begin{bmatrix} d_x + J_{\mathcal{A}, x}(x, y) d_y + \beta \nabla_{xy}^2 g(x, y) \nabla_y g(x, y) \\ J_{\mathcal{A}, y}(x, y) d_y + \beta \nabla_{yy}^2 g(x, y) \nabla_y g(x, y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \mathcal{D}_f(x, \mathcal{A}(x, y)) \right\}, \tag{3.3}$$

we have the following proposition characterizing the property of $\mathcal{D}_h$.

Proposition 3.10. *$\mathcal{D}_h(x, y)$ is a convex-valued conservative field that admits $h(x, y)$ as its potential.*

Proof. Since $\mathcal{A}$ is continuously differentiable, it holds that $\mathcal{A}$ is a potential mapping for its Jacobian $[J_{\mathcal{A}, x}(x, y), J_{\mathcal{A}, y}(x, y)]^\top$. As a result, by the chain rule and sum rule in Proposition 2.13 and Proposition 2.14, $\mathcal{D}_h$ is a conservative field that admits $h(x, y)$ as its potential function. □

Proposition 3.11. *For any given $(x, y) \in \mathcal{M}$, (x, y) is a $\mathcal{D}_f$-stationary point of BLO if and only if (x, y) is a $\mathcal{D}_h$-stationary point of CDB.*

From Definition 2.21, any (x, y) satisfying $0 \in \mathcal{D}_h(x, y)$ is called a $\mathcal{D}_h$-stationary point of CDB. Then Proposition 3.11 directly follows from the expression of $\mathcal{D}_h$, and we omit its proof for simplicity.

Theorem 3.12. *Suppose $\beta \geq \frac{2 Q_g M_f}{\mu^3}$, then $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ is a $\mathcal{D}_f$-stationary point of BLO if and only if (x, y) is a $\mathcal{D}_h$-stationary point of CDB.*

Proof. For any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ and any $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x, y))$, the inclusion $0 \in \mathcal{D}_h(x, y)$ implies that there exists $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x, y))$ such that

$$0 = d_x + J_{\mathcal{A},x}(x, y)d_y + \beta \nabla_{xy}^2 g(x, y) \nabla_y g(x, y),$$
$$0 = J_{\mathcal{A},y}(x, y)d_y + \beta \nabla_{yy}^2 g(x, y) \nabla_y g(x, y).$$

From (3.2) and Assumption 2.15, it holds that $\left\| J_{\mathcal{A},y}(x, y)d_y \right\| \leq \frac{Q_g M_f}{\mu^2} \left\| \nabla_y g(x, y) \right\|$. Then we obtain that

$$
\begin{aligned}
0 &= \left\| J_{\mathcal{A},y}(x, y)d_y + \beta \nabla_{yy}^2 g(x, y) \nabla_y g(x, y) \right\| \\
&\geq \left\| \beta \nabla_{yy}^2 g(x, y) \nabla_y g(x, y) \right\| - \left\| J_{\mathcal{A},y}(x, y)d_y \right\| \\
&\geq \left(\mu\beta - \frac{Q_g M_f}{\mu^2} \right) \left\| \nabla_y g(x, y) \right\| \geq \frac{\mu\beta}{2} \left\| \nabla_y g(x, y) \right\|,
\end{aligned}
$$

which shows that $\nabla_y g(x, y) = 0$. Hence $(x, y) \in \mathcal{M}$. Therefore, Proposition 3.11 illustrates that (x, y) is a $\mathcal{D}_f$-stationary point of BLO.

On the other hand, when (x, y) is a $\mathcal{D}_f$-stationary point of BLO, Proposition 3.11 shows that $0 \in \mathcal{D}_h(x, y)$. Hence (x, y) is a $\mathcal{D}_h$-stationary point of CDB by Definition 2.21. $\qquad \square$

As illustrated in Remark 2.22, the most important example for us is $\mathcal{D}_f = \partial f$, and it is usually easy to compute an approximation for $\partial f(x, y)$ in practice through some randomized approaches [7, 8, 19, 56, 45, 9, 39]. Therefore, we present the following corollary to illustrate the equivalence between BLO and CDB when we choose $\mathcal{D}_f$ as ∂f in (3.3).

Corollary 3.13. *Suppose $\beta \geq \frac{2Q_g M_f}{\mu^3}$ and $\mathcal{D}_h$ is chosen by (3.3) with $\mathcal{D}_f = \partial f$. Then $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ is a first-order stationary point of BLO if and only if (x, y) is a $\mathcal{D}_h$-stationary point of CDB.*

Proof. When $\mathcal{D}_f = \partial f$ in (3.3), the corresponding conservative field $\mathcal{D}_h$ is set as

$$
\mathcal{D}_h(x, y) = \left\{ \begin{bmatrix} d_x + J_{\mathcal{A},x}(x,y)d_y + \beta\nabla_{xy}^2 g(x,y)\nabla_y g(x,y) \\ J_{\mathcal{A},y}(x,y)d_y + \beta\nabla_{yy}^2 g(x,y)\nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \partial_f(x, \mathcal{A}(x, y)) \right\}.
$$

Then from Theorem 3.12 and Definition 2.18, whenever (x, y) is a $\mathcal{D}_h$-stationary point of CDB with $\mathcal{D}_f = \partial f$ in (3.3), (x, y) is a ∂f-stationary point, and hence it is a first-order stationary point of BLO.

On the other hand, when (x, y) is a first-order stationary point of BLO, Proposition 3.11 directly shows that $0 \in \mathcal{D}_h(x, y)$, thus we complete the proof. $\qquad \square$

4 Algorithmic Design

Subgradient method and its variants play important roles in minimizing nonsmooth functions that are not necessarily regular, particularly in training deep neural networks involving nonsmooth activation functions. Recently, [16] shows the global convergence for applying subgradient methods in minimizing nonsmooth functions based on their Clarke subdifferentials. Moreover, [4] introduces the concept of conservative field, which overcomes the limitations of Clarke subdifferential, and further explains the behavior of stochastic subgradient methods when they are applied to train nonsmooth neural networks with automatic differentiation algorithms. Furthermore, [4, 10, 3] establish the convergence properties for some subgradient methods that are developed from the conservative field of the objective function, as they are implemented in practice.

In this section, we aim to design subgradient methods to solve CDB based on the formulation of $\mathcal{D}_h$. In Proposition 3.10, we show that $\mathcal{D}_h$ is a conservative field that admits h as the potential function. Then various existing subgradient approaches [4, 10, 3] can be directly applied to CDB from the explicit formulation of $\mathcal{D}_h$. However, it may be expensive to calculate the $\nabla_{xyy}^3 g$ and $\nabla_{yyy}^3 g$ in practice, hence computing $\mathcal{D}_h(x, y)$ exactly may be expensive and impractical.

To this end, we first propose a general framework for applying subgradient methods to solve CDB, which enables the inexact evaluation of $\mathcal{D}_h$. Then we propose several different set-valued mappings $\hat{\mathcal{D}}_h$, $\hat{\mathcal{D}}_p$ and $\hat{\mathcal{D}}_s$, all of which approximates $\mathcal{D}_h$ and avoid computing the third-order derivatives of g. Based on these set-valued mappings, we design several subgradient methods that adopt inexact evaluations to achieve better efficiency. Moreover, we demonstrate that the global convergence for these subgradient-based methods directly follows from the proposed framework in Section 4.1.

4.1 A unified framework for subgradient-based methods

In this subsection, we utilize the conservative field $\mathcal{D}_h$ to develop a framework for applying sub-gradient methods to solve CDB. We first consider the iteration sequence $\{(x_k, y_k)\}$ generated by the following updating scheme that generalizes the subgradient methods,

$$x_{k+1} = x_k - \eta_k (u_{x,k} + \xi_{x,k}), \quad \text{and} \quad y_{k+1} = y_k - \eta_k \left(u_{y,k} + \xi_{y,k} \right). \tag{4.1}$$

Here $\eta_k > 0$ refers to the stepsize, $(u_{x,k}, u_{y,k})$ should be thought as an approximate descent direction for $h(x, y)$ at (x_k, y_k), Moreover, $\xi_{x,k}$ and $\xi_{y,k}$ denote the "errors" introduced by stochasticity and inexact evaluation. Similar to [16], we stipulate the following assumptions on (4.1).

Assumption 4.1. *(a) The generated iterates $\{(x_k, y_k)\}$ are uniformly bounded: $\sup_{k>0} \|x_k\| + \|y_k\| < +\infty$.*

(b) The stepsizes are nonnegative, square summable, but not summable:

$$\eta_k > 0, \quad \sum_{k=0}^{+\infty} \eta_k = +\infty, \quad \text{and} \quad \sum_{k=0}^{+\infty} \eta_k^2 < +\infty.$$

(c) The series of weighted noise is convergent. That is, there exists $v_x \in \mathbb{R}^n$ and $v_y \in \mathbb{R}^p$, such that

$$\lim_{N \to +\infty} \sum_{k=0}^{N} \eta_k \xi_{x,k} = v_x \text{ and } \lim_{N \to +\infty} \sum_{k=0}^{N} \eta_k \xi_{y,k} = v_y.$$

(d) There exists a set-valued mapping $\mathcal{D}(x, y)$ that has closed graph and compact convex values. Moreover, $\mathcal{D}$ has the property that for any sequence $\{(x_{k_j}, y_{k_j})\}$ that converges to a point $(\tilde{x}, \tilde{y})$ and any unbounded increasing sequence $\{k_j\}$, it holds that $\lim_{N \to +\infty} \text{dist} \left(\frac{1}{N} \sum_{j=1}^{N} \begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \mathcal{D}(\tilde{x}, \tilde{y}) \right) = 0.$

(e) The set $\{h(x, y) : 0 \in \mathcal{D}(x, y)\}$ has empty interior, i.e. its complementary is dense in $\mathbb{R}$.

(f) There exists a constant $\delta > 0$ such that for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$ and any $w \in \mathcal{D}(x, y)$, it holds that $\sup_{\zeta \in \mathcal{D}_h(x,y)} \zeta^\top w \geq \delta \|w\|^2.$

Assumption 4.1(a)-(b) are common assumptions in various existing works [16, 4, 10]. Assumption 4.1(c) is a mild assumption that controls the growth of the noise sequence $\{(\xi_{x,k}, \xi_{y,k})\}$ as the stepsize decreases, which can be satisfied by the stochastic subgradient method described in [16]. Moreover, Assumption 4.1(d) illustrates how $(u_{x,k}, u_{y,k})$ approximates $\mathcal{D}(x_k, y_k)$. Assumption 4.1(e) is the weak Sard's condition [16, Assumption B(1), Assumption F(1)], which holds whenever h is definable and $\mathcal{D} = \partial h$ [16, Lemma 5.7]. Furthermore, Assumption 4.1(f) implies the descent condition in [16, Assumption B(2)], as illustrated in the following proposition.

Proposition 4.2. *Suppose Assumption 4.1 holds. Let $\gamma : \mathbb{R}_+ \to \mathbb{R}^n \times \mathbb{R}^p$ be any absolutely continuous path such that the differential inclusion $\gamma'(t) \in -\mathcal{D}(\gamma(t))$ holds for a.e. $t \in \mathbb{R}_+$. Then the following inequality holds for any $t > 0$,*

$$h(\gamma(t)) - h(\gamma(0)) \leq -\delta \int_0^t \text{dist} \left(0, \mathcal{D}(\gamma(\tau)) \right)^2 d\tau.$$

Proof. Notice that h is the potential function of the conservative field $\mathcal{D}_h$. Therefore, it follows from Definition 2.10 that

$$h(\gamma(t)) - h(\gamma(0)) = \int_0^t \inf_{\zeta \in \mathcal{D}_h(\gamma(t))} \langle \zeta, -\gamma'(t) \rangle \, d\tau \leq -\delta \int_0^t \text{dist} \, (0, \mathcal{D}(\gamma(\tau)))^2 \, d\tau,$$

and this completes the proof. $\square$

Proposition 4.3. *For any set-valued mapping $\mathcal{D} : \mathbb{R}^m \rightrightarrows \mathbb{R}^s$, suppose $\mathcal{D}$ is compact and convex valued and has closed graph, then for any $\tilde{w} \in \mathbb{R}^m$, any sequence $\{w_k\}$ that converges to $\tilde{w}$ and any $\{u_k\}$ that satisfies $\lim_{k \to +\infty} \text{dist} \, (u_k, \mathcal{D}(w_k)) = 0$, it holds that*

$$\lim_{N \to +\infty} \text{dist} \left(\frac{1}{N} \sum_{k=1}^N u_k, \mathcal{D}(\tilde{w}) \right) = 0.$$

Proof. We first assume that the argument to be proved is not true. Then there exists a constant $\varepsilon_0 > 0$, a sequence $\{w_k\}$ converging to $\tilde{w}$, a sequence $\{u_k\}$ that satisfies $\lim_{k \to +\infty} \text{dist} \, (u_k, \mathcal{D}(w_k)) = 0$ and a sequence $\{N_j\} \subset \mathcal{N}$ satisfying $N_j \to +\infty$, such that

$$\text{dist} \left(\frac{1}{N_j} \sum_{k=1}^{N_j} u_k, \mathcal{D}(\tilde{w}) \right) \geq \varepsilon_0. \tag{4.2}$$

From the convexity of $\mathcal{D}(\tilde{w})$, we conclude that for any $j \geq 0$, there exists an index $k_j \leq N_j$ such that

$$\text{dist} \left(u_{k_j}, \mathcal{D}(\tilde{w}) \right) \geq \frac{\varepsilon_0}{2}. \tag{4.3}$$

We claim that we can always choose a sequence $\{k_j\}$ such that $k_j \to +\infty$. Otherwise, for any $N > \sup_{j \geq 0} k_j + \left(\frac{2}{\varepsilon_0} \right) \sum_{i=1}^{\sup_{j \geq 0} k_j} \text{dist} \, (u_i, \mathcal{D}(\tilde{w}))$, it holds that

$$\text{dist} \left(\frac{1}{N} \sum_{k=1}^N u_k, \mathcal{D}(\tilde{w}) \right) \leq \frac{1}{N} \sum_{k=1}^N \text{dist} \, (u_k, \mathcal{D}(\tilde{w})) \leq \frac{1}{N} \sum_{k=1}^{\sup_{j \geq 0} k_j} \text{dist} \, (u_k, \mathcal{D}(\tilde{w})) + \frac{\varepsilon_0}{2} < \varepsilon_0,$$

which contradicts (4.2) and further verifies our claim.

Therefore, for the selected sequence of indices $\{k_j\}$, it holds that $\lim_{j \to +\infty} w_{k_j} = \tilde{w}$, and

$$\lim_{j \to +\infty} \text{dist} \left((w_{k_j}, u_{k_j}), \text{graph}(\mathcal{D}) \right) = 0.$$

Since $\mathcal{D}$ has closed graph, any cluster point of $\{u_{k_j}\}$ lies in $\mathcal{D}(\tilde{w})$, which further leads to

$$\liminf_{j \to +\infty} \text{dist} \left(u_{k_j}, \mathcal{D}(\tilde{w}) \right) = 0.$$

But this contradicts (4.3). Thus the proof is completed by contradiction. $\square$

Theorem 4.4. *Suppose Assumption 4.1 holds. Then for the sequence $\{(x_k, y_k)\}$ generated from (4.1), all its limit point lies in $\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^p : 0 \in \mathcal{D}(x, y)\}$. Moreover, the sequence of function values $\{h(x_k, y_k)\}$ converges.*

Proof. Assumption 4.1(a)-(d) imply the validity of Assumption A in [16]. Moreover, Assumption 4.1(e)-(f) and Proposition 4.2 show that the Assumption B in [16] holds. Then the proof directly follows from Theorem 3.2 in [16]. $\square$

4.2 Basic subgradient methods

In this subsection, we first propose a set-valued mapping $\hat{\mathcal{D}}_h(x,y)$ that has compact values and satisfies Assumption 4.1(f). Based on $\hat{\mathcal{D}}_h(x,y)$, we develop a subgradient method as illustrated in Algorithm 1, where the update direction in each iteration is approximately chosen from $\hat{\mathcal{D}}_h(x,y)$. Then we establish the global convergence of Algorithm 1 directly from our proposed framework.

Definition 4.5. *For any given* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, *we define the set-valued mapping* $\hat{\mathcal{D}}_h : \mathbb{R}^n \times \mathbb{R}^p \rightrightarrows$ $\mathbb{R}^n \times \mathbb{R}^p$ *as*

$$\hat{\mathcal{D}}_h(x,y) := \left\{ \begin{bmatrix} d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \\ \beta \nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \mathcal{D}_f(x, \mathcal{A}(x,y)) \right\}.$$

It is easy to verify that $\hat{\mathcal{D}}_h$ has closed graph. Moreover, compared with $\mathcal{D}_h$, the formulation of $\hat{\mathcal{D}}_h$ avoids the third-order derivatives of g. Therefore, computing an element from $\hat{\mathcal{D}}_h$ can be potentially more efficient than directly computing one from $\mathcal{D}_h$.

Proposition 4.6. *Suppose* $\beta \geq \frac{2M_f Q_g}{\mu^3}$. *Then for any given* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$, *it is a* $\mathcal{D}_h$-*stationary point of CDB if and only if* $0 \in \hat{\mathcal{D}}_h(x,y)$.

Proof. When $0 \in \hat{\mathcal{D}}_h(x,y)$, we first conclude that $\nabla_y g(x,y) = 0$, which results in the inclusion $(x,y) \in \mathcal{M}$. Moreover, $0 \in \hat{\mathcal{D}}_h(x,y)$ implies that there exists $(d_x, d_y) \in \mathcal{D}_f(x,y)$ such that $d_x - \nabla_{xy}^2 g(x,y)\nabla_{yy}^2 g(x,y)^{-1} d_y = 0$. Therefore, it follows from Definition 2.20 that (x,y) is a $\mathcal{D}_h$-stationary point of CDB.

On the other hand, when (x,y) is a $\mathcal{D}_h$-stationary point of CDB, from Theorem 3.12, it holds that $(x,y) \in \mathcal{M}$. Therefore, from the expression of $\mathcal{D}_h(x,y)$ and $\hat{\mathcal{D}}_h(x,y)$, we obtain that $0 \in \mathcal{D}_h(x,y) = \hat{\mathcal{D}}_h(x,y)$ and the proof is completed. $\qquad\square$

Proposition 4.7. *Suppose* $\beta \geq \frac{2M_f Q_g}{\mu^3}$. *Then for any given* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ *and* $w \in \hat{\mathcal{D}}_h(x,y)$, *it holds that*

$$\sup_{\zeta \in \mathcal{D}_h(x,y)} \langle w, \zeta \rangle \geq \min \left\{ 1, \frac{(2 - \sqrt{2})\mu}{2} \right\} \|w\|^2.$$

Proof. For any $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x,y))$, let

$$w = \begin{bmatrix} d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \\ \beta \nabla_y g(x,y) \end{bmatrix} \in \hat{\mathcal{D}}_h(x,y),$$

and define

$$z_1 = \begin{bmatrix} d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \\ \beta \nabla_{yy}^2 g(x,y)\nabla_y g(x,y) \end{bmatrix}, z_2 = \begin{bmatrix} \nabla_{xyy}^3 g(x,y)[\nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)]\nabla_{yy}^2 g(x,y)^{-1} d_y \\ \nabla_{yyy}^3 g(x,y)[\nabla_{yy}^2 g(x,y)^{-1}\nabla_y g(x,y)]\nabla_{yy}^2 g(x,y)^{-1} d_y \end{bmatrix}.$$

Then from the expression of $\mathcal{D}_h$, we have $z_1 + z_2 \in \mathcal{D}_h(x,y)$. Moreover, the expression of w and Lemma 3.3 implies $\|w\| \geq \beta \|\nabla_y g(x,y)\|$ and $\|z_2\| \leq \frac{\sqrt{2}Q_g M_f}{\mu^2} \|\nabla_y g(x,y)\|$. As a result, we obtain

$$\langle w, z_1 + z_2 \rangle \geq \left\| d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \right\|^2$$
$$+ \beta^2 \mu \|\nabla_y g(x,y)\|^2 - \frac{\sqrt{2}Q_g M_f}{\mu^2} \|\nabla_y g(x,y)\| \|w\|$$
$$\geq \left\| d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \right\|^2 + \frac{(2 - \sqrt{2})\beta^2 \mu}{2} \|\nabla_y g(x,y)\|^2$$
$$\geq \min \left\{ 1, \frac{(2 - \sqrt{2})\mu}{2} \right\} \|w\|^2,$$

and this completes the proof. $\qquad\square$

Algorithm 1 Basic subgradient method for solving CDB.

Require: Function f, g, initial point x_0, y_0.
1: **for** k = 1,2,... **do**
2:     Compute w_k by approximately evaluating $\nabla_{yy}^2 g(x_k, y_k)^{-1} \nabla_y g(x_k, y_k)$ such that $\left\| \nabla_{yy}^2 g(x_k, y_k) w_k - \nabla_y g(x_k, y_k) \right\| \leq \varepsilon_{1,k}$.
3:     Choose $(d_{x,k}, d_{y,k})$ as an approximated evaluation of $\mathcal{D}_f(x_k, y_k - w_k)$.
4:     Compute v_k such that $\left\| \nabla_{yy}^2 g(x_k, y_k) v_k - d_{y,k} \right\| \leq \varepsilon_{2,k}$.
5:     Update x_k and y_k by

$$x_{k+1} = x_k - \eta_k \left(d_{x,k} - \nabla_{xy}^2 g(x_k, y_k) \left(v_k - \beta \nabla_y g(x_k, y_k) \right) \right),$$
$$y_{k+1} = y_k - \eta_k \beta \nabla_y g(x_k, y_k).$$

6: **end for**
7: Return x_k and y_k.

With the definition of $\hat{\mathcal{D}}_h$, Proposition 4.6 and Proposition 4.7, we can now present a basic subgradient method for solving CDB in Algorithm 1. We observe that in Algorithm 1, the search direction $\begin{bmatrix} d_{x,k} - \nabla_{xy}^2 g(x_k, y_k) \left(v_k - \beta \nabla_y g(x_k, y_k) \right) \\ \nabla_y g(x_k, y_k) \end{bmatrix}$ is an element that is approximately in $\hat{\mathcal{D}}_h(x_k, y_k)$.

To establish the convergence of Algorithm 1, we need the following assumption.

Assumption 4.8. *In Algorithm 1, we assume*

(a) *The iterates are uniformly bounded:* $\sup_{k>0} \|x_k\| + \|y_k\| < +\infty$.

(b) *The stepsize is nonnegative, square summable, but not summable:*

$$\eta_k \geq 0, \quad \sum_{k=0}^{+\infty} \eta_k = +\infty, \quad and \quad \sum_{k=0}^{+\infty} \eta_k^2 < +\infty. \tag{4.4}$$

(c) *The set* $\{f(x,y) : (x,y) \text{ is a } \mathcal{D}_f\text{-stationary point of BLO}\}$ *has empty interior.*

(d) *Let the filtration* $\{\mathcal{F}_k\}$ *be the collection of the increasing* σ*-fields, i.e.,*

$$\mathcal{F}_k := \sigma((x_j, y_j, d_{x,j}, d_{y,j}) : j < k).$$

There exists a constant M_σ *such that the approximated evaluation* $(d_{x,k}, d_{y,k})$ *satisfies the following inequalities,*

$$\mathbb{E}\left[\left\| \left(d_{x,k} - \mathbb{E}[d_{x,k}|\mathcal{F}_k], \ d_{y,k} - \mathbb{E}[d_{y,k}|\mathcal{F}_k] \right) \right\|^2 \Big| \mathcal{F}_k \right] \leq M_\sigma, \quad \text{for any } k \geq 1,$$
$$\lim_{k \to +\infty} \text{dist}\left(\mathcal{D}_f(x_k, y_k - w_k), \left(\mathbb{E}[d_{x,k}|\mathcal{F}_k], \ \mathbb{E}[d_{y,k}|\mathcal{F}_k] \right) \right) = 0.$$

Assumption 4.8(a)-(b) is the same as Assumption 4.1(a)-(b). Moreover, Assumption 4.8(c) holds whenever both f and $\mathcal{M}$ are definable, and $\mathcal{D}_f = \partial f$ [16, Corollary 6.4], hence it is mild in practice. In addition, Assumption 4.8(d) characterizes the way in which $(d_{x,k}, d_{y,k})$ is an approximated evaluation of $\mathcal{D}_f(x_k, y_k - w_k)$ in the sense of conditional expectation.

Proposition 4.9. *Suppose* $\{\chi_k\}$ *is a series of random variables such that for each* $k \geq 1$, χ_k *is* $\mathcal{F}_{k+1}$*-measurable,* $\mathbb{E}[\chi_k|\mathcal{F}_k] = 0$, $\mathbb{E}[\|\chi_k\|] < +\infty$, *and* $\sup_{k>1} \mathbb{E}[\|\chi_k\|^2] < +\infty$. *Then for any* $\{\eta_k\}$ *satisfying* (4.4), $\sum_{j=1}^{k} \eta_j \chi_j$ *converges to a finite limit almost surely.*

Proof. Let $\tau_k := \sum_{j=1}^{k} \eta_j \chi_j$. From the definition of τ_k, we can conclude that for each $k \geq 1$, $\mathbb{E}[\tau_k | \mathcal{F}_k] = \tau_{k-1}$ and $\mathbb{E}[\|\tau_k\|] \leq \sum_{j=1}^{k} \eta_j \mathbb{E}[\|\chi_j\|] < +\infty$. Then $\{\tau_k\}$ is a martingale with respect to the filtration $\{\mathcal{F}_k\}$ [17, Definition 5.1.4]. Moreover, since $\mathbb{E}[\chi_k | \mathcal{F}_k] = 0$, it holds that

$$\mathbb{E}[\|\tau_k\|^2] = \mathbb{E}[\mathbb{E}[\|\tau_{k-1} + \eta_k \chi_k\|^2 \,|\mathcal{F}_k]] \leq \eta_k^2 \mathbb{E}[\|\chi_k\|^2] + \mathbb{E}[\|\tau_{k-1}\|^2],$$

Therefore, $\sup_{k>1} \mathbb{E}[\|\tau_k\|^2] < +\infty$, hence $\{\tau_k\}$ is an L^2-martingale. Then from [17, Theorem 5.3.33], we can conclude that τ_k converges to a finite limit almost surely. □

Theorem 4.10. *Suppose Assumption 4.8 holds, $\beta \geq \frac{2M_f Q_g}{\mu^3}$ and the tolerances $\varepsilon_{1,k}$ and $\varepsilon_{2,k}$ satisfy $\lim_{k \to +\infty} \varepsilon_{1,k} = 0$ and $\sum_{k=0}^{+\infty} \varepsilon_{2,k} \eta_k < +\infty$. Then almost surely, every limit point of $\{(x_k, y_k)\}$ in Algorithm 1 is a $\mathcal{D}_f$-stationary point of BLO and $\{h(x_k, y_k)\}$ converges.*

Proof. Consider the following auxiliary set-valued mapping $\mathcal{D}_{temp} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p \rightrightarrows \mathbb{R}^n \times \mathbb{R}^p$,

$$\mathcal{D}_{temp}(x, y, z) := \left\{ \begin{bmatrix} d_x - \nabla_{xy}^2 g(x,y)\left(\nabla_{yy}^2 g(x,y)^{-1} d_y - \beta \nabla_y g(x,y)\right) \\ \beta \nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \mathcal{D}_f(x, z) \right\}.$$

It is easy to verify that $\mathcal{D}_{temp}$ has closed graph. Moreover, $\hat{\mathcal{D}}_h(x, y) = \mathcal{D}_{temp}(x, y, \mathcal{A}(x, y))$ holds for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$.

Assumption 4.8(a) and 4.8(b) imply that Assumption 4.1(a) and 4.1(b) hold. Let $\tilde{d}_{x,k} = \mathbb{E}[d_{x,k} | \mathcal{F}_k]$, $\tilde{d}_{y,k} = \mathbb{E}[d_{y,k} | \mathcal{F}_k]$, and

$$u_{x,k} = \tilde{d}_{x,k} - \nabla_{xy}^2 g(x_k, y_k) \left(\nabla_{yy}^2 g(x_k, y_k)^{-1} \tilde{d}_{y,k} - \beta \nabla_y g(x_k, y_k)\right), \quad u_{y,k} = \beta \nabla_y g(x_k, y_k),$$

$$\chi_{x,k} = (d_{x,k} - \tilde{d}_{x,k}) - \nabla_{xy}^2 g(x_k, y_k) \nabla_{yy}^2 g(x_k, y_k)^{-1} (d_{y,k} - \tilde{d}_{y,k}),$$

$$\xi_{x,k} = \nabla_{xy}^2 g(x_k, y_k)(\nabla_{yy}^2 g(x_k, y_k)^{-1} d_{y,k} - v_k) + \chi_{x,k}, \quad \xi_{y,k} = 0.$$

Then from Step 4 in Algorithm 1 we obtain $\|\xi_{x,k}\| \leq \mu^{-1} L_g \varepsilon_{2,k}$. As a result, Assumption 4.8(b) shows that $\frac{L_g}{\mu} \sum_{k=0}^{+\infty} \varepsilon_{2,k} \eta_k < +\infty$. Moreover, Proposition 4.9 illustrates that $\sum_{k=1}^{+\infty} \eta_k \chi_{x,k} < +\infty$. Thus $\sum_{k=0}^{+\infty} \|\eta_k \xi_{x,k}\| < +\infty$ and Assumption 4.1(c) holds.

Assumption 4.8(d) illustrates that $\lim_{k \to +\infty} \text{dist}\left((u_{x,k}, u_{y,k}), \mathcal{D}_{temp}(x_k, y_k, y_k - w_k)\right) = 0$, and Step 2 in Algorithm 1 shows that $\lim_{k \to +\infty} \left\|w_k - (\nabla_{yy}^2 g(x_k, y_k))^{-1} \nabla_y g(x_k, y_k)\right\| = 0$. For any sequence $\{k_j\} \subset \mathbb{N}$ such that $\lim_{j \to +\infty}(x_{k_j}, y_{k_j}) = (\tilde{x}, \tilde{y})$, it holds from Step 2 in Algorithm 1 that $\lim_{j \to +\infty}(y_{k_j} - w_{k_j}) = \mathcal{A}(\tilde{x}, \tilde{y})$. Then Proposition 4.3 illustrates that

$$\lim_{N \to +\infty} \text{dist}\left(\frac{1}{N}\sum_{j=1}^{N}\begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \hat{\mathcal{D}}_h(\tilde{x}, \tilde{y})\right) = \lim_{N \to +\infty} \text{dist}\left(\frac{1}{N}\sum_{j=1}^{N}\begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \mathcal{D}_{temp}(\tilde{x}, \tilde{y}, \mathcal{A}(\tilde{x}, \tilde{y}))\right) = 0,$$

which guarantees Assumption 4.1(d).

Furthermore, Assumption 4.1(e) directly follows from Assumption 4.8(c) and Proposition 4.6, and Assumption 4.1(f) is implied by Proposition 4.7. Therefore, Assumption 4.1 holds for Algorithm 1.

As a result, based on Theorem 4.4 and Theorem 3.12, we obtain that any cluster point of the sequence $\{(x_k, y_k)\}$ generated by Algorithm 1 is a $\mathcal{D}_f$-stationary point of BLO, and the sequence $\{h(x_k, y_k)\}$ converges. □

As illustrated in Remark 2.22, when we choose the $(d_{x,k}, d_{y,k})$ in Algorithm 1 by random sampling of gradient approaches, i.e., $(d_{x,k}, d_{y,k}) \in \partial_{\eta_k} f(x_k, y_k - w_k)$. It holds from Proposition 2.5 that $\text{dist}\left((d_{x,k}, d_{y,k}), \partial f(x_k, y_k - w_k)\right) \to 0$. Hence $(d_{x,k}, d_{y,k})$ satisfies Assumption 4.8(d) with $\mathcal{D}_f = \partial f$.

Similarly, if we choose $(d_{x,k}, d_{y,k})$ by randomized smoothing approaches, [39, Theorem 3.1, Lemma D.1] illustrates that Assumption 4.8(d) is satisfied with $\mathcal{D}_f = \partial f$. Then we immediately have the

following corollary illustrating that $\{(x_k, y_k)\}$ in Algorithm 1 converges to a first-order stationary point of BLO.

Corollary 4.11. *Suppose Assumption 4.8 holds with $\mathcal{D}_f = \partial f$, the tolerances $\varepsilon_{1,k}$ and $\varepsilon_{2,k}$ satisfy $\lim\limits_{k \to +\infty} \varepsilon_{1,k} = 0$ and $\sum\limits_{k=0}^{+\infty} \varepsilon_{2,k} \eta_k < +\infty$. Moreover, suppose $(d_{x,k}, d_{y,k})$ in Algorithm 1 is generated by one of the following schemes in each iteration k,*

- $(d_{x,k}, d_{y,k}) \in \partial_{\eta_k} f(x_k, y_k - w_k)$;

- $(d_{x,k}, d_{y,k}) = \tilde{\partial}_{\eta_k} f(x_k, y_k - w_k; \zeta_{x,k}, \zeta_{y,k})$, where $(\zeta_{x,k}, \zeta_{y,k})$ is uniformly sampled over $\mathbb{B}_\delta(0)$ and independent of $\mathcal{F}_k$.

Then every limit point of $\{(x_k, y_k)\}$ in Algorithm 1 is a first-order stationary point of BLO and $\{h(x_k, y_k)\}$ converges.

4.3 A modified subgradient method

Recently, an efficient single-loop algorithm, named TTSA, is proposed by [27] for BLO with smooth f. The deterministic version of TTSA follows the following updating schemes,

$$
\begin{aligned}
x_{k+1} &= x_k - \eta_k \left(\nabla_x f(x_k, y_k) - \nabla^2_{xy} g(x_k, y_k) \nabla^2_{yy} g(x_k, y_k)^{-1} \nabla_y f(x_k, y_k) \right), \\
y_{k+1} &= y_k - \tau_k \nabla_y g(x_k, y_k).
\end{aligned}
\tag{4.5}
$$

The x-variable in TTSA is updated along an approximate gradient direction of $\Phi(x)$, while the y-variable is updated by taking a gradient descent step for the lower-level problem of BLO. [27] proves the global convergence of TTSA under a two-timescale condition, i.e., the ratio of stepsizes η_k/τ_k tends to zero as the maximum number of iterations goes to infinity. Very recently, [35] proposes another single-loop algorithm named SUSTAIN, which can be regarded as a momentum-accelerated version of TTSA and waives the two-timescale condition in TTSA. However, the analysis for TTSA and SUSTAIN is based on the Lipschitz smoothness of f. To our best knowledge, the methodologies employed in [27, 35] cannot be applied to the nonsmooth bilevel problem (BLO).

In this subsection, we first consider the following set-valued mapping with a prefixed constant $\hat{\beta} > 0$,

$$
\hat{\mathcal{D}}_s(x, y) := \left\{ \begin{bmatrix} d_x - \nabla^2_{xy} g(x,y) \nabla^2_{yy} g(x,y)^{-1} d_y \\ \hat{\beta} \nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \mathcal{D}_f(x, \mathcal{A}(x,y)) \right\},
$$

which yields a subgradient method as presented in Algorithm 2. Moreover, based on our proposed framework in Section 4.1, we prove the convergence properties of Algorithm 2 and discuss its relationship with the TTSA algorithm in Remark 4.16.

In the next two propositions, we establish some properties of $\hat{\mathcal{D}}_s$.

Proposition 4.12. *Suppose $\beta \geq \frac{2 Q_g M_f}{\mu^3}$. Then for any given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, $0 \in \mathcal{D}_h(x, y)$ if and only if $0 \in \hat{\mathcal{D}}_s(x, y)$.*

The proof is similar to Proposition 4.6, hence we omit its proof for simplicity.

Proposition 4.13. *Suppose $\beta \geq \frac{4 Q_g M_f}{\mu^3}$, and $\hat{\beta} \geq \beta \max \left\{ \frac{8 L_g^2}{\mu}, \frac{1}{4\mu}, \frac{\mu}{4} \right\}$. Then for any given $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, and for any $w \in \hat{\mathcal{D}}_s(x, y)$, it holds that*

$$
\sup_{z \in \mathcal{D}_h(x,y)} \langle \xi, z \rangle \geq \min \left\{ \frac{1}{4}, \frac{\beta^2}{16 \hat{\beta}^2} \right\} \|w\|^2.
$$

Proof. For any $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x,y))$, let

$$z_1 = \begin{bmatrix} d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y \\ \beta\nabla^2_{yy}g(x,y)\nabla_y g(x,y) \end{bmatrix}, \quad z_2 = \begin{bmatrix} \beta\nabla^2_{xy}g(x,y)\nabla_y g(x,y) \\ 0 \end{bmatrix},$$

$$z_3 = \begin{bmatrix} \nabla^3_{xyy}g(x,y)[\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)]\nabla^2_{yy}g(x,y)^{-1}d_y \\ \nabla^3_{yyy}g(x,y)[\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)]\nabla^2_{yy}g(x,y)^{-1}d_y \end{bmatrix}, \quad w = \begin{bmatrix} d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y \\ \hat{\beta}\nabla_y g(x,y) \end{bmatrix}. \tag{4.6}$$

Then it holds that $z_1 + z_2 + z_3 \in \mathcal{D}_h(x,y)$ and $w \in \hat{\mathcal{D}}_s(x,y)$. From the expression of w and the Lipschitz continuity of $\nabla^2_{yy}g(x,y)$, we have

$$\|w\| \le \hat{\beta}\left\|\nabla_y g(x,y)\right\| + \left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|, \|z_3\| \le \frac{Q_g M_f}{\mu^2}\left\|\nabla_y g(x,y)\right\|,$$

which further implies that

$$\langle z_3, w\rangle$$
$$\ge -\frac{Q_g M_f}{\mu^2}\left\|\nabla_y g(x,y)\right\|\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\| - \frac{\hat{\beta}Q_g M_f}{\mu^2}\left\|\nabla_y g(x,y)\right\|^2$$
$$\ge -\frac{1}{4}\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|^2 - \left(\frac{\hat{\beta}Q_g M_f}{\mu^2} + \frac{Q_g^2 M_f^2}{\mu^4}\right)\left\|\nabla_y g(x,y)\right\|^2.$$

As a result, we obtain

$$\langle w, z_1 + z_2 + z_3\rangle \ge \left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|^2 + \beta\hat{\beta}\mu\left\|\nabla_y g(x,y)\right\|^2$$
$$- 2L_g\beta\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|\left\|\nabla_y g(x,y)\right\| + \langle z_3, w\rangle$$
$$\ge \frac{1}{2}\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|^2 + \left(\beta\hat{\beta}\mu - 2L_g^2\beta^2\right)\left\|\nabla_y g(x,y)\right\|^2 + \langle z_3, w\rangle$$
$$\ge \frac{1}{4}\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|^2 + \left(\beta\hat{\beta}\mu - 2L_g^2\beta^2 - \frac{\hat{\beta}Q_g M_f}{\mu^2} - \frac{Q_g^2 M_f^2}{\mu^4}\right)\left\|\nabla_y g(x,y)\right\|^2$$
$$\ge \frac{1}{4}\left\|d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y\right\|^2 + \frac{\beta^2}{16}\left\|\nabla_y g(x,y)\right\|^2 \ge \min\left\{\frac{1}{4}, \frac{\beta^2}{16\hat{\beta}^2}\right\}\|w\|^2,$$

and the proof is completed. $\qquad\square$

With Propositions 4.13 and (4.6), we are now ready to present our modified subgradient method for solving CDB in Algorithm 2, and establish its convergence.

Theorem 4.14. *Suppose Assumption 4.8 holds, $\beta \ge \frac{4Q_g M_f}{\mu^3}$, $\hat{\beta} \ge \beta \cdot \max\left\{\frac{8L_g^2}{\mu}, \frac{1}{4\mu}, \frac{\mu}{4}\right\}$ and the tolerance $\varepsilon_{1,k}$ and $\varepsilon_{2,k}$ satisfy $\lim_{k\to+\infty}\varepsilon_{1,k} = 0$, $\sum_{k=0}^{+\infty}\varepsilon_{2,k}\eta_k < +\infty$. Then every limit point of $\{(x_k, y_k)\}$ generated by Algorithm 2 is a $\mathcal{D}_f$-stationary point of BLO and $\{h(x_k, y_k)\}$ converges.*

Proof. Consider the auxiliary set-valued mapping $\mathcal{D}_{temp} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p \rightrightarrows \mathbb{R}^n \times \mathbb{R}^p$ that is defined as

$$\mathcal{D}_{temp}(x,y,z) := \left\{\begin{bmatrix} d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y \\ \hat{\beta}\nabla_y g(x,y) \end{bmatrix} : \begin{bmatrix} d_x \\ d_y \end{bmatrix} \in \mathcal{D}_f(x,z)\right\}.$$

Then it is easy to verify that $\mathcal{D}_{temp}$ has closed graph.

Assumption 4.8(a) and 4.8(b) implies that Assumption 4.1(a) and 4.1(b) hold. Let $\tilde{d}_{x,k} = \mathbb{E}[d_{x,k}|\mathcal{F}_k]$, $\tilde{d}_{y,k} = \mathbb{E}[d_{y,k}|\mathcal{F}_k]$, and

$$u_{x,k} = \tilde{d}_{x,k} - \nabla^2_{xy}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}\tilde{d}_{y,k}, \quad u_{y,k} = \hat{\beta}\nabla_y g(x_k, y_k),$$
$$\chi_{x,k} = (d_{x,k} - \tilde{d}_{x,k}) - \nabla^2_{xy}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}(d_{y,k} - \tilde{d}_{y,k}),$$
$$\xi_{x,k} = \nabla^2_{xy}g(x_k, y_k)(\nabla^2_{yy}g(x_k, y_k)^{-1}d_{y,k} - v_k) + \chi_{x,k}, \quad \xi_{y,k} = 0.$$

Algorithm 2 A modified subgradient method for solving CDB.

Require: Function f, g, initial point x_0, y_0.

1: **for** k = 1,2,... **do**
2:     Set the tolerance $\varepsilon_{1,k}$ and $\varepsilon_{2,k}$.
3:     Compute $\nabla_y g(x_k, y_k)$.
4:     Compute an approximated evaluation w_k for $\nabla^2_{yy} g(x_k, y_k)^{-1} \nabla_y g(x_k, y_k)$ that satisfies $\left\| \nabla^2_{yy} g(x_k, y_k) w_k - \nabla_y g(x_k, y_k) \right\| \leq \varepsilon_{1,k}$.
5:     Choose $(d_{x,k}, d_{y,k})$ as an approximated evaluation of $\mathcal{D}_f(x_k, y_k - w_k)$.
6:     Compute v_k such that $\left\| \nabla^2_{yy} g(x_k, y_k) v_k - d_{y,k} \right\| \leq \varepsilon_{2,k}$.
7:     Update x_k and y_k by

$$x_{k+1} = x_k - \eta_k \left(d_{x,k} - \nabla^2_{xy} g(x_k, y_k) v_k \right),$$
$$y_{k+1} = y_k - \eta_k \hat{\beta} \nabla_y g(x_k, y_k).$$

8: **end for**
9: Return x_k and y_k.

Then from Step 6 in Algorithm 2 we obtain $\left\| \xi_{x,k} \right\| \leq \mu^{-1} L_g \varepsilon_{2,k}$. As a result, Assumption 4.1(b) shows that $\mu^{-1} L_g \sum_{k=1}^{+\infty} \varepsilon_{2,k} \eta_k < +\infty$. Furthermore, from Proposition 4.9 it holds that $\sum_{k=1}^{+\infty} \chi_{x,k}$ converges to a finite limit almost surely. Therefore, $\sum_{k=0}^{+\infty} \eta_k \xi_{x,k}$ converges to a finite limit almost surely, and hence Assumption 4.1(c) holds.

Notice that $\text{dist}\left((u_{x,k}, u_{y,k}), \mathcal{D}_{temp}(x_k, y_k, y_k - w_k) \right) \to 0$. Moreover, Step 4 in Algorithm 1 shows that $\lim_{k \to +\infty} \left\| w_k - \nabla^2_{yy} g(x_k, y_k) \nabla_y g(x_k, y_k) \right\| = 0$. For any sequence $\{k_j\} \subset \mathbb{N}$ such that $\lim_{j \to +\infty} (x_{k_j}, y_{k_j}) = (\tilde{x}, \tilde{y})$, $y_{k_j} - w_{k_j} \to \mathcal{A}(x_{k_j}, y_{k_j})$. Then Proposition 4.3 illustrates that

$$\lim_{N \to +\infty} \text{dist}\left(\frac{1}{N} \sum_{j=1}^{N} \begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \hat{\mathcal{D}}_s(\tilde{x}, \tilde{y}) \right) = \lim_{N \to +\infty} \text{dist}\left(\frac{1}{N} \sum_{j=1}^{N} \begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \mathcal{D}_{temp}(\tilde{x}, \tilde{y}, \mathcal{A}(\tilde{x}, \tilde{y})) \right) = 0,$$

which guarantees Assumption 4.1(d).

Furthermore, Assumption 4.1(e) directly follows from Assumption 4.8(c) and Proposition 4.12, and Assumption 4.1(f) is implied by Proposition 4.13. From Theorem 4.4 and Theorem 3.12, we can conclude that for the sequence $\{(x_k, y_k)\}$ generated by Algorithm 2, any cluster point of $\{(x_k, y_k)\}$ is a $\mathcal{D}_f$-stationary point of BLO, and the sequence $\{h(x_k, y_k)\}$ converges. $\square$

Similar to Corollary 4.11, the following corollary illustrates that when $(d_{x,k}, d_{y,k})$ in Step 5 of Algorithm 2 is generated by the randomized approaches mentioned in Remark 2.22, the yielded sequence $\{(x_k, y_k)\}$ converges to a first-order stationary point of BLO.

Corollary 4.15. *Suppose Assumption 4.8 holds with $\mathcal{D}_f = \partial f$, $\beta \geq \frac{4 Q_g M_f}{\mu^3}$, $\hat{\beta} \geq \beta \cdot \max\left\{ \frac{8 L_g^2}{\mu}, \frac{1}{4\mu}, \frac{\mu}{4} \right\}$, and the tolerance $\varepsilon_{1,k}$ and $\varepsilon_{2,k}$ satisfy $\lim_{k \to +\infty} \varepsilon_{1,k} = 0$, $\sum_{k=0}^{+\infty} \varepsilon_{2,k} \eta_k < +\infty$. Moreover, suppose $(d_{x,k}, d_{y,k})$ in Algorithm 2 is generated by one of the following schemes in each iteration k,*

- *$(d_{x,k}, d_{y,k}) \in \partial_{\eta_k} f(x_k, y_k - w_k)$;*

- *$(d_{x,k}, d_{y,k}) = \tilde{\partial}_{\eta_k} f(x_k, y_k - w_k; \zeta_{x,k}, \zeta_{y,k})$, where $(\zeta_{x,k}, \zeta_{y,k})$ is uniformly sampled over $\mathbb{B}_\delta(0)$ and independent of $\mathcal{F}_k$.*

Then almost surely, every limit point of $\{(x_k, y_k)\}$ generated by Algorithm 2 is a first-order stationary point of BLO and $\{h(x_k, y_k)\}$ converges.

Remark 4.16. *When f is assumed to be Lipschitz smooth over $\mathbb{R}^n \times \mathbb{R}^p$, Algorithm 2 coincides with (4.5), which can be regarded as the deterministic version of the TTSA algorithm in [27], and the SUSTAIN algorithm with $\eta_t^g = \eta_t^f = 1$ in [35, Equation (13)-(14)] (i.e. SUSTAIN algorithm without momentum accelerations). Therefore, the deterministic version of TTSA can be interpreted as an approximated gradient descent algorithm that minimizes CDB over $\mathbb{R}^n \times \mathbb{R}^p$, while SUSTAIN can be regarded as a momentum-accelerated (stochastic) gradient method for solving (CDB). Moreover, as illustrated in Algorithm 2, we can extend the deterministic version of these algorithms to handle nonsmooth bilevel optimization problems based on our proposed framework.*

4.4 An inexact subgradient method

Recently, another efficient single-loop approach named STABLE [11], is proposed for nonconvex-strongly-convex bilevel optimization problems where the objective functions are assumed to be Lipschitz smooth over $\mathbb{R}^n \times \mathbb{R}^p$. The deterministic version of STABLE algorithm employs the following updating schemes,

$$
\begin{aligned}
x_{k+1} &= x_k - \eta_k \left(\nabla_x f(x_k, y_k) - \nabla^2_{xy} g(x_k, y_k) \nabla^2_{yy} g(x_k, y_k)^{-1} \nabla_y f(x_k, y_k) \right), \\
y_{k+1} &= y_k - \tau_k \nabla_y g(x_k, y_k) + \eta_k \nabla^2_{yx} g(x_k, y_k) \nabla^2_{yy} g(x_k, y_k)^{-1} \Big(\nabla_x f(x_k, y_k) \\
&\quad - \nabla^2_{xy} g(x_k, y_k) \nabla^2_{yy} g(x_k, y_k)^{-1} \nabla_y f(x_k, y_k) \Big).
\end{aligned} \tag{4.7}
$$

Here the x-variable takes an approximated gradient descent step for $\Phi(x)$. However, the updating schemes of y-variable can be hard to understand by regarding STABLE algorithm as an approximated gradient descent algorithm for minimizing $\Phi(x)$.

In this subsection, we propose an inexact subgradient method based on our proposed framework with the following set-valued mapping $\hat{\mathcal{D}}_p(x, y)$,

$$
\hat{\mathcal{D}}_p(x, y) = W(x, y)^\top W(x, y) \mathcal{D}_f(x, \mathcal{A}(x, y)) + \begin{bmatrix} 0 \\ \beta \nabla_y g(x, y) \end{bmatrix}, \tag{4.8}
$$

where $W(x, y) \in \mathbb{R}^{n \times (n+p)}$ is defined by $W(x, y) = \left[I_n, -\nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y)^{-1} \right]$.

We first prove that $\hat{\mathcal{D}}_p(x, y)$ has compact and convex values, and satisfies the Assumption 4.1(f). Moreover, based on $\hat{\mathcal{D}}_p(x, y)$, we propose a subgradient method as presented in Algorithm 3 and show its global convergence properties directly from our proposed framework. A discussion on how to understand STABLE algorithm based on CDB is presented at the end of this subsection.

In the next two propositions, we establish some properties of $\hat{\mathcal{D}}_p(x, y)$.

Proposition 4.17. *Suppose $\beta \geq \frac{2M_f Q_g}{\mu^3}$. Then for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, $0 \in \mathcal{D}_h(x, y)$ if and only if $0 \in \hat{\mathcal{D}}_p(x, y)$.*

Proof. When $0 \in \mathcal{D}_h(x, y)$, it holds from Theorem 3.12 that $(x, y) \in \mathcal{M}$ and there exists $(d_x, d_y) \in \mathcal{D}_f(x, y)$ such that $0 = d_x - \nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y)^{-1} d_y \in W(x, y) \mathcal{D}_f(x, y)$. Therefore, we conclude that

$$
0 \in W(x, y)^\top W(x, y) \mathcal{D}_f(x, y) + \begin{bmatrix} 0 \\ \beta \nabla_y g(x, y) \end{bmatrix} = \hat{\mathcal{D}}_p(x, y).
$$

On the other hand, suppose $0 \in \hat{\mathcal{D}}_p(x, y)$, then there exists $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x, y))$ such that

$$
\begin{aligned}
&d_x - \nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y)^{-1} d_y = 0 \\
&- \nabla^2_{yy} g(x, y)^{-1} \nabla_{yx} g(x, y) \left(d_x - \nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y)^{-1} d_y \right) + \beta \nabla_y g(x, y) = 0.
\end{aligned}
$$

As a result, it holds that $\nabla_y g(x, y) = 0$ and hence $(x, y) \in \mathcal{M}$. Together with (4.7) and Definition 2.20, we obtain that (x, y) is a $\mathcal{D}_h$-stationary point of BLO. $\qquad \square$

Proposition 4.18. *Suppose* $\beta \geq \max\left\{\frac{8M_f Q_g}{\mu^3}, \frac{4M_f Q_g L_g}{\mu^{3.5}}\right\}$. *Then for any* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^p$ *and any* $w \in \hat{\mathcal{D}}_p(x,y)$, *it holds that*

$$\sup_{z \in \mathcal{D}_h(x,y)} \langle z, w \rangle \geq \min\left\{\frac{\mu^2}{4L_g^2}, \frac{\mu}{4}\right\} \|w_1 + w_2\|^2.$$

Proof. For any $(d_x, d_y) \in \mathcal{D}_f(x, \mathcal{A}(x,y))$, let z_1, z_2 and z_3 be defined as

$$z_1 = \begin{bmatrix} d_x - \nabla^2_{xy}g(x,y)\nabla^2_{yy}g(x,y)^{-1}d_y \\ 0 \end{bmatrix}, \quad z_2 = \begin{bmatrix} \nabla^3_{xyy}g(x,y)[\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)]\nabla^2_{yy}g(x,y)^{-1}d_y \\ \nabla^3_{yyy}g(x,y)[\nabla^2_{yy}g(x,y)^{-1}\nabla_y g(x,y)]\nabla^2_{yy}g(x,y)^{-1}d_y \end{bmatrix},$$

$$z_3 = \begin{bmatrix} \beta\nabla^2_{xy}g(x,y)\nabla_y g(x,y) \\ \beta\nabla^2_{yy}g(x,y)\nabla_y g(x,y) \end{bmatrix}, \quad w_1 = W(x,y)^\top W(x,y)\begin{bmatrix} d_x \\ d_y \end{bmatrix}, \quad w_2 = \begin{bmatrix} 0 \\ \beta\nabla_y g(x,y) \end{bmatrix}.$$

Then it holds that $z_1 + z_2 + z_3 \in \mathcal{D}_h(x,y)$, and $w_1 + w_2 \in \hat{\mathcal{D}}_p(x,y)$. Moreover, as $\|z_2\| \leq \frac{2M_f Q_g}{\mu^2}\|\nabla_y g(x,y)\|$, we obtain the following inequalities through simple calculations,

$$\langle z_1, w_1 \rangle = \|z_1\|^2, \quad \langle z_2, w_1 \rangle \geq -\left(\frac{2M_f Q_g}{\mu^2}\right)\|\nabla_y g(x,y)\|\,\|w_1\|, \quad \langle z_3, w_1 \rangle = 0,$$

$$\langle z_1, w_2 \rangle = 0, \quad \langle z_2, w_2 \rangle \geq -\frac{2M_f Q_g \beta}{\mu^2}\|\nabla_y g(x,y)\|^2, \quad \langle z_3, w_2 \rangle \geq \mu\beta^2\|\nabla_y g(x,y)\|^2.$$

By Cauchy's inequality, it holds from $\beta \geq \frac{4M_f Q_g L_g}{\mu^{3.5}}$ that

$$\frac{\mu^2}{4L_g^2}\|w_1\|^2 + \frac{\mu\beta^2}{4}\|\nabla_y g(x,y)\|^2 \geq \frac{2M_f Q_g}{\mu^2}\|\nabla_y g(x,y)\|\,\|w_1\|.$$

Therefore, we get

$$\langle z_1 + z_2 + z_3, w_1 + w_2 \rangle$$

$$\geq \|z_1\|^2 + \mu\beta^2\|\nabla_y g(x,y)\|^2 - \left(\frac{2M_f Q_g}{\mu^2}\right)\|\nabla_y g(x,y)\|\,\|w_1\| - \frac{2M_f Q_g \beta}{\mu^2}\|\nabla_y g(x,y)\|^2$$

$$\geq \frac{\mu^2}{2L_g^2}\|w_1\|^2 + \mu\beta^2\|\nabla_y g(x,y)\|^2 - \left(\frac{2M_f Q_g}{\mu^2}\right)\|\nabla_y g(x,y)\|\,\|w_1\| - \frac{\mu\beta^2}{4}\|\nabla_y g(x,y)\|^2$$

$$\geq \frac{\mu^2}{4L_g^2}\|w_1\|^2 + \frac{\mu\beta^2}{4}\|\nabla_y g(x,y)\|^2 \geq \min\left\{\frac{\mu^2}{4L_g^2}, \frac{\mu}{4}\right\}\|w_1 + w_2\|^2,$$

and this completes the proof. $\qquad\square$

With Propositions 4.17 and 4.18, we can now present our inexact subgradient method for solving CDB in Algorithm 3 and establish its convergence.

Theorem 4.19. *Suppose Assumption 4.8 holds,* $\beta \geq \max\left\{\frac{8M_f Q_g}{\mu^3}, \frac{4M_f Q_g L_g}{\mu^{3.5}}\right\}$ *and* $\lim_{k \to +\infty} \varepsilon_{1,k} = 0$. *Then every limit point of* $\{(x_k, y_k)\}$ *generated by Algorithm 3 is a* $\mathcal{D}_f$*-stationary points of BLO and* $\{h(x_k, y_k)\}$ *converges.*

Proof. Assumption 4.8(a) and 4.8(b) imply that Assumption 4.1(a) and 4.1(b) hold. Let $\hat{\mathcal{D}}_{temp}(x,y,z) := W(x,y)^\top W(x,y)\mathcal{D}_f(x,z) + \begin{bmatrix} 0 \\ \beta\nabla_y g(x,y) \end{bmatrix}$, $\tilde{d}_{x,k} = \mathbb{E}[d_{x,k}|\mathcal{F}_k]$, $\tilde{d}_{y,k} = \mathbb{E}[d_{y,k}|\mathcal{F}_k]$, and

$$u_{x,k} = \tilde{d}_{x,k} - \nabla^2_{xy}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}\tilde{d}_{y,k},$$

$$\hat{\chi}_{x,k} = (d_{x,k} - \tilde{d}_{x,k}) - \nabla^2_{xy}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}(d_{y,k} - \tilde{d}_{y,k}),$$

$$u_{y,k} = \beta\nabla_y g(x_k, y_k) - \nabla^2_{yx}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}u_{x,k},$$

$$\xi_{x,k} = -\nabla^2_{yx}g(x_k, y_k)\nabla^2_{yy}g(x_k, y_k)^{-1}\hat{\chi}_{x,k}, \quad \xi_{y,k} = 0.$$

Algorithm 3 Inexact subgradient method for solving CDB.

Require: Function f, g, initial point x_0, y_0.

1: **for** k = 1,2,... **do**

2:   Compute w_k by approximately evaluating $\nabla_{yy}^2 g(x_k, y_k)^{-1} \nabla_y g(x_k, y_k)$ such that $\left\| \nabla_{yy}^2 g(x_k, y_k) w_k - \nabla_y g(x_k, y_k) \right\| \le \varepsilon_{1,k}$.

3:   Choose $(d_{x,k}, d_{y,k})$ as an approximated evaluation of $\mathcal{D}_f(x_k, y_k - w_k)$.

4:   Compute $p_{x,k} = d_{x,k} - \nabla_{xy}^2 g(x_k, y_k) \nabla_{yy}^2 g(x_k, y_k)^{-1} d_{y,k}$.

5:   Update x_k and y_k by

$$x_{k+1} = x_k - \eta_k p_{x,k},$$
$$y_{k+1} = y_k - \eta_k \left(\beta \nabla_y g(x_k, y_k) - \nabla_{yx}^2 g(x_k, y_k) \nabla_{yy}^2 g(x_k, y_k)^{-1} p_{x,k} \right).$$

6: **end for**

7: Return x_k and y_k.

It is easy to verify the validity of Assumption 4.1(c) from Proposition 4.9. Moreover, $\hat{\mathcal{D}}_{temp}(x, y, \mathcal{A}(x,y)) = \hat{\mathcal{D}}_p(x,y)$ holds for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^p$, and

$$\lim_{k \to +\infty} \text{dist}\left((u_{x,k}, u_{y,k}), \hat{\mathcal{D}}_{temp}(x_k, y_k, y_k - w_k) \right) = 0.$$

Furthermore, notice that $\lim_{k \to +\infty} \varepsilon_{1,k} = 0$. Then for any subsequence $\{(x_{k_j}, y_{k_j})\}$ that converges to $\{(\tilde{x}, \tilde{y})\}$, it holds that $(x_k, y_k, y_k - w_k)$ converges to $(\tilde{x}, \tilde{y}, \mathcal{A}(\tilde{x}, \tilde{y}))$. Then Proposition 4.3 illustrates that $\lim_{N \to +\infty} \text{dist}\left(\frac{1}{N} \sum_{j=1}^{N} \begin{bmatrix} u_{x,k_j} \\ u_{y,k_j} \end{bmatrix}, \hat{\mathcal{D}}_p(\tilde{x}, \tilde{y}) \right) = 0$, which verifies the validity of Assumption 4.1(d). Additionally, Assumption 4.8(c) implies Assumption 4.1(e), and Proposition 4.18 guarantees the validity of Assumption 4.1(f). Then from Theorem 4.4, we can conclude that $\{h(x_k, y_k)\}$ converges and any cluster point of $\{(x_k, y_k)\}$ yielded by Algorithm 3 is a $\mathcal{D}_f$-stationary point of BLO. $\qquad\square$

Similar to Corollary 4.11 and Corollary 4.15, we have the following corollary illustrating that $\{(x_k, y_k)\}$ weakly converges to first-order stationary points of BLO when $(d_{x,k}, d_{y,k})$ is generated by some randomized approaches mentioned in Remark 2.22.

Corollary 4.20. *Suppose Assumption 4.8 holds with $\mathcal{D}_f = \partial f$, $\lim_{k \to +\infty} \varepsilon_{1,k} = 0$, and $\beta \ge \max\left\{ \frac{8 M_f Q_g}{\mu^3}, \frac{4 M_f Q_g L_g}{\mu^{3.5}} \right\}$. Moreover, suppose $(d_{x,k}, d_{y,k})$ in Algorithm 2 is generated by one of the following schemes in each iteration k,*

- *$(d_{x,k}, d_{y,k}) \in \partial_{\eta_k} f(x_k, y_k - w_k)$;*

- *$(d_{x,k}, d_{y,k}) = \tilde{\partial}_{\eta_k} f(x_k, y_k - w_k; \zeta_{x,k}, \zeta_{y,k})$, where $(\zeta_{x,k}, \zeta_{y,k})$ is uniformly sampled over $\mathbb{B}_\delta(0)$ and independent of $\mathcal{F}_k$.*

Then every limit point of $\{(x_k, y_k)\}$ generated by Algorithm 3 is a first-order stationary point of BLO and $\{h(x_k, y_k)\}$ converges.

Remark 4.21. *When f is assumed to be Lipschitz smooth over $\mathbb{R}^n \times \mathbb{R}^p$, Algorithm 3 coincides with the updating schemes (4.7) of the deterministic version of the STABLE algorithm. As illustrated in Proposition 4.18, the deterministic version of STABLE can be regarded as a descent algorithm for h in CDB in each iteration. This provides a clear understanding of the convergence properties of the STABLE algorithm, and demonstrates the efficiency of Algorithm 3. Moreover, according to Step 4 in Algorithm 3, the stepsizes η_k and τ_k in (4.7) should satisfy $\tau_k = \beta \eta_k$, which further explains the different theoretical bounds for η_k and τ_k suggested in [11, Theorem 2]. Therefore, we can conclude that CDB exhibits its ability in interpreting the STABLE algorithm and allows great flexibility in employing advanced theoretical analysis developed for unconstrained optimization.*

5 Conclusion

In this paper, we propose an unconstrained optimization problem CDB for the bilevel optimization problem BLO. We prove that under mild conditions, BLO and CDB have the same stationary points over $\mathbb{R}^n \times \mathbb{R}^p$ in the sense of both Clarke subdifferential and conservative field. Moreover, CDB has explicit formulation, and its function value and corresponding conservative field can be easily calculated in the presence of $\mathcal{D}_f$ and the derivatives of g. Therefore, various prior arts for unconstrained nonsmooth optimization can be directly employed to solve BLO through the unconstrained optimization problem CDB.

We propose a unified framework for developing subgradient methods, which further inspires several subgradient-based methods for solving BLO through CDB. In addition, we show that the proposed framework provides simple interpretations for some existing single-loop algorithms. Specifically, we show that the TTSA, SUSTAIN and STABLE algorithm can be regarded as approximated first-order methods for minimizing CDB when f is assumed to be Lipschitz smooth. Based on our proposed framework, we can straightforwardly extend these algorithms to nonsmooth cases and establish their global convergence properties.

Furthermore, suppose the objective functions f and g in BLO are expressed as the expectation of some random variables, i.e.

$$f(x,y) = \mathbb{E}_\xi[f_\xi(x,y)], \quad g(x,y) = \mathbb{E}_\theta[g_\theta(x,y)],$$

where $f_\xi : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and $g_\theta : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ are continuous functions that depend on the random variables ξ and θ, respectively. Then the corresponding CDB can be formulated as

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^p} \tilde{h}(x,y) := \mathbb{E}_\xi \left[f_\xi \left(x, y - \mathbb{E}_\theta \left[\nabla_{yy}^2 g_\theta(x,y) \right]^{-1} \mathbb{E}_\theta \left[\nabla_y g_\theta(x,y) \right] \right) \right] + \frac{\beta}{2} \left\| \mathbb{E}_\theta [\nabla_y g_\theta(x,y)] \right\|^2,$$

which can be categorized as a special case of unconstrained conditional stochastic optimization [30]. Therefore, we can directly apply some existing advanced approaches [37, 31, 12, 23] to solve CDB when f is Lipschitz smooth. Moreover, their theoretical properties, including global convergence, iteration complexity and sample complexity, directly follow the results from these existing works. We leave the discussion on how to design efficient algorithms to minimize $\tilde{h}$ over $\mathbb{R}^n \times \mathbb{R}^p$ for future investigation.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[3] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, pages 1–31, 2022.

[4] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

[5] Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 2020.

[7] James V Burke, Adrian S Lewis, and Michael L Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002.

[8] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *Siam Journal on Optimization*, 15(3):751–779, 2005.

[9] James V Burke, Frank E Curtis, Adrian S Lewis, Michael L Overton, and Lucas EA Simões. Gradient sampling methods for nonsmooth optimization. *Numerical Nonsmooth Optimization*, pages 201–225, 2020.

[10] Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial Newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31, 2021.

[11] Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

[12] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

[13] Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. SIAM, 1990.

[14] Benoıt Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

[15] Frank E Curtis and Michael L Overton. A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization. *SIAM Journal on Optimization*, 22(2):474–500, 2012.

[16] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[17] Amir Dembo. Probability theory: STAT310/MATH230; September 12, 2010. Available at https://web.stanford.edu/class/stats310a/lnotes.pdf. 2010.

[18] Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

[19] John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[21] 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.

[22] 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.

[23] Hongchang Gao and Heng Huang. Fast training method for stochastic compositional optimization problems. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[25] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[26] Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

[27] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[28] Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

[29] Xiaoyin Hu, Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. A constraint dissolving approach for nonsmooth optimization over the Stiefel manifold. *arXiv preprint arXiv:2205.10500*, 2022.

[30] Yifan Hu, Xin Chen, and Niao He. Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020.

[31] Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic gradient descent for conditional stochastic optimization. *arXiv preprint arXiv:2002.10790*, 2020.

[32] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523. Springer, 2011.

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

[34] Frank Jones. Honors calculus: MATH 221/222; August 26, 2004. Available at `https://www.owlnet.rice.edu/~fjones/`. 2004.

[35] Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[37] Xiangru Lian, Mengdi Wang, and Ji Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pages 1159–1167. PMLR, 2017.

[38] Renjie Liao, Yuwen Xiong, Ethan Fetaya, Lisa Zhang, KiJung Yoon, Xaq Pitkow, Raquel Urtasun, and Richard Zemel. Reviving and improving recurrent back-propagation. In *International Conference on Machine Learning*, pages 3082–3091. PMLR, 2018.

[39] Tianyi Lin, Zeyu Zheng, and Michael I Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *arXiv preprint arXiv:2209.05045*, 2022.

[40] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[41] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

[42] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA, 2013.

[43] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[44] Gregory M Moore. *Bilevel programming algorithms for machine learning model selection*. Rensselaer Polytechnic Institute, 2010.

[45] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[46] Jiawang Nie, Li Wang, Jane J Ye, and Suhan Zhong. A lagrange multiplier expression method for bilevel polynomial optimization. *SIAM Journal on Optimization*, 31(3):2368–2395, 2021.

[47] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

[48] Kegan GG Samuel and Marshall F Tappen. Learning optimized map estimates in continuously-valued mrf models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 477–484. IEEE, 2009.

[49] Marshall F Tappen, Kegan GG Samuel, Craig V Dean, and David M Lyle. The logistic random field—a convenient graphical model for learning parameters for mrf-based labeling. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[50] Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

[51] Alex J Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

[52] Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. Constraint dissolving approaches for Riemannian optimization. *arXiv preprint arXiv:2203.10319*, 2022.

[53] Mengwei Xu, Jane J Ye, and Liwei Zhang. Smoothing SQP methods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs. *SIAM Journal on Optimization*, 25(3):1388–1410, 2015.

[54] Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[55] Jane J Ye. Constraint qualifications and KKT conditions for bilevel programming problems. *Mathematics of Operations Research*, 31(4):811–824, 2006.

[56] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

[57] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Ali Jadbabaie, and Suvrit Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.