

SUPPLEMENT FOR “EFFICIENT BI-LEVEL OPTIMIZATION FOR NON-SMOOTH OPTIMIZATION”

Anonymous authors

Paper under double-blind review

1 CONVERGENCE ANALYSIS

In this section, we give the detailed convergence analysis. First, we give several definitions used in our analysis,

Definition 1 (Tangent cone Clarke (1990).) For a nonempty closed convex subset \mathcal{X} of \mathbb{R}^d , the tangent cone to \mathcal{W} at $\mathbf{w} \in \mathcal{W}$, denoted as $\mathcal{T}_{\mathcal{W}}(\mathbf{w})$, is the set consisting of all tangent vectors, where we call a vector $\mathbf{v} \in \mathbb{R}^d$ a tangent vector to \mathcal{W} at \mathbf{w} , if there are a sequence $\{\mathbf{w}^k\}$ in \mathcal{W} converging to \mathbf{w} and a sequence $\{\tau^k\}$ of positive numbers converging to 0 such that

$$\mathbf{v} = \lim_{k \rightarrow \infty} \frac{\mathbf{w}^k - \mathbf{w}}{\tau^k}. \quad (1)$$

Definition 2 (Normal cone Clarke (1990).) We define the normal cone to \mathcal{W} at \mathbf{w} by polarity with $\mathcal{T}_{\mathcal{W}}(\mathbf{w})$:

$$\mathcal{N}_{\mathcal{W}}(\mathbf{w}) = \{\boldsymbol{\xi} \in \mathcal{W}^* : \langle \boldsymbol{\xi}, \mathbf{v} \rangle \leq 0 \text{ for all } \mathbf{v} \in \mathcal{T}_{\mathcal{W}}(\mathbf{w})\}, \quad (2)$$

where \mathcal{W}^* denotes the dual space of \mathcal{W} .

Definition 3 (Clarke generalized directional derivative Clarke (1990).) For the Lipschitz continuous function ϕ over \mathcal{W} , the Clarke generalized directional derivative at $\bar{\mathbf{w}}$ is defined as

$$\phi^\circ(\bar{\mathbf{w}}; \mathbf{v}; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \bar{\mathbf{w}}, \mathbf{w} \in \mathcal{W} \\ t \downarrow 0, \mathbf{y} + t\mathbf{v} \in \mathcal{W}}} \frac{\phi(\mathbf{w} + t\mathbf{v}) - \phi(\mathbf{w})}{t} \quad (3)$$

Definition 4 (Interior set.) The interior of a subset \mathcal{W} , denoted by $\text{int}(\mathcal{W})$, is defined as the set of all interior points of \mathcal{W} , where the point \mathbf{w} is said to a interior point if there exists an open ball centered at \mathbf{w} which is completely contained in \mathcal{W} .

To solve the non-Lipschitz continuous term in our original problem, for a fixed point $\bar{\mathbf{w}} \in \mathcal{W}$, we find out which h_i is not Lipschitz continuous at $\mathbf{D}_i^T \bar{\mathbf{w}}$ and use a Lipschitz continuous function to replace it according to Bian & Chen (2017). Specifically, define the index set where h_i is not Lipschitz continuous at $\mathbf{D}_i^T \bar{\mathbf{w}}$ as follows

$$\mathcal{I}_{\bar{\mathbf{w}}} = \{i \in \{1, 2, \dots, n\} : h_i \text{ is not Lipschitz continuous at } \mathbf{D}_i^T \bar{\mathbf{w}}\}, \quad (4)$$

and define a new function

$$h_i^{\bar{\mathbf{w}}}(\mathbf{D}_i^T \mathbf{w}) := \begin{cases} h_i(\mathbf{D}_i^T \mathbf{w}) & i \notin \mathcal{I}_{\bar{\mathbf{w}}} \\ h_i(\mathbf{D}_i^T \bar{\mathbf{w}}) & i \in \mathcal{I}_{\bar{\mathbf{w}}} \end{cases}, \quad (5)$$

which is Lipschitz continuous at $\mathbf{D}_i^T \bar{\mathbf{w}}$, $i = 1, 2, \dots, n$. Then, we have a new function $h_{\bar{\mathbf{w}}}(\mathbf{w}) := (h_1^{\bar{\mathbf{w}}}(\mathbf{D}_1^T \mathbf{w}), h_2^{\bar{\mathbf{w}}}(\mathbf{D}_2^T \mathbf{w}), \dots, h_n^{\bar{\mathbf{w}}}(\mathbf{D}_n^T \mathbf{w}))$, which has the same value as $h(\mathbf{w})$ but different property. For convenience, we define $\phi_{\bar{\mathbf{w}}}(\mathbf{w}) = \varphi(h_{\bar{\mathbf{w}}}(\mathbf{w}))$ and $\phi(\mathbf{w}) = \varphi(h(\mathbf{w}))$. Besides, we define the vector set for the non-Lipschitz continuous index set as follows,

$$\mathcal{V}_{\bar{\mathbf{w}}} = \{\mathbf{v} : \mathbf{D}_i^T \mathbf{v} = \mathbf{0}, i \in \mathcal{I}_{\bar{\mathbf{w}}}\}, \quad (6)$$

which means that \mathbf{v} is perpendicular to all column vectors in \mathbf{D}_i , $i \in \mathcal{I}_{\bar{\mathbf{w}}}$. Besides, we define $\text{rint}(\mathcal{T}_{\mathcal{W}}(\mathbf{w})) = \text{int}(\mathcal{T}_{\mathcal{W}_1}(\mathbf{w})) \cap \mathcal{T}_{\mathcal{W}_2}(\mathbf{w})$.

1.1 STATIONARY POINTS OF THE SINGLE-LEVEL PROBLEM

Now we prove that the conditions in definition 1 is necessary conditions for the stationary point of a certain single-level problem.

First, we give the exact single-level problem of the original non-smooth problem. Recently, W. Bian and X. Chen (2017) proposed a necessary condition of the general non-smooth non-convex, even non-Lipschitz problem, which is shown as the lower-level problem in our original bi-level problem. Specifically, if \mathbf{w}^* is the local minimizer and $\text{rint}(\mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)) \cap \mathcal{V}_{\mathbf{w}^*} \neq \emptyset$, we have

$$\nabla_{\mathbf{w}}g(\mathbf{w}^*, \bar{\lambda})^T \mathbf{v} + \exp(\lambda_1)\phi^\circ(\mathbf{w}^*; \mathbf{v}; \mathcal{W}) \geq 0, \forall \mathbf{v} \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*} \quad (7)$$

where $\phi^\circ(\mathbf{w}^*; \mathbf{v}; \mathcal{W})$ denotes the Clarke generalized directional derivative of $\varphi(h(\mathbf{w}))$ at the local minimizer \mathbf{w}^* . Then, we can replace the original non-smooth bi-level problem with above condition and obtain the following single-level problem,

$$\min_{\mathbf{w}, \lambda} f(\mathbf{w}, \lambda) \quad (8)$$

$$s.t. c(\mathbf{w}, \lambda) = \nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v} + \exp(\lambda_1)\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) \geq 0 \quad \forall \mathbf{v} \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}) \cap \mathcal{V}_{\mathbf{w}}$$

if $\text{rint}(\mathcal{T}_{\mathcal{W}}(\mathbf{w})) \cap \mathcal{V}_{\mathbf{w}} \neq \emptyset$ and $\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})$ denotes the Clarke generalized directional derivative of $\varphi(h(\mathbf{w}))$ at point \mathbf{w} .

Based on our assumptions, we first give the following lemma.

Lemma 1 $c(\mathbf{w}, \lambda)$ is Lipschitz continuous on both \mathbf{w} and λ .

Proof 1 We have

$$\begin{aligned} & |c(\mathbf{w}, \lambda) - c(\mathbf{w}', \lambda)| \\ &= |(\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v} + \exp(\lambda_1)\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})) - (\nabla_{\mathbf{w}}g(\mathbf{w}', \bar{\lambda})^T \mathbf{v} + \exp(\lambda_1)\phi^\circ(\mathbf{w}'; \mathbf{v}; \mathcal{W}))| \\ &\leq |\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v} - \nabla_{\mathbf{w}}g(\mathbf{w}', \bar{\lambda})^T \mathbf{v}| + \exp(\lambda_1)|\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) - \phi^\circ(\mathbf{w}'; \mathbf{v}; \mathcal{W})| \end{aligned} \quad (9)$$

According to Proposition 2.1.1 in Clarke (1990) and g is L_g^w -Lipschitz continuous, where L_g^w is Lipschitz constant for \mathbf{w} , we have $\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v}$ is L_g^w -Lipschitz continuous. Besides, we have

$$\begin{aligned} & |\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) - \phi^\circ(\mathbf{w}'; \mathbf{v}; \mathcal{W})| \\ &= |\phi_{\bar{\mathbf{w}}}^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) - \phi_{\bar{\mathbf{w}}}^\circ(\mathbf{w}'; \mathbf{v}; \mathcal{W})|. \end{aligned} \quad (10)$$

Since $\phi_{\bar{\mathbf{w}}}$ is $L_{\phi}^{\bar{\mathbf{w}}}$ -Lipschitz continuous, where $L_{\phi}^{\bar{\mathbf{w}}}$ is the Lipschitz constant for \mathbf{w} , we have $\phi_{\bar{\mathbf{w}}}^\circ$ is $L_{\phi}^{\bar{\mathbf{w}}}$ -Lipschitz continuous according to Proposition 2.1.1 in Clarke (1990). Namely, we have

$$|\phi_{\bar{\mathbf{w}}}^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) - \phi_{\bar{\mathbf{w}}}^\circ(\mathbf{w}'; \mathbf{v}; \mathcal{W})| \leq L_{\phi}^{\bar{\mathbf{w}}} \|\mathbf{w} - \mathbf{w}'\|_2 \quad (11)$$

Therefore, we have

$$|c(\mathbf{w}, \lambda) - c(\mathbf{w}', \lambda)| \leq (L_g^w + L_{\phi_{\bar{\mathbf{w}}}}^w) \|\mathbf{w} - \mathbf{w}'\|_2, \quad (12)$$

which means that $c(\mathbf{w}, \lambda)$ is Lipschitz continuous on \mathbf{w} .

Then we prove c is Lipschitz continuous on λ . We have

$$\begin{aligned} & |c(\mathbf{w}, \lambda) - c(\mathbf{w}, \lambda')| \\ &= |(\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v} + \exp(\lambda_1)\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})) - (\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda}')^T \mathbf{v} + \exp(\lambda'_1)\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}))| \\ &= |\nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda})^T \mathbf{v} - \nabla_{\mathbf{w}}g(\mathbf{w}, \bar{\lambda}')^T \mathbf{v}| + |\exp(\lambda_1) - \exp(\lambda'_1)| \cdot |\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})| \end{aligned} \quad (13)$$

According to the Proposition 2.1.1 in Clarke (1990), we have

$$\begin{aligned} & |\phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})| \\ &= |\phi_{\bar{\mathbf{w}}}^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W})| \\ &\leq L_{\phi_{\bar{\mathbf{w}}}} \|\mathbf{v}\|_2 \end{aligned} \quad (14)$$

Thus, we obtain

$$|c(\mathbf{w}, \lambda) - c(\mathbf{w}, \lambda')| \leq C \|\lambda - \lambda'\|_2. \quad (15)$$

where C is a constant. That completes the proof.

Since the objective $f(\mathbf{w}, \boldsymbol{\lambda})$ and the constraint $c(\mathbf{w}, \boldsymbol{\lambda})$ are both Lipschitz continuous, we have the following lemma,

Lemma 2 (Necessary condition.) *Let $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ be a stationary point of problem (8). Then $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ together with $\xi^* \geq 0$ satisfy the following conditions,*

$$\nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_2 - (\mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + \exp(\lambda_1^*) \phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W})) \xi^* \geq 0 \quad (16)$$

$$\nabla_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_3 - (\bar{\mathbf{v}}_3^T \nabla_{\boldsymbol{\lambda}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + v_3^1 \exp(\lambda_1^*) \phi^{\circ}(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})) \xi^* \geq 0 \quad (17)$$

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^{\circ}(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0 \quad (18)$$

for all $\mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$, $\mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$, $\mathbf{v}_3 \in \mathcal{T}_{\mathcal{U}}(\boldsymbol{\lambda}^*)$ if $\text{rint}(\mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)) \cap \mathcal{V}_{\mathbf{w}^*} \neq \emptyset$, where $\mathbf{v}_3 = [v_3^1, \bar{\mathbf{v}}_3^T]^T$ and

$$\phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^{\circ}(\mathbf{w} + \mathbf{v}_2 s; \mathbf{v}_1; \mathcal{W}) - \phi^{\circ}(\mathbf{w}; \mathbf{v}_1; \mathcal{W})}{s} \quad (19)$$

Proof 2 Since $c(\mathbf{w}, \boldsymbol{\lambda})$ and $f(\mathbf{w}, \boldsymbol{\lambda})$ are Lipschitz continuous, based on the theorem 1 in Clarke (1976), $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ is the stationary point of the following Lagrange function of problem (8)

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}, \xi) = f(\mathbf{w}) - c(\mathbf{w}, \boldsymbol{\lambda})^T \xi \quad (20)$$

where $\xi \geq 0$. Thus, we have

$$\mathbf{0} \in \partial_{\mathbf{w}} \mathcal{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*) + \mathcal{N}_{\mathcal{W}}(\mathbf{w}^*) \quad (21)$$

$$\mathbf{0} \in \partial_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*) + \mathcal{N}_{\mathcal{U}}(\boldsymbol{\lambda}^*) \quad (22)$$

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \lambda_1^* \phi^{\circ}(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0 \quad (23)$$

where $\mathcal{N}_{\mathcal{W}}(\mathbf{w}^*)$ and $\mathcal{N}_{\mathcal{U}}(\boldsymbol{\lambda}^*)$ are the normal cones of \mathcal{W} and \mathcal{U} .

Since $c(\mathbf{w}, \boldsymbol{\lambda})$ and $f(\mathbf{w}, \boldsymbol{\lambda})$ are Lipschitz continuous, we have

$$\begin{aligned} & |\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}, \xi) - \mathcal{L}(\mathbf{w}', \boldsymbol{\lambda}, \xi)| \\ & \leq |f(\mathbf{w}, \boldsymbol{\lambda}) - f(\mathbf{w}', \boldsymbol{\lambda})| + \xi |c(\mathbf{w}, \boldsymbol{\lambda}) - c(\mathbf{w}', \boldsymbol{\lambda})| \\ & \leq (L_f^{\mathbf{w}} + L_g^{\mathbf{w}} + L_{\phi_{\bar{\mathbf{w}}}}^{\mathbf{w}}) \|\mathbf{w} - \mathbf{w}'\|_2, \end{aligned} \quad (24)$$

where $L_f^{\mathbf{w}}$ is the Lipschitz constant of f on \mathbf{w} . This means $\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}, \xi)$ is Lipschitz continuous. Thus, condition 21 is equivalent to

$$\mathcal{L}_{\mathbf{w}}^{\circ}(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*; \mathbf{v}_2; \mathcal{W}) \geq 0, \forall \mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \quad (25)$$

where $\mathcal{L}_{\mathbf{w}}^{\circ}(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*; \mathbf{v}_2; \mathcal{W})$ is the generalized clarke directional derivative at \mathbf{w}^* defined as follows

$$\mathcal{L}_{\mathbf{w}}^{\circ}(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*; \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\mathcal{L}(\mathbf{w} + s\mathbf{v}_2, \boldsymbol{\lambda}^*, \xi^*) - \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}^*, \xi^*)}{s}. \quad (26)$$

Then, for all $\mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$, we have,

$$\begin{aligned}
& \mathcal{L}_w^\circ(\mathbf{w}^*, \boldsymbol{\lambda}^*, \xi^*; \mathbf{v}_2; \mathcal{W}) \\
&= \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\mathcal{L}(\mathbf{w} + s\mathbf{v}_2, \boldsymbol{\lambda}^*, \xi^*) - \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}^*, \xi^*)}{s} \\
&= \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{f(\mathbf{w} + s\mathbf{v}_2, \boldsymbol{\lambda}^*) - f(\mathbf{w}, \boldsymbol{\lambda}^*, \xi^*)}{s} \\
&\quad - \xi^* \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{g(\mathbf{w} + s\mathbf{v}_1, \boldsymbol{\lambda}^*)\mathbf{v}_2 - g(\mathbf{w}, \boldsymbol{\lambda}^*)\mathbf{v}_1}{s} \\
&\quad - \exp(\lambda_1^*)\xi^* \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^\circ(\mathbf{w} + s\mathbf{v}_2; \mathbf{v}_1; \mathcal{W}) - \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})}{s} \\
&= \nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_2 - (\mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + \exp(\lambda_1^*) \phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W})) \xi^* \\
&\geq 0
\end{aligned} \tag{27}$$

where

$$\phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^\circ(\mathbf{w} + s\mathbf{v}_2; \mathbf{v}_1; \mathcal{W}) - \phi^\circ(\mathbf{w}; \mathbf{v}_1; \mathcal{W})}{s} \tag{28}$$

Using the same method, we can obtain condition (22) is equivalent to

$$\nabla_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_3 - (\bar{v}_3 \nabla_{\bar{\boldsymbol{\lambda}}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + v_3^1 \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})) \xi^* \geq 0 \tag{29}$$

for all $\mathbf{v}_3 \in \mathcal{T}_{\mathcal{U}}(\boldsymbol{\lambda})$.

Therefore, conditions (16)-(18) is equivalent to conditions (21)-(23).

Then, based on Lemma 2, we have the following corollary ,

Corollary 1 (Sufficient condition.) $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ is said to be a stationary point of problem (8), if it satisfies the following conditions for all $\mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$, $\mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$ and $\mathbf{v}_3 \in \mathcal{T}_{\mathcal{U}}(\boldsymbol{\lambda}^*)$

$$\nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_2 - (\mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + \exp(\lambda_1^*) \phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W})) \xi^* \geq 0 \tag{30}$$

$$\nabla_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_3 - (\bar{v}_3 \nabla_{\bar{\boldsymbol{\lambda}}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + v_3^1 \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})) \xi^* \geq 0 \tag{31}$$

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0 \tag{32}$$

where $\xi^* \geq 0$ and $\mathbf{v}_3 = [v_3^1, \bar{\mathbf{v}}_3^T]^T$ and

$$\phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^\circ(\mathbf{w} + \mathbf{v}_2 s; \mathbf{v}_1; \mathcal{W}) - \phi^\circ(\mathbf{w}; \mathbf{v}_1; \mathcal{W})}{s} \tag{33}$$

According to Dempe & Zemkoho (2020), if the lower problem is convex, then the bilevel optimization problem

$$\min_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}) \quad \text{s.t. } \mathbf{w}^* \in \arg \min_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}) + \exp(\lambda_1) \varphi(h(\mathbf{w})), \tag{34}$$

is equivalent to the single level problem

$$\min_{\boldsymbol{\lambda}, \mathbf{w}} f(\mathbf{w}, \boldsymbol{\lambda}) \quad \text{s.t. } \mathbf{0} \in \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}) + \exp(\lambda_1) \partial_{\mathbf{w}} \varphi(h(\mathbf{w})), \tag{35}$$

where $\partial_{\mathbf{w}} \varphi(h(\mathbf{w}))$ is the subgradient. Then, according to Bian & Chen (2017), we have

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0, \tag{36}$$

is equivalent to

$$\mathbf{0} \in \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}) + \exp(\lambda_1) \partial_{\mathbf{w}} \varphi(h(\mathbf{w})). \quad (37)$$

This means that if a point is the stationary point of the problem

$$\min_{\mathbf{w} \in \mathcal{W}, \boldsymbol{\lambda} \in \mathcal{U}} f(\mathbf{w}, \boldsymbol{\lambda}) \quad (38)$$

$$s.t. \ c(\mathbf{w}, \boldsymbol{\lambda}) = \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}})^T \mathbf{v} + \exp(\lambda_1) \phi^\circ(\mathbf{w}; \mathbf{v}; \mathcal{W}) \geq 0 \ \forall \mathbf{v} \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}) \cap \mathcal{V}_{\mathbf{w}}, \quad (39)$$

then, it is a stationary point of the above problem 35. Thus, it means that such a point is the stationary point of the original nonsmooth bilevel problem if the lower level problem is convex and nonsmooth. In addition, if the lower-level problem is nonconvex, the conditions 16-18 are the necessary conditions of the original bilevel problem. Therefore, we have the following definition,

Definition 5 $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ is said to be a stationary point of problem (??), if it satisfies the following conditions for all $\mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$, $\mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$ and $\mathbf{v}_3 \in \mathcal{T}_{\mathcal{U}}(\boldsymbol{\lambda}^*)$, where $\mathcal{U} = \mathbb{R}^m$ and the lower-level problem is convex,

$$\nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_2 - (\mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + \exp(\lambda_1^*) \phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W})) \xi^* \geq 0 \quad (40)$$

$$\nabla_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_3 - (\bar{\mathbf{v}}_3^T \nabla_{\boldsymbol{\lambda}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 + \mathbf{v}_3^T \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})) \xi^* \geq 0 \quad (41)$$

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0 \quad (42)$$

where $\xi^* \geq 0$, $\mathbf{v}_3 = [v_3^1, \bar{\mathbf{v}}_3^T]^T$, $\phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^\circ(\mathbf{w} + \mathbf{v}_2 s; \mathbf{v}_1; \mathcal{W}) - \phi^\circ(\mathbf{w}; \mathbf{v}_1; \mathcal{W})}{s}$. and $\phi^\circ(\mathbf{w}; \mathbf{v}_1; \mathcal{W}) = \limsup_{\substack{\mathbf{w}' \mapsto \mathbf{w}, \mathbf{w}' \in \mathcal{W} \\ t \downarrow 0, \mathbf{w}' + t\mathbf{v} \in \mathcal{W}}} \frac{\varphi(h(\mathbf{w}' + t\mathbf{v})) - \varphi(h(\mathbf{w}'))}{t}$ denotes the Clarke generalized directional derivative of $\varphi(h(\mathbf{w}))$ at point \mathbf{w} . Note if the lower-level problem is nonconvex, conditions 16-18 is the necessary conditions of the original nonsmooth, perhaps non-Lipschitz bilevel problem.

1.2 CONVERGE TO THE STATIONARY POINT

In this subsection, we show our method will finally converge to the stationary point of problem (8).

Theorem 1 Suppose $\{\epsilon_{i,k}\}_{k=1}^\infty$ ($i = 1, 2, 3$) are positive and convergent ($\lim_{k \rightarrow \infty} \epsilon_{i,k} = 0$) sequences, $\{\mu^k\}_{k=1}^\infty$ is a positive and convergent ($\lim_{k \rightarrow \infty} \mu^k = 0$) sequence, and β^k is increasing and divergent ($\beta^1 < \beta^2 < \dots$). Then any limit point of the sequence points generated by SPNBO satisfies the conditions (16)-(18).

Proof 3 In our method, we solve the following smoothed single level problem for each given μ^k

$$\min_{\mathbf{w} \in \mathcal{W}, \boldsymbol{\lambda} \in \mathcal{U}} f(\mathbf{w}, \boldsymbol{\lambda}) \quad (43)$$

$$s.t. \ c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda}) := \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}) + \exp(\lambda_1) \nabla_{\mathbf{w}} \varphi(\tilde{h}(\mathbf{w}, \mu^k)) = \mathbf{0},$$

where $\nabla_{\mathbf{w}} \varphi(\tilde{h}(\mathbf{w}, \mu^k)) = \varphi'(z)_{h(\mathbf{w}, \mu^k)} \nabla_{\mathbf{w}} h(\mathbf{w}, \mu^k)$ and μ^k is the smoothing parameter.

Its corresponding augmented Lagrangian function can be rewritten as follows,

$$\hat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \beta, \mu) = f(\mathbf{w}, \boldsymbol{\lambda}) + \frac{1}{d} \sum_{j=1}^d \left(\alpha_j c_j^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda}) + \frac{\beta^k}{2} c_j^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})^2 \right) \quad (44)$$

where $\boldsymbol{\alpha}$ denotes the Lagrangian multiplier, $\beta^k > 0$ denotes the penalty parameter and α_j and $c_j^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})$ is the j -th element of $\boldsymbol{\alpha}$ and $c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})$.

Based on the tolerance condition, we have

$$\frac{1}{d} \|c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})\|_2^2 \leq \epsilon_{3,k}^2 \quad (45)$$

Multiplying $\|\mathbf{v}_1\|_2^2$, where $\mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$, on both sides of above equality, we have

$$\|c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})\|_2 \|\mathbf{v}_1\|_2 \leq \epsilon_{3,k} \sqrt{d} \|\mathbf{v}_1\|_2 \quad (46)$$

According to $|ab| \leq \|a\|_2 \|b\|_2$, we have

$$|c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})^T \mathbf{v}_1| \leq \epsilon_{3,k} \sqrt{d} \|\mathbf{v}_1\|_2. \quad (47)$$

Thus, we have

$$c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})^T \mathbf{v}_1 \geq -\epsilon_{3,k} \sqrt{d} \|\mathbf{v}_1\|_2. \quad (48)$$

Then, taking the limit on both sides (i.e., $k \rightarrow \infty$), we have

$$\lim_{k \rightarrow \infty} c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})^T \mathbf{v}_1 \geq 0 \quad (49)$$

Then, according to the Theorem 2 in Bian & Chen (2017), we have

$$\lim_{k \rightarrow \infty} c^{\mu^k}(\mathbf{w}, \boldsymbol{\lambda})^T \mathbf{v}_1 = \nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \quad (50)$$

Thus, we have

$$\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \geq 0 \quad (51)$$

Let $\hat{\boldsymbol{\alpha}}^k = -\boldsymbol{\alpha}^k - \beta c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k)$. Then, we have

$$|c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \hat{\boldsymbol{\alpha}}^k| \leq \epsilon_{2,k}^2 \quad (52)$$

Assume the limitation of c^{μ^k} exists. Taking the limit on both sides, we have

$$\lim_{k \rightarrow \infty} |c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \hat{\boldsymbol{\alpha}}^k| \leq 0 \quad (53)$$

It means that

$$\lim_{k \rightarrow \infty} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \hat{\boldsymbol{\alpha}}^k = 0. \quad (54)$$

Therefore, we have $\hat{\boldsymbol{\alpha}}^* \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$. If $\hat{\boldsymbol{\alpha}}^* \notin \mathcal{V}_{\mathbf{w}^*}$, the limitation doesn't exist, since ϕ is not lipschitz continuous. This means that we have $\hat{\boldsymbol{\alpha}}^* \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$. Let $\hat{\boldsymbol{\alpha}}^* = \xi^* \mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$ and $\xi^* \geq 0$, such that

$$\begin{aligned} & \lim_{k \rightarrow \infty} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \hat{\boldsymbol{\alpha}}^k \\ &= c^{\mu^*}(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \xi^* \mathbf{v}_1 \\ &= \xi^* (\nabla_{\mathbf{w}} g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*)^T \mathbf{v}_1 + \exp(\lambda_1^*) \phi^\circ(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W})) \\ &= 0 \end{aligned} \quad (55)$$

We also have

$$\|\nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \nabla_{\mathbf{w}} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k\|_2^2 \leq \epsilon_{1,k}^2 \quad (56)$$

Assume that we have a vector $\mathbf{v}_2 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*)$. Multiplying $\|\mathbf{v}_2\|_2^2$ on the both side of the above inequality, we have

$$\|\nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \nabla_{\mathbf{w}} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k\|_2 \|\mathbf{v}_2\|_2 \leq \epsilon_{1,k} \|\mathbf{v}_2\|_2. \quad (57)$$

According to $|\langle \mathbf{a}, \mathbf{b} \rangle| \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$, we have

$$|\langle \nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \nabla_{\mathbf{w}} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k, \mathbf{v}_2 \rangle| \leq \epsilon_{1,k} \|\mathbf{v}_2\|_2 \quad (58)$$

Obviously, we have

$$\langle \nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \nabla_{\mathbf{w}} c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k, \mathbf{v}_2 \rangle \geq -\epsilon_{1,k} \|\mathbf{v}_2\|_2 \quad (59)$$

Besides, we have

$$\begin{aligned} & \nabla_{\mathbf{w}} \left(f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k \right) \\ &= \nabla_{\mathbf{w}} \left(f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \langle \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}), \hat{\boldsymbol{\alpha}}^k \rangle - \exp(\lambda_1) \langle \nabla \psi(\mathbf{z})_{\mathbf{z}=\tilde{h}(\mathbf{w}^k, \mu^k)}, \nabla_{\mathbf{w}} \tilde{h}(\mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k \rangle \right) \end{aligned} \quad (60)$$

where

$$\begin{aligned} & \nabla_{\mathbf{w}} \tilde{h}(\mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k \\ &= \left(\nabla_{\mathbf{w}} \tilde{h}_1(\mathbf{D}_1 \mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k, \nabla_{\mathbf{w}} \tilde{h}_2(\mathbf{D}_2 \mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k, \dots, \nabla_{\mathbf{w}} \tilde{h}_n(\mathbf{D}_n \mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k \right)^T \end{aligned} \quad (61)$$

Let $\hat{\boldsymbol{\alpha}}^k = \xi^k \mathbf{v}_1 \in \mathcal{V}_{\mathbf{w}^k}$ and $\xi \geq 0$. For $i \in \mathcal{I}_{\mathbf{w}^k}$, we obtain $\mathbf{D}_i^T \hat{\boldsymbol{\alpha}}^k = \mathbf{0}$, then $\nabla_{\mathbf{w}} \tilde{h}_i(\mathbf{D}_i \mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k = \nabla_{\mathbf{z}} \tilde{h}_i(\mathbf{z}, \mu^k)^T_{\mathbf{z}=\mathbf{D}_i^T \mathbf{w}^k} \mathbf{D}_i^T \hat{\boldsymbol{\alpha}}^k = \mathbf{0}$. Besides, we define the smoothing function

$$\tilde{h}_i^{\bar{\mathbf{w}}}(\mathbf{D}_i \mathbf{w}, \mu) := \begin{cases} \tilde{h}_i(\mathbf{D}_i^T \mathbf{w}, \mu) & i \notin \mathcal{I}_{\bar{\mathbf{w}}} \\ \tilde{h}_i(\mathbf{D}_i^T \bar{\mathbf{w}}, \mu) & i \in \mathcal{I}_{\bar{\mathbf{w}}} \end{cases} \quad (62)$$

and $\tilde{h}_{\bar{\mathbf{w}}}(\mathbf{w}, \mu) = \left(\tilde{h}_1^{\bar{\mathbf{w}}}(\mathbf{D}_1 \mathbf{w}, \mu), \tilde{h}_2^{\bar{\mathbf{w}}}(\mathbf{D}_2 \mathbf{w}, \mu), \dots, \tilde{h}_n^{\bar{\mathbf{w}}}(\mathbf{D}_n \mathbf{w}, \mu) \right)^T$. We can obtain the smoothing function ϕ and $\phi_{\bar{\mathbf{w}}}$ as $\tilde{\phi}(\mathbf{w}^k, \mu^k) = \varphi(h_{\bar{\mathbf{w}}}(\mathbf{w}, \mu^k))$ and $\tilde{\phi}_{\bar{\mathbf{w}}}(\mathbf{w}^k, \mu^k) = \varphi(h_{\bar{\mathbf{w}}}(\mathbf{w}, \mu^k))$. Then, we have

$$\nabla_{\mathbf{w}} \tilde{h}(\mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k = \nabla_{\mathbf{w}} \tilde{h}_{\bar{\mathbf{w}}}(\mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k \quad (63)$$

Thus, coming back to (50), we obtain

$$\begin{aligned} & f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k \\ &= f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \langle \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}), \hat{\boldsymbol{\alpha}}^k \rangle - \exp(\lambda_1) \langle \nabla \psi(\mathbf{z})_{\mathbf{z}=\tilde{h}(\mathbf{w}^k, \mu^k)}, \nabla_{\mathbf{w}} \tilde{h}_{\bar{\mathbf{w}}}(\mathbf{w}^k, \mu^k)^T \hat{\boldsymbol{\alpha}}^k \rangle \\ &= f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \langle \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}), \xi^k \mathbf{v}_1 \rangle - \exp(\lambda_1) \langle \nabla \psi(\mathbf{z})_{\mathbf{z}=\tilde{h}(\mathbf{w}^k, \mu^k)}, \nabla_{\mathbf{w}} \tilde{h}_{\bar{\mathbf{w}}}(\mathbf{w}^k, \mu^k)^T \xi^k \mathbf{v}_1 \rangle \\ &= f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - \langle \nabla_{\mathbf{w}} g(\mathbf{w}, \bar{\boldsymbol{\lambda}}), \xi^k \mathbf{v}_1 \rangle - \exp(\lambda_1) \xi^k \tilde{\phi}_{\bar{\mathbf{w}}}^{\circ}(\mathbf{w}^k, \mu^k; \mathbf{v}_1; \mathcal{W}) \end{aligned}$$

where $\tilde{\phi}_{\bar{\mathbf{w}}}^{\circ}(\mathbf{w}^k, \mu^k; \mathbf{v}_1; \mathcal{W})$ is the directional derivative of $\tilde{\phi}_{\bar{\mathbf{w}}}(\mathbf{w}^k, \mu^k)$. Then, we can calculate directional derivative on \mathbf{v}_2 as follows,

$$\begin{aligned} & \left\langle \nabla_{\mathbf{w}} \left(f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k \right), \mathbf{v}_2 \right\rangle \\ &= \nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \mathbf{v}_2 - \xi^k \mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^k, \bar{\boldsymbol{\lambda}}^k) \mathbf{v}_1 \\ & \quad - \lambda_1 \xi^k \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^k, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\tilde{\phi}_{\bar{\mathbf{w}}}^{\circ}(\mathbf{w} + \mathbf{v}_2 s, \mu^k; \mathbf{v}_1; \mathcal{W}) - \tilde{\phi}_{\bar{\mathbf{w}}}^{\circ}(\mathbf{w}, \mu^k; \mathbf{v}_1; \mathcal{W})}{s} \\ &= \nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \mathbf{v}_2 - \xi^k \mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^k, \bar{\boldsymbol{\lambda}}^k) \mathbf{v}_1 \\ & \quad - \lambda_1 \xi^k \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^k, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\tilde{\phi}^{\circ}(\mathbf{w} + \mathbf{v}_2 s, \mu^k; \mathbf{v}_1; \mathcal{W}) - \tilde{\phi}^{\circ}(\mathbf{w}, \mu^k; \mathbf{v}_1; \mathcal{W})}{s} \end{aligned} \quad (64)$$

Let $k \rightarrow \infty$ and $\mathbf{v}_1 \in \mathcal{T}_{\mathcal{W}}(\mathbf{w}^*) \cap \mathcal{V}_{\mathbf{w}^*}$, we obtain

$$\begin{aligned} & \lim_{k \rightarrow \infty} \left\langle \nabla_{\mathbf{w}} \left(f(\mathbf{w}^k, \boldsymbol{\lambda}^k) - c^{\mu^k}(\mathbf{w}^k, \boldsymbol{\lambda}^k) \hat{\boldsymbol{\alpha}}^k \right), \mathbf{v}_2 \right\rangle \\ &= \lim_{k \rightarrow \infty} \left(\nabla_{\mathbf{w}} f(\mathbf{w}^k, \boldsymbol{\lambda}^k)^T \mathbf{v}_2 - \xi^k \mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^k, \bar{\boldsymbol{\lambda}}^k) \mathbf{v}_1 \right. \\ & \quad \left. - \exp(\lambda_1) \xi^k \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^k, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\tilde{\phi}^{\circ}(\mathbf{w} + \mathbf{v}_2 s, \mu^k; \mathbf{v}_1; \mathcal{W}) - \tilde{\phi}^{\circ}(\mathbf{w}, \mu^k; \mathbf{v}_1; \mathcal{W})}{s} \right) \\ &= \nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*)^T \mathbf{v}_2 - \xi^* \mathbf{v}_2^T \nabla_{\mathbf{w}}^2 g(\mathbf{w}^*, \bar{\boldsymbol{\lambda}}^*) \mathbf{v}_1 - \exp(\lambda_1) \xi^* \phi^{\circ\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) \\ &\geq 0 \end{aligned} \quad (65)$$

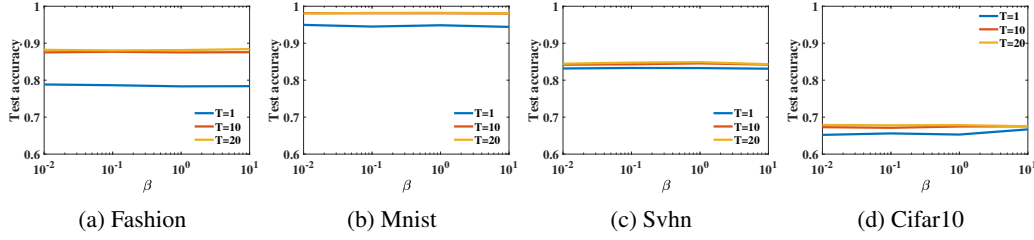


Figure 1: Test accuracy of data re-weight on dataset with different β . (Note the inner iteration number T is fixed.)

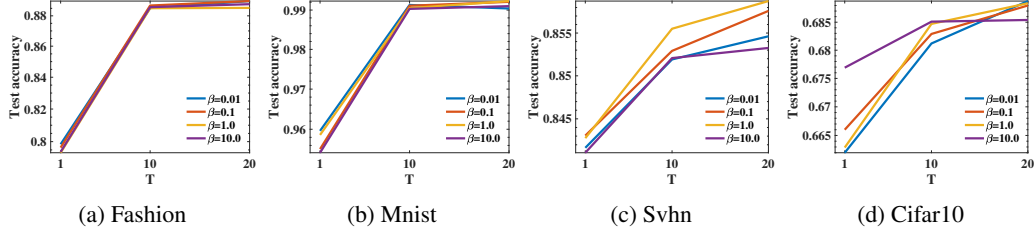


Figure 2: Test accuracy of data re-weight with different T . (Note the penalty parameter β is fixed.)

where

$$\phi^{\circ}(\mathbf{w}^*; \mathbf{v}_1, \mathbf{v}_2; \mathcal{W}) = \limsup_{\substack{\mathbf{w} \mapsto \mathbf{w}^*, \mathbf{w} \in \mathcal{W} \\ s \downarrow 0, \mathbf{w} + s\mathbf{v}_2 \in \mathcal{W}}} \frac{\phi^{\circ}(\mathbf{w} + \mathbf{v}_2 s; \mathbf{v}_1; \mathcal{W}) - \phi^{\circ}(\mathbf{w}; \mathbf{v}_1; \mathcal{W})}{s} \quad (66)$$

By using the same method, we can obtain

$$\nabla_{\lambda} f(\mathbf{w}^*, \lambda^*)^T \mathbf{v}_3 - \left(\bar{v}_3 \nabla_{\mathbf{w} \bar{\lambda}}^2 g(\mathbf{w}^*, \bar{\lambda}^*) \mathbf{v}_1 + v_3^1 e^{\lambda^*} \phi^{\circ}(\mathbf{w}^*; \mathbf{v}_1; \mathcal{W}) \right) \xi^* \geq 0 \quad (67)$$

for all $\mathbf{v}_3 \in \mathcal{T}_{\mathcal{U}}(\lambda)$.

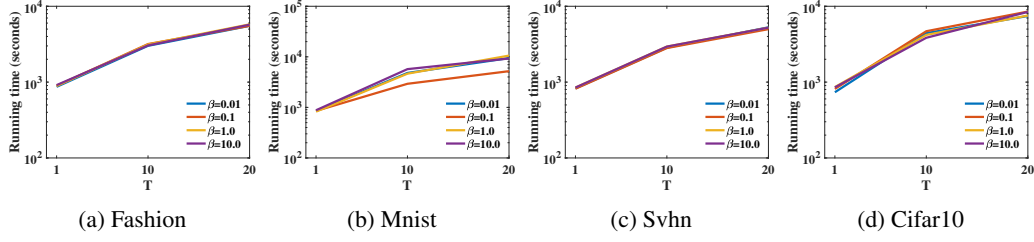
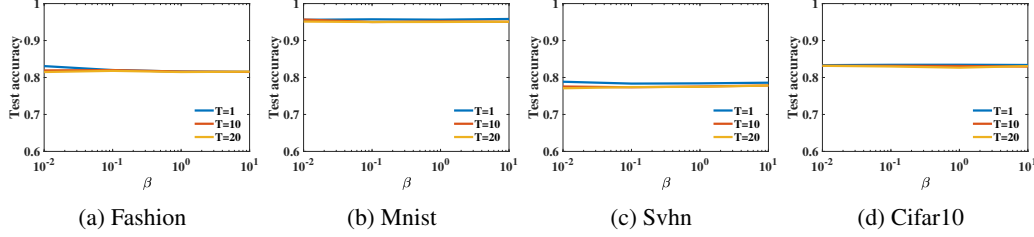
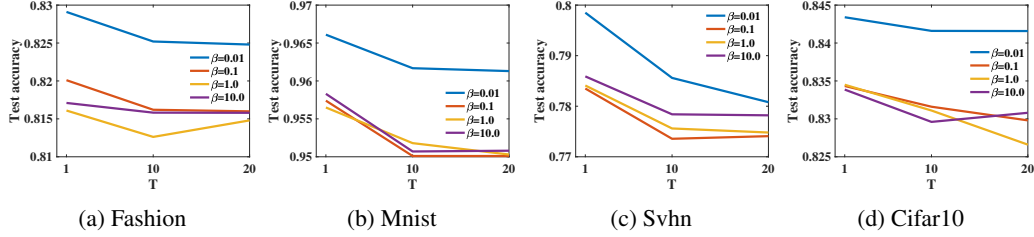
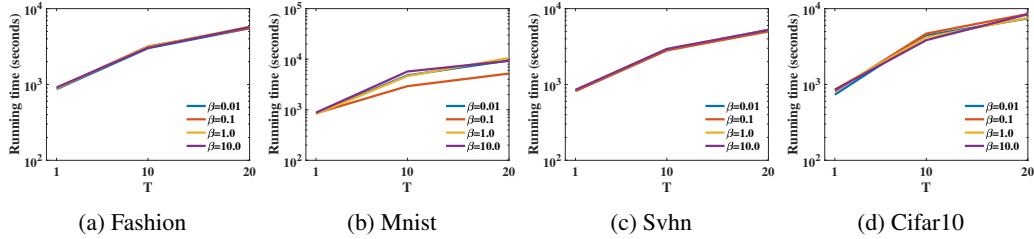
That completes the proof.

This means that our method can converge to the stationary point the single-level problem 8. Meanwhile, it also means that our method will finally converge to the stationary point of the original bi-level problem.

2 IMPACT OF PENALTY PARAMETER AND INNER ITERATION NUMBER

In this section, we evaluate the impact of different initial values for penalty parameter β and the inner iteration number T in three applications. In data re-weight and training data poisoning, we use the deep neural network which has 3 convolution-maxpooling-relu layers and 3 dense layers. We randomly sample 3 layers to update \mathbf{w} and λ and data batch is fixed at 128. We run our method for 10 epochs. For meta-learning, we run our method for 1000 iterations. We present all the results in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9. From Figure 1, Figure 4 and Figure 7, we can find that the results do not change much when β is different and T is fixed. This means that our method is not insensitive to the initial value of β . We also present the running time and test accuracy in Figure 2, Figure 3, Figure 5, Figure 6, Figure 8 and Figure 9. In most cases, using a larger inner iteration number, we will get a better result. However, it needs a long running time for the large inner iteration number. In addition, when the number of inner iterations reaches a certain level, the accuracy improvement brought by increasing the number of iterations decreases.

Here we also discuss the effect of the different initial values of η_{λ} in three applications. In data re-weight and training data poisoning attack, we use the deep neural network which has 3 convolution-maxpooling-relu layers and 3 dense layers. We randomly sample 3 layers and 64 data samples to

Figure 3: Running time of data re-weight with different T . (Note the penalty parameter β is fixed.)Figure 4: Test accuracy of training data poisoning with different β . (Note the inner iteration number T is fixed.)Figure 5: Test accuracy of training data poisoning with different T . (Note the penalty parameter β is fixed.)Figure 6: Running time of training data poisoning with different T . (Note the penalty parameter β is fixed.)

calculate the stochastic gradient. We run our method for 20 epochs. For meta-learning, we run our method for 1000 iterations. We set $\mu = 0.001$, $\eta_w = 0.0001$, $T = 10$ and $\beta = 0.01$. The results of using different η_λ are given in the following Tables 1, 2 and 3. Here we discuss the effect of different initial values of μ in three applications. We set $\eta_\lambda = 0.1$, $\eta_w = 0.0001$, $T = 10$ and $\beta = 0.01$. The results of using different μ are given in the following Tables 4, 6 and 5.

From our results, we can find that our method is not sensitive to the initial values of the learning rate η_λ or the smoothing parameter μ .

Table 1: Results of all the methods in data-reweighgt with different η_λ .

Name	0.1	0.01	0.001
Cifar10	0.683	0.682	0.685
Svhn	0.844	0.845	0.842
Mnist	0.984	0.983	0.984
Fashion	0.881	0.879	0.878

Table 2: Results of all the methods in attack with different η_λ .

Name	0.1	0.01	0.001
Cifar10	0.494	0.503	0.493
Svhn	0.777	0.773	0.774
Mnist	0.959	0.961	0.959
Fashion	0.829	0.828	0.831

Table 3: Results of all the methods in data-reweighgt with different η_λ .

Name	Setting	0.1	0.01	0.001
Omnglot	5way-1shot	0.961	0.963	0.963
Miniimagenet	5way-1shot	0.448	0.447	0.448

Table 4: Results of all the methods in data reweight with different μ .

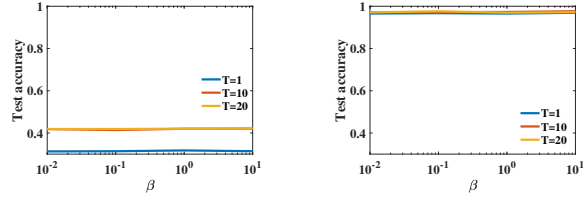
Name	0.01	0.001	0.0001
Cifar10	0.683	0.683	0.685
Svhn	0.842	0.844	8.842
Mnist	0.983	0.984	0.982
Fashion	0.878	0.881	0.878

Table 5: Results of all the methods in data meta-learning with different μ .

Name	Setting	0.01	0.001	0.0001
Omnglot	5way-1shot	0.962	0.961	0.963
Miniimagenet	5way-1shot	0.449	0.448	0.448

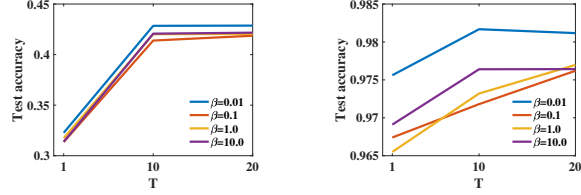
Table 6: Results of all the methods in data attack with different μ .

Name	0.01	0.001	0.0001
Cifar10	0.501	0.494	0.503
Svhn	0.788	0.777	0.795
Mnist	0.958	0.959	0.958
Fashion	0.827	0.829	0.829



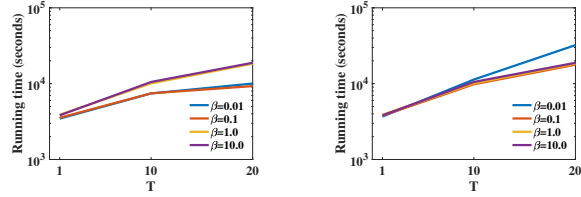
(a) Miniimagenet 5-way 1-shot (b) Omniglot 5-way 1-shot

Figure 7: Test accuracy of meta-learning with different β . (Note the inner iteration number T is fixed.)



(a) Miniimagenet 5-way 1-shot (b) Omniglot 5-way 1-shot

Figure 8: Test accuracy of meta-learning with different T . (Note the penalty parameter β is fixed.)



(a) Miniimagenet 5-way 1-shot (b) Omniglot 5-way 1-shot

Figure 9: Running time of meta-learning with different T . (Note the penalty parameter β is fixed.)

REFERENCES

- Wei Bian and Xiaojun Chen. Optimality and complexity for constrained optimization problems with nonconvex regularization. *Mathematics of Operations Research*, 42(4):1063–1084, 2017.
- Frank H Clarke. A new approach to lagrange multipliers. *Mathematics of Operations Research*, 1(2): 165–174, 1976.
- Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.
- Stephan Dempe and Alain Zemkoho. *Bilevel optimization*. Springer, 2020.