Supplementary Materials for "A Value-Function-based Interior-point Method for Non-convex Bilevel Optimization"

Risheng Liu [1 2] **Xuan Liu** [1 2] **Xiaoming Yuan** [3] **Shangzhi Zeng** [3] **Jin Zhang** [4]

This supplementary material is organized as follows. In Section A, we present the detailed proofs of all the theoretical results. Section B presents the further details of the experiments in Section 6 and additional results.

A. Detailed Proofs

We first recall an equivalent definition of epiconvergence given in (Bonnans & Shapiro, 2013)[page 41].

Definition 1. *$\varphi_k \xrightarrow{e} \varphi$ iff for all $\mathbf{x} \in \mathbb{R}^m$ the following two conditions hold:*

1. *For any sequence $\{\mathbf{x}_k\}$ converging to $\mathbf{x}$,*
$$\liminf_{k\to\infty} \varphi_k(\mathbf{x}_k) \geq \varphi(\mathbf{x}). \tag{1}$$
2. *There is a sequence $\{\mathbf{x}_k\}$ converging to $\mathbf{x}$ such that*
$$\limsup_{k\to\infty} \varphi_k(\mathbf{x}_k) \leq \varphi(\mathbf{x}). \tag{2}$$

For a given function $f(\mathbf{x}, \mathbf{y})$, we state the property that it is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$ in the following definition.

Definition 2. *Given a function $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, if for a point $\bar{\mathbf{x}} \in \mathcal{X} \subseteq \mathbb{R}^m$, for any $c \in \mathbb{R}$, there exist $\delta > 0$ along with a bounded set $\mathcal{B} \in \mathbb{R}^m$, such that*
$$\{\mathbf{y} \in \mathbb{R}^n | f(\mathbf{x}, \mathbf{y}) \leq c\} \subseteq \mathcal{B}, \ \ \forall \mathbf{x} \in \mathcal{B}_\delta(\bar{\mathbf{x}}) \cap \mathcal{X}, \tag{3}$$
then we call $f(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\bar{\mathbf{x}} \in \mathcal{X}$. If the above property holds for each $\bar{\mathbf{x}} \in \mathcal{X}$, we further call $f(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$.

The convergence results are given under following standing assumptions:

Assumption 1. *We take the following as our blanket assumption*

1. *$\mathcal{S}(\mathbf{x})$ is nonempty for $\mathbf{x} \in \mathcal{X}$.*
2. *Both $F(\mathbf{x}, \mathbf{y})$ and $f(\mathbf{x}, \mathbf{y})$ are jointly continuous and continuously differentiable.*
3. *Either $F(\mathbf{x}, \mathbf{y})$ or $f(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$.*

Proposition 1. *Suppose $F(\mathbf{x}, \mathbf{y})$ and $f(\mathbf{x}, \mathbf{y})$ are continuously differentiable, given $\mathbf{x} \in \mathcal{X}$ and $\mu, \theta, \tau > 0$, when*
$$\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} F(\mathbf{x}, \mathbf{y}) + \frac{\theta}{2}\|\mathbf{y}\|^2 - \tau_k \ln(f^*_\mu(\mathbf{x}) - f(\mathbf{x}, \mathbf{y})),$$
and
$$\mathbf{z}^*_\mu(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}) + \frac{\mu_1}{2}\|\mathbf{y}\|^2 + \mu_2,$$
are unique, then $\varphi_{\mu,\theta,\tau}$ is differentiable and
$$\frac{\partial \varphi_{\mu,\theta,\tau}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial F(\mathbf{x}, \mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} + G(\mathbf{x}),$$
where
$$G(\mathbf{x}) = \frac{\tau\left(\frac{\partial f(\mathbf{x}, \mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} - \frac{\partial f(\mathbf{x}, \mathbf{z}^*_\mu(\mathbf{x}))}{\partial \mathbf{x}}\right)}{f^*_\mu(\mathbf{x}) - f(\mathbf{x}, \mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))},$$
and $f^*_\mu(\mathbf{x}) = f(\mathbf{x}, \mathbf{z}^*_\mu(\mathbf{x})) + \frac{\mu_1}{2}\|\mathbf{z}^*_\mu(\mathbf{x})\|^2 + \mu_2$.

Proof. Since $\operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}) + \frac{\mu_1}{2}\|\mathbf{y}\|^2 + \mu_2$ is a singleton, it follows from (Bonnans & Shapiro, 2013)[Theorem 4.13, Remark 4.14] that
$$\frac{\partial f^*_\mu(\mathbf{x})}{\partial \mathbf{x}} = \left.\frac{\partial\left(f(\mathbf{x}, \mathbf{y}) + \frac{\mu_1}{2}\|\mathbf{y}\|^2 + \mu_2\right)}{\partial \mathbf{x}}\right|_{\mathbf{y}=\mathbf{z}^*_\mu(\mathbf{x})} = \frac{\partial f(\mathbf{x}, \mathbf{z}^*_\mu(\mathbf{x}))}{\partial \mathbf{x}}.$$
As $\operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^n} F(\mathbf{x}, \mathbf{y}) + \frac{\theta}{2}\|\mathbf{y}\|^2 - \tau \ln\left(f^*_\mu(\mathbf{x}) - f(\mathbf{x}, \mathbf{y})\right)$ is a singleton, (Bonnans & Shapiro, 2013)[Theorem 4.13,

Remark 4.14] shows that

$$\begin{aligned}
&\frac{\partial \varphi_{\mu,\theta,\tau}(\mathbf{x})}{\partial \mathbf{x}} = \\
&\frac{\partial \left(F(\mathbf{x},\mathbf{y}) + \frac{\theta}{2}\|\mathbf{y}\|^2 - \tau \ln\left(f^*_\mu(\mathbf{x}) - f(\mathbf{x},\mathbf{y})\right)\right)}{\partial \mathbf{x}} |_{\mathbf{y}=\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x})} \\
&= \frac{\partial F(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} + \frac{\tau\left(\frac{\partial f(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} - \frac{\partial f^*_\mu(\mathbf{x})}{\partial \mathbf{x}}\right)}{f^*_\mu(\mathbf{x}) - f(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))} \\
&= \frac{\partial F(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} + \frac{\tau\left(\frac{\partial f(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}{\partial \mathbf{x}} - \frac{\partial f(\mathbf{x},\mathbf{z}^*_{\mu}(\mathbf{x}))}{\partial \mathbf{x}}\right)}{f^*_\mu(\mathbf{x}) - f(\mathbf{x},\mathbf{y}^*_{\mu,\theta,\tau}(\mathbf{x}))}.
\end{aligned}$$

□

Lemma 1. *Let $\{\mu_k\}$ be a positive sequence such that $\mu_k \to 0$. Then for any sequence $\{\mathbf{x}_k\}$ converging to $\bar{\mathbf{x}}$,*

$$\limsup_{k\to\infty} f^*_{\mu_k}(\mathbf{x}_k) \le f^*(\bar{\mathbf{x}}).$$

Proof. For any $\epsilon > 0$, there exists $\bar{\mathbf{y}} \in \mathbb{R}^n$ such that $f(\bar{\mathbf{x}},\bar{\mathbf{y}}) \le f^*(\bar{\mathbf{x}}) + \epsilon$. And as $\mu_k \to 0$, we can find $k_1 > 0$ such that

$$f(\bar{\mathbf{x}},\bar{\mathbf{y}}) + \frac{\mu_{k,1}}{2}\|\bar{\mathbf{y}}\|^2 + \mu_{k,2} \le f^*(\mathbf{x}) + 2\epsilon$$

for all $k \ge k_1$. Next, as $\{\mathbf{x}_k\}$ converging to $\bar{\mathbf{x}}$, it follows from the continuity of $f(\mathbf{x},\mathbf{y})$ that there exists $k_2 \ge k_1$ such that $f(\mathbf{x}_k,\bar{\mathbf{y}}) \le f(\bar{\mathbf{x}},\bar{\mathbf{y}}) + \epsilon$ for any $k \ge k_2$ and thus

$$f^*_{\mu_k}(\mathbf{x}_k) \le f(\mathbf{x}_k,\bar{\mathbf{y}}) + \frac{\mu_{k,1}}{2}\|\bar{\mathbf{y}}\|^2 + \mu_{k,2} \le f^*(\mathbf{x}) + 3\epsilon$$

for all $k \ge k_2$. By letting $k \to \infty$, we obtain

$$\limsup_{k\to\infty} f^*_{\mu_k}(\mathbf{x}_k) \le f^*(\bar{\mathbf{x}}) + 3\epsilon,$$

and taking $\epsilon \to 0$ in the above inequality yields the conclusion. □

Let $\psi_\mu(\mathbf{x})$ denote the value function of following relaxed problem:

$$\psi_\mu(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^n} F(\mathbf{x},\mathbf{y}), \text{s.t.} -1 \le f(\mathbf{x},\mathbf{y}) - f^*_\mu(\mathbf{x}) \le 0.$$

Lemma 2. *Given $\mathbf{x} \in \mathcal{X}$, suppose either $F(\mathbf{x},\mathbf{y})$ or $f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x}$. Let $\{\mu_k\}$ be a positive sequence such that $\mu_k \to 0$, and then for any sequence $\{\mathbf{x}_k\}$ converging to $\mathbf{x}$,*

$$\liminf_{k\to\infty} \psi_{\mu_k}(\mathbf{x}_k) \ge \varphi(\mathbf{x}). \tag{4}$$

Proof. We assume to arrive a contradiction that there exists $\bar{\mathbf{x}} \in \mathbb{R}^m$ satisfying $\mathbf{x}_k \to \bar{\mathbf{x}}$ as $k \to \infty$ with following inequality

$$\liminf_{k\to\infty} \psi_{\mu_k}(\mathbf{x}_k) < \varphi(\bar{\mathbf{x}}).$$

Then, there exist $\epsilon > 0$ and sequences $\mathbf{x}_k \to \bar{\mathbf{x}}$ and $\{\mathbf{y}_k\}$ satisfying

$$-1 \le f(\mathbf{x}_k,\mathbf{y}_k) - f^*_{\mu_k}(\mathbf{x}_k) \le 0 \tag{5}$$

$$F(\mathbf{x}_k,\mathbf{y}_k) \le \psi_{\mu_k}(\mathbf{x}_k) + \epsilon < \varphi(\bar{\mathbf{x}}) - \epsilon. \tag{6}$$

Then it follows from Lemma 1 that

$$\limsup_{k\to\infty} f(\mathbf{x}_k,\mathbf{y}_k) \le \limsup_{k\to\infty} f^*_{\mu_k}(\mathbf{x}_k) \le f^*(\bar{\mathbf{x}}). \tag{7}$$

Since either $F(\mathbf{x},\mathbf{y})$ or $f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\bar{\mathbf{x}}$, we have that $\{\mathbf{y}_k\}$ is bounded. Take a subsequence $\{\mathbf{y}_t\}$ of $\{\mathbf{y}_k\}$ such that $\mathbf{y}_t \to \hat{\mathbf{y}}$. Then, it follows from Eq. (5), Eq. (7) and the continuity of $f(\mathbf{x},\mathbf{y})$ that

$$f(\bar{\mathbf{x}},\hat{\mathbf{y}}) \le \limsup_{t\to\infty} f(\mathbf{x}_t,\mathbf{y}_t) \le f^*(\bar{\mathbf{x}}),$$

and thus

$$\hat{\mathbf{y}} \in \mathcal{S}(\bar{\mathbf{x}}).$$

Then, Eq. (6) yields that

$$\varphi(\bar{\mathbf{x}}) \le F(\bar{\mathbf{x}},\hat{\mathbf{y}}) \le \limsup_{k\to\infty} F(\mathbf{x}_k,\mathbf{y}_k) \le \varphi(\bar{\mathbf{x}}) - \epsilon,$$

which implies a contradiction. Thus we get the conclusion. □

Lemma 3. *Let $\{(\mu_k,\theta_k,\tau_k)\}$ be a positive sequence such that $(\mu_k,\theta_k,\tau_k) \to 0$ and $\tau_k \ln \mu_{k,2} \to 0$. Then for any $\mathbf{x} \in \mathcal{X}$,*

$$\limsup_{k\to\infty} \varphi_k(\mathbf{x}) \le \varphi(\mathbf{x}).$$

Proof. Given any $\mathbf{x} \in \mathcal{X}$, for any $\epsilon > 0$, there exists $\bar{\mathbf{y}} \in \mathbb{R}^n$ satisfying $f(\mathbf{x},\bar{\mathbf{y}}) \le f^*(\mathbf{x})$ and $F(\mathbf{x},\bar{\mathbf{y}}) \le \varphi(\mathbf{x}) + \epsilon$. As $f^*(\mathbf{x}) + \mu_{k,2} \le f^*_{\mu_k}(\mathbf{x})$, and by the definition of φ_k, we have

$$\begin{aligned}
\varphi_k(\mathbf{x}) &\le F(\mathbf{x},\bar{\mathbf{y}}) + \frac{\theta_k}{2}\|\bar{\mathbf{y}}\|^2 - \tau_k \ln\left(f^*_{\mu_k}(\mathbf{x}) - f(\mathbf{x},\bar{\mathbf{y}})\right) \\
&\le \varphi(\mathbf{x}) + \epsilon + \frac{\theta_k}{2}\|\bar{\mathbf{y}}\|^2 - \tau_k \ln \mu_{k,2}.
\end{aligned}$$

By taking $k \to \infty$ in above inequality, as $\theta_k \to 0$ and $\tau_k \ln \mu_{k,2} \to 0$, we have

$$\limsup_{k\to\infty} \varphi_k(\mathbf{x}) \le \varphi(\mathbf{x}) + \epsilon.$$

Then, we get the conclusion by letting $\epsilon \to 0$. □

Proposition 2. *Suppose either $F(\mathbf{x},\mathbf{y})$ or $f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$. Let $\{(\mu_k,\theta_k,\tau_k)\}$ be a positive sequence such that $(\mu_k,\theta_k,\tau_k) \to 0$ and $\tau_k \ln \mu_{k,2} \to 0$, and then*

$$\varphi_k(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}) \xrightarrow{e} \varphi(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}). \tag{8}$$

Proof. To prove the epiconvergence of φ_k to φ, we just need to verify that sequence $\{\varphi_k\}$ satisfies the two conditions given in Definition 2. Considering any sequence $\{\mathbf{x}_k\}$ converging to $\bar{\mathbf{x}}$, since

$$F(\mathbf{x},\mathbf{y}) \le F(\mathbf{x},\mathbf{y}) + \frac{\theta_k}{2}\|\mathbf{y}\|^2 - \tau_k \ln\left(f^*_{\mu_k}(\mathbf{x}) - f(\mathbf{x},\mathbf{y})\right),$$

for any $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n$ satisfying $-1 \le f(\mathbf{x},\mathbf{y}) - f^*_\mu(\mathbf{x}) < 0$, then we have

$$\psi_{\mu_k}(\mathbf{x}_k) \le \varphi_k(\mathbf{x}_k), \quad \forall k.$$

By taking $k \to \infty$ in above inequality, we obtain from Lemma 2 that when $\bar{\mathbf{x}} \in \mathcal{X}$,

$$\begin{aligned}\varphi(\bar{\mathbf{x}}) + \delta_{\mathcal{X}}(\bar{\mathbf{x}}) &= \varphi(\bar{\mathbf{x}}) \\ &\le \liminf_{k\to\infty} \psi_{\mu_k}(\mathbf{x}_k) \\ &\le \liminf_{k\to\infty} \varphi_k(\mathbf{x}_k) \\ &\le \liminf_{k\to\infty} \varphi_k(\mathbf{x}_k) + \delta_{\mathcal{X}}(\mathbf{x}_k).\end{aligned}$$

We have $\liminf_{k\to\infty} \varphi_k(\mathbf{x}_k) + \delta_{\mathcal{X}}(\mathbf{x}_k) = +\infty$ when $\bar{\mathbf{x}} \notin \mathcal{X}$, as $\mathcal{X}$ is closed. And thus condition 1 in Definition 2 is satisfied. Next, for any $\mathbf{x} \in \mathbb{R}^m$, if $\mathbf{x} \in \mathcal{X}$, then it follows from Lemma 3 that

$$\limsup_{k\to\infty} \varphi_k(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}) \le \varphi(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}).$$

When $\mathbf{x} \notin \mathcal{X}$, we have $\varphi(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}) = +\infty$. Thus condition 2 in Definition 2 is satisfied. Therefore, we get the conclusion immediately from Definition 2. □

Theorem 1. *Suppose either $F(\mathbf{x},\mathbf{y})$ or $f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$. Let $\{(\mu_k,\theta_k,\tau_k)\}$ be a positive sequence such that $(\mu_k,\theta_k,\tau_k) \to 0$ and $\tau_k \ln \mu_{k,2} \to 0$. Then*

1. *We have the following inequality*

$$\limsup_{k\to\infty} \left(\inf_{\mathbf{x}\in\mathcal{X}} \varphi_k(\mathbf{x})\right) \le \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}). \tag{9}$$

2. *If $\mathbf{x}_\ell \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \varphi_\ell(\mathbf{x})$, for some sequence $\{\ell\} \subset \mathbb{N}$ and $\mathbf{x}_\ell$ converges to $\tilde{\mathbf{x}}$, then $\tilde{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$ and*

$$\lim_{\ell\to\infty} \left(\inf_{\mathbf{x}\in\mathcal{X}} \varphi_\ell(\mathbf{x})\right) = \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}). \tag{10}$$

Proof. According to Proposition 2, we know that

$$\varphi_k(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}) \xrightarrow{e} \varphi(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}).$$

Then the conclusion follows from (Bonnans & Shapiro, 2013)[Proposition 4.6]. □

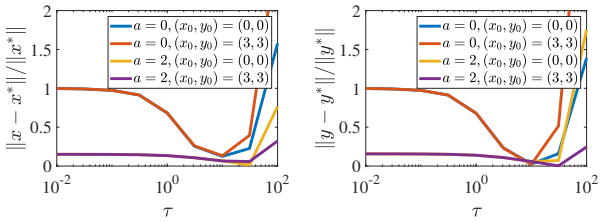

Figure 1. Convergence results for different regularization coefficients in different initialization settings.

B. Experiments

We use PyTorch 1.6 as our computational framework and base our implementation on (Grefenstette et al., 2019; Grazzi et al., 2020). In all the experiments, we use the Adam method for accelerating the gradient descent of $\mathbf{x}$. We conducted these experiments on a PC with Intel Core i7-9700F CPU, 32GB RAM and an NVIDIA RTX 2060S 8GB GPU.

B.1. Numerical Experiment

In numerical experiment, we set $T = 100$ for explicit method RHG, $T = 100$, $J = 20$ for implicit method CG, and $\mu_2 = f(x,y) + 1$, $(\mu_{k,1},\theta_k,\tau_k) = (1.0,1.0,1.0)/1.01^k$, step sizes s_1, s_2 and α all equal to 0.01, $T_z = 50$, $T_y = 25$, and $L = 1$ in BVFIM.

We can see that our method has a weaker convergence in LL problem than the existing method under proper initialization (i.e., the initial point is within a locally convex neighborhood of the global optimal point). This is because the main purpose of our experiment is to compare the convergence behavior between different methods and scenarios, so we have not carefully adjusted the regularization coefficients of our methods in order to better show the differences. To verify that our method can also converge as well as the existing methods under proper initialization, we show how to obtain better convergence performance by adjusting τ in Figure 1. It can be seen that an appropriate τ can greatly improve the convergence behavior. In addition, we validate this with a larger LLC problem in the section B.3.

B.2. Hyperparameter Optimization

In hyperparameter optimization, we set $T = 100$ for explicit method RHG, TRHG and BDA, $T = 100$, $J = 20$ for implicit method CG and Neumann, and $\mu_2 = f(\mathbf{x},\mathbf{y})$, $(\mu_{k,1},\theta_k,\tau_k) = (1.0,1.0,1.0)/1.01^k$, step sizes s_1, s_2 and α all equal to 0.01, $T_{\mathbf{z}} = 50$, $T_{\mathbf{y}} = 25$, and $L = 1$ for BVFIM. We let TRHG truncate at $T/2$ and $\alpha_k = 0.5 \times 0.999^k$ in BDA.

We set the training set, validation set, and test set as class bal-

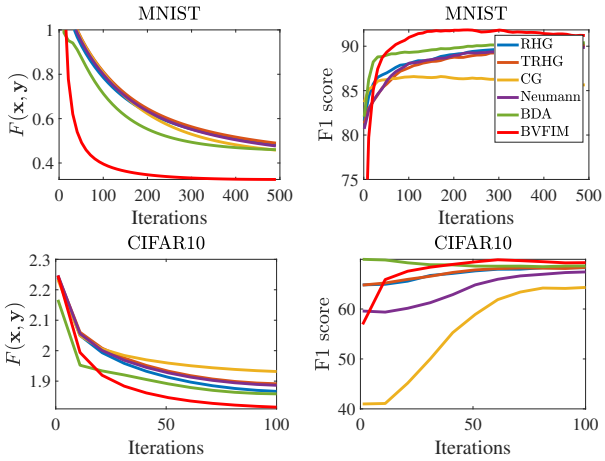

Figure 2. $F(\mathbf{x}, \mathbf{y})$ and F1 score between existing methods and BVFIM. The curves are based on the MNIST and CIFAR10 experiment. The legend is only plotted in the second subfigure.

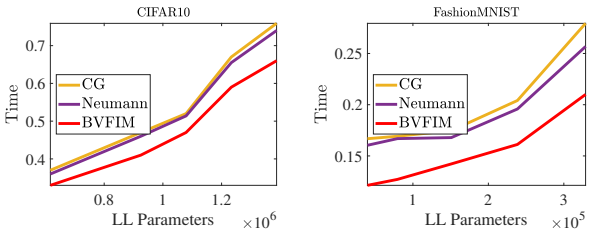

Figure 3. Comparison of calculation time of IGBMs and BVFIM under different LL parameter quantities.

anced. For each contaminated training sample, we randomly replace its label with a label different from the original one with equal probability. In the calculation of F1 score, if $\mathbf{x}_i \leq 0$, we marks the sample u_i as contaminated. In the CIFAR10 experiment, we used an early stop strategy to avoid over-fitting and report the best results achieved. Since $T_{\mathbf{y}}$ gradient descent require $\frac{\partial F}{\partial \mathbf{y}}$ and $\frac{\partial f}{\partial \mathbf{y}}$ separately, we set $T_{\mathbf{z}} + 2 \times T_{\mathbf{y}} = T$ to fairly compare the time consumed by the algorithm. The UL objective and F1 scores of BVFIM and compared methods on the MNIST and CIFAR10 dataset are plotted in Figure 2.

In addition, we verify the computation time variation of BVFIM and existing gradient-based methods under different LL variable dimensions on the FashionMNIST and CIFAR10 dataset. In order to show the comparison results more clearly, we compared with IGBMs which are faster in the existing gradient-based methods. Figure 3 shows the computation time with different LL variable parameter quantities. It can be seen that BVFIM is faster than IGBMs at different parameter quantities, and this advantage becomes more significant as the number of LL parameters increases.

Table 1. The averaged few-shot classification accuracy on Omniglot and MiniImageNet (M=1)

Alg.	Omniglot 5-way	Omniglot 20-way	MiniImagenet 5-way
RHG	98.60	95.50	48.89
TRHG	98.74	95.82	47.67
BDA	**99.04**	96.50	49.08
BVFIM	98.85	**95.55**	**49.28**

B.3. Additional LLC Experiments

To verify the validity of the BVFIM method in conventional LLC problems, we supplemented a meta-learning experiment. The goal of meta-learning is to learn an algorithm that can handle new tasks well. In particular, we consider the few-shot learning problem, where each task is a N-way classification and it aim to learn the hyperparameter $\mathbf{x}$ so that each task can be solved by only M training samples. (i.e. N-way M-shot)

Similar to recent work (Franceschi et al., 2018; Liu et al., 2020), we modeled the network in two parts: a four-layer convolution network $\mathbf{x}$ as a common feature extraction layer between different tasks, and logical regression layer $\mathbf{y} = \mathbf{y}^i$ as separate classifier for each task. We also set dataset as $\mathcal{D} = \{\mathcal{D}^j\}$, where $\mathcal{D}^i = \mathcal{D}^i_{\mathtt{tr}} \cup \mathcal{D}^i_{\mathtt{val}}$ is for the i-th task. Then we set the loss function of the j-th task to $\mathtt{CE}(\mathbf{x}, \mathbf{y}^i; \mathcal{D}^i_{\mathtt{tr}})$ for the LL problem, thus the LL objective can be defined as

$$f(\mathbf{x}, \mathbf{y}) = \sum_i \mathtt{CE}(\mathbf{x}, \mathbf{y}^i; D^i_{\mathtt{tr}})$$

As for the UL objective, we also utilize cross-entropy function but define it based on $\{D^i_{\mathtt{val}}\}$ as

$$F(\mathbf{x}, \mathbf{y}) = \sum_i \mathtt{CE}(\mathbf{x}, \mathbf{y}^i; D^i_{\mathtt{val}})$$

Our experiment was performed on two widely used benchmark datasets: Omniglot (Lake et al., 2015), which contains examples of 1623 different handwritten characters from 50 alphabets and MiniImagenet (Vinyals et al., 2016), which is a subset of ImageNet (Deng et al., 2009) that contains 60000 downsampled images from 100 different classes. We compared our BVFIM to several approaches, such as RHG, TRHG and BDA (Liu et al., 2020).

For RHG, TRHG and BDA, we follow the settings in (Liu et al., 2020). For BVFIM, we set $T_{\mathbf{z}} = 5$, $T_{\mathbf{y}} = 10$, $\mu_2 = f(\mathbf{x}, \mathbf{y})$, $(\mu_{k,1}, \theta_k, \tau_k) = (1, 0.1, 10)/k$, step sizes $(s_1, s_2, \alpha) = (0.01, 0.01, 0.001)$, $L = 1$ and $K = 50000$. We set meta-batch size of 16 episodes for Omniglot dataset and of 4 episodes for Miniimagenet dataset. We set $\mathbf{y}^0_{k,l} = \mathbf{z}^{T_{\mathbf{z}}}_{k,l}$ to warm up.

It can be seen in Table 1 that BVFIM can get slightly poorer performance than existing methods on Omniglot dataset and get the best performance on MiniimageNet dataset, which proves that our method can also obtain competitive results for LLC problems.

References

Bonnans, J. F. and Shapiro, A. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 2009.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, volume 80, pp. 1563–1572, 2018.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *ICML*. PMLR, 2020.

Grefenstette, E., Amos, B., Yarats, D., Htut, P. M., Molchanov, A., Meier, F., Kiela, D., Cho, K., and Chintala, S. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*, 2019.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350, 2015.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *ICML*. PMLR, 2020.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *NeurIPS*, 2016.