Supplementary Materials

A. Further Specifications on Meta-Learning Experiments

A.1. Datasets and Model Architectures

FC100 (Oreshkin et al., 2018) is a dataset derived from CIFAR-100 (Krizhevsky & Hinton, 2009), and contains 100 classes with each class consisting of 600 images of size 32. Following Oreshkin et al. 2018, these 100 classes are split into 60 classes for meta-training, 20 classes for meta-validation, and 20 classes for meta-testing. For all comparison algorithms, we use a 4-layer convolutional neural networks (CNN) with four convolutional blocks, in which each convolutional block contains a 3×3 convolution (padding $= 1$, stride $= 2$), batch normalization, ReLU activation, and 2×2 max pooling. Each convolutional layer has 64 filters.

The miniImageNet dataset (Vinyals et al., 2016) is generated from ImageNet (Russakovsky et al., 2015), and consists of 100 classes with each class containing 600 images of size 84×84. Following the repository (Arnold et al., 2019), we partition these classes into 64 classes for meta-training, 16 classes for meta-validation, and 20 classes for meta-testing. Following the repository (Arnold et al., 2019), we use a four-layer CNN with four convolutional blocks, where each block sequentially consists of a 3×3 convolution, batch normalization, ReLU activation, and 2×2 max pooling. Each convolutional layer has 32 filters.

A.2. Implementations and Hyperparameter Settings

We adopt the existing implementations in the repository (Arnold et al., 2019) for ANIL and MAML. For all algorithms, we adopt Adam (Kingma & Ba, 2014) as the optimizer for the outer-loop update.

Parameter selection for the experiments in Figure 2(a): For ANIL and MAML, we adopt the suggested hyperparameter selection in the repository (Arnold et al., 2019). In specific, for ANIL, we choose the inner-loop stepsize as 0.1, the outer-loop (meta) stepsize as 0.002, the task sampling size as 32, and the number of inner-loop steps as 5. For MAML, we choose the inner-loop stepsize as 0.5, the outer-loop stepsize as 0.003, the task sampling sizeas 32, and the number of inner-loop steps as 3. For ITD-BiO, AID-BiO-constant and AID-BiO-increasing, we use a grid search to choose the inner-loop stepsize from $\{0.01, 0.1, 1, 10\}$, the task sampling size from $\{32, 128, 256\}$, and the outer-loop stepsize from $\{10^i, i = -3, -2, -1, 0, 1, 2, 3\}$, where values that achieve the lowest loss after a fixed running time are selected. For ITD-BiO and AID-BiO-constant, we choose the number of inner-loop steps from $\{5, 10, 15, 20, 50\}$, and for AID-BiO-increasing, we choose the number of inner-loop steps as $\lceil c(k+1)^{1/4} \rceil$ as adopted by the analysis in Ghadimi & Wang 2018, where we choose c from $\{0.5, 2, 5, 10, 50\}$. For both AID-BiO-constant and AID-BiO-increasing, we choose the number N of CG steps for solving the linear system from $\{5, 10, 15\}$.

Parameter selection for the experiments in Figure 2(b): For ANIL and MAML, we adopt the suggested hyperparameter selection in the repository (Arnold et al., 2019). Specifically, for ANIL, we choose the inner-loop stepsize as 0.1, the outer-loop (meta) stepsize as 0.001, the task sampling size as 32 and the number of inner-loop steps as 10. For MAML, we choose the inner-loop stepsize as 0.5, the outer-loop stepsize as 0.001, the task samling size as 32, and the number of inner-loop steps as 3. For ITD-BiO, AID-BiO-constant and AID-BiO-increasing, we adopt the same procedure as in the experiments in Figure 2(a).

Parameter selection for the experiments in Figure 3: For the experiments in Figure 3(a), we choose the inner-loop stepsize as 0.05, the outer-loop (meta) stepsize as 0.002, the mini-batch size as 32, and the number T of inner-loop steps as 10 for both ANIL and ITD-BiO. For the experiments in Figure 3(b), we choose the inner-loop stepsize as 0.1, the outer-loop (meta) stepsize as 0.001, the mini-batch size as 32, and the number T of inner-loop steps as 20 for both ANIL and ITD-BiO.

B. Further Specifications on Hyperparameter Optimization Experiments

We demonstrate the effectiveness of the proposed stocBiO algorithm on two experiments: data hyper-cleaning and logistic regression, as introduced below.

Logistic Regression on 20 Newsgroup: We compare the performance of our algorithm **stocBiO** with the existing baseline algorithms **reverse, AID-FP, AID-CG and HOAG** over a logistic regression problem on 20 Newsgroup dataset (Grazzi et al., 2020). The objective function of such a problem is given by

$$\begin{aligned}&\min_{\lambda} E(\lambda, w^*) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{val}}} L(x_i w^*, y_i)\\ &\text{s.t.} \quad w^* = \arg\min_{w\in\mathbb{R}^{p\times c}} \Big(\frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{tr}}} L(x_i w, y_i) + \frac{1}{cp}\sum_{i=1}^{c}\sum_{j=1}^{p} \exp(\lambda_j) w_{ij}^2 \Big),\end{aligned}$$

where L is the cross-entropy loss, $c = 20$ is the number of topics, and $p = 101631$ is the feature dimension. Following Grazzi et al. 2020, we use SGD as the optimizer for the outer-loop update for all algorithms. For reverse, AID-FP, AID-CG, we use the suggested and well-tuned hyperparameter setting in their implementations `https://github.com/prolearner/hypertorch` on this application. In specific, they choose the inner- and outer-loop stepsizes as 100, the number of inner loops as 10, the number of CG steps as 10. For HOAG, we use the same parameters as reverse, AID-FP, AID-CG. For stocBiO, we use the same parameters as reverse, AID-FP, AID-CG, and choose $\eta = 0.5, Q = 10$. We use stocBiO-B as a shorthand of stocBiO with a batch size of B.

Data Hyper-Cleaning on MNIST. We compare the performance of our proposed algorithm stocBiO with other baseline algorithms BSA, TTSA, HOAG on a hyperparameter optimization problem: data hyper-cleaning (Shaban et al., 2019) on a dataset derived from MNIST (LeCun et al., 1998), which consists of 20000 images for training, 5000 images for validation, and 10000 images for testing. Data hyper-cleaning is to train a classifier in a corrupted setting where each label of training data is replaced by a random class number with a probability p (i.e., the corruption rate). The objective function is given by

$$\begin{aligned}&\min_{\lambda} E(\lambda, w^*) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{val}}} L(w^* x_i, y_i)\\ &\text{s.t.} \quad w^* = \arg\min_{w} \mathcal{L}(w,\lambda) := \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{tr}}} \sigma(\lambda_i) L(w x_i, y_i) + C_r \|w\|^2,\end{aligned}$$

where L is the cross-entropy loss, $\sigma(\cdot)$ is the sigmoid function, C_r is a regularization parameter. Following Shaban et al. 2019, we choose $C_r = 0.001$. All results are averaged over 10 trials with different random seeds. We adopt Adam (Kingma & Ba, 2014) as the optimizer for the outer-loop update for all algorithms. For stochastic algorithms, we set the batch size as 50 for stocBiO, and 1 for BSA and TTSA because they use the single-sample data sampling. For all algorithms, we use a grid search to choose the inner-loop stepsize from $\{0.01, 0.1, 1, 10\}$, the outer-loop stepsize from $\{10^i, i = -4, -3, -2, -1, 0, 1, 2, 3, 4\}$, and the number D of inner-loop steps from $\{1, 10, 50, 100, 200, 1000\}$, where values that achieve the lowest loss after a fixed running time are selected. For stocBiO, BSA, and TTSA, we choose η from $\{0.5 \times 2^i, i = -3, -2, -1, 0, 1, 2, 3\}$, and Q from $\{3 \times 2^i, i = 0, 1, 2, 3\}$.

C. Supporting Lemmas

In this section, we provide some auxiliary lemmas used for proving the main convergence results.

First note that the Lipschitz properties in Assumption 2 imply the following lemma.

Lemma 1. *Suppose Assumption 2 holds. Then, the stochastic derivatives $\nabla F(z;\xi)$, $\nabla G(z;\xi)$, $\nabla_x\nabla_y G(z;\xi)$ and $\nabla_y^2 G(z;\xi)$ have bounded variances, i.e., for any z and ξ,*

- $\mathbb{E}_\xi \left\|\nabla F(z;\xi) - \nabla f(z)\right\|^2 \leq M^2$.
- $\mathbb{E}_\xi \left\|\nabla_x\nabla_y G(z;\xi) - \nabla_x\nabla_y g(z)\right\|^2 \leq L^2$.
- $\mathbb{E}_\xi \left\|\nabla_y^2 G(z;\xi) - \nabla_y^2 g(z)\right\|^2 \leq L^2$.

Recall that $\Phi(x) = f(x, y^*(x))$ in eq. (2). Then, we use the following lemma to characterize the Lipschitz properties of $\nabla\Phi(x)$, which is adapted from Lemma 2.2 in Ghadimi & Wang 2018.

Lemma 2. *Suppose Assumptions 1, 2 and 3 hold. Then, we have, for any $x, x' \in \mathbb{R}^p$,*

$$\|\nabla\Phi(x) - \nabla\Phi(x')\| \leq L_\Phi \|x - x'\|,$$

where the constant L_Φ is given by

$$L_\Phi = L + \frac{2L^2 + \tau M^2}{\mu} + \frac{\rho LM + L^3 + \tau ML}{\mu^2} + \frac{\rho L^2 M}{\mu^3}. \tag{12}$$

D. Proof of Propositions 1 and 2

In this section, we provide the proofs for Proposition 1 and Proposition 2 in Section 2.

D.1. Proof of Proposition 1

Using the chain rule over the gradient $\nabla\Phi(x_k) = \frac{\partial f(x_k, y^*(x_k))}{\partial x_k}$, we have

$$\nabla\Phi(x_k) = \nabla_x f(x_k, y^*(x_k)) + \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y f(x_k, y^*(x_k)). \tag{13}$$

Based on the optimality of $y^*(x_k)$, we have $\nabla_y g(x_k, y^*(x_k)) = 0$, which, using the implicit differentiation w.r.t. x_k, yields

$$\nabla_x \nabla_y g(x_k, y^*(x_k)) + \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y^2 g(x_k, y^*(x_k)) = 0. \tag{14}$$

Let v_k^* be the solution of the linear system $\nabla_y^2 g(x_k, y^*(x_k))v = \nabla_y f(x_k, y^*(x_k))$. Then, multiplying v_k^* at the both sides of eq. (14), yields

$$-\nabla_x \nabla_y g(x_k, y^*(x_k)) v_k^* = \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y^2 g(x_k, y^*(x_k)) v_k^* = \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y f(x_k, y^*(x_k)),$$

which, in conjunction with eq. (13), completes the proof.

D.2. Proof of Proposition 2

Based on the iterative update of line 5 in Algorithm 1, we have $y_k^D = y_k^0 - \alpha \sum_{t=0}^{D-1} \nabla_y g(x_k, y_k^t)$, which, combined with the fact that $\nabla_y g(x_k, y_k^t)$ is differentiable w.r.t. x_k, indicates that the inner output y_k^T is differentiable w.r.t. x_k. Then, based on the chain rule, we have

$$\frac{\partial f(x_k, y_k^D)}{\partial x_k} = \nabla_x f(x_k, y_k^D) + \frac{\partial y_k^D}{\partial x_k} \nabla_y f(x_k, y_k^D). \tag{15}$$

Based on the iterative updates that $y_k^t = y_k^{t-1} - \alpha \nabla_y g(x_k, y_k^{t-1})$ for $t = 1, ..., D$, we have

$$\begin{aligned}
\frac{\partial y_k^t}{\partial x_k} =& \frac{\partial y_k^{t-1}}{\partial x_k} - \alpha \nabla_x \nabla_y g(x_k, y_k^{t-1}) - \alpha \frac{\partial y_k^{t-1}}{\partial x_k} \nabla_y^2 g(x_k, y_k^{t-1}) \\
=& \frac{\partial y_k^{t-1}}{\partial x_k} (I - \alpha \nabla_y^2 g(x_k, y_k^{t-1})) - \alpha \nabla_x \nabla_y g(x_k, y_k^{t-1}).
\end{aligned}$$

Telescoping the above equality over t from 1 to D yields

$$\begin{aligned}
\frac{\partial y_k^D}{\partial x_k} =& \frac{\partial y_k^0}{\partial x_k} \prod_{t=0}^{D-1} (I - \alpha \nabla_y^2 g(x_k, y_k^t)) - \alpha \sum_{t=0}^{D-1} \nabla_x \nabla_y g(x_k, y_k^t) \prod_{j=t+1}^{D-1} (I - \alpha \nabla_y^2 g(x_k, y_k^j)) \\
\overset{(i)}{=}& - \alpha \sum_{t=0}^{D-1} \nabla_x \nabla_y g(x_k, y_k^t) \prod_{j=t+1}^{D-1} (I - \alpha \nabla_y^2 g(x_k, y_k^j)).
\end{aligned} \tag{16}$$

where (i) follows from the fact that $\frac{\partial y_k^0}{\partial x_k} = 0$. Combining eq. (15) and eq. (16) finishes the proof.

E. Proof of Theorem 1

For notation simplification, we define the following quantities.

$$\begin{aligned}\Gamma =&3L^2+\frac{3\tau^2M^2}{\mu^2}+6L^2\big(1+\sqrt{\kappa}\big)^2\big(\kappa+\frac{\rho M}{\mu^2}\big)^2,\ \delta_{D,N}=\Gamma(1-\alpha\mu)^D+6L^2\kappa\big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\big)^{2N}\\ \Omega =&8\Big(\beta\kappa^2+\frac{2\beta ML}{\mu^2}+\frac{2\beta LM\kappa}{\mu^2}\Big)^2,\ \Delta_0=\|y_0-y^*(x_0)\|^2+\|v_0^*-v_0\|^2. \end{aligned} \tag{17}$$

We first provide some supporting lemmas. The following lemma characterizes the Hypergradient estimation error $\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|$, where $\widehat{\nabla}\Phi(x_k)$ is given by eq. (3) via implicit differentiation.

Lemma 3. *Suppose Assumptions 1, 2 and 3 hold. Then, we have*

$$\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2 \leq\Gamma(1-\alpha\mu)^D\|y^*(x_k)-y_k^0\|^2+6L^2\kappa\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_k^*-v_k^0\|^2.$$

where Γ is given by eq. (17).

Proof of Lemma 3. Based on the form of $\nabla\Phi(x_k)$ given by Proposition 1, we have

$$\begin{aligned}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2 \leq&3\|\nabla_x f(x_k,y^*(x_k))-\nabla_x f(x_k,y_k^D)\|^2+3\|\nabla_x\nabla_y g(x_k,y_k^D)\|^2\|v_k^*-v_k^N\|^2\\ &+3\|\nabla_x\nabla_y g(x_k,y^*(x_k))-\nabla_x\nabla_y g(x_k,y_k^D)\|^2\|v_k^*\|^2,\end{aligned}$$

which, in conjunction with Assumptions 1, 2 and 3, yields

$$\begin{aligned}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2 \leq&3L^2\|y^*(x_k)-y_k^D\|^2+3L^2\|v_k^*-v_k^N\|^2+3\tau^2\|v_k^*\|^2\|y_k^D-y^*(x_k)\|^2\\ \overset{(i)}{\leq}&3L^2\|y^*(x_k)-y_k^D\|^2+3L^2\|v_k^*-v_k^N\|^2+\frac{3\tau^2M^2}{\mu^2}\|y_k^D-y^*(x_k)\|^2.\end{aligned} \tag{18}$$

where (i) follows from the fact that $\|v_k^*\|\leq\|(\nabla_y^2 g(x_k,y^*(x_k)))^{-1}\|\|\nabla_y f(x_k,y^*(x_k))\|\leq\frac{M}{\mu}$.

For notation simplification, let $\widehat{v}_k=(\nabla_y^2 g(x_k,y_k^D))^{-1}\nabla_y f(x_k,y_k^D)$. We next upper-bound $\|v_k^*-v_k^N\|$ in eq. (18). Based on the convergence result of CG for the quadratic programing, e.g., eq. (17) in Grazzi et al. 2020, we have $\|v_k^N-\widehat{v}_k\|\leq\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_k^0-\widehat{v}_k\|$. Based on this inequality, we further have

$$\begin{aligned}\|v_k^*-v_k^N\|\leq&\|v_k^*-\widehat{v}_k\|+\|v_k^N-\widehat{v}_k\|\leq\|v_k^*-\widehat{v}_k\|+\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_k^0-\widehat{v}_k\|\\ \leq&\Big(1+\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\Big)\|v_k^*-\widehat{v}_k\|+\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_k^*-v_k^0\|.\end{aligned} \tag{19}$$

Next, based on the definitions of v_k^* and $\widehat{v}_k$, we have

$$\begin{aligned}\|v_k^*-\widehat{v}_k\|=&\|(\nabla_y^2 g(x_k,y_k^D))^{-1}\nabla_y f(x_k,y_k^D)-(\nabla_y^2 g(x_k,y^*(x_k))^{-1}\nabla_y f(x_k,y^*(x_k))\|\\ \leq&\Big(\kappa+\frac{\rho M}{\mu^2}\Big)\|y_k^D-y^*(x_k)\|.\end{aligned} \tag{20}$$

Combining eq. (18), eq. (19), eq. (20) yields

$$\begin{aligned}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2 \leq&\Big(3L^2+\frac{3\tau^2M^2}{\mu^2}\Big)\|y^*(x_k)-y_k^D\|^2+6L^2\kappa\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_k^*-v_k^0\|^2\\ &+6L^2\Big(1+\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\Big)^2\Big(\kappa+\frac{\rho M}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2,\end{aligned}$$

which, in conjunction with $\|y_k^D-y^*(x_k)\|\leq(1-\alpha\mu)^{\frac{D}{2}}\|y_k^0-y^*(x_k)\|$ and the notations in eq. (17), finishes the proof. □

Lemma 4. *Suppose Assumptions 1, 2 and 3 hold. Choose*

$$\begin{aligned}D \geq& \log\big(36\kappa(\kappa+\frac{\rho M}{\mu^2})^2+16(\kappa^2+\frac{4LM\kappa}{\mu^2})^2\beta^2\Gamma\big)/\log\frac{1}{1-\alpha}=\Theta(\kappa)\\ N \geq& \frac{1}{2}\log(8\kappa+48(\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2})^2\beta^2L^2\kappa)/\log\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}=\Theta(\sqrt{\kappa}),\end{aligned} \tag{21}$$

where Γ is given by eq. (17). Then, we have

$$\|y_k^0-y^*(x_k)\|^2+\|v_k^*-v_k^0\|^2 \leq \Big(\frac{1}{2}\Big)^k\Delta_0+\Omega\sum_{j=0}^{k-1}\Big(\frac{1}{2}\Big)^{k-1-j}\|\nabla\Phi(x_j)\|^2, \tag{22}$$

where Ω and Δ_0 are given by eq. (17).

Proof of Lemma 4. Recall that $y_k^0 = y_{k-1}^D$. Then, we have

$$\begin{aligned}\|y_k^0-y^*(x_k)\|^2 \leq& 2\|y_{k-1}^D-y^*(x_{k-1})\|^2+2\|y^*(x_k)-y^*(x_{k-1})\|^2\\ \overset{(i)}{\leq}& 2(1-\alpha\mu)^D\|y_{k-1}^0-y^*(x_{k-1})\|^2+2\kappa^2\beta^2\|\widehat{\nabla}\Phi(x_{k-1})\|^2\\ \leq& 2(1-\alpha\mu)^D\|y_{k-1}^0-y^*(x_{k-1})\|^2+4\kappa^2\beta^2\|\nabla\Phi(x_{k-1})-\widehat{\nabla}\Phi(x_{k-1})\|^2\\ &+4\kappa^2\beta^2\|\nabla\Phi(x_{k-1})\|^2\\ \overset{(ii)}{\leq}& \big(2(1-\alpha\mu)^D+4\kappa^2\beta^2\Gamma(1-\alpha\mu)^D\big)\|y^*(x_{k-1})-y_{k-1}^0\|^2\\ &+24\kappa^4L^2\beta^2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_{k-1}^*-v_{k-1}^0\|^2+4\kappa^2\beta^2\|\nabla\Phi(x_{k-1})\|^2,\end{aligned} \tag{23}$$

where (i) follows from Lemma 2.2 in Ghadimi & Wang 2018 and (ii) follows from Lemma 3. In addition, note that

$$\begin{aligned}\|v_k^*-v_k^0\|^2 =& \|v_k^*-v_{k-1}^N\|^2 \leq 2\|v_{k-1}^*-v_{k-1}^N\|^2+2\|v_k^*-v_{k-1}^*\|^2\\ \overset{(i)}{\leq}& 4\Big(1+\sqrt{\kappa}\Big)^2\Big(\kappa+\frac{\rho M}{\mu^2}\Big)^2(1-\alpha\mu)^D\|y_{k-1}^0-y^*(x_{k-1})\|^2\\ &+4\kappa\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_{k-1}^*-v_{k-1}^0\|^2+2\|v_k^*-v_{k-1}^*\|^2,\end{aligned} \tag{24}$$

where (i) follows from eq. (19). Combining eq. (24) with $\|v_k^*-v_{k-1}^*\| \leq (\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2})\|x_k-x_{k-1}\|$, we have

$$\begin{aligned}\|v_k^*-v_k^0\|^2 \overset{(i)}{\leq}& \Big(16\kappa\Big(\kappa+\frac{\rho M}{\mu^2}\Big)^2+4\Big(\kappa^2+\frac{4LM\kappa}{\mu^2}\Big)^2\beta^2\Gamma\Big)(1-\alpha\mu)^D\|y_{k-1}^0-y^*(x_{k-1})\|^2\\ &+\Big(4\kappa+48\Big(\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2}\Big)^2\beta^2L^2\kappa\Big)\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_{k-1}^*-v_{k-1}^0\|^2\\ &+4\Big(\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2}\Big)^2\beta^2\|\nabla\Phi(x_{k-1})\|^2,\end{aligned} \tag{25}$$

where (i) follows from Lemma 3. Combining eq. (23) and eq. (25) yields

$$\begin{aligned}&\|y_k^0-y^*(x_k)\|^2+\|v_k^*-v_k^0\|^2\\ &\quad\leq\Big(18\kappa\Big(\kappa+\frac{\rho M}{\mu^2}\Big)^2+8\Big(\kappa^2+\frac{4LM\kappa}{\mu^2}\Big)^2\beta^2\Gamma\Big)(1-\alpha\mu)^D\|y_{k-1}^0-y^*(x_{k-1})\|^2\\ &\qquad+\Big(4\kappa+24\Big(\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2}\Big)^2\beta^2L^2\kappa\Big)\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2N}\|v_{k-1}^*-v_{k-1}^0\|^2\\ &\qquad+8\Big(\kappa^2+\frac{2ML}{\mu^2}+\frac{2LM\kappa}{\mu^2}\Big)^2\beta^2\|\nabla\Phi(x_{k-1})\|^2,\end{aligned}$$

which, in conjunction with eq. (21), yields

$$\begin{aligned}\|y_k^0 - y^*(x_k)\|^2 + \|v_k^* - v_k^0\|^2 \le & \frac{1}{2}(\|y_{k-1}^0 - y^*(x_{k-1})\|^2 + \|v_{k-1}^* - v_{k-1}^0\|^2) \\ & + 8\Big(\beta\kappa^2 + \frac{2\beta ML}{\mu^2} + \frac{2\beta LM\kappa}{\mu^2}\Big)^2 \|\nabla\Phi(x_{k-1})\|^2. \end{aligned} \tag{26}$$

Telescoping eq. (26) over k and using the notations in eq. (17), we finish the proof. □

Lemma 5. *Under the same setting as in Lemma 4, we have*

$$\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\|^2 \le \delta_{D,N}\Big(\frac{1}{2}\Big)^k \Delta_0 + \delta_{D,N}\Omega\sum_{j=0}^{k-1}\Big(\frac{1}{2}\Big)^{k-1-j}\|\nabla\Phi(x_j)\|^2.$$

where $\delta_{T,N}$, Ω and Δ_0 are given by eq. (17).

Proof of Lemma 5. Based on Lemma 3, eq. (17) and using $ab + cd \le (a+c)(b+d)$ for any positive a, b, c, d, we have

$$\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\|^2 \le \delta_{D,N}(\|y^*(x_k) - y_k^0\|^2 + \|v_k^* - v_k^0\|^2),$$

which, in conjunction with Lemma 4, finishes the proof. □

E.1. Proof of Theorem 1

In this subsection, provide the proof for Theorem 1. Based on the smoothness of the function $\Phi(x)$ established in Lemma 2, we have

$$\begin{aligned}\Phi(x_{k+1}) \le & \Phi(x_k) + \langle\nabla\Phi(x_k), x_{k+1} - x_k\rangle + \frac{L_\Phi}{2}\|x_{k+1} - x_k\|^2 \\ \le & \Phi(x_k) - \beta\langle\nabla\Phi(x_k), \widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\rangle - \beta\|\nabla\Phi(x_k)\|^2 + \beta^2 L_\Phi\|\nabla\Phi(x_k)\|^2 \\ & + \beta^2 L_\Phi \|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2 \\ \le & \Phi(x_k) - \Big(\frac{\beta}{2} - \beta^2 L_\Phi\Big)\|\nabla\Phi(x_k)\|^2 + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2, \end{aligned} \tag{27}$$

which, combined with Lemma 5, yields

$$\begin{aligned}\Phi(x_{k+1}) \le & \Phi(x_k) - \Big(\frac{\beta}{2} - \beta^2 L_\Phi\Big)\|\nabla\Phi(x_k)\|^2 + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\delta_{D,N}\Big(\frac{1}{2}\Big)^k \Delta_0 \\ & + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\delta_{D,N}\Omega\sum_{j=0}^{k-1}\Big(\frac{1}{2}\Big)^{k-1-j}\|\nabla\Phi(x_j)\|^2. \end{aligned} \tag{28}$$

Telescoping eq. (28) over k from 0 to $K-1$ yields

$$\begin{aligned}\Big(\frac{\beta}{2} - \beta^2 L_\Phi\Big)\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2 \le & \Phi(x_0) - \inf_x \Phi(x) + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\delta_{D,N}\Delta_0 \\ & + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\delta_{D,N}\Omega\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\Big(\frac{1}{2}\Big)^{k-1-j}\|\nabla\Phi(x_j)\|^2, \end{aligned}$$

which, using the fact that $\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\|\nabla\Phi(x_j)\|^2 \le \sum_{k=0}^{K-1}\frac{1}{2^k}\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2 \le 2\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2$, yields

$$\begin{aligned}&\Big(\frac{\beta}{2} - \beta^2 L_\Phi - \big(\beta\Omega + 2\Omega\beta^2 L_\Phi\big)\delta_{D,N}\Big)\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2 \\ &\qquad \le \Phi(x_0) - \inf_x \Phi(x) + \Big(\frac{\beta}{2} + \beta^2 L_\Phi\Big)\delta_{D,N}\Delta_0. \end{aligned} \tag{29}$$

Choose N and D such that

$$\big(\Omega+2\Omega\beta L_\Phi\big)\delta_{D,N}\leq\frac{1}{4},\quad \delta_{D,N}\leq 1. \tag{30}$$

Note that based on the definition of $\delta_{D,N}$ in eq. (17), it suffices to choose $D\geq\Theta(\kappa)$ and $N\geq\Theta(\sqrt{\kappa})$ to satisfy eq. (30). Then, substituting eq. (30) into eq. (29) yields

$$\Big(\frac{\beta}{4}-\beta^2L_\Phi\Big)\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2\leq\Phi(x_0)-\inf_x\Phi(x)+\Big(\frac{\beta}{2}+\beta^2L_\Phi\Big)\Delta_0,$$

which, in conjunction with $\beta\leq\frac{1}{8L_\Phi}$, yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2\leq\frac{64L_\Phi(\Phi(x_0)-\inf_x\Phi(x))+5\Delta_0}{K}. \tag{31}$$

In order to achieve an ϵ-accurate stationary point, we obtain from eq. (31) that AID-BiO requires at most the total number $K=\mathcal{O}(\kappa^3\epsilon^{-1})$ of outer iterations. Then, based on eq. (3), we have the following complexity results.

- Gradient complexity:
$$\text{Gc}(f,\epsilon)=2K=\mathcal{O}(\kappa^3\epsilon^{-1}),\text{Gc}(g,\epsilon)=KD=\mathcal{O}\big(\kappa^4\epsilon^{-1}\big).$$
- Jacobian- and Hessian-vector product complexities:
$$\text{JV}(g,\epsilon)=K=\mathcal{O}\left(\kappa^3\epsilon^{-1}\right),\text{HV}(g,\epsilon)=KN=\mathcal{O}\left(\kappa^{3.5}\epsilon^{-1}\right).$$

Then, the proof is complete.

F. Proof of Theorem 2

We first characterize an important estimation property of the outer-loop gradient estimator $\frac{\partial f(x_k,y_k^D)}{\partial x_k}$ in ITD-BiO for approximating the true gradient $\nabla\Phi(x_k)$ based on Proposition 2.

Lemma 6. *Suppose Assumptions 1, 2 and 3 hold. Choose $\alpha\leq\frac{1}{L}$. Then, we have*

$$\Big\|\frac{\partial f(x_k,y_k^D)}{\partial x_k}-\nabla\Phi(x_k)\Big\|\leq\Big(\frac{L(L+\mu)(1-\alpha\mu)^{\frac{D}{2}}}{\mu}+\frac{2M\left(\tau\mu+L\rho\right)}{\mu^2}(1-\alpha\mu)^{\frac{D-1}{2}}\Big)\|y_k^0-y^*(x_k)\|+\frac{LM(1-\alpha\mu)^D}{\mu}.$$

Lemma 6 shows that the gradient estimation error $\big\|\frac{\partial f(x_k,y_k^D)}{\partial x_k}-\nabla\Phi(x_k)\big\|$ decays exponentially w.r.t. the number D of the inner-loop steps. We note that Grazzi et al. 2020 proved a similar result via a fixed point based approach. As a comparison, our proof of Lemma 6 directly characterizes the rate of the sequence $\big(\frac{\partial y_k^t}{\partial x_k},t=0,...,D\big)$ converging to $\frac{\partial y^*(x_k)}{\partial x_k}$ via the differentiation over all corresponding points along the inner-loop GD path as well as the optimality of the point $y^*(x_k)$.

Proof of Lemma 6. Using $\nabla\Phi(x_k)=\nabla_xf(x_k,y^*(x_k))+\frac{\partial y^*(x_k)}{\partial x_k}\nabla_yf(x_k,y^*(x_k))$ and eq. (15) , and using the triangle inequality, we have

$$\begin{aligned}\Big\|&\frac{\partial f(x_k,y_k^D)}{\partial x_k}-\nabla\Phi(x_k)\Big\|\\&=\|\nabla_xf(x_k,y_k^D)-\nabla_xf(x_k,y^*(x_k))\|+\Big\|\frac{\partial y_k^D}{\partial x_k}-\frac{\partial y^*(x_k)}{\partial x_k}\Big\|\|\nabla_yf(x_k,y_k^D)\|\\&\quad+\Big\|\frac{\partial y^*(x_k)}{\partial x_k}\Big\|\|\nabla_yf(x_k,y_k^D)-\nabla_yf(x_k,y^*(x_k))\|\\&\overset{(i)}{\leq}L\|y_k^D-y^*(x_k)\|+M\Big\|\frac{\partial y_k^D}{\partial x_k}-\frac{\partial y^*(x_k)}{\partial x_k}\Big\|+L\Big\|\frac{\partial y^*(x_k)}{\partial x_k}\Big\|\|y_k^D-y^*(x_k)\|,\end{aligned} \tag{32}$$

where (i) follows from Assumption 2. Our next step is to upper-bound $\left\|\frac{\partial y_k^D}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k}\right\|$ in eq. (32).

Based on the updates $y_k^t = y_k^{t-1} - \alpha \nabla_y g(x_k, y_k^{t-1})$ for $t = 1, ..., D$ in ITD-BiO and using the chain rule, we have

$$\frac{\partial y_k^t}{\partial x_k} = \frac{\partial y_k^{t-1}}{\partial x_k} - \alpha \left(\nabla_x \nabla_y g(x_k, y_k^{t-1}) + \frac{\partial y_k^{t-1}}{\partial x_k} \nabla_y^2 g(x_k, y_k^{t-1}) \right). \tag{33}$$

Based on the optimality of $y^*(x_k)$, we have $\nabla_y g(x_k, y^*(x_k)) = 0$, which, in conjunction with the implicit differentiation theorem, yields

$$\nabla_x \nabla_y g(x_k, y^*(x_k)) + \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y^2 g(x_k, y^*(x_k)) = 0. \tag{34}$$

Substituting eq. (34) into eq. (33) yields

$$\begin{aligned}
\frac{\partial y_k^t}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} =& \frac{\partial y_k^{t-1}}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} - \alpha \left(\nabla_x \nabla_y g(x_k, y_k^{t-1}) + \frac{\partial y_k^{t-1}}{\partial x_k} \nabla_y^2 g(x_k, y_k^{t-1}) \right) \\
& + \alpha \left(\nabla_x \nabla_y g(x_k, y^*(x_k)) + \frac{\partial y^*(x_k)}{\partial x_k} \nabla_y^2 g(x_k, y^*(x_k)) \right) \\
=& \frac{\partial y_k^{t-1}}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} - \alpha \left(\nabla_x \nabla_y g(x_k, y_k^{t-1}) - \nabla_x \nabla_y g(x_k, y^*(x_k)) \right) \\
& - \alpha \left(\frac{\partial y_k^{t-1}}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right) \nabla_y^2 g(x_k, y_k^{t-1}) \\
& + \alpha \frac{\partial y^*(x_k)}{\partial x_k} \left(\nabla_y^2 g(x_k, y^*(x_k)) - \nabla_y^2 g(x_k, y_k^{t-1}) \right).
\end{aligned} \tag{35}$$

Combining eq. (34) and Assumption 2 yields

$$\left\| \frac{\partial y^*(x_k)}{\partial x_k} \right\| = \left\| \nabla_x \nabla_y g(x_k, y^*(x_k)) \left[\nabla_y^2 g(x_k, y^*(x_k)) \right]^{-1} \right\| \leq \frac{L}{\mu}. \tag{36}$$

Then, combining eq. (35) and eq. (36) yields

$$\begin{aligned}
\left\| \frac{\partial y_k^t}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| \overset{(i)}{\leq} & \left\| I - \alpha \nabla_y^2 g(x_k, y_k^{t-1}) \right\| \left\| \frac{\partial y_k^{t-1}}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| \\
& + \alpha \left(\tau + \frac{L\rho}{\mu} \right) \| y_k^{t-1} - y^*(x_k) \| \\
\overset{(ii)}{\leq} & (1 - \alpha\mu) \left\| \frac{\partial y_k^{t-1}}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| + \alpha \left(\tau + \frac{L\rho}{\mu} \right) \| y_k^{t-1} - y^*(x_k) \|,
\end{aligned} \tag{37}$$

where (i) follows from Assumption 3 and (ii) follows from the strong-convexity of $g(x, \cdot)$. Based on the strong-convexity of the lower-level function $g(x, \cdot)$, we have

$$\| y_k^{t-1} - y^*(x_k) \| \leq (1 - \alpha\mu)^{\frac{t-1}{2}} \| y_k^0 - y^*(x_k) \|. \tag{38}$$

Substituting eq. (38) into eq. (37) and telecopting eq. (37) over t from 1 to D, we have

$$\begin{aligned}
\left\| \frac{\partial y_k^D}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| \leq & (1 - \alpha\mu)^D \left\| \frac{\partial y_k^0}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| \\
& + \alpha \left(\tau + \frac{L\rho}{\mu} \right) \sum_{t=0}^{D-1} (1 - \alpha\mu)^{D-1-t} (1 - \alpha\mu)^{\frac{t}{2}} \| y_k^0 - y^*(x_k) \| \\
=& (1 - \alpha\mu)^D \left\| \frac{\partial y_k^0}{\partial x_k} - \frac{\partial y^*(x_k)}{\partial x_k} \right\| + \frac{2(\tau\mu + L\rho)}{\mu^2} (1 - \alpha\mu)^{\frac{D-1}{2}} \| y_k^0 - y^*(x_k) \| \\
\leq & \frac{L(1 - \alpha\mu)^D}{\mu} + \frac{2(\tau\mu + L\rho)}{\mu^2} (1 - \alpha\mu)^{\frac{D-1}{2}} \| y_k^0 - y^*(x_k) \|,
\end{aligned} \tag{39}$$

where the last inequality follows from $\frac{\partial y_k^0}{\partial x_k} = 0$ and eq. (36). Then, combining eq. (32), eq. (36), eq. (38) and eq. (39) completes the proof. □

F.1. Proof of Theorem 2

Based on the characterization on the estimation error of the gradient estimate $\frac{\partial f(x_k,y_k^D)}{\partial x_k}$ in Lemma 6, we now prove Theorem 2.

Recall the notation that $\widehat{\nabla}\Phi(x_k)=\frac{\partial f(x_k,y_k^D)}{\partial x_k}$. Using an approach similar to eq. (27), we have

$$\Phi(x_{k+1}) \leq \Phi(x_k) - \Big(\frac{\beta}{2}-\beta^2 L_\Phi\Big)\|\nabla\Phi(x_k)\|^2 + \Big(\frac{\beta}{2}+\beta^2 L_\Phi\Big)\|\nabla\Phi(x_k)-\widehat{\nabla}\Phi(x_k)\|^2, \tag{40}$$

which, in conjunction with Lemma 6 and using $\|y_k^0-y^*(x_k)\|^2\leq\Delta$, yields

$$\begin{aligned}\Phi(x_{k+1}) \leq &\Phi(x_k) - \Big(\frac{\beta}{2}-\beta^2 L_\Phi\Big)\|\nabla\Phi(x_k)\|^2\\ &+3\Delta\Big(\frac{\beta}{2}+\beta^2 L_\Phi\Big)\Big(\frac{L^2(L+\mu)^2}{\mu^2}(1-\alpha\mu)^D+\frac{4M^2\left(\tau\mu+L\rho\right)^2}{\mu^4}(1-\alpha\mu)^{D-1}\Big)\\ &+3\Big(\frac{\beta}{2}+\beta^2 L_\Phi\Big)\frac{L^2M^2(1-\alpha\mu)^{2D}}{\mu^2}.\end{aligned} \tag{41}$$

Telescoping eq. (41) over k from 0 to $K-1$ yields

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\Big(\frac{1}{2}-\beta L_\Phi\Big)\|\nabla\Phi(x_k)\|^2 \leq &\frac{\Phi(x_0)-\inf_x\Phi(x)}{\beta K}+3\Big(\frac{1}{2}+\beta L_\Phi\Big)\frac{L^2M^2(1-\alpha\mu)^{2D}}{\mu^2}\\ &+3\Delta\Big(\frac{1}{2}+\beta L_\Phi\Big)\Big(\frac{L^2(L+\mu)^2}{\mu^2}(1-\alpha\mu)^D+\frac{4M^2\left(\tau\mu+L\rho\right)^2}{\mu^4}(1-\alpha\mu)^{D-1}\Big).\end{aligned} \tag{42}$$

Substuting $\beta=\frac{1}{4L_\Phi}$ and $D=\log\Big(\max\big\{\frac{3LM}{\mu},9\Delta L^2(1+\frac{L}{\mu})^2,\frac{36\Delta M^2(\tau\mu+L\rho)^2}{(1-\alpha\mu)\mu^4}\big\}\frac{9}{2\epsilon}\Big)/\log\frac{1}{1-\alpha\mu}=\Theta(\kappa\log\frac{1}{\epsilon})$ in eq. (42) yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2\leq\frac{16L_\Phi(\Phi(x_0)-\inf_x\Phi(x))}{K}+\frac{2\epsilon}{3}. \tag{43}$$

In order to achieve an ϵ-accurate stationary point, we obtain from eq. (43) that ITD-BiO requires at most the total number $K=\mathcal{O}(\kappa^3\epsilon^{-1})$ of outer iterations. Then, based on the gradient form given by Proposition 2, we have the following complexity results.

- Gradient complexity: $\mathrm{Gc}(f,\epsilon)=2K=\mathcal{O}(\kappa^3\epsilon^{-1}),\mathrm{Gc}(g,\epsilon)=KD=\mathcal{O}\left(\kappa^4\epsilon^{-1}\log\frac{1}{\epsilon}\right)$.
- Jacobian- and Hessian-vector product complexities:

$$\mathrm{JV}(g,\epsilon)=KD=\mathcal{O}\left(\kappa^4\epsilon^{-1}\log\frac{1}{\epsilon}\right),\mathrm{HV}(g,\epsilon)=KD=\mathcal{O}\left(\kappa^4\epsilon^{-1}\log\frac{1}{\epsilon}\right).$$

Then, the proof is complete.

G. Proofs of Proposition 3 and Theorem 3

In this section, we provide the proofs for the convergence and complexity results of the proposed algorithm stocBiO for the stochastic case.

G.1. Proof of Proposition 3

Based on the definition of v_Q in eq. (5) and conditioning on x_k, y_k^D, we have

$$\begin{aligned}
\mathbb{E}v_Q =&\mathbb{E}\eta\sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))\nabla_yF(x_k,y_k^D;\mathcal{D}_F),\\
&=\eta\sum_{q=0}^{Q}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\nabla_yf(x_k,y_k^D)\\
&=\eta\sum_{q=0}^{\infty}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\nabla_yf(x_k,y_k^D)-\eta\sum_{q=Q+1}^{\infty}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\nabla_yf(x_k,y_k^D)\\
&=\eta(\eta\nabla_y^2g(x_k,y_k^D))^{-1}\nabla_yf(x_k,y_k^D)-\eta\sum_{q=Q+1}^{\infty}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\nabla_yf(x_k,y_k^D),
\end{aligned}$$

which, in conjunction with the strong-convexity of function $g(x,\cdot)$, yields

$$\left\|\mathbb{E}v_Q-[\nabla_y^2g(x_k,y_k^D)]^{-1}\nabla_yf(x_k,y_k^D)\right\|\leq\eta\sum_{q=Q+1}^{\infty}(1-\eta\mu)^qM\leq\frac{(1-\eta\mu)^{Q+1}M}{\mu}. \tag{44}$$

This finishes the proof for the estimation bias. We next prove the variance bound. Note that

$$\begin{aligned}
&\mathbb{E}\left\|\eta\sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))\nabla_yF(x_k,y_k^D;\mathcal{D}_F)-(\nabla_y^2g(x_k,y_k^D))^{-1}\nabla_yf(x_k,y_k^D)\right\|^2\\
&\overset{(i)}{\leq}2\mathbb{E}\left\|\eta\sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))-(\nabla_y^2g(x_k,y_k^D))^{-1}\right\|^2M^2+\frac{2M^2}{\mu^2D_f}\\
&\leq4\mathbb{E}\left\|\eta\sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))-\eta\sum_{q=0}^{Q}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\right\|^2M^2\\
&\quad+4\mathbb{E}\left\|\eta\sum_{q=0}^{Q}(I-\eta\nabla_y^2g(x_k,y_k^D))^q)-(\nabla_y^2g(x_k,y_k^D))^{-1}\right\|^2M^2+\frac{2M^2}{\mu^2D_f}\\
&\overset{(ii)}{\leq}4\eta^2\mathbb{E}\left\|\sum_{q=0}^{Q}\prod_{j=Q+1-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))-\sum_{q=0}^{Q}(I-\eta\nabla_y^2g(x_k,y_k^D))^q\right\|^2M^2+\frac{4(1-\eta\mu)^{2Q+2}M^2}{\mu^2}+\frac{2M^2}{\mu^2D_f}\\
&\overset{(iii)}{\leq}4\eta^2M^2Q\mathbb{E}\sum_{q=0}^{Q}\underbrace{\left\|\prod_{j=Q+1-q}^{Q}(I-\eta\nabla_y^2G(x_k,y_k^D;\mathcal{B}_j))-(I-\eta\nabla_y^2g(x_k,y_k^D))^q\right\|^2}_{M_q}\\
&\quad+\frac{4(1-\eta\mu)^{2Q+2}M^2}{\mu^2}+\frac{2M^2}{\mu^2D_f}
\end{aligned}\tag{45}$$

where (i) follows from Lemma 1, (ii) follows from eq. (44), and (iii) follows from the Cauchy-Schwarz inequality.

Our next step is to upper-bound M_q in eq. (45). For simplicity, we define a general quantity M_i for by replacing q in M_q

with i. Then, we have

$$\begin{aligned}
\mathbb{E}M_i =&\mathbb{E}\Big\|(I-\eta\nabla_y^2 g(x_k,y_k^D))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))-(I-\eta\nabla_y^2 g(x_k,y_k^D))^i\Big\|^2\\
&+\mathbb{E}\Big\|\eta(\nabla_y^2 g(x_k,y_k^D)-\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_{Q+1-i}))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))\Big\|^2\\
&+2\mathbb{E}\Big\langle(I-\eta\nabla_y^2 g(x_k,y_k^D))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))-(I-\eta\nabla_y^2 g(x_k,y_k^D))^i,\\
&\quad\eta(\nabla_y^2 g(x_k,y_k^D)-\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_{Q+1-i}))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))\Big\rangle\\
\overset{(i)}{=}&\mathbb{E}\Big\|(I-\eta\nabla_y^2 g(x_k,y_k^D))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))-(I-\eta\nabla_y^2 g(x_k,y_k^D))^i\Big\|^2\\
&+\mathbb{E}\Big\|\eta(\nabla_y^2 g(x_k,y_k^D)-\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_{Q+1-i}))\prod_{j=Q+2-i}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))\Big\|^2\\
\overset{(ii)}{\leq}&(1-\eta\mu)^2\mathbb{E}M_{i-1}+\eta^2(1-\eta\mu)^{2i-2}\mathbb{E}\|\nabla_y^2 g(x_k,y_k^D)-\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_{Q+1-i})\|^2\\
\overset{(iii)}{\leq}&(1-\eta\mu)^2\mathbb{E}M_{i-1}+\eta^2(1-\eta\mu)^{2i-2}\frac{L^2}{|\mathcal{B}_{Q+1-i}|},
\end{aligned}\tag{46}$$

where (i) follows from the fact that $\mathbb{E}_{\mathcal{B}_{Q+1-i}}\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_{Q+1-i})=\nabla_y^2 g(x_k,y_k^D)$, (ii) follows from the strong-convexity of function $G(x,\cdot;\xi)$, and (iii) follows from Lemma 1.

Then, telescoping eq. (46) over i from 2 to q yields

$$\mathbb{E}M_q\leq L^2\eta^2(1-\eta\mu)^{2q-2}\sum_{j=1}^{q}\frac{1}{|\mathcal{B}_{Q+1-j}|},$$

which, in conjunction with the choice of $|\mathcal{B}_{Q+1-j}|=BQ(1-\eta\mu)^{j-1}$ for $j=1,...,Q$, yields

$$\begin{aligned}
\mathbb{E}M_q\leq&\eta^2(1-\eta\mu)^{2q-2}\sum_{j=1}^{q}\frac{L^2}{BQ}\Big(\frac{1}{1-\eta\mu}\Big)^{j-1}\\
=&\frac{\eta^2L^2}{BQ}(1-\eta\mu)^{2q-2}\frac{\left(\frac{1}{1-\eta\mu}\right)^{q-1}-1}{\frac{1}{1-\eta\mu}-1}\leq\frac{\eta L^2}{(1-\eta\mu)\mu}\frac{1}{BQ}(1-\eta\mu)^q.
\end{aligned}\tag{47}$$

Substituting eq. (47) into eq. (45) yields

$$\begin{aligned}
&\mathbb{E}\Big\|\eta\sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\eta\nabla_y^2 G(x_k,y_k^D;\mathcal{B}_j))\nabla_y F(x_k,y_k^D;\mathcal{D}_F)-(\nabla_y^2 g(x_k,y_k^D))^{-1}\nabla_y f(x_k,y_k^D)\Big\|^2\\
&\leq4\eta^2M^2Q\sum_{q=0}^{Q}\frac{\eta L^2}{(1-\eta\mu)\mu}\frac{1}{BQ}(1-\eta\mu)^q+\frac{4(1-\eta\mu)^{2Q+2}M^2}{\mu^2}+\frac{2M^2}{\mu^2D_f}\\
&\leq\frac{4\eta^2L^2M^2}{\mu^2}\frac{1}{B}+\frac{4(1-\eta\mu)^{2Q+2}M^2}{\mu^2}+\frac{2M^2}{\mu^2D_f},
\end{aligned}\tag{48}$$

where the last inequality follows from the fact that $\sum_{q=0}^{S}x^q\leq\frac{1}{1-x}$. Then, the proof is complete.

G.2. Auxiliary Lemmas for Proving Theorem 3

We first use the following lemma to characterize the first-moment error of the gradient estimate $\widehat{\nabla}\Phi(x_k)$, whose form is given by eq. (6).

Lemma 7. *Suppose Assumptions 1, 2 and 3 hold. Then, conditioning on x_k and y_k^D, we have*

$$\left\|\mathbb{E}\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\right\|^2\leq 2\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2+\frac{2L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}.$$

Proof of Lemma 7. To simplify notations, we define

$$\widetilde{\nabla}\Phi_D(x_k)=\nabla_x f(x_k,y_k^D)-\nabla_x\nabla_y g(x_k,y_k^D)\big[\nabla_y^2 g(x_k,y_k^D)\big]^{-1}\nabla_y f(x_k,y_k^D). \tag{49}$$

Based on the definition of $\widehat{\nabla}\Phi(x_k)$ in eq. (6) and conditioning on x_k and y_k^D, we have

$$\begin{aligned}\mathbb{E}\widehat{\nabla}\Phi(x_k)=&\nabla_x f(x_k,y_k^D)-\nabla_x\nabla_y g(x_k,y_k^D)\mathbb{E}v_Q\\=&\widetilde{\nabla}\Phi_D(x_k)-\nabla_x\nabla_y g(x_k,y_k^D)(\mathbb{E}v_Q-[\nabla_y^2 g(x_k,y_k^D)]^{-1}\nabla_y f(x_k,y_k^D)),\end{aligned}$$

which further implies that

$$\begin{aligned}&\left\|\mathbb{E}\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\right\|^2\\&\leq 2\mathbb{E}\|\widetilde{\nabla}\Phi_D(x_k)-\nabla\Phi(x_k)\|^2+2\|\mathbb{E}\widehat{\nabla}\Phi(x_k)-\widetilde{\nabla}\Phi_D(x_k)\|^2\\&\leq 2\mathbb{E}\|\widetilde{\nabla}\Phi_D(x_k)-\nabla\Phi(x_k)\|^2+2L^2\|\mathbb{E}v_Q-[\nabla_y^2 g(x_k,y_k^D)]^{-1}\nabla_y f(x_k,y_k^D)\|^2\\&\leq 2\mathbb{E}\|\widetilde{\nabla}\Phi_D(x_k)-\nabla\Phi(x_k)\|^2+\frac{2L^2M^2(1-\eta\mu)^{2Q+2}}{\mu^2},\end{aligned} \tag{50}$$

where the last inequality follows from Proposition 3. Our next step is to upper-bound the first term at the right hand side of eq. (50). Using the fact that $\left\|\nabla_y^2 g(x,y)^{-1}\right\|\leq\frac{1}{\mu}$ and based on Assumptions 2 and 3, we have

$$\begin{aligned}\|\widetilde{\nabla}\Phi_D(x_k)-\nabla\Phi(x_k)\|\leq&\|\nabla_x f(x_k,y_k^D)-\nabla_x f(x_k,y^*(x_k))\|\\&+\frac{L^2}{\mu}\|y_k^D-y^*(x_k)\|+\frac{M\tau}{\mu}\|y_k^D-y^*(x_k)\|\\&+LM\left\|\nabla_y^2 g(x_k,y_k^D)^{-1}-\nabla_y^2 g(x_k,y^*(x_k))^{-1}\right\|\\\leq&\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)\|y_k^D-y^*(x_k)\|,\end{aligned} \tag{51}$$

where the last inequality follows from the inequality $\|M_1^{-1}-M_2^{-1}\|\leq\|M_1^{-1}M_2^{-1}\|\|M_1-M_2\|$ for any two matrices M_1 and M_2. Combining eq. (50) and eq. (51) yields

$$\left\|\mathbb{E}\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\right\|^2\leq 2\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2+\frac{2L^2M^2(1-\eta\mu)^{2Q}}{\mu^2},$$

which completes the proof. □

Then, we use the following lemma to characterize the variance of the estimator $\widehat{\nabla}\Phi(x_k)$.

Lemma 8. *Suppose Assumptions 1, 2 and 3 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2\leq&\frac{4L^2M^2}{\mu^2D_g}+\Big(\frac{8L^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}\\&+\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\mathbb{E}\|y_k^D-y^*(x_k)\|^2.\end{aligned}$$

Proof of Lemma 8. Based on the definitions of $\nabla\Phi(x_k)$ and $\widetilde{\nabla}\Phi_D(x_k)$ in eq. (4) and eq. (49) and conditioning on x_k and y_k^D, we have

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2\\
&\overset{(i)}{=}\mathbb{E}\|\widehat{\nabla}\Phi(x_k)-\widetilde{\nabla}\Phi_D(x_k)\|^2+\|\widetilde{\nabla}\Phi_D(x_k)-\nabla\Phi(x_k)\|^2\\
&\overset{(ii)}{\leq}2\mathbb{E}\big\|\nabla_x\nabla_y G(x_k,y_k^D;\mathcal{D}_G)v_Q-\nabla_x\nabla_y g(x_k,y_k^D)\big[\nabla_y^2 g(x_k,y_k^D)\big]^{-1}\nabla_y f(x_k,y_k^D)\big\|^2+\frac{2M^2}{D_f}\\
&\quad+\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2\\
&\overset{(iii)}{\leq}\frac{4M^2}{\mu^2}\mathbb{E}\|\nabla_x\nabla_y G(x_k,y_k^D;\mathcal{D}_G)-\nabla_x\nabla_y g(x_k,y_k^D)\|^2+4L^2\mathbb{E}\|v_Q-\big[\nabla_y^2 g(x_k,y_k^D)\big]^{-1}\nabla_y f(x_k,y_k^D)\|^2\\
&\quad+\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2+\frac{2M^2}{D_f},
\end{aligned}
\tag{52}
$$

where (i) follows from the fact that $\mathbb{E}_{\mathcal{D}_G,\mathcal{D}_H,\mathcal{D}_F}\widehat{\nabla}\Phi(x_k)=\widetilde{\nabla}\Phi_D(x_k)$, (ii) follows from Lemma 1 and eq. (51), and (iii) follows from the Young's inequality and Assumption 2.

Using Lemma 1 and Proposition 3 in eq. (52), yields

$$
\begin{aligned}
\mathbb{E}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2\leq&\frac{4L^2M^2}{\mu^2D_g}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16(1-\eta\mu)^{2Q}L^2M^2}{\mu^2}+\frac{8L^2M^2}{\mu^2D_f}\\
&+\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\|y_k^D-y^*(x_k)\|^2+\frac{2M^2}{D_f},
\end{aligned}
\tag{53}
$$

which, unconditioning on x_k and y_k^D, completes the proof. □

It can be seen from Lemmas 7 and 8 that the upper bounds on both the estimation error and bias depend on the tracking error $\|y_k^D-y^*(x_k)\|^2$. The following lemma provides an upper bound on such a tracking error $\|y_k^D-y^*(x_k)\|^2$.

Lemma 9. *Suppose Assumptions 1, 2 and 4 hold. Define constants*

$$
\begin{aligned}
\lambda=&\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\Big(2+\frac{4\beta^2L^2}{\mu^2}\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\Big)\\
\Delta=&\frac{4L^2M^2}{\mu^2D_g}+\Big(\frac{8L^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}\\
\omega=&\frac{4\beta^2L^2}{\mu^2}\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}.
\end{aligned}
\tag{54}
$$

Choose D such that $\lambda<1$ and set inner-loop stepsize $\alpha=\frac{2}{L+\mu}$. Then, we have

$$
\begin{aligned}
&\mathbb{E}\|y_k^D-y^*(x_k)\|^2\\
&\leq\lambda^k\left(\left(\frac{L-\mu}{L+\mu}\right)^{2D}\|y_0-y^*(x_0)\|^2+\frac{\sigma^2}{L\mu S}\right)+\omega\sum_{j=0}^{k-1}\lambda^{k-1-j}\mathbb{E}\|\nabla\Phi(x_j)\|^2+\frac{\omega\Delta+\frac{\sigma^2}{L\mu S}}{1-\lambda}.
\end{aligned}
$$

Proof of Lemma 9. First note that for an integer $t\leq D$

$$
\begin{aligned}
\|y_k^{t+1}-y^*(x_k)\|^2=&\|y_k^{t+1}-y_k^t\|^2+2\langle y_k^{t+1}-y_k^t,y_k^t-y^*(x_k)\rangle+\|y_k^t-y^*(x_k)\|^2\\
=&\alpha^2\|\nabla_y G(x_k,y_k^t;\mathcal{S}_t)\|^2-2\alpha\langle\nabla_y G(x_k,y_k^t;\mathcal{S}_t),y_k^t-y^*(x_k)\rangle+\|y_k^t-y^*(x_k)\|^2.
\end{aligned}
\tag{55}
$$

Conditioning on y_k^t and taking expectation in eq. (55), we have

$$\begin{aligned}
&\mathbb{E}\|y_k^{t+1}-y^*(x_k)\|^2\\
&\overset{(i)}{\leq}\alpha^2\Big(\frac{\sigma^2}{S}+\|\nabla_y g(x_k,y_k^t)\|^2\Big)-2\alpha\langle\nabla_y g(x_k,y_k^t),y_k^t-y^*(x_k)\rangle\\
&\quad+\|y_k^t-y^*(x_k)\|^2\\
&\overset{(ii)}{\leq}\frac{\alpha^2\sigma^2}{S}+\alpha^2\|\nabla_y g(x_k,y_k^t)\|^2-2\alpha\left(\frac{L\mu}{L+\mu}\|y_k^t-y^*(x_k)\|^2+\frac{\|\nabla_y g(x_k,y_k^t)\|^2}{L+\mu}\right)\\
&\quad+\|y_k^t-y^*(x_k)\|^2\\
&=\frac{\alpha^2\sigma^2}{S}-\alpha\left(\frac{2}{L+\mu}-\alpha\right)\|\nabla_y g(x_k,y_k^t)\|^2+\left(1-\frac{2\alpha L\mu}{L+\mu}\right)\|y_k^t-y^*(x_k)\|^2
\end{aligned}\tag{56}$$

where (i) follows from the third item in Assumption 2, and (ii) follows from the strong-convexity and smoothness of the function g. Since $\alpha=\frac{2}{L+\mu}$, we obtain from eq. (56) that

$$\mathbb{E}\|y_k^{t+1}-y^*(x_k)\|^2\leq\left(\frac{L-\mu}{L+\mu}\right)^2\|y_k^t-y^*(x_k)\|^2+\frac{4\sigma^2}{(L+\mu)^2S}.\tag{57}$$

Unconditioning on y_k^t in eq. (57) and telescoping eq. (57) over t from 0 to $D-1$ yield

$$\begin{aligned}
\mathbb{E}\|y_k^D-y^*(x_k)\|^2&\leq\left(\frac{L-\mu}{L+\mu}\right)^{2D}\mathbb{E}\|y_k^0-y^*(x_k)\|^2+\frac{\sigma^2}{L\mu S}\\
&=\left(\frac{L-\mu}{L+\mu}\right)^{2D}\mathbb{E}\|y_{k-1}^D-y^*(x_k)\|^2+\frac{\sigma^2}{L\mu S},
\end{aligned}\tag{58}$$

where the last inequality follows from Algorithm 2 that $y_k^0=y_{k-1}^D$. Note that

$$\begin{aligned}
\mathbb{E}\|y_{k-1}^D-y^*(x_k)\|^2\leq&2\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2+2\mathbb{E}\|y^*(x_{k-1})-y^*(x_k)\|^2\\
\overset{(i)}{\leq}&2\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2+\frac{2L^2}{\mu^2}\mathbb{E}\|x_k-x_{k-1}\|^2\\
\leq&2\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2+\frac{2\beta^2L^2}{\mu^2}\mathbb{E}\|\widehat{\nabla}\Phi(x_{k-1})\|^2\\
\leq&2\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2+\frac{4\beta^2L^2}{\mu^2}\mathbb{E}\|\nabla\Phi(x_{k-1})\|^2\\
&+\frac{4\beta^2L^2}{\mu^2}\mathbb{E}\|\widehat{\nabla}\Phi(x_{k-1})-\nabla\Phi(x_{k-1})\|^2,
\end{aligned}\tag{59}$$

where (i) follows from Lemma 2.2 in Ghadimi & Wang 2018. Using Lemma 8 in eq. (59) yields

$$\begin{aligned}
&\mathbb{E}\|y_{k-1}^D-y^*(x_k)\|^2\\
&\leq\left(2+\frac{4\beta^2L^2}{\mu^2}\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\right)\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2+\frac{4\beta^2L^2}{\mu^2}\mathbb{E}\|\nabla\Phi(x_{k-1})\|^2\\
&\quad+\frac{4\beta^2L^2}{\mu^2}\left(\frac{4L^2M^2}{\mu^2D_g}+\Big(\frac{8L^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}\right).
\end{aligned}\tag{60}$$

Combining eq. (58) and eq. (60) yields

$$\begin{aligned}
&\mathbb{E}\|y_k^D-y^*(x_k)\|^2\\
&\leq\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\Big(2+\frac{4\beta^2L^2}{\mu^2}\Big(L+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{LM\rho}{\mu^2}\Big)^2\Big)\mathbb{E}\|y_{k-1}^D-y^*(x_{k-1})\|^2\\
&\quad+\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\frac{4\beta^2L^2}{\mu^2}\left(\frac{4L^2M^2}{\mu^2D_g}+\Big(\frac{8L^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2L^4M^2}{\mu^2}\frac{1}{B}+\frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2}\right)\\
&\quad+\frac{4\beta^2L^2}{\mu^2}\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\mathbb{E}\|\nabla\Phi(x_{k-1})\|^2+\frac{\sigma^2}{L\mu S}.
\end{aligned}\tag{61}$$

Based on the definitions of λ, ω, Δ in eq. (54), we obtain from eq. (61) that

$$\mathbb{E}\|y_k^D - y^*(x_k)\|^2 \leq \lambda \mathbb{E}\|y_{k-1}^D - y^*(x_{k-1})\|^2 + \omega\Delta + \frac{\sigma^2}{L\mu S} + \omega \mathbb{E}\|\nabla\Phi(x_{k-1})\|^2. \tag{62}$$

Telescoping eq. (62) over k yields

$$\begin{aligned}
&\mathbb{E}\|y_k^D - y^*(x_k)\|^2 \\
&\leq \lambda^k \mathbb{E}\|y_0^D - y^*(x_0)\|^2 + \omega \sum_{j=0}^{k-1} \lambda^{k-1-j} \mathbb{E}\|\nabla\Phi(x_j)\|^2 + \frac{\omega\Delta + \frac{\sigma^2}{L\mu S}}{1-\lambda} \\
&\leq \lambda^k \left(\left(\frac{L-\mu}{L+\mu}\right)^{2D} \|y_0 - y^*(x_0)\|^2 + \frac{\sigma^2}{L\mu S} \right) + \omega \sum_{j=0}^{k-1} \lambda^{k-1-j} \mathbb{E}\|\nabla\Phi(x_j)\|^2 + \frac{\omega\Delta + \frac{\sigma^2}{L\mu S}}{1-\lambda},
\end{aligned}$$

which completes the proof. □

G.3. Proof of Theorem 3

In this subsection, we provide the proof for Theorem 3, based on the supporting lemmas we develop in Appendix G.2.

Based on the smoothness of the function $\Phi(x)$ in Lemma 2, we have

$$\begin{aligned}
\Phi(x_{k+1}) &\leq \Phi(x_k) + \langle \nabla\Phi(x_k), x_{k+1} - x_k \rangle + \frac{L_\Phi}{2}\|x_{k+1} - x_k\|^2 \\
&\leq \Phi(x_k) - \beta\langle \nabla\Phi(x_k), \widehat{\nabla}\Phi(x_k)\rangle + \beta^2 L_\Phi \|\nabla\Phi(x_k)\|^2 + \beta^2 L_\Phi \|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2.
\end{aligned}$$

For simplicity, let $\mathbb{E}_k = \mathbb{E}(\cdot \,|\, x_k, y_k^D)$. Note that we choose $\beta = \frac{1}{4L_\phi}$. Then, taking expectation over the above inequality, we have

$$\begin{aligned}
\mathbb{E}\Phi(x_{k+1}) \leq& \mathbb{E}\Phi(x_k) - \beta\mathbb{E}\langle \nabla\Phi(x_k), \mathbb{E}_k\widehat{\nabla}\Phi(x_k)\rangle + \beta^2 L_\Phi \mathbb{E}\|\nabla\Phi(x_k)\|^2 \\
&+ \beta^2 L_\Phi \mathbb{E}\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2 \\
\overset{(i)}{\leq}& \mathbb{E}\Phi(x_k) + \frac{\beta}{2}\mathbb{E}\|\mathbb{E}_k\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\|^2 - \frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2 \\
\overset{(ii)}{\leq}& \mathbb{E}\Phi(x_k) - \frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \frac{\beta L^2 M^2 (1-\eta\mu)^{2Q}}{\mu^2} \\
&+ \frac{\beta}{4}\left(\frac{4L^2M^2}{\mu^2 D_g} + \Big(\frac{8L^2}{\mu^2} + 2\Big)\frac{M^2}{D_f} + \frac{16\eta^2 L^4 M^2}{\mu^2}\frac{1}{B} + \frac{16L^2M^2(1-\eta\mu)^{2Q}}{\mu^2} \right) \\
&+ \frac{5\beta}{4}\Big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2 \mathbb{E}\|y_k^D - y^*(x_k)\|^2
\end{aligned} \tag{63}$$

where (i) follows from Cauchy-Schwarz inequality, and (ii) follows from Lemma 7 and Lemma 8. For simplicity, let

$$\nu = \frac{5}{4}\Big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2. \tag{64}$$

Then, applying Lemma 9 in eq. (63) and using the definitions of ω, Δ, λ in eq. (54), we have

$$\begin{aligned}
\mathbb{E}\Phi(x_{k+1}) \leq& \mathbb{E}\Phi(x_k) - \frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \frac{\beta L^2 M^2 (1-\eta\mu)^{2Q}}{\mu^2} \\
&+ \frac{\beta}{4}\Delta + \beta\nu\lambda^k \left(\left(\frac{L-\mu}{L+\mu}\right)^{2D} \|y_0 - y^*(x_0)\|^2 + \frac{\sigma^2}{L\mu S} \right) \\
&+ \beta\nu\omega \sum_{j=0}^{k-1} \lambda^{k-1-j}\mathbb{E}\|\nabla\Phi(x_j)\|^2 + \frac{\beta\nu(\omega\Delta + \frac{\sigma^2}{L\mu S})}{1-\lambda}.
\end{aligned}$$

Telescoping the above inequality over k from 0 to $K-1$ yields

$$\begin{aligned}\mathbb{E}\Phi(x_K) \leq \Phi(x_0) -& \frac{\beta}{4}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \beta\nu\omega\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\lambda^{k-1-j}\mathbb{E}\|\nabla\Phi(x_j)\|^2 \\ &+ \frac{K\beta\Delta}{4} + \Big(\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\|y_0 - y^*(x_0)\|^2 + \frac{\sigma^2}{L\mu S}\Big)\frac{\beta\nu}{1-\lambda} \\ &+ \frac{K\beta L^2M^2(1-\eta\mu)^{2Q}}{\mu^2} + \frac{K\beta\nu(\omega\Delta + \frac{\sigma^2}{L\mu S})}{1-\lambda},\end{aligned}$$

which, using the fact that

$$\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\lambda^{k-1-j}\mathbb{E}\|\nabla\Phi(x_j)\|^2 \leq \left(\sum_{k=0}^{K-1}\lambda^k\right)\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 < \frac{1}{1-\lambda}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2,$$

yields

$$\begin{aligned}&\Big(\frac{1}{4} - \frac{\nu\omega}{1-\lambda}\Big)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 \\ &\leq \frac{\Phi(x_0) - \inf_x\Phi(x)}{\beta K} + \frac{\nu\big((\frac{L-\mu}{L+\mu})^{2D}\|y_0 - y^*(x_0)\|^2 + \frac{\sigma^2}{L\mu S}\big)}{K(1-\lambda)} + \frac{\Delta}{4} + \frac{L^2M^2(1-\eta\mu)^{2Q}}{\mu^2} \\ &\quad + \frac{\nu(\omega\Delta + \frac{\sigma^2}{L\mu S})}{1-\lambda}.\end{aligned} \tag{65}$$

We choose the number D of inner-loop steps as

$$D \geq \max\left\{\frac{\log\big(12 + \frac{48\beta^2L^2}{\mu^2}(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2})^2\big)}{2\log(\frac{L+\mu}{L-\mu})}, \frac{\log\big(\sqrt{\beta}(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2})\big)}{\log(\frac{L+\mu}{L-\mu})}\right\}.$$

Then, since $\beta = \frac{1}{4L_\Phi}$ and $D \geq \frac{\log\big(12 + \frac{48\beta^2L^2}{\mu^2}(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2})^2\big)}{2\log(\frac{L+\mu}{L-\mu})}$, we have $\lambda \leq \frac{1}{6}$, and eq. (65) is further simplified to

$$\begin{aligned}&\Big(\frac{1}{4} - \frac{6}{5}\nu\omega\Big)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 \\ &\leq \frac{\Phi(x_0) - \inf_x\Phi(x)}{\beta K} + \frac{2\nu\big((\frac{L-\mu}{L+\mu})^{2D}\|y_0 - y^*(x_0)\|^2 + \frac{\sigma^2}{L\mu S}\big)}{K} + \frac{\Delta}{4} + \frac{L^2M^2(1-\eta\mu)^{2Q}}{\mu^2} \\ &\quad + 2\nu\Big(\omega\Delta + \frac{\sigma^2}{L\mu S}\Big).\end{aligned} \tag{66}$$

By the definitions of ω in eq. (54) and ν in eq. (64) and $D \geq \frac{\log\big(12 + \frac{48\beta^2L^2}{\mu^2}(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2})^2\big)}{2\log(\frac{L+\mu}{L-\mu})}$, we have

$$\begin{aligned}\nu\omega =& \frac{5\beta^2L^2}{\mu^2}\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\Big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2 \\ <& \frac{\frac{5\beta^2L^2}{\mu^2}\Big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2}{12 + \frac{48\beta^2L^2}{\mu^2}(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2})^2} \leq \frac{5}{48}.\end{aligned} \tag{67}$$

In addition, since $D > \frac{\log\big(\sqrt{\beta}\big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\big)\big)}{\log(\frac{L+\mu}{L-\mu})}$, we have

$$\nu\Big(\frac{L-\mu}{L+\mu}\Big)^{2D} = \frac{5}{4}\Big(\frac{L-\mu}{L+\mu}\Big)^{2D}\Big(L + \frac{L^2}{\mu} + \frac{M\tau}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2 < \frac{5}{4\beta}. \tag{68}$$

Substituting eq. (67) and eq. (68) in eq. (66) yields

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 \leq& \frac{8(\Phi(x_0)-\inf_x\Phi(x)+\frac{5}{2}\|y_0-y^*(x_0)\|^2)}{\beta K}+\Big(1+\frac{1}{K}\Big)\frac{16\nu\sigma^2}{L\mu S} \\ &+\frac{11}{3}\Delta+\frac{8L^2M^2}{\mu^2}(1-\eta\mu)^{2Q},\end{aligned}$$

which, in conjunction with eq. (54) and eq. (64), yields eq. (9) in Theorem 3.

Then, based on eq. (9), in order to achieve an ϵ-accurate stationary point, i.e., $\mathbb{E}\|\nabla\Phi(\bar{x})\|^2\leq\epsilon$ with $\bar{x}$ chosen from $x_0,...,x_{K-1}$ uniformly at random, it suffices to choose

$$\begin{aligned}K =&\frac{32L_\Phi(\Phi(x_0)-\inf_x\Phi(x)+\frac{5}{2}\|y_0-y^*(x_0)\|^2)}{\epsilon}=\mathcal{O}\Big(\frac{\kappa^3}{\epsilon}\Big), D=\Theta(\kappa) \\ Q =&\kappa\log\frac{\kappa^2}{\epsilon}, S=\mathcal{O}\Big(\frac{\kappa^5}{\epsilon}\Big), D_g=\mathcal{O}\left(\frac{\kappa^2}{\epsilon}\right), D_f=\mathcal{O}\left(\frac{\kappa^2}{\epsilon}\right), B=\mathcal{O}\left(\frac{\kappa^2}{\epsilon}\right).\end{aligned}$$

Note that the above choices of Q and B satisfy the condition that $B\geq\frac{1}{Q(1-\eta\mu)^{Q-1}}$ required in Proposition 3.

Then, the gradient complexity is given by $\mathrm{Gc}(F,\epsilon)=KD_f=\mathcal{O}(\kappa^5\epsilon^{-2})$, $\mathrm{Gc}(G,\epsilon)=KDS=\mathcal{O}(\kappa^9\epsilon^{-2})$. In addition, the Jacobian- and Hessian-vector product complexities are given by $\mathrm{JV}(G,\epsilon)=KD_g=\mathcal{O}(\kappa^5\epsilon^{-2})$ and

$$\mathrm{HV}(G,\epsilon)=K\sum_{j=1}^{Q}BQ(1-\eta\mu)^{j-1}=\frac{KBQ}{\eta\mu}\leq\mathcal{O}\left(\frac{\kappa^6}{\epsilon^2}\log\frac{\kappa^2}{\epsilon}\right).$$

Then, the proof is complete.

H. Proof of Theorem 4

To prove Theorem 4, we first establish the following lemma to characterize the estimation variance $\mathbb{E}_{\mathcal{B}}\big\|\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}-\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}\big\|^2$, where $\widetilde{w}_k^D$ is the output of D inner-loop steps of gradient descent at the k^{th} outer loop.

Lemma 10. *Suppose Assumptions 2 and 3 are satisfied and suppose each task loss $\mathcal{L}_{\mathcal{S}_i}(\phi,w_i)$ is μ-strongly-convex w.r.t. w_i. Then, we have*

$$\mathbb{E}_{\mathcal{B}}\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}-\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}\Big\|^2\leq\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{|\mathcal{B}|}.$$

Proof. Let $\widetilde{w}_k^D=(w_{1,k}^D,...,w_{m,k}^D)$ be the output of D inner-loop steps of gradient descent at the k^{th} outer loop. Using Proposition 2, we have, for task $\mathcal{T}_i$,

$$\begin{aligned}\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i,k}^D)}{\partial\phi_k}\Big\|\leq&\|\nabla_\phi\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i,k}^D)\| \\ &+\Big\|\alpha\sum_{t=0}^{D-1}\nabla_\phi\nabla_{w_i}\mathcal{L}_{\mathcal{S}_i}(\phi_k,w_{i,k}^t)\prod_{j=t+1}^{D-1}(I-\alpha\nabla_{w_i}^2\mathcal{L}_{\mathcal{S}_i}(\phi_k,w_{i,k}^j))\nabla_{w_i}\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i.k}^D)\Big\| \\ &\overset{(i)}{\leq}M+\alpha LM\sum_{t=0}^{D-1}(1-\alpha\mu)^{D-t-1}=M+\frac{LM}{\mu},\end{aligned} \tag{69}$$

where (i) follows from Assumptions 2 and strong-convexity of $\mathcal{L}_{\mathcal{S}_i}(\phi,\cdot)$. Then, using the definition of $\mathcal{L}_{\mathcal{D}}(\phi,\widetilde{w};\mathcal{B})=$

$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathcal{L}_{\mathcal{D}_i}(\phi,w_i)$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{B}}\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}-\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}\Big\|^2 =&\frac{1}{|\mathcal{B}|}\mathbb{E}_i\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i,k}^D)}{\partial\phi_k}-\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}\Big\|^2\\ \overset{(i)}{\leq}&\frac{1}{|\mathcal{B}|}\mathbb{E}_i\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i,k}^D)}{\partial\phi_k}\Big\|^2\\ \overset{(ii)}{\leq}&\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{|\mathcal{B}|}. \end{aligned}\tag{70}$$

where (i) follows from $\mathbb{E}_i\frac{\partial\mathcal{L}_{\mathcal{D}_i}(\phi_k,w_{i,k}^D)}{\partial\phi_k}=\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}$ and (ii) follows from eq. (69). Then, the proof is complete. □

Proof of Theorem 4. Recall $\Phi(\phi):=\mathcal{L}_{\mathcal{D}}(\phi,\widetilde{w}^*(\phi))$ be the objective function, and let $\widehat{\nabla}\Phi(\phi_k)=\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)}{\partial\phi_k}$. Using an approach similar to eq. (40), we have

$$\begin{aligned}\Phi(\phi_{k+1})\leq&\Phi(\phi_k)+\langle\nabla\Phi(\phi_k),\phi_{k+1}-\phi_k\rangle+\frac{L_\Phi}{2}\|\phi_{k+1}-\phi_k\|^2\\ \leq&\Phi(\phi_k)-\beta\Big\langle\nabla\Phi(\phi_k),\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}\Big\rangle+\frac{\beta^2L_\Phi}{2}\Big\|\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}\Big\|^2.\end{aligned}\tag{71}$$

Taking the expectation of eq. (71) yields

$$\begin{aligned}\mathbb{E}\Phi(\phi_{k+1})\overset{(i)}{\leq}&\mathbb{E}\Phi(\phi_k)-\beta\mathbb{E}\big\langle\nabla\Phi(\phi_k),\widehat{\nabla}\Phi(\phi_k)\big\rangle+\frac{\beta^2L_\Phi}{2}\mathbb{E}\|\widehat{\nabla}\Phi(\phi_k)\|^2\\ &+\frac{\beta^2L_\Phi}{2}\mathbb{E}\Big\|\widehat{\nabla}\Phi(\phi_k)-\frac{\partial\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})}{\partial\phi_k}\Big\|^2\\ \overset{(ii)}{\leq}&\mathbb{E}\Phi(\phi_k)-\beta\mathbb{E}\big\langle\nabla\Phi(\phi_k),\widehat{\nabla}\Phi(\phi_k)\big\rangle+\frac{\beta^2L_\Phi}{2}\mathbb{E}\|\widehat{\nabla}\Phi(\phi_k)\|^2+\frac{\beta^2L_\Phi}{2}\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{|\mathcal{B}|}\\ \leq&\mathbb{E}\Phi(\phi_k)-\Big(\frac{\beta}{2}-\beta^2L_\Phi\Big)\mathbb{E}\|\nabla\Phi(\phi_k)\|^2+\Big(\frac{\beta}{2}+\beta^2L_\Phi\Big)\mathbb{E}\|\nabla\Phi(\phi_k)-\widehat{\nabla}\Phi(\phi_k)\|^2\\ &+\frac{\beta^2L_\Phi}{2}\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{|\mathcal{B}|},\end{aligned}\tag{72}$$

where (i) follows from $\mathbb{E}_{\mathcal{B}}\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D;\mathcal{B})=\mathcal{L}_{\mathcal{D}}(\phi_k,\widetilde{w}_k^D)$ and (ii) follows from Lemma 10. Using Lemma 6 in eq. (72) and rearranging the terms, we have

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\Big(\frac{1}{2}-\beta L_\Phi\Big)\mathbb{E}\|\nabla\Phi(\phi_k)\|^2\\ &\leq\frac{\Phi(\phi_0)-\inf_\phi\Phi(\phi)}{\beta K}+3\Big(\frac{1}{2}+\beta L_\Phi\Big)\frac{L^2M^2(1-\alpha\mu)^{2D}}{\mu^2}+\frac{\beta L_\Phi}{2}\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{|\mathcal{B}|}\\ &\quad+3\Delta\Big(\frac{1}{2}+\beta L_\Phi\Big)\Big(\frac{L^2(L+\mu)^2}{\mu^2}(1-\alpha\mu)^D+\frac{4M^2\left(\tau\mu+L\rho\right)^2}{\mu^4}(1-\alpha\mu)^{D-1}\Big),\end{aligned}$$

where $\Delta=\max_k\|\widetilde{w}_k^0-\widetilde{w}^*(\phi_k)\|^2<\infty$. Choose the same parameters β,D as in Theorem 2. Then, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\phi_k)\|^2\leq\frac{16L_\Phi(\Phi(\phi_0)-\inf_\phi\Phi(\phi))}{K}+\frac{2\epsilon}{3}+\Big(1+\frac{L}{\mu}\Big)^2\frac{M^2}{8|\mathcal{B}|}.$$

Then, the proof is complete. □