A Near-Optimal Algorithm for Stochastic Bilevel Optimization via Double-Momentum

Prashant Khanduri
University of Minnesota
khand095@umn.edu

Siliang Zeng
University of Minnesota
zeng0176@umn.edu

Mingyi Hong[*]
University of Minnesota
mhong@umn.edu

Hoi-To Wai
CUHK
htwai@se.cuhk.edu.hk

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Zhuoran Yang
Princeton University
zy6@princeton.edu

Abstract

This work proposes a new algorithm – the Single-timescale Double-momentum Stochastic Approximation (SUSTAIN) – for tackling stochastic unconstrained bilevel optimization problems. We focus on bilevel problems where the lower level subproblem is strongly-convex and the upper level objective function is smooth. Unlike prior works which rely on *two-timescale* or *double loop* techniques, we design a stochastic momentum-assisted gradient estimator for both the upper and lower level updates. The latter allows us to control the error in the stochastic gradient updates due to inaccurate solution to both subproblems. If the upper objective function is smooth but possibly non-convex, we show that SUSTAIN requires $\mathcal{O}(\epsilon^{-3/2})$ iterations (each using $\mathcal{O}(1)$ samples) to find an ϵ-stationary solution. The ϵ-stationary solution is defined as the point whose squared norm of the gradient of the outer function is less than or equal to ϵ. The total number of stochastic gradient samples required for the upper and lower level objective functions match the best-known complexity for single-level stochastic gradient algorithms. We also analyze the case when the upper level objective function is strongly-convex.

1 Introduction

Many learning and inference problems take a “hierarchical" form, wherein the optimal solution of one problem affects the objective function of others [27]. Bilevel optimization is often used to model problems of this kind with two levels of hierarchy [27, 8], where the variables of an *upper level* problem depend on the optimizer of certain *lower level* problem. In this work, we consider unconstrained bilevel optimization problems of the form:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_{\mathsf{up}}}} \ \ell(x) = f(x, y^*(x)) := \mathbb{E}_\xi[f(x, y^*(x); \xi)] \\ &\text{s.t.} \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d_{\mathsf{lo}}}} \big\{g(x,y) := \mathbb{E}_\zeta[g(x,y;\zeta)]\big\}, \end{aligned} \tag{1}$$

where $f, g : \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}} \to \mathbb{R}$ with $x \in \mathbb{R}^{d_{\mathsf{up}}}$ and $y \in \mathbb{R}^{d_{\mathsf{lo}}}$; $f(x, y; \xi)$ with $\xi \sim \pi_f$ (resp. $g(x, y; \zeta)$ with $\zeta \sim \pi_g$) represents a stochastic sample of the upper level objective (resp. lower level objective). Note here that the *upper level objective* f depends on the minimizer of the *lower level objective* g, and we refer to $\ell(x)$ as the *outer function*. Throughout this paper, $g(x, y)$ is assumed to be strongly-convex in y, which implies that $\ell(x)$ is smooth but possibly non-convex.

The applications of (1) include many machine learning problems that have a hierarchical structure. Examples are meta learning [13, 31], data hyper-cleaning [35], hyper-parameter optimization [12,

[*] Corresponding Author: Mingyi Hong.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Algorithm	Sample (Upper, Lower)	Implementation	Batch Size	Per-Iteration Complexity
BSA [14]	$\mathcal{O}(\epsilon^{-2}),\ \mathcal{O}(\epsilon^{-3})$	Double loop	$\mathcal{O}(1)$	$\mathcal{O}(d_{\sf lo}^2 \cdot \log T)$
stocBiO [19]	$\mathcal{O}(\epsilon^{-2}),\ \mathcal{O}(\epsilon^{-2})$	Double loop	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(d_{\sf lo}^2 \cdot \log T)$
TTSA [18]	$\mathcal{O}(\epsilon^{-5/2}),\ \mathcal{O}(\epsilon^{-5/2})$	Single loop	$\mathcal{O}(1)$	$\mathcal{O}(d_{\sf lo}^2 \cdot \log T)$
STABLE [5]	$\mathcal{O}(\epsilon^{-2}),\ \mathcal{O}(\epsilon^{-2})$	Single loop	$\mathcal{O}(1)$	$\mathcal{O}(d_{\sf lo}^3)$
SVRB [17]	$\mathcal{O}(\epsilon^{-3/2}),\ \mathcal{O}(\epsilon^{-3/2})$	Single loop	$\mathcal{O}(1)$	$\mathcal{O}(d_{\sf lo}^3)$
SUSTAIN (this work)	$\mathcal{O}(\epsilon^{-3/2}),\ \mathcal{O}(\epsilon^{-3/2})$	Single loop	$\mathcal{O}(1)$	$\mathcal{O}(d_{\sf lo}^2 \cdot \log T)$

Table 1: Comparison of the number of upper and lower level gradient samples required to achieve an ϵ-stationary point in Definition 1.1. For the algorithms with $\mathcal{O}(d_{\sf lo}^2 \cdot \log T)$ per-iteration dependence, the Hessian inverse can be computed via matrix vector products; algorithms with $\mathcal{O}(d_{\sf lo}^3)$ dependency requires Hessian inverses and Hessian projections, which incur heavy computational cost.

13, 29], and reinforcement learning [22], etc.. To better contextualize our study, below we describe examples on meta-learning problem and data hyper-cleaning problem:

Example 1: Meta learning. The meta learning problem aims to learn task specific parameters that generalize to a diverse set of tasks [30]. Suppose we have M tasks $\{\mathcal{T}_i, i = 1, \dots, M\}$ and each task has a corresponding loss function $L(x, y_i; \xi_i)$ with ξ_i representing a data sample for task $\mathcal{T}_i$, $x \in \mathbb{R}^{d_{\sf up}}$ the model parameters shared among tasks, and $y_i \in \mathbb{R}^{d_{\sf lo}^i}$ the task specific parameters. The goal of meta learning is then to solve the following problem:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_{\sf up}}} \ \Big\{L_{\sf ts}(x,\bar{y}^*(x)) := \tfrac{1}{M}\textstyle\sum_{i=1}^M \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[L(x,y_i^*(x);\xi_i)]\Big\} \\ &\text{s.t. } \ \bar{y}^*(x) \in \arg\min_{\bar{y}\in\mathbb{R}^{\sum_{i=1}^M d_{\sf lo}^i}} L_{\sf tr}(x,\bar{y}) := \tfrac{1}{M}\textstyle\sum_{i=1}^M \big(\mathbb{E}_{\zeta_i\sim\mathcal{S}_i}[L(x,y_i;\zeta_i)] + \mathcal{R}(y_i)\big), \end{aligned} \tag{2}$$

where $\bar{y} = [y_1^T, \dots, y_M^T]^T$, $\mathcal{R}(\cdot)$ is a strongly convex regularizer while $\mathcal{S}_i$ and $\mathcal{D}_i$ are the training and testing datasets for task $\mathcal{T}_i$. Compared to the number of tasks, the dataset sizes are usually small for meta-learning problems, so the stochasticity in tackling (2) results from the fact that at each iteration we can only sample a subset m out of M tasks. Note that this problem is a special case of (1). □

Example 2: Data hyper-cleaning. The data hyper-cleaning is a hyperparameter optimization problem that aims to train a classifier model with a dataset of randomly corrupted labels [35]. The optimization problem is formulated below:

$$\begin{aligned} \min_{x\in\mathbb{R}^{d_{\sf up}}} \ & \ell(x) := \textstyle\sum_{i\in\mathcal{D}_{\sf val}} L(a_i^\top y^*(x), b_i) \\ \text{s.t. } \ & y^*(x) = \arg\min_{y\in\mathbb{R}^{d_{\sf lo}}} \big\{c\|y\|^2 + \textstyle\sum_{i\in\mathcal{D}_{\sf tr}} \sigma(x_i) L(a_i^\top y, b_i)\big\}. \end{aligned} \tag{3}$$

In this problem, we have $d_{\sf up} = |\mathcal{D}_{\rm tr}|$ and $d_{\sf lo}$ is the dimension of the classifier. Moreover, (a_i, b_i) is the ith data point; $L(\cdot)$ is the loss function, with y being the model parameter; x_i is the parameter that determines the weight for the ith data sample, and $\sigma : \mathbb{R} \to \mathbb{R}_+$ is the weight function; $c > 0$ is a regularization parameter; $\mathcal{D}_{\rm val}$ and $\mathcal{D}_{\rm tr}$ are validation and training sets, respectively. Clearly, (3) is a special case of (1) where the lower level problem finds the classifier $y^*(x)$ with the training set $\mathcal{D}_{\sf tr}$, and the upper level problem finds the best weights x with respect to the validation set $\mathcal{D}_{\sf val}$. □

A natural approach to tackling (1) is to apply alternating stochastic gradient (SG) updates. Let $\beta, \alpha > 0$ be some step sizes, one performs the recursion

$$y^+ \leftarrow y - \beta \hat{\nabla}_y g(x,y), \quad x^+ \leftarrow x - \alpha \hat{\nabla}_x \hat{\ell}(x;y) \tag{4}$$

such that $\hat{\nabla}_y g(x,y)$, $\hat{\nabla}_x \hat{\ell}(x;y)$ are stochastic estimates of $\nabla_y g(x,y)$, $\nabla \ell(x)$, respectively. Notice that (4) is significantly different from the standard alternating primal-dual gradient algorithm for saddle point problems. Particularly, the design of $\hat{\nabla}_x \hat{\ell}(x;y)$ is crucial to the SG scheme in (4). Observe that $\nabla \ell(x)$ can be computed using the implicit function theorem, and its evaluation requires $f(\cdot,\cdot)$ and $y^\star(x)$, the minimizer of $g(x,y)$ given x (cf. (5)). This gives rise to a unique challenge to bilevel optimization, where $y^\star(x)$ can only be *approximated* by y obtained in the first relation of (4).

In light of the above observations, previous endeavors have considered two approaches to improve the estimate of $y^\star(x)$ while $\hat{\nabla}_x \hat{\ell}(x;y)$ is used as a *biased* approximation of $\nabla \ell(x)$. The first approach is to apply the *double-loop* algorithms. For example, [14] proposed to repeat the y^+ update for multiple times to obtain a better estimate of $y^\star(x)$ before performing the x^+ update, [19] proposed to take a large batch size to estimate $\nabla_y g(x, y)$. While simple to analyze, these algorithms may suffer from a poor sample complexity for the inner problem. The second approach is to apply *single-loop* algorithms where the y^+-updates are performed simultaneously with the x^+-updates. Instead, advanced techniques are utilized that allows y^+ to accurately track $y^\star(x)$. For example, [18] suggested to tune the step size schedule with $\beta \gg \alpha$, [5, 17] proposed single-timescale algorithms with advanced variance reduction techniques. However, the latter two algorithms require Hessian projections onto a compact set along with Hessian matrices inversion which scales poorly with dimension (i.e., in $\mathcal{O}(d_{\sf lo}^3)$). We summarize and compare the complexity results of the state-of-the-art algorithms in Table 1.

A careful inspection on the above results reveals a gap in the iteration/sample complexity compared to *single-level* stochastic optimization. For instance, an optimal stochastic gradient algorithm finds an ϵ-stationary solution [cf. Definition 1.1] to $\min_x \mathbb{E}_\xi[\ell(x; \xi)]$ in $\mathcal{O}(\epsilon^{-3/2})$ iterations [10, 7, 37, 42]. For bilevel optimization, the fastest rate available is only $\mathcal{O}(\epsilon^{-2})$ to the best of the authors' knowledge. In comparison, the proposed algorithm achieves a rate of $\mathcal{O}(\epsilon^{-3/2})$. During the preparation of the current paper, a preprint [17] has appeared which extended [5], and achieves an improved rate of $\mathcal{O}(\epsilon^{-3/2})$. We remark that the latter work follows a different design philosophy from ours and maybe less efficient; see the detailed discussion at the end of Sec. 3.

Contributions. In this paper, we depart from the prior developments which focused on finding better inner solutions $y^*(x)$ to approximate $\hat{\nabla}_x \hat{\ell}(x; y) \approx \nabla \ell(x)$. Our idea is to exploit the gradient estimates from prior iterations to improve the quality of the current gradient estimation. This leads to *momentum*-assisted stochastic gradient estimators for *both* $\nabla_y g(x, y)$ and $\nabla \ell(x)$ using similar techniques in [7, 37] for single-level stochastic optimization. The resultant algorithm only requires $O(1)$ samples at each update, and updates x and y using step sizes of the same order, hence the name $\underline{\text{s}}$ingle-timescale do$\underline{\text{u}}$ble-momentum $\underline{\text{s}}$tochastic $\underline{\text{a}}$pprox$\underline{\text{i}}$matio$\underline{\text{n}}$(SUSTAIN) algorithm. Additionally, it is worth noting that our algorithm has a $\mathcal{O}(d_{\sf lo}^2)$ per iteration complexity, compared to the $\mathcal{O}(d_{\sf lo}^3)$ complexity of STABLE [5] and SVRB [17]. That is, the SUSTAIN algorithm is both *sample and computation* efficient. Our specific contributions are:

- We propose the SUSTAIN algorithm for bilevel problems which matches the best complexity bounds as the optimal SGD algorithms for single-level stochastic optimization. That is, it requires $\mathcal{O}(\epsilon^{-3/2})$ [resp. $\mathcal{O}(\epsilon^{-1})$] samples to find an ϵ-stationary solution for non-convex (resp. strongly-convex) bilevel problems; see Table 1. Furthermore, the algorithm utilizes a single-loop update with step sizes of the same order for both upper and lower level problems. Such complexity bounds match the optimal sample complexity of stochastic gradient algorithms for single-level problems.
- By developing the Lipschitz continuous property of the (biased) stochastic estimates of $\nabla \ell(x)$, we show that obtaining a good estimate of $\nabla \ell(x)$ does not require explicit (sampled) Hessian inversion. This key result ensures that our algorithm depends favorably on the problem dimension.
- Comparing with prior works such as TTSA [18], BSA [14], STABLE [5] and SVRB [17], our analysis reveals that improving the gradient estimation quality for *both* $\nabla_y g(x, y)$ and $\nabla \ell(x)$ is the key to obtain a sample and computation efficient stochastic algorithm for bilevel optimization.

Related works. The study of the bilevel problem (1) can be traced to that of game theory [36] and was formally introduced in [2–4]. It is also related to the broader class of problems of Mathematical Programming with Equilibrium Constraints [26]. Related algorithms include approximate descent [9, 38], and penalty-based methods [40]; see [6] and [25] for a comprehensive survey.

In addition to the works cited in Table 1, recent works on bilevel optimization have focused on algorithms with provable convergence rates. In [34], the authors proposed BigSAM algorithm for solving simple bilevel problems (with a single variable) with convex lower level problem. Subsequently, the works [24, 23] utilized BigSAM and developed algorithms for a general bilevel problem for the cases when the solution of the lower level problem is not a singleton. Note that all the aforementioned works [34, 24, 23] assumed the upper level problem to be strongly-convex with convex lower level problem. In a separate line of work, backpropagation based algorithms have been proposed to approximately solve bilevel problems [12, 35, 16, 15]. However, the major focus of these works was to develop efficient gradient estimators rather than on developing efficient optimization algorithms.

Notation. For any $x \in \mathbb{R}^d$, we denote $\|x\|$ as the standard Euclidean norm; as for $X \in \mathbb{R}^{n \times d}$, $\|X\|$ is induced by the Euclidean norm. For a multivariate function $f(x, y)$, the notation $\nabla_x f(x, y)$ [resp. $\nabla_y f(x, y)$] refers to the partial gradient taken with respect to (w.r.t.) x [resp. y]. For some $\mu >$

0, a function $f(x, y)$ is said to be μ-strongly-convex in x if $f(x, y) - \frac{\mu}{2}\|x\|^2$ is convex in x. For some $L > 0$, the map $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is said to be L-Lipschitz continuous if $\|\mathcal{A}(x) - \mathcal{A}(y)\| \le L\|x - y\|$ for any $x, y \in \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth if its gradient is L-Lipschitz continuous. Uniform distribution over a discrete set $\{1, \ldots, T\}$ is represented by $\mathcal{U}\{1, \ldots, T\}$.

Finally, we state the following definitions for the optimality criteria of (1).

Definition 1.1 (ϵ-Stationary Point)**.** A point x is called ϵ-stationary if $\|\nabla \ell(x)\|^2 \le \epsilon$. A stochastic algorithm is said to achieve an ϵ-stationary point in t iterations if $\mathbb{E}[\|\nabla \ell(x_t)\|^2] \le \epsilon$, where the expectation is over the stochasticity of the algorithm until time instant t.

Definition 1.2 (ϵ-Optimal Point)**.** A point x is called ϵ-optimal if $\ell(x) - \ell^* \le \epsilon$, where $\ell^* := \min_{x \in \mathbb{R}^{d_{\mathsf{up}}}} \ell(x)$. A stochastic algorithm is said to achieve an ϵ-optimal point in t iterations if $\mathbb{E}[\ell(x_t) - \ell^*] \le \epsilon$, where the expectation is over the stochasticity of the algorithm until time instant t.

2 Preliminaries

We discuss the assumptions on (1) to specify the problem class of interest. We also preface the proposed algorithm by describing a practical procedure for estimating the stochastic gradients.

Assumption 1 (Upper Level Function)**.** $f(x, y)$ satisfies the following conditions:

(i) $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ are Lipschitz continuous w.r.t. $(x, y) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$, and with constants $L_{f_x} \ge 0$ and $L_{f_y} \ge 0$, respectively.

(ii) For any $(x, y) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$, we have $\|\nabla_y f(x, y)\| \le C_{f_y}$, for some $C_{f_y} \ge 0$.

Assumption 2 (Lower level Function)**.** $g(x, y)$ satisfies the following conditions:

(i) For any $x \in \mathbb{R}^{d_{\mathsf{up}}}$ and $y \in \mathbb{R}^{d_{\mathsf{lo}}}$, $g(x, y)$ is twice continuously differentiable in (x, y).

(ii) $\nabla_y g(x, y)$ is Lipschitz continuous w.r.t. $(x, y) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$, and with constant $L_g \ge 0$.

(iii) For any $x \in \mathbb{R}^{d_{\mathsf{up}}}$, $g(x, \cdot)$ is μ_g-strongly-convex in y for some $\mu_g > 0$.

(iv) $\nabla^2_{xy} g(x, y)$ and $\nabla^2_{yy} g(x, y)$ are Lipschitz continuous w.r.t. $(x, y) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$, and with constants $L_{g_{xy}} \ge 0$ and $L_{g_{yy}} \ge 0$, respectively.

(v) For any $(x, y) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$, we have $\|\nabla^2_{xy} g(x, y)\|^2 \le C_{g_{xy}}$ for some $C_{g_{xy}} > 0$.

Assumption 3 (Stochastic Functions)**.** Assumptions 1 and 2 hold for $f(x, y; \xi)$ and $g(x, y; \zeta)$, for all $\xi \in \operatorname{supp}(\pi_f)$ and $\zeta \in \operatorname{supp}(\pi_g)$ where $\operatorname{supp}(\pi)$ is the support of π. Moreover, we assume the following variance bounds.

$$\mathbb{E}\big[\|\nabla_x f(x, y) - \nabla_x f(x, y; \xi)\|^2\big] \le \sigma^2_{f_x}, \quad \mathbb{E}\|\nabla_y f(x, y) - \nabla_y f(x, y; \xi)\|^2 \le \sigma^2_{f_y},$$
$$\mathbb{E}\|\nabla^2_{xy} g(x, y) - \nabla^2_{xy} g(x, y; \xi)\|^2 \le \sigma^2_{g_{xy}} \text{ for some } \sigma_{f_x} \ge 0, \sigma_{f_y} \ge 0 \text{ and } \sigma_{g_{xy}} \ge 0.$$

These assumptions are standard in the analysis of bilevel optimization [14]. For example, they are satisfied by a range of applications such as the meta learning problem (2), data hypercleaning problem (3) with linear classifier. Notice that under these assumptions, the gradient $\nabla \ell(\cdot)$ is well-defined. By utilizing Assumption 2–(i) and (ii) along with the implicit function theorem [33], it is easy to show that for a given $\bar{x} \in \mathbb{R}^{d_{\mathsf{up}}}$, the following holds [14, Lemma 2.1]:

$$\nabla \ell(\bar{x}) = \nabla_x f(\bar{x}, y^*(\bar{x})) - \nabla^2_{xy} g(\bar{x}, y^*(\bar{x}))[\nabla^2_{yy} g(\bar{x}, y^*(\bar{x}))]^{-1} \nabla_y f(\bar{x}, y^*(\bar{x})). \tag{5}$$

Obtaining $y^*(x)$ in closed-form is usually a challenging task, so it is natural to use the following gradient surrogate. At any $(\bar{x}, \bar{y}) \in \mathbb{R}^{d_{\mathsf{up}} \times d_{\mathsf{lo}}}$, define:

$$\bar{\nabla} f(\bar{x}, \bar{y}) = \nabla_x f(\bar{x}, \bar{y}) - \nabla^2_{xy} g(\bar{x}, \bar{y})[\nabla^2_{yy} g(\bar{x}, \bar{y})]^{-1} \nabla_y f(\bar{x}, \bar{y}). \tag{6}$$

Evaluating (6) requires computing the exact gradients and Hessian inverse which can be non-trivial. Below, we describe a practical procedure from [14] to generate a *biased* estimate of $\bar{\nabla} f(\bar{x}, \bar{y})$.

Stochastic gradient estimator for $\nabla \ell(x)$. The estimator requires a parameter $K \in \mathbb{N}$ and is based on a collection of $K+3$ independent samples $\bar{\xi} := \{\xi, \zeta^{(0)}, ..., \zeta^{(K)}, \mathsf{k}(K)\}$, where $\xi \sim \mu$, $\zeta^{(i)} \sim \pi_g$, $i = 0, ..., K$, and $\mathsf{k}(K) \sim \mathcal{U}\{0, ..., K-1\}$. We set

$$\bar{\nabla} f(x, y; \bar{\xi}) = \nabla_x f(x, y; \xi) - \frac{K}{L_g} \nabla^2_{xy} g(x, y; \zeta^{(0)}) \prod_{i=1}^{\mathsf{k}(K)} \Big(I - \frac{\nabla^2_{yy} g(x, y; \zeta^{(i)})}{L_g} \Big) \nabla_y f(x, y; \xi), \tag{7}$$

where we have used the convention $\prod_{i=1}^{j} A_i = I$ if $j = 0$. It has been shown in [14, 18] that the bias with the gradient estimator (7) decays exponentially fast with K, as summarized below:

Lemma 2.1. *Under Assumptions 1, 2. For any $K \geq 1$, the gradient estimator in (7) satisfies*

$$\|\bar{\nabla} f(x,y) - \mathbb{E}_{\bar{\xi}}[\bar{\nabla} f(x,y;\bar{\xi})]\| \leq \frac{C_{g_{xy}} C_{f_y}}{\mu_g} \left(1 - \frac{\mu_g}{L_g}\right)^K, \ \forall\, (x,y) \in \mathbb{R}^{d_{\sf up}} \times \mathbb{R}^{d_{\sf lo}}. \tag{8}$$

The detailed statement of the above lemma is included in Appendix C. We remark that each computation of $\bar{\nabla} f(x,y;\bar{\xi})$ requires at most K Hessian-vector products, and later we will show that setting $K = \mathcal{O}(\log(T))$ is necessary for the proposed algorithm. Since $\nabla_{yy}^2 g(x,y;\zeta)$ is of size $d_{\sf lo} \times d_{\sf lo}$, the total complexity of this step is $\mathcal{O}(\log(T) d_{\sf lo}^2)$. On the contrary, STABLE [5] and SVRB [17] require $\mathcal{O}(d_{\sf lo}^3)$ to estimate the Hessian inverse, which is more computationally expensive when $d_{\sf lo} \gg 1$. Indeed, it has been explicitly mentioned in [5] that "*our algorithm (STABLE) is preferable in the regime where the sampling is more costly than computation or the dimension d is relatively small*".

Notice that (7) is not the only option for estimating the gradient surrogate $\bar{\nabla} f(x,y)$. For ease of presentation, below we abstract out the conditions on the stochastic estimates of $\nabla_y g$, $\bar{\nabla} f$ required by our analysis as the following assumption:

Assumption 4 (Stochastic Gradients)**.** For any $(x,y) \in \mathbb{R}^{d_{\sf up}} \times \mathbb{R}^{d_{\sf lo}}$, there exists constants $\sigma_f, \sigma_g \geq 0$ such that the estimates $\nabla_y g(x,y;\zeta)$, $\bar{\nabla} f(x,y;\bar{\xi})$ satisfy:

(i) The gradient estimate of the upper level objective satisfies:

$$\mathbb{E}_{\bar{\xi}}\left[\|\bar{\nabla} f(x,y;\bar{\xi}) - \bar{\nabla} f(x,y) - B(x,y)\|^2\right] \leq \sigma_f^2, \tag{9}$$

where $B(x,y) = \mathbb{E}_{\bar{\xi}}[\bar{\nabla} f(x,y;\bar{\xi})] - \bar{\nabla} f(x,y)$ is the bias in estimating $\bar{\nabla} f(x,y)$.

(ii) The gradient estimate of the lower level objective satisfies

$$\mathbb{E}_{\zeta}\left[\|\nabla_y g(x,y;\zeta) - \nabla_y g(x,y)\|^2\right] \leq \sigma_g^2. \tag{10}$$

As observed from Lemma 2.1, the gradient estimator (7) satisfies Assumption 4(i).

Lastly, the approximate gradient defined in (6), the true gradient (5), as well as the optimal solution of the lower level problem are Lipschitz continuous, as proven below:

Lemma 2.2. *[14, Lemma 2.2] Under Assumptions 1, 2 and 3, we have*

$$\begin{gathered}\|\bar{\nabla} f(x,y) - \nabla \ell(x)\| \leq L\|y^*(x) - y\|, \quad \|y^*(x_1) - y^*(x_2)\| \leq L_y\|x_1 - x_2\|, \\ \|\nabla \ell(x_1) - \nabla \ell(x_2)\| \leq L_f\|x_1 - x_2\|,\end{gathered} \tag{11}$$

for all $x, x_1, x_2 \in \mathbb{R}^{d_{\sf up}}$ and $y \in \mathbb{R}^{d_{\sf lo}}$. The above Lipschitz constants are defined as:

$$L = L_{f_x} + \frac{L_{f_y} C_{g_{xy}}}{\mu_g} + C_{f_y}\left(\frac{L_{g_{xy}}}{\mu_g} + \frac{L_{g_{yy}} C_{g_{xy}}}{\mu_g^2}\right), \quad L_f = L + \frac{L C_{g_{xy}}}{\mu_g}, \quad L_y = \frac{C_{g_{xy}}}{\mu_g}. \tag{12}$$

The first result in (11) reveals that $\bar{\nabla} f(x,y)$ approximates $\nabla \ell(x)$ when $y \approx y^*(x)$. This suggests that a *double-loop* algorithm which solves the strongly-convex lower level problem to sufficient accuracy can be applied to tackle (1). Such approach has been pursued in [14, 19]. Next, we propose an algorithm which rely on *single-loop* updates with improved sample efficiency.

3 The proposed SUSTAIN algorithm

Equipped with a practical stochastic gradient estimator for $\nabla \ell(x)$ [cf. (7)], our next endeavor is to develop a *single-loop* algorithm to tackle (1) through drawing $\mathcal{O}(1)$ samples for upper and lower level problems at each iteration. Our main idea is to adopt the recursive momentum techniques developed in [7, 37]. Notice that these works utilize *unbiased* stochastic gradients evaluated at consecutive iterates to construct a variance reduced gradient estimate for single-level stochastic optimization.

In the context of bilevel stochastic optimization (1), a few key challenges are in order:

Algorithm 1 The Proposed SUSTAIN Algorithm

```
**Input**: Parameters: $\{\beta_t\}_{t=0}^{T-1}$, $\{\alpha_t\}_{t=0}^{T-1}$, $\{\eta_t^f\}_{t=0}^{T-1}$, and $\{\eta_t^g\}_{t=0}^{T-1}$ with $\eta_0^f = \eta_0^g = 1$
**Initialize**: $x_0$, $y_0$; set $x_{-1} = y_{-1} = h_{-1}^f = h_{-1}^g = 0$
**for** $t = 0$ to $T-1$ **do**
    ($y$-update) Compute the gradient estimator $h_t^g$ by (13) and set $y_{t+1} = y_t - \beta_t h_t^g$.
    ($x$-update) Compute the gradient estimator $h_t^f$ by (14) and set $x_{t+1} = x_t - \alpha_t h_t^f$.
**end for**
**Return:** $x_{a(T)}$ where $a(T) \sim \mathcal{U}\{1, ..., T\}$.
```

- Recall from Lemma 2.1 that obtaining an unbiased estimator for the outer gradient $\nabla \ell(x)$ requires using $K \to \infty$ samples in (7), this calls for the new techniques to control the bias arising from approximating $\nabla \ell(x)$.
- The gradient estimator (7) has a more complicated structure than a plain gradient estimator, as it involves up to three different stochastic vectors/matrices related to $\nabla_x f(x, y)$, $\nabla_y f(x, y)$, $\nabla_{xy} g(x, y)$, and one stochastic inversion that is related to $[\nabla_{yy} g(x, y)]^{-1}$. It is not clear which are the most important objects for which variance reduction shall be applied.

Our key innovation is to develop a useful estimate of $\bar\nabla f(x, y)$ by using a novel *double-momentum* technique. First, we build a recursive momentum estimator for $\nabla_y g(x, y)$, based upon which the variable y gets updated. Then, with such a "stabilized" inner iteration, we compute an estimate of $\bar\nabla f(x, y)$ as given in (7), by using the four stochastic vectors/matrices mentioned above but without performing any variance reduction. Such a stochastic estimator will then be used to construct a recursive momentum estimator for $\bar\nabla f(x, y)$. The intuition is that as long as y is *accurate enough*, then the stochastic terms in (7) are also accurate enough, so they can be used to construct the estimator for the outer gradient. Our approach only tracks two vector estimators, while still being able to leverage the low-complexity sample-based Hessian inversion as given in (7).

The SUSTAIN algorithm is summarized in Algorithm 1. Define $\eta_t^g \in [0, 1]$, $\eta_t^f \in [0, 1]$. For the lower level problem involving y, it utilizes the following momentum-assisted gradient estimator, $h_t^g \in \mathbb{R}^{d_{\mathsf{lo}}}$, defined recursively as

$$h_t^g = \eta_t^g \nabla_y g(x_t, y_t; \zeta_t) + (1 - \eta_t^g)\big(h_{t-1}^g + \nabla_y g(x_t, y_t; \zeta_t) - \nabla_y g(x_{t-1}, y_{t-1}; \zeta_t)\big); \tag{13}$$

For the upper level problem involving x, we utilize a similar estimate, $h_t^f \in \mathbb{R}^{d_{\mathsf{up}}}$, defined as

$$h_t^f = \eta_t^f \bar\nabla f(x_t, y_t; \bar\xi_t) + (1 - \eta_t^f)\big(h_{t-1}^f + \bar\nabla f(x_t, y_t; \bar\xi_t) - \bar\nabla f(x_{t-1}, y_{t-1}; \bar\xi_t)\big). \tag{14}$$

The gradient estimators h_t^g and h_t^f are computed from the current and past gradient estimates $\nabla_y g(x_t, y_t; \zeta_t)$, $\nabla_y g(x_{t-1}, y_{t-1}; \zeta_t)$ and $\bar\nabla f(x_t, y_t; \bar\xi_t)$, $\bar\nabla f(x_{t-1}, y_{t-1}; \bar\xi_t)$. Note that the stochastic gradients at two consecutive iterates are computed using the same sample sets ζ_t for h_t^g and $\bar\xi_t$ for h_t^f.

Both x and y-update steps mark a major departure of the SUSTAIN algorithm from existing algorithms on bilevel optimization [14, 18, 19]. The latter works apply the direct gradient estimator $\bar\nabla f(x_t, y_{t+1}; \bar\xi_t)$ [cf. (7)] to serve as an estimate to $\bar\nabla f(x, y)$ [and subsequently $\nabla \ell(x)$]. To guarantee convergence, these works focused on improving the *tracking performance* of $y_{t+1} \approx y^\star(x_t)$ by employing double-loop updates, e.g., by repeatedly applying SG step multiple times for the inner problem; or a sophisticated two-timescale design for the step sizes, e.g., by setting $\beta_t/\alpha_t \to \infty$.

A recent preprint [17] suggested the SVRB algorithm which applies a similar recursive momentum technique as SUSTAIN. However, SVRB is different from SUSTAIN as the momentum estimator is applied exhaustively to *all* the individual random quantities involved in (7) and requires Hessian projection. As a result, the SVRB algorithm entails a high complexity in storage and computation as the latter has to store matrix variables of size $d_{\mathsf{lo}} \times d_{\mathsf{lo}}$ and computes a matrix inverse for each iteration. In comparison, the SUSTAIN algorithm only requires storing the gradient estimators h_t^g, h_t^f of size $d_{\mathsf{lo}}, d_{\mathsf{up}}$, respectively, and the computation complexity is only $\mathcal{O}(d_{\mathsf{lo}}^2 K)$ for each iteration.

3.1 Convergence analysis

In the following, we present the convergence analysis for the SUSTAIN algorithm when $\ell(\cdot)$ is a smooth function [cf. consequence of Assumptions 1, 2 and 3]. Before proceeding to the main results, we present a lemma about the Lipschitzness of the gradient estimate $\bar{\nabla} f(x, y; \bar{\xi})$ given in (7):

Lemma 3.1. *Under Assumptions 1, 2 and 3, we have for any* $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{d_{\mathsf{up}}} \times \mathbb{R}^{d_{\mathsf{lo}}}$,

$$\mathbb{E}_{\bar{\xi}} \|\bar{\nabla} f(x_1, y_1; \bar{\xi}) - \bar{\nabla} f(x_2, y_2; \bar{\xi})\| \leq L_K^2 \big\{\|x_1 - x_2\| + \|y_1 - y_2\|\big\}^2,$$

where

$$L_K = \sqrt{2L_{f_x}^2 + \frac{6C_{g_{xy}}^2 L_{f_y}^2 K}{2\mu_g L_g - \mu_g^2} + \frac{6C_{f_y}^2 L_{g_{xy}}^2 K}{2\mu_g L_g - \mu_g^2} + \frac{6C_{g_{xy}}^2 C_{f_y}^2 L_{g_{yy}}^2 K^3}{(L_g - \mu_g)^2(2\mu_g L_g - \mu_g^2)}}, \tag{15}$$

and K is the number of samples required to construct the stochastic gradient estimate given in (7).

The detailed proof can be found in Appendix C. We remark that the above result is crucial for analyzing the error of the gradient estimate h_t^f defined in (14). To see this, let us first define the errors of the gradient estimates for the outer and inner functions as follows

$$e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t) - B_t, \quad e_t^g := h_t^g - \bar{\nabla}_y g(x_t, y_t), \tag{16}$$

where $B_t := B(x_t, y_t)$ denotes the bias. Rewriting e_t^f using (14) gives the following recursion:

$$\begin{aligned} e_t^f &= (1 - \eta_t^f) e_{t-1}^f + (1 - \eta_t^f)\big\{\bar{\nabla} f(x_t, y_t; \bar{\xi}_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\xi}_t) \\ &\quad - (\bar{\nabla} f(x_t, y_t) + B_t - \bar{\nabla} f(x_{t-1}, y_{t-1}) - B_{t-1})\big\} + \eta_t^f \big(\bar{\nabla} f(x_t, y_t; \bar{\xi}_t) - \bar{\nabla} f(x_t, y_t) - B_t\big). \end{aligned}$$

Lemma 3.1 allows us to control the variance of the second term in the above relation as $\mathcal{O}(\alpha_t^2 \|h_{t-1}^f\|^2 + \beta_t^2 \|h_{t-1}^g\|^2)$. This subsequently leads to a reduced error magnitude for $\mathbb{E}[\|e_t^f\|^2]$. Similarly, we can show a reduced error magnitude for $\mathbb{E}[\|e_t^g\|^2]$ for the inner gradient estimate.

The above discussion suggests that we can track the gradient $\nabla \ell(x)$ using only stochastic gradient estimates (7), without needing to track each component stochastic vectors/matrices. This allows us to avoid costly Hessian inversions. In contrast, [5, 17] track the individual stochastic vectors/matrices of (7), and then combine them together to yield an estimate of $\nabla \ell(x)$. This approach is unable to utilize the cheap stochastic estimates of Hessian and have to invert it directly.

Turning back to the convergence analysis of the SUSTAIN algorithm, the main idea of our analysis is to demonstrate reduction of a properly constructed potential function across iterations. For smooth (possibly non-convex) objective function, this potential function consists of a linear combination of the norms of the error terms $\mathbb{E}[\|e_t^f\|^2]$ and $\mathbb{E}[\|e_t^g\|^2]$ along with the outer objective function $\ell(x_t)$ and the inner optimality gap $\|y_t - y^*(x_t)\|^2$. We obtain:

Theorem 3.2. *Under Assumptions 1–4. Fix $T \geq 1$ as the maximum iteration number. Set the number of samples used for the gradient estimator in* (7) *as* $K = (L_g/\mu_g) \log\big(C_{g_{xy}} C_{f_y} T/\mu_g\big)$ *and*

$$\alpha_t = \frac{1}{(w + t)^{1/3}}, \quad \beta_t = c_\beta \alpha_t, \quad \eta_t^f = c_{\eta_f} \alpha_t^2, \quad \eta_t^g = c_{\eta_g} \alpha_t^2, \tag{17}$$

where $w, c_\beta, c_{\eta_f}, c_{\eta_g}$ are defined in (29) *of appendix. The iterates generated by Algorithm 1 satisfy*

$$\mathbb{E}\|\nabla \ell(x_{a(T)})\|^2 = \mathcal{O}\bigg(\frac{\ell(x_0) - \ell^*}{T^{2/3}} + \frac{\|y_0 - y^*(x_0)\|^2}{T^{2/3}} + \frac{\log(T)\sigma_f^2}{T^{2/3}} + \frac{\log(T)\sigma_g^2}{T^{2/3}}\bigg). \tag{18}$$

Details of the constants in the theorem and its proof can be found in Appendix D. The above result shows that to reach an ϵ-stationary point, the SUSTAIN algorithm requires $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ (omitting logarithmic factors) samples of stochastic gradients from both the upper and lower level functions.

This sample complexity matches the best complexity bounds for single-level stochastic optimization like SPIDER [10], STORM [7], SNVRG [42] and Hybrid SGD [37]. We claim that this is a *near-optimal* sample complexity for bilevel stochastic optimization since for example, we have imposed additional smoothness conditions on the Hessian of the lower level problem. We will leave this as an open question to investigate the lower bound complexity for bilevel stochastic optimization.

Strongly-convex $\ell(x)$. We also discuss the case when in addition to smoothness, $\ell(\cdot)$ is μ_f-strongly-convex. Here, a stronger guarantee can be obtained:

Theorem 3.3. *Under Assumptions 1–4, and suppose $\ell(x)$ is μ_f-strongly-convex. Fix any $T \geq 1$, set the number of samples for the gradient estimator (7) as $K = (L_g/2\mu_g)\log\left(C^2_{g_{xy}} C^2_{f_y} T/\mu_g^2\right)$ and*

$$\alpha_t \equiv \alpha \leq \left\{ \frac{1}{\mu_f + 1}, \frac{1}{2\mu_g \hat{c}_\beta}, \frac{\mu_g}{\hat{c}_\beta L_g^2}, \frac{1}{8L_K^2 + L_f}, \frac{L^2 + 2L_y^2}{4L_K^2 L_g^2 \hat{c}_\beta^2} \right\}, \;\; \eta_t^f \equiv (\mu_f + 1)\alpha, \;\; \beta_t \equiv \hat{c}_\beta \alpha,$$

where $\eta_t^g \equiv 1$, $\hat{c}_\beta = 8L_y^2 + 8L^2 + 2\mu_f/\mu_g$ and L_K is defined in (15). The iterates generated by Algorithm 1 satisfy for any $t \geq 1$ that:

$$\mathbb{E}[\ell(x_t) - \ell^*] \leq (1 - \mu_f \alpha)^t \bar{\Delta}_0 + \frac{1}{\mu_f}\left\{ \frac{2}{T} + \left[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2 L_K^2)\sigma_g^2 + 2(\mu_f + 1)^2 \sigma_f^2\right]\alpha \right\}, \tag{19}$$

where $\bar{\Delta}_0 := \ell(x_0) - \ell^* + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2$.

The detailed proof can be found in Appendix E. For large T, setting $\alpha \asymp 1/T$ shows that the bound in (19) decreases at the rate of $\mathcal{O}(1/T)$.

Theorem 3.3 shows that to reach an ϵ-optimal point, the SUSTAIN algorithm requires $\widetilde{\mathcal{O}}(\epsilon^{-1})$ stochastic gradient samples from the upper and lower level problems, also see the detailed calculations in Appendix E. This improves over TTSA [18] which requires $\widetilde{\mathcal{O}}(\epsilon^{-1.5})$ samples, and BSA [14] which requires $\widetilde{\mathcal{O}}(\epsilon^{-1})$, $\mathcal{O}(\epsilon^{-2})$ samples for the upper and lower level problems, respectively. Again, we achieve similar sample complexity as SGD applied on strongly-convex single-level optimization.

Interestingly, in Theorem 3.3, we have selected $\eta_t^g \equiv 1$ where the momentum term in the lower level gradient vanishes. In this way, the SUSTAIN algorithm is reduced into a *single-momentum* algorithm where the recursive momentum acceleration is only applied to the upper level gradient. Similarly, in Theorem 3.2, if SUSTAIN utilizes only the upper level momentum, i.e., $\eta_t^g \equiv 1$, then with appropriate choice of parameters, we get $\mathbb{E}\|\nabla \ell(x_{a(T)})\|^2 \leq \mathcal{O}(1/\sqrt{T})$ (please see [20] for further details). This implies that to achieve an ϵ-stationary solution SUSTAIN with only upper level momentum requires $\mathcal{O}(\epsilon^{-2})$ stochastic samples for both the upper and the lower level functions. Note that this improves over TTSA [18] which utilizes a vanilla SGD update for both the upper and the lower level problems, i.e., $\eta_t^f \equiv 1$ and $\eta_t^g \equiv 1$ and requires $\mathcal{O}(\epsilon^{-5/2})$ stochastic samples for both upper and lower level functions.

4 Numerical experiments

In this section, we evaluate the performance of the SUSTAIN algorithm on two popular machine learning tasks: hyperparameter optimization and meta learning.

Hyperparameter optimization. We consider the data hyper-cleaning task (3), and compare SUSTAIN with several algorithms such as stocBiO [19] for different batch size choices, and the HOAG algorithm in [29]. Note that in [19], the authors shown that stocBio exhibits better practical performance compared with other bilevel optimization algorithms.

We consider problem (3) with $L(\cdot)$ being the cross-entropy loss (i.e., a data cleaning problem for logistic regression); $\sigma(x) := \frac{1}{1+\exp(-x)}$; $c = 0.001$; see [35]. The problem is trained on the `FashionMNIST` dataset [41] with 50k, 10k, and 10k image samples allocated for training, validation and testing purposes, respectively. The step sizes for different algorithms are chosen according to their theoretically suggested values. Let the outer iteration be indexed by t, for SUSTAIN we choose $\alpha_t = \beta_t = 0.1/(1+t)^{1/3}$ and tune for c_{η_f} and c_{η_g} (see Theorem 3.2), for stocBiO and HOAG we select $\alpha_t = d_\alpha$, $\beta_t = d_\beta$ and tune for parameters d_α and d_α in the range $[0, 1]$.

In Figure 1, we compare the performance of different algorithms when the dataset has a corruption probability of 0.3. As observed, SUSTAIN outperforms stocBiO and HOAG. We remark that HOAG is a deterministic algorithm and hence requires full batch gradient computations at each iteration. Similarly, stocBio relies on large batch gradients which results in relatively slow convergence. This fast convergence of SUSTAIN results form the single timescale update with reduced variance resulting from the double-momentum variance reduced updates.

Meta learning. We consider a few-shot meta learning problem [11, 30] (cf. (2)) and compare the performance of SUSTAIN to ITD-BiO [19] and ANIL [30]. The task of interest is 5-way 5-shot

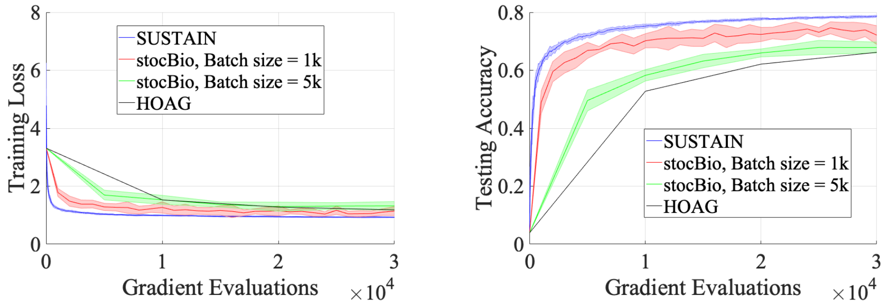

Figure 1: **Hyperparameter optimization**: Data hyper-cleaning task on the `FashionMNIST` dataset. We plot the training loss and testing accuracy against the number of gradients evaluated with corruption rate $p = 0.3$.

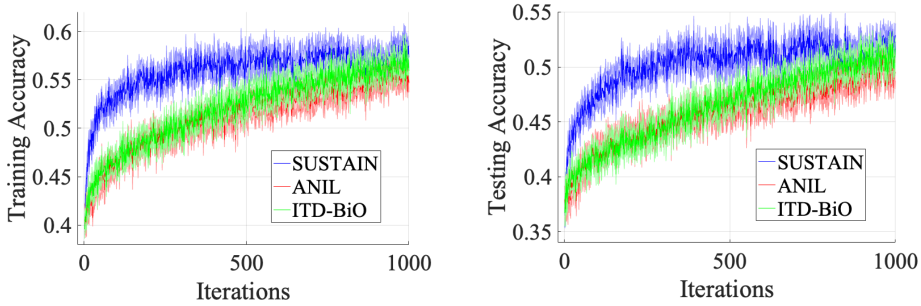

Figure 2: **Meta learning**: 5-way 5-shot learning task on the `miniImageNet` dataset. We plot the training and testing accuracy against the number of iterations.

learning and we conduct experiments on the `miniImageNet` dataset [39, 32] with 100 classes and 600 images per class. We apply `learn2learn` [1] (available: `https://github.com/learnables/learn2learn`) to partition the 100 classes from `miniImageNet` into subsets of 64, 16 and 20 for meta training, meta validation and meta testing, respectively. Similar to [1, 19], we implement a 4-layer convolutional neural network (CNN) with ReLU activation for the learning task. At each iteration, we sample a batch of 32 tasks from a set of 20000 tasks allocated for training and 600 each for validation and testing. For each algorithm, we implement 10 inner and 1 outer update. The performance is averaged over 10 Monte Carlo runs.

For ANIL and ITD-BiO, we use the parameter selection suggested in [1, 19]. Specifically, for ANIL, we use inner-loop stepsize of 0.1 and the outer-loop (meta) stepsize as 0.002. For ITD-BiO, we choose the inner-loop stepsize as 0.05 and the outer-loop stepsize to be 0.005. For SUSTAIN, we choose the outer-loop stepsize α_t as $\kappa/(1+t)^{1/3}$ and choose $\kappa \in [0.1, 1]$, we choose the momentum parameter η_t as $\bar{c}\alpha_t^2/\kappa^2$ and tune for $\bar{c} \in \{2, 5, 10, 15, 20\}$, finally, we fix the inner stepsize as 0.05. For the outer loop update ANIL and ITD-BiO utilize SGD optimizer whereas SUSTAIN uses the hybrid gradient estimator.

From Figure 2 which compares the training and testing accuracy against the iteration number, we observe that SUSTAIN achieves a better performance compared to ANIL and ITD-BiO on the meta learning task. Also, notice that in the initial iterations SUSTAIN converges faster but then converges probably as a consequence of diminishing stepsizes (and momentum parameter). In contrast, ANIL and ITD-BiO slowly improve in performance and catch up with SUSTAIN's performance. In the appendix, we show that the SUSTAIN algorithm requires less computation time to achieve better performance compared to the ANIL and ITD-BiO.

For further evaluation of the performance of SUSTAIN, we have included additional experiments on hyperparameter optimization and meta learning on different datasets in the supplementary material.

Conclusions and limitations

We have developed the SUSTAIN algorithm for unconstrained bilevel optimization with strongly convex lower level subproblems. The proposed algorithm executes on a single-timescale, without the need to use either two-timescale updates, large batch gradients, or double-loop algorithm. We showed that SUSTAIN is both *sample* and *computation* efficient, because it matches the best-known sample complexity guarantees for single-level problems with non-convex and strongly convex objective (smooth) functions, while matching the best-known per-iteration computational complexity for the same class of bi-level problems. In the future, we plan to rigorously show the sample complexity lower bounds for the considered class of bilevel problems. Further, we plan to develop sample and communication efficient algorithms for a more general class of bilevel problems, such as those with constraints in the lower level problems.

Acknowledgement

We thank the anonymous reviewers for their valuable comments and suggestions. The work of Prashant Khanduri, Siliang Zeng, and Mingyi Hong was supported by the National Science Foundation (NSF) through grant CIF-1910385. The work of Mingyi Hong was also supported by an IBM Faculty Research award. The work of Hoi-To Wai was supported by CUHK Direct Grant #4055113. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their support. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

[1] S. M. R. Arnold, P. Mahajan, D. Datta, I. Bunner, and K. S. Zarkias. learn2learn: A library for Meta-Learning research. *CoRR*, Aug. 2020.

[2] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973. ISSN 0030364X, 15265463.

[3] J. Bracken and J. T. McGill. Defense applications of mathematical programs with optimization problems in the constraints. *Operations Research*, 22(5):1086–1096, 1974. ISSN 0030364X, 15265463.

[4] J. Bracken, J. E. Falk, and J. T. McGill. Technical note—the equivalence of two mathematical programs with optimization problems in the constraints. *Operations Research*, 22(5):1102–1104, 1974. doi: 10.1287/opre.22.5.1102.

[5] T. Chen, Y. Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

[6] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153:235–256, 2007.

[7] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32*, pages 15236–15245. Curran Associates, Inc., 2019.

[8] S. Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[9] J. E. Falk and J. Liu. On bilevel programming, part I: General nonlinear cases. *Mathematical Programming volume*, 70:47–72, 1995.

[10] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[11] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[12] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1165–1173, 2017.

[13] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *arXiv preprint arXiv:1806.04910*, 2018.

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

[15] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. *arXiv preprint arXiv:2006.16218*, 2020.

[16] R. Grazzi, M. Pontil, and S. Salzo. Convergence properties of stochastic hypergradients. *arXiv preprint arXiv:2011.07122*, 2020.

[17] Z. Guo and T. Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

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

[19] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Nonasymptotic analysis and faster algorithms. *arXiv preprint arXiv:2010.07962*, 2020.

[20] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization. *arXiv preprint arXiv:2102.07367*, 2021.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

[23] J. Li, B. Gu, and H. Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.

[24] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. *arXiv preprint arXiv:2006.04045*, 2020.

[25] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *arXiv preprint arXiv:2101.11517*, 2021.

[26] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511983658.

[27] A. Migdalas, P. M. Pardalos, and P. Värbrand. *Multilevel optimization: algorithms and applications*, volume 20. Springer Science & Business Media, 2013.

[28] B. N. Oreshkin, P. Rodriguez, and A. Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

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

[30] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *ICLR*, 2019.

[31] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, pages 113–124. Curran Associates, Inc., 2019.

[32] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[33] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill New York, 3d ed. edition, 1976. ISBN 007054235.

[34] S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM J. Optim.*, 27(2):640–660, 2017.

[35] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. *arXiv preprint arXiv:1810.10667*, 2019.

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

[37] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

[38] L. Vicente, , G. Savard, and J. Júdice. Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, pages 379–399, 1994.

[39] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[40] D. J. White and G. Anandalingam. A penalty function approach for solving bi-level linear programs. *Journal of Global Optimization*, 3:397–419, 1993.

[41] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[42] D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *arXiv preprint arXiv:1806.07811*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the conclusion & limitations section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the experiment section and the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] The experiments were conducted on local machines.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Additional experiments

In this section, we supplement the numerical results presented in Section 4 with additional experiments on real datasets. We demonstrate the efficacy of SUSTAIN for the meta learning and hyperparameter optimization tasks. Furthermore, we examine the performance of SUSTAIN when combined with an Adam-like update rule [cf. see Algorithm 2].

Meta learning. In Figure 2 of Section 4 in the main paper, we established that when SUSTAIN, ITD-BiO and ANIL utilize vanilla SG direction for the outer level update, SUSTAIN outperforms rest of the algorithms for the meta learning problem. Specifically, we compared the training and testing performance of the algorithms with the number of iterations (i.e., the outer update t in Algorithm 1). In each iteration, all the algorithms access the same number of samples while SUSTAIN requiring twice the number of gradient computations (cf. (14)). As observed from Figure 2, SUSTAIN requires the smallest number of iterations (samples) and gradient computations to achieve a given training/testing accuracy on the benchmarked dataset.

We conduct additional experiments for meta learning and demonstrate the following: (1) for the outer level update we can adapt Adam [21] optimizer with the SUSTAIN framework to achieve better performance, (2) the outer gradient estimate (14) for SUSTAIN can be designed with only one gradient computation per iteration (instead of two) without compromising performance, and (3) SUSTAIN outperforms MAML [11], ANIL [30] and ITD-BiO [19] when all algorithms implement Adam for the outer level update. Next, we discuss the datasets and the parameter settings.

We consider meta learning problem with miniImageNet [39, 32] and FC100 [28] datasets. Both datasets consist of 100 classes with each class containing 600 images. For the miniImageNet, we consider the same setting as in Section 4. For FC100, we follow the setting of [28, 19] where 100 classes are split into 60, 20 and 20 classes for meta-training, meta-validation and meta-testing, respectively. For both datasets, we consider a 5-way 5-shot learning task where the algorithm aims to classify samples into 5 unseen classes using only 5 available samples. We implement the solver using a 4-layer CNN (with different width for each dataset). We compare heuristic versions of SUSTAIN with MAML [11], ANIL [30] and recently proposed ITD-BiO [19], where these algorithms all utilize the Adam [21] solver for the outer problem's update. These heuristic algorithms are also used in [19] when comparing performance of the bilevel algorithms for meta-learning tasks. Note that these Adam-based bilevel algorithms for meta learning do not have any theoretical performance guarantees. Nevertheless, in the following we show that they perform well in practice [19].

We first discuss the parameter setting for the meta learning task using miniImageNet dataset. All the algorithms sample 32 tasks in each iteration. For the Adam versions of MAML, ANIL and ITD-BiO, we choose the parameters as suggested in [1, 19]. For all the algorithms, we execute 10 update steps in the inner loop followed by a single outer update step. Each update step is counted as a

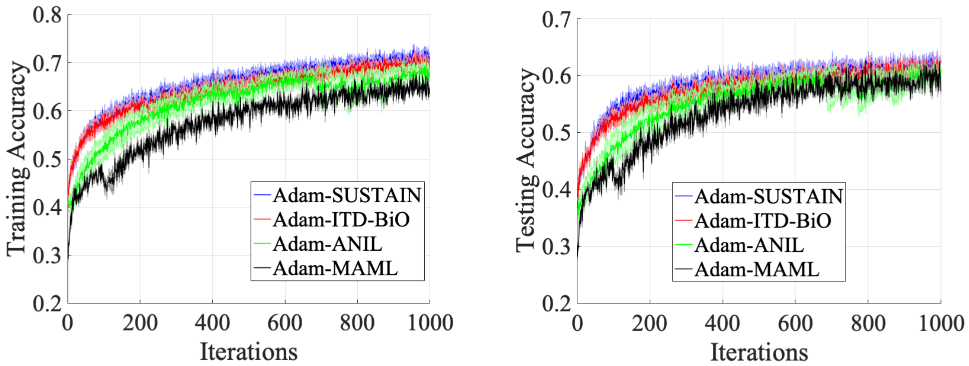

Figure 3: **Meta learning**: 5-way 5-shot learning task on the miniImageNet dataset. We plot the training and testing accuracy against the number of iterations with each iteration representing one outer level update step. All the algorithms utilize Adam [21] optimizer for the outer loop update.

Algorithm 2 Update direction for Adam-SUSTAIN (also see footnote[1])

```
**Parameters**: $\gamma_1 = 0.9$, $\gamma_2 = 0.999$, $m_0 = 0$, $v_0 = 0$, $\epsilon = 10^{-8}$ and $\eta_t^f$
**for** $t = 1, \cdots, T$ **do**
    **Input**: $(x_t, y_t)$, $(x_{t-1}, y_{t-1})$ from Algorithm 1.
    Compute the gradient estimator $\bar{h}_t^f$ using Option I or II in (20)
    Update first moment estimate: $m_t \leftarrow \gamma_1 \cdot m_{t-1} + (1 - \gamma_1)\bar{h}_t^f$
    Bias-correction for first moment estimate: $m_t \leftarrow m_t/(1 - (\gamma_1)^t)$
    Update second moment estimate: $v_t \leftarrow \gamma_2 \cdot v_{t-1} + (1 - \gamma_2)(\bar{h}_t^f)^2$
    Bias-correction for second moment estimate: $v_t \leftarrow v_t/(1 - (\gamma_2)^t)$
    Use the update direction: $h_t^f \leftarrow m_t/(\sqrt{v_t} + \epsilon)$
**end for**
**Return:** $h_t^f$
```

single iteration. The implementation of MAML and ANIL is adopted from existing implementations in [1]. For MAML, we choose the inner loop stepsize to be 0.5 and the outer loop stepsize to be 0.003. For ANIL we utilize inner loop stepsize of 0.1 and outer loop stepsize of 0.002. Both ITD-BiO and SUSTAIN utilize gradient descent with stepsize of 0.05 as the inner optimizer. For the outer update ITD-BiO uses a stepsize of 0.002 (the parameters for ITD-BiO are selected based the repository `https://github.com/JunjieYang97/stocBiO`). For SUSTAIN we set the outer stepsize as $\alpha_t = 0.005$ and tune for the momentum parameter $\eta_t^f = \bar{c}/\kappa^2(1+t)^{2/3}$ with fixed $\kappa = 0.005$ by choosing $\bar{c} \in \{0.25, 2.5, 5, 10\}$. In contrast to other algorithms, SUSTAIN applies Adam [21] to the hybrid stochastic gradient estimator used for the outer update (14). For detailed steps please see Algorithm 2[2]. Moreover, it is worth noting that the direction update rule Option II given in (20) is a modification of the original update given in (14) (or equivalently Option I in (20)). Such a rule requires just a single (mini-batch) gradient computation per iteration (which is the same as MAML, ANIL and ITD-BiO), and in practice, its performance is very close to that of Option I. Our results below uses Option II as the update direction.

$$\bar{h}_t^f = \begin{cases} \bar{\nabla} f(x_t, y_t; \bar{\xi}_t) + (1 - \eta_t^f)\big(\bar{h}_{t-1}^f - \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\xi}_t)\big) & \text{Option I} \\ \bar{\nabla} f(x_t, y_t; \bar{\xi}_t) + (1 - \eta_t^f)\big(\bar{h}_{t-1}^f - \underbrace{\bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\xi}_{t-1})}_{\text{Previous SG}}\big) & \text{Option II} \end{cases} \tag{20}$$

In Figure 3, we plot the training and testing performance against the number of iterations for the Adam version SUSTAIN with other algorithms for 5-way 5-shot learning task on `miniImageNet` dataset. Note from the discussion above, we know that in each iteration all the algorithms access

[2] Note that the vector division and exponent operations in the Algorithm are implemented element wise. The values of the parameters chosen for Adam are default values used by the PyTorch library.

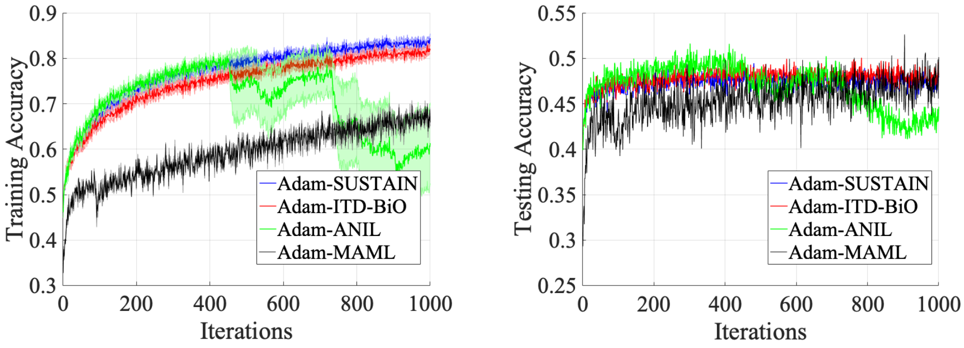

Figure 4: **Meta learning**: 5-way 5-shot learning task on the `FC100` dataset. We plot the training and testing accuracy against the number of iterations with each iteration representing one outer level update. All the algorithms utilize Adam [21] optimizer for the outer loop update.

the same number of sample, and spend the same amount of (mini-batch) gradient computation efforts. Consequently, Figure 3 implies that SUSTAIN outperforms ITD-BiO, ANIL and MAML as it requires fewest iteration (thus samples and gradient computation) to achieve the improved performance. Importantly, these Adam-based algorithms significantly outperform their vanilla version (cf. Figure 2 for performance with SGD), in terms of both accuracy and speed.

Next, we compare the performance of SUSTAIN with other algorithms for the meta learning task using `FC100` dataset. For this task all the algorithms sample 32 tasks in each iteration. In contrast to the previous dataset, for this task we execute 20 update steps in the inner loop followed by a single outer update step. Similar to `miniImageNet` dataset, we adopt existing implementations of MAML and ANIL from [1] and ITD-BiO from [19]. For MAML, we choose inner loop stepsize of 0.5 and the outer loop stepsize of 0.001. For ANIL we utilize inner loop stepsize of 0.1 and outer loop stepsize of 0.001. In the inner loop, both ITD-BiO and SUSTAIN utilize gradient descent with a stepsize of 0.1. For the outer update ITD-BiO uses a stepsize of 0.001 ((the parameters for ITD-BiO are selected based the repository `https://github.com/JunjieYang97/stocBiO`)). For the outer update SUSTAIN utilizes the same setting as required for `miniImageNet` dataset and the Adam based outer update direction as computed in Algorithm 2.

In Figure 4, we plot the training and testing performance with the number of iterations for SUSTAIN and other algorithms for 5-way 5-shot learning task on `FC100` dataset. Note that SUSTAIN outperforms rest of the algorithms on the training task and performs on par with other algorithms with respect to the testing performance. Moreover, note that initially ANIL performs better but since the number of inner steps are relatively large (20 in this case), ANIL's performance degrades after a certain number of iterations. Similar behavior was noted for ANIL in the results of [19].

The above set of experiments showed that the Adam [21] optimizer can be incorporated with SUSTAIN and other algorithms to achieve improved performance compared to vanilla SG based algorithms. We also showed that the gradient estimator for SUSTAIN can be modified to require only single (batch) gradient evaluation per iteration (cf. (20)) without comprising performance of the algorithm. In this section, we use an additional set of results to demonstrate that under most settings SUSTAIN outperforms other state-of-the-art algorithms.

Hyperparameter optimization. For data hyperparameter optimization problem, we consider the hyper-cleaning task as discussed earlier in Section 4 and benchmark the performance of SUSTAIN against stocBiO [19] and HOAG [29]. Importantly, in this section we demonstrate that SUSTAIN performs well under (relatively) high level of data corruption.

Here we consider an additional set of results for the hyper-cleaning task on Fashion-MNIST dataset [41]. All the parameter settings are the same as in Section 4, except that we use a higher level 40% corruption rate. Note that HOAG is a deterministic algorithm and requires full gradient computation at each iteration. In contrast, stocBiO is a stochastic algorithm but it relies on large batch gradient computations. We conduct experiments for two settings where stocBiO uses a batch size of 5000 and 1000 (for both inner and outer updates). Our algorithm SUSTAIN is purely a stochastic algorithm and does not rely on large batch gradient computations. Specifically, SUSTAIN computes two gradients (on a single sample) in each iteration for both inner and outer updates (cf. (13) and (14)). Since at

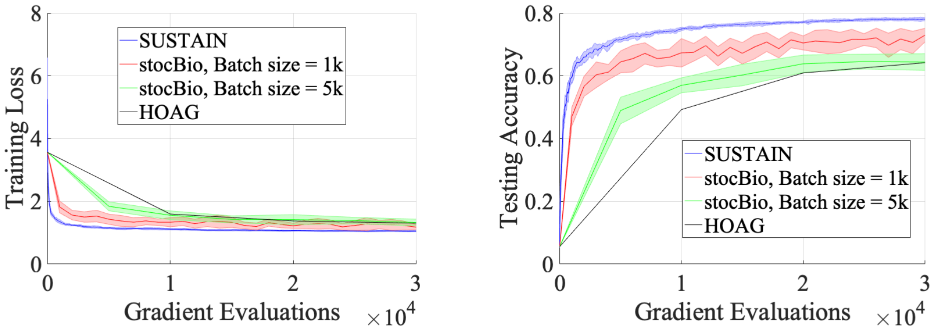

Figure 5: Data hyperparameter optimization: Training loss and testing accuracy against the number of gradients evaluated with corruption rate $p = 0.4$.

each outer iteration, the sample sizes (and gradient computations) accessed by each algorithms are very different, so it is no longer fair to compare the per-iteration performance for different algorithms (this is different compared with the meta learning example in the previous section). Therefore, in this section we compare the training and testing performance of the competing algorithms using the number of total outer gradient computations (which is same as the inner gradient computations) across iterations. Note that for HOAG and stocBiO, the number of samples accessed is same as the number of gradient evaluations, whereas for SUSTAIN we compute two gradients for each sample accessed (cf. (14))[3]. The experiments in Figure 5 establish that SUSTAIN outperforms HOAG and stocBiO, in terms of the total number of gradient evaluations as well as the number of samples, even under high corruption rate.

[3] Note that this requirement can be easily relaxed without compromising performance via using the gradient construction (20).

Proofs of Theoretical Results

Now we present the proofs of the theoretical results.

B Useful lemmas

Lemma B.1. *Consider a collection of functions $\Phi_i : \mathbb{R}^n \to \mathcal{Z}$ with $i = \{1, 2, \ldots, k\}$ and $\mathcal{Z} \subseteq \mathbb{R}^{n\times n}$, which satisfy the following assumptions:*

(i) There exist $L_i > 0,\ i \in [k]$, such that

$$\|\Phi_i(x) - \Phi_i(y)\| \le L_i\|x - y\|,\ \forall\, i \in [k],\ x, y \in \mathbb{R}^n.$$

(ii) For each $i \in [N]$ and $k \in \mathbb{N}$ we have $\|\Phi_i(x)\| \le M_i$ for all $x \in \mathbb{R}^n$.

Then the following holds for all $x, y \in \mathbb{R}^n$:

$$\Big\|\prod_{i=1}^{k}\Phi_i(x) - \prod_{i=1}^{k}\Phi_i(y)\Big\|^2 \le k\sum_{i=1}^{k}\Big(\prod_{j=1, j\neq i}^{k} M_j\Big)^2 L_i^2\|x - y\|^2. \tag{21}$$

Moreover, if k is generated uniformly at random from $\{0, 1, \ldots, K - 1\}$, then the following holds for all $x, y \in \mathbb{R}^n$:

$$\mathbb{E}_k\Big\|\prod_{i=1}^{k}\Phi_i(x) - \prod_{i=1}^{k}\Phi_i(y)\Big\|^2 \le K\sum_{i=1}^{K}\mathbb{E}_k\Big[\Big(\prod_{j=1, j\neq i}^{k} M_j\Big)^2\Big] L_i^2\|x - y\|^2. \tag{22}$$

Here we use the convention that $\prod_{i=1}^{k}\Phi_i(x) = I$ if $k = 0$.

Proof. We first prove (21). To do so we will first show that the following holds for all $x, y \in \mathbb{R}^n$ and $k \in \mathbb{N}$:

$$\Big\|\prod_{i=1}^{k}\Phi_i(x) - \prod_{i=1}^{k}\Phi_i(y)\Big\| \le \sum_{i=1}^{k}\Big(\prod_{j=1, j\neq i}^{k} M_j\Big) L_i\|x - y\|, \tag{23}$$

Then by combining the above result with the identity that

$$\|z_1 + z_2 + \ldots + z_k\|^2 \le k\|z_1\|^2 + k\|z_2\|^2 + \ldots + k\|z_k\|^2,\ \text{for all } z,\ k \in \mathbb{N}, \tag{24}$$

we can conclude the first statement.

To show (23), we use an induction argument. The base case for $k = 1$ holds because of the Lipschitz assumption (i) given in the statement of the lemma. Then assuming claim (23) holds for arbitrary k, we have for $k + 1$

$$\begin{aligned}
\Big\|\prod_{i=1}^{k+1}\Phi_i(x) - \prod_{i=1}^{k+1}\Phi_i(y)\Big\| &= \Big\|\prod_{i=1}^{k+1}\Phi_i(x) - \prod_{i=1}^{k}\Phi_i(x)\Phi_{k+1}(y) + \prod_{i=1}^{k}\Phi_i(x)\Phi_{k+1}(y) - \prod_{i=1}^{k+1}\Phi_i(y)\Big\| \\
&\overset{(a)}{\le} \Big\|\prod_{i=1}^{k}\Phi_i(x)\Big\|\,\big\|\Phi_{k+1}(x) - \Phi_{k+1}(y)\big\| + \big\|\Phi_{k+1}(y)\big\|\,\Big\|\prod_{i=1}^{k}\Phi_i(x) - \prod_{i=1}^{k}\Phi_i(y)\Big\| \\
&\overset{(b)}{\le} \Big(\prod_{j=1}^{k} M_j\Big) L_{k+1}\|x - y\| + \sum_{i=1}^{k}\Big(\prod_{j=1, j\neq i}^{k+1} M_j\Big) L_i\|x - y\| \\
&\overset{(c)}{\le} \sum_{i=1}^{k+1}\Big(\prod_{j=1, j\neq i}^{k+1} M_j\Big) L_i\|x - y\|.
\end{aligned}$$

where (a) follows from the application of the triangle inequality and the Cauchy-Schwartz inequality; the first expression in (b) results from the application of Cauchy-Schwartz inequality and Assumption (i) and (ii) of the statement of the lemma; the second expression in (b) follows from the assumption that claim (23) holds for k; (c) follows from combining the two expressions. We conclude that (23) holds for all $k \in \mathbb{N}$.

Now consider the case when k is chosen uniformly at random from $k \in \{0, 1, \dots, K-1\}$. First, note from the definition that for $k = 0$ we have $\prod_{i=1}^{k} \Phi_i(x) = I$. This implies that (21) is also satisfied if we have $k = 0$. We then have

$$\begin{aligned}
\mathbb{E}_k \left\| \prod_{i=1}^{k} \Phi_i(x) - \prod_{i=1}^{k} \Phi_i(y) \right\|^2 &\overset{(a)}{\leq} \mathbb{E}_k \left[k \sum_{i=1}^{k} \Big(\prod_{j=1, j\neq i}^{k} M_j \Big)^2 \|\Phi_i(x) - \Phi_i(y)\|^2 \right] \\
&\overset{(b)}{\leq} K \sum_{i=1}^{K} \mathbb{E}_k \left[\Big(\prod_{j=1, j\neq i}^{k} M_j \Big)^2 \right] \|\Phi_i(x) - \Phi_i(y)\|^2 \\
&\overset{(c)}{\leq} K \sum_{i=1}^{K} \mathbb{E}_k \left[\Big(\prod_{j=1, j\neq i}^{k} M_j \Big)^2 \right] L_i^2 \|x - y\|^2.
\end{aligned}$$

where (a) uses the fact that (21) holds for all $k \in \{0, 1, \dots, K-1\}$ almost surely; (b) follows from the fact that $k \leq K$ almost surely; (c) results from Assumption (i) of the lemma. □

C Proofs of preliminary lemmas

C.1 Estimation of the stochastic gradient

We construct the stochastic gradient $\bar{\nabla} f(x, y; \bar{\xi})$ as [14, 18]:

1. For $K \in \mathbb{N}$, choose $k \in \{0, 1, \dots, K-1\}$ uniformly at random.
2. Compute unbiased Hessian approximations $\nabla^2_{xy} g(x, y; \zeta^{(0)})$ and $\nabla^2_{yy} g(x, y; \zeta^{(i)})$ for $i \in \{1, \dots, k\}$, where $\{\zeta^{(i)}\}_{i=0}^{k}$ are chosen independently.
3. Compute unbiased gradient approximations $\nabla_x f(x, y; \xi)$ and $\nabla_y f(x, y; \xi)$ where ξ is chosen independently of $\{\zeta^{(i)}\}_{i=0}^{k}$.
4. Construct the stochastic gradient estimate $\bar{\nabla} f(x, y; \bar{\xi})$ with $\bar{\xi}$ denoted as $\bar{\xi} = \{\xi, \{\zeta^{(i)}\}_{i=0}^{k}\}$:
$$\begin{aligned}
&\bar{\nabla} f(x, y; \bar{\xi}) \\
&= \nabla_x f(x, y; \xi) - \nabla^2_{xy} g(x, y; \zeta^{(0)}) \left[\frac{K}{L_g} \prod_{i=1}^{k} \left(I - \frac{1}{L_g} \nabla^2_{yy} g(x, y; \zeta^{(i)}) \right) \right] \nabla_y f(x, y; \xi),
\end{aligned} \tag{25}$$
with $\prod_{i=1}^{k} \left(I - \frac{1}{L_g} \nabla^2_{yy} g(x, y; \zeta^{(i)}) \right) = I$ if $k = 0$.

Next, we state the result showing that the bias of the stochastic gradient estimate of the upper level objective defined in (7) decays linearly with the number of samples K chosen to approximate the Hessian inverse.

Lemma C.1. *[18, Lemma 11] Under Assumptions 1, 2 and 3 the stochastic gradient estimate of the upper level objective defined in (25), satisfies*

$$\|B(x, y)\| = \|\bar{\nabla} f(x, y) - \mathbb{E}[\bar{\nabla} f(x, y; \bar{\xi})]\| \leq \frac{C_{g_{xy}} C_{f_y}}{\mu_g} \left(1 - \frac{\mu_g}{L_g} \right)^K,$$

where $B(x, y)$ is the bias of the stochastic gradient estimate and K is the number of samples chosen to approximate the Hessian inverse in (25). Moreover, we have

$$\mathbb{E}_{\bar{\xi}}\left[\left\| \bar{\nabla} f(x, y) - \mathbb{E}_{\bar{\xi}}[\bar{\nabla} f(x, y; \bar{\xi})] \right\|^2 \right] \leq \sigma_{f_x}^2 + \frac{3}{\mu_g^2} \left[(\sigma_{f_y}^2 + C_{f_y}^2)(\sigma_{g_{xy}}^2 + 2 C_{g_{xy}}^2) + \sigma_{f_y}^2 C_{g_{xy}}^2 \right].$$

Lemma C.1 implies that the bias $B(x, y)$ can be made to satisfy $\|B(x, y)\| \leq \epsilon$ with only

$$K = (L_g / \mu_g) \log(C_{g_{xy}} C_{f_y} / \mu_g \epsilon)$$

stochastic Hessian samples of $\nabla^2_{yy} g(x, y)$.

C.2 Lipschitz continuity of gradient estimate

Lemma C.2 (Lipschitzness of Stochastic Gradient Estimate)**.** *If the stochastic functions $f(x,y;\xi)$ and $g(x,y;\zeta)$ satisfy Assumptions 1, 2 and 3, then we have*

(i) For a fixed $y \in \mathbb{R}^{d_{\mathsf{up}}}$

$$\mathbb{E}_{\bar{\xi}}\|\bar{\nabla} f(x_1,y;\bar{\xi}) - \bar{\nabla} f(x_2,y;\bar{\xi})\|^2 \le L_K^2\|x_1 - x_2\|^2, \; \forall\, x_1, x_2 \in \mathbb{R}^{d_{\mathsf{up}}}.$$

(ii) For a fixed $x \in \mathbb{R}^{d_{\mathsf{up}}}$

$$\mathbb{E}_{\bar{\xi}}\|\bar{\nabla} f(x,y_1;\bar{\xi}) - \bar{\nabla} f(x,y_2;\bar{\xi})\|^2 \le L_K^2\|y_1 - y_2\|^2, \; \forall\, y_1, y_2 \in \mathbb{R}^{d_{\mathsf{up}}}.$$

In the above expressions, $L_K > 0$ is defined as:

$$L_K^2 = 2L_{f_x}^2 + 6C_{g_{xy}}^2 L_{f_y}^2\left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) + 6C_{f_y}^2 L_{g_{xy}}^2\left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) + 6C_{g_{xy}}^2 C_{f_y}^2 \frac{K^3 L_g^2}{(L_g - \mu_g)^2(2\mu_g L_g - \mu_g^2)},$$

and where K is the number of samples required to construct the stochastic approximation of $\bar{\nabla} f$ (see (25) above).

Proof. We prove only statement (i) of the lemma, the proof of (ii) follows from a similar argument. From the definition of $\bar{\nabla} f(x_1,y;\bar{\xi})$ we have for $x_1, x_2 \in \mathbb{R}^{d_{\mathsf{up}}}$ and $y \in \mathbb{R}^{d_{\mathsf{up}}}$

$$\begin{aligned}
&\|\bar{\nabla} f(x_1,y;\bar{\xi}) - \bar{\nabla} f(x_2,y;\bar{\xi})\|^2 \\
&\overset{(a)}{\le} 2\big\|\nabla_x f(x_1,y;\xi) - \nabla_x f(x_2,y;\xi)\big\|^2 \\
&+ 2\bigg\|\nabla_{xy}^2 g(x_1,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)\right]\nabla_y f(x_1,y;\xi) \\
&\qquad - \nabla_{xy}^2 g(x_2,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right]\nabla_y f(x_2,y;\xi)\bigg\|^2 \\
&\overset{(b)}{\le} 2L_{f_x}^2\big\|x_1 - x_2\big\|^2 \\
&+ 2\bigg\|\nabla_{xy}^2 g(x_1,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)\right]\nabla_y f(x_1,y;\xi) \\
&\qquad - \nabla_{xy}^2 g(x_2,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right]\nabla_y f(x_2,y;\xi)\bigg\|^2,
\end{aligned} \tag{26}$$

where inequality (a) follows from the definition of $\bar{\nabla} f(x_1,y;\bar{\xi})$ and (24); inequality (b) follows from the Lipschitz-ness Assumption 1–(ii) made for stochastic upper level objective. The variable $k \in \{0,\ldots,K-1\}$ above is a random variable define in Section C.1 above. Let us consider the second term of (26) above, we have

$$\begin{aligned}
&\bigg\|\nabla_{xy}^2 g(x_1,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)\right]\nabla_y f(x_1,y;\xi) \\
&\qquad - \nabla_{xy}^2 g(x_2,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^{k}\left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right]\nabla_y f(x_2,y;\xi)\bigg\|^2 \\
&\overset{(a)}{\le} 3C_{g_{xy}}^2\frac{K^2}{L_g^2}\left(1 - \frac{\mu_g}{L_g}\right)^{2k}\|\nabla_y f(x_1,y;\xi) - \nabla_y f(x_2,y;\xi)\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 3C_{f_y}^2 \frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}\|\nabla_{xy}^2 g(x_1,y;\zeta^{(0)}) - \nabla_{xy}^2 g(x_2,y;\zeta^{(0)})\|^2 \\
&+ 3C_{g_{xy}}^2 C_{f_y}^2 \left\| \frac{K}{L_g}\prod_{i=1}^k \left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right) - \frac{K}{L_g}\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right\|^2 \\
\overset{(b)}{\le}\ & 3C_{g_{xy}}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k} L_{f_y}^2\|x_1-x_2\|^2 + 3C_{f_y}^2 \frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{g_{xy}}^2\|x_1-x_2\|^2 \\
&\quad + 3C_{g_{xy}}^2 C_{f_y}^2\frac{K^2}{L_g^2}\left\|\prod_{i=1}^k \left(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right) - \prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right\|^2,
\end{aligned}
$$

where inequality (a) follows from (21) in Lemma B.1, Assumption 1–(iii) and Assumption 2–(ii)(iii)(vi); inequality (b) follows from the Lipschitz continuity Assumption 1–(ii) and Assumption 2–(v) made for the stochastic upper and lower level objectives. On both sides taking expectation w.r.t k, we get:

$$
\begin{aligned}
&\mathbb{E}_k\Bigg\|\nabla_{xy}^2 g(x_1,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)\right]\nabla_y f(x_1,y;\xi) \\
&\qquad\qquad - \nabla_{xy}^2 g(x_2,y;\zeta^{(0)})\left[\frac{K}{L_g}\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right]\nabla_y f(x_2,y;\xi)\Bigg\|^2 \\
&\le 3C_{g_{xy}}^2\frac{K^2}{L_g^2}\mathbb{E}_k\left[\left(1-\frac{\mu_g}{L_g}\right)^{2k}\right]L_{f_y}^2\|x_1-x_2\|^2 + 3C_{f_y}^2\frac{K^2}{L_g^2}\mathbb{E}_k\left[\left(1-\frac{\mu_g}{L_g}\right)^{2k}\right]L_{g_{xy}}^2\|x_1-x_2\|^2 \\
&\quad + 3C_{g_{xy}}^2 C_{f_y}^2\frac{K^2}{L_g^2}\mathbb{E}_k\left\|\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)-\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right\|^2 \\
&\overset{(a)}{\le} 3C_{g_{xy}}^2 L_{f_y}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right)\|x_1-x_2\|^2 + 3C_{f_y}^2 L_{g_{xy}}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right)\|x_1-x_2\|^2 \\
&\quad + 3C_{g_{xy}}^2 C_{f_y}^2\frac{K^2}{L_g^2}\mathbb{E}_k\left\|\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)-\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right\|^2,
\end{aligned} \tag{27}
$$

where (a) follows from the fact that we have:

$$
\mathbb{E}_k\left[\left(1-\frac{\mu_g}{L_g}\right)^{2k}\right] = \frac{1}{K}\sum_{k=0}^{K-1}\left(1-\frac{\mu_g}{L_g}\right)^{2k} \le \frac{1}{K}\left(\frac{L_g^2}{2\mu_g L_g-\mu_g^2}\right),
$$

where the first equality above follows from the fact that $k \in \{0,1,\ldots,K-1\}$ is chosen uniformly at random and the second equality results from the sum of a geometric progression.

Finally, considering the last term of (27), we have

$$
\begin{aligned}
&\mathbb{E}_k\left\|\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})\right)-\prod_{i=1}^k\left(I-\frac{1}{L_g}\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\right)\right\|^2 \\
&\qquad\overset{(a)}{\le} K\sum_{i=1}^K\mathbb{E}_k\left[\left(1-\frac{\mu_g}{L_g}\right)^{2(k-1)}\right]\frac{1}{L_g^2}\|\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})-\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\|^2 \\
&\qquad\overset{(b)}{\le}\left(\frac{L_g^2}{(L_g-\mu_g)^2}\right)\left(\frac{1}{2\mu_g L_g-\mu_g^2}\right)\sum_{i=1}^K\|\nabla_{yy}^2 g(x_1,y;\zeta^{(i)})-\nabla_{yy}^2 g(x_2,y;\zeta^{(i)})\|^2 \\
&\qquad\overset{(c)}{\le}\frac{KL_g^2 L_{g_{yy}}^2}{(L_g-\mu_g)^2(2\mu_g L_g-\mu_g^2)}\|x_1-x_2\|^2,
\end{aligned} \tag{28}
$$

where (a) follows from the application of (22) in Lemma B.1 along with Assumption 2–(ii)(iii); inequality (b) utilizes

$$\mathbb{E}_k\left[\left(1-\frac{\mu_g}{L_g}\right)^{2(k-1)}\right]=\frac{1}{K}\sum_{k=0}^{K-1}\left(1-\frac{\mu_g}{L_g}\right)^{2(k-1)}\leq\frac{1}{K}\left(\frac{L_g^2}{(L_g-\mu_g)^2}\right)\left(\frac{L_g^2}{2\mu_g L_g-\mu_g^2}\right),$$

where the first equality above again utilizes the fact that $k\in\{0,1,\dots,K-1\}$ is chosen uniformly at random and the second equality results from the sum of a geometric progression; inequality (c) utilizes Assumption 2–(v) made for stochastic lower level objective.

Finally, taking expectation in (26) and substituting the expressions obtained in (27) and (28) in (26), we obtain

$$\mathbb{E}\|\bar{\nabla}f(x_1,y;\bar{\xi})-\bar{\nabla}f(x_2,y;\bar{\xi})\|^2\leq L_K^2\|x_1-x_2\|^2,$$

where L_K^2 defined as:

$$L_K^2:=2L_{f_x}^2+6C_{g_{xy}}^2L_{f_y}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right)+6C_{f_y}^2L_{g_{xy}}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right)$$
$$+6C_{g_{xy}}^2C_{f_y}^2\frac{K^3L_{g_{yy}}^2}{(L_g-\mu_g)^2(2\mu_g L_g-\mu_g^2)}.$$

Statement (i) of the Lemma is proved.

The proof of the statement (ii) follows the same procedure, so it is omitted. □

D Proof of Theorem 3.2: smooth (possibly non-convex) outer objective

First, we consider the descent achieved by the outer objective in consecutive iterates generated by the Algorithm 1 when the outer problem is smooth and is possibly non-convex. We define the following constants for the stepsize parameters:

$$w=\max\left\{2,\ 27L_f^3,\ 8L_{\mu_g}^3c_\beta^3,\ (\mu_g+L_g)^3c_\beta^3,\ c_{\eta_f}^{3/2},\ c_{\eta_g}^{3/2}\right\},\quad c_\beta=\frac{6\sqrt{2}L_yL}{L_{\mu_g}},$$
$$c_{\eta_f}=\frac{1}{3L_f}+\max\left\{36L_K^2,\frac{4L_K^2L_{\mu_g}(\mu_g+L_g)c_\beta^2}{L^2}\right\},\tag{29}$$
$$c_{\eta_g}=\frac{1}{3L_f}+8L_g^2c_\beta^2+\left[\frac{8L^2}{L_{\mu_g}^2}+\frac{2L^2}{L_{\mu_g}(\mu_g+L_g)}\right]\max\left\{36L_g^2,\frac{4L_g^2L_{\mu_g}(\mu_g+L_g)c_\beta^2}{L^2}\right\},$$

where we have defined $L_{\mu_g}=\frac{\mu_g L_g}{\mu_g+L_g}$.

D.1 Descent in the function value

Lemma D.1. *For non-convex and smooth $\ell(\cdot)$, with e_t^f defined as: $e_t^f:=h_t^f-\bar{\nabla}f(x_t,y_t)-B_t$, the consecutive iterates of Algorithm 1 satisfy:*

$$\mathbb{E}[\ell(x_{t+1})]\leq\mathbb{E}\Big[\ell(x_t)-\frac{\alpha_t}{2}\|\nabla\ell(x_t)\|^2-\frac{\alpha_t}{2}(1-\alpha_tL_f)\|h_t^f\|^2+\alpha_t\|e_t^f\|^2$$
$$+2\alpha_tL^2\|y_t-y^*(x_t)\|^2+2\alpha_t\|B_t\|^2\Big].$$

for all $t\in\{0,1,\dots,T-1\}$, where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Using the Lipschitz smoothness of the objective function from Lemma 2.2 we have:

$$\begin{aligned}\ell(x_{t+1})&\leq\ell(x_t)+\langle\nabla\ell(x_t),x_{t+1}-x_t\rangle+\frac{L_f}{2}\|x_{t+1}-x_t\|^2\\&\overset{(a)}{=}\ell(x_t)-\alpha_t\langle\nabla\ell(x_t),h_t^f\rangle+\frac{\alpha_t^2L_f}{2}\|h_t^f\|^2\\&\overset{(b)}{=}\ell(x_t)-\frac{\alpha_t}{2}\|\nabla\ell(x_t)\|^2-\frac{\alpha_t}{2}(1-\alpha_tL_f)\|h_t^f\|^2+\frac{\alpha_t}{2}\|h_t^f-\nabla\ell(x_t)\|^2.\end{aligned}\tag{30}$$

where (a) results from Step 7 of Algorithm 1 and (b) uses $\langle a, b\rangle = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a-b\|^2$. Next, we bound the term $\|h_t^f - \nabla \ell(x_t)\|^2$ as follows

$$\begin{aligned}
\|h_t^f - \nabla \ell(x_t)\|^2 &= \|h_t^f - \bar{\nabla} f(x_t, y_t) - B_t + \bar{\nabla} f(x_t, y_t) + B_t - \nabla \ell(x_t)\|^2 \\
&\overset{(c)}{\leq} 2\|h_t^f - \bar{\nabla} f(x_t, y_t) - B_t\|^2 + 4\|\bar{\nabla} f(x_t, y_t) - \nabla \ell(x_t)\|^2 + 4\|B_t\|^2 \\
&\overset{(d)}{\leq} 2\|e_t^f\|^2 + 4L^2\|y_t - y^*(x_t)\|^2 + 4\|B_t\|^2,
\end{aligned}$$

where inequality (c) uses (24) and (d) results from the definition of $e_t^f \coloneqq h_t^f - \bar{\nabla} f(x_t, y_t) - B_t$ and (11) in Lemma 2.2. Substituting the above in (30) and taking expectation w.r.t. the stochasticity of the algorithm we get the statement of the lemma. □

D.2 Descent in the iterates of the lower level problem

Lemma D.2. *Define $e_t^g \coloneqq h_t^g - \nabla_y g(x_t, y_t)$. then the iterates of the inner problem generated according to Algorithm 1, satisfy*

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\leq (1+\gamma_t)(1+\delta_t)\left(1 - 2\beta_t \frac{\mu_g L_g}{\mu_g + L_g}\right)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 \\
&\quad - (1+\gamma_t)(1+\delta_t)\left(\frac{2\beta_t}{\mu_g + L_g} - \beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t, y_t)\| + (1+\gamma_t)\left(1 + \frac{1}{\delta_t}\right)\beta_t^2 \mathbb{E}\|e_t^g\|^2.
\end{aligned}$$

for all $t \in \{0, \ldots, T-1\}$ with some $\gamma_t, \delta_t > 0.$, where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Consider the term $\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 &\overset{(a)}{\leq} (1+\gamma_t)\mathbb{E}\|y_{t+1} - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right)\mathbb{E}\|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\overset{(b)}{=} (1+\gamma_t)\mathbb{E}\|y_t - \beta_t h_t^g - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \mathbb{E}\|x_{t+1} - x_t\|^2 \\
&\overset{(c)}{\leq} (1+\gamma_t)(1+\delta_t)\mathbb{E}\|y_t - \beta_t \nabla_y g(x_t, y_t) - y^*(x_t)\|^2 \\
&\quad + (1+\gamma_t)\left(1 + \frac{1}{\delta_t}\right)\beta_t^2 \|h_t^g - \nabla_y g(x_t, y_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2.
\end{aligned} \tag{31}$$

where (a) results from the Young's inequality; (b) uses Step 5 of Algorithm 1 and Lipschitzness of $y^*(\cdot)$ in Lemma 2.2 and (c) again utilizes Young's inequality and Step 7 of Algorithm 1. Next, we consider the first term of the above equation we have

$$\begin{aligned}
&\|y_t - \beta_t \nabla_y g(x_t, y_t) - y^*(x_t)\|^2 \\
&\qquad = \|y_t - y^*(x_t)\|^2 + \beta_t^2 \|\nabla_y g(x_t, y_t)\|^2 - 2\beta_t \langle \nabla_y g(x_t, y_t), y_t - y^*(x_t)\rangle \\
&\qquad \overset{(d)}{\leq} \left(1 - 2\beta_t \frac{\mu_g L_g}{\mu_g + L_g}\right)\|y_t - y^*(x_t)\|^2 - \left(\frac{2\beta_t}{\mu_g + L_g} - \beta_t^2\right)\|\nabla_y g(x_t, y_t)\|^2,
\end{aligned}$$

where inequality (d) above results from the strong convexity of g, which implies

$$\langle \nabla_y g(x_t, y_t), y_t - y^*(x_t)\rangle \geq \frac{\mu_g L_g}{\mu_g + L_g}\|y_t - y^*(x_t)\|^2 + \frac{1}{\mu_g + L_g}\|\nabla_y g(x_t, y_t)\|^2.$$

Substituting in (31) and using the definition $e_t^g \coloneqq h_t^g - \nabla_y g(x_t, y_t)$ we get the statement of the lemma. □

D.3 Descent in the gradient estimation error of the outer function

Before presenting the descent in the gradient estimation error of the outer function we define $\mathcal{F}_t = \sigma\{y_0, x_0, \dots, y_t, x_t\}$ as the sigma algebra generated by the sequence of iterates up to the tth iteration of SUSTAIN.

Lemma D.3. *Define $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t) - B_t$. Then the consecutive iterates of Algorithm 1 satisfy:*

$$\begin{aligned}\mathbb{E}\|e_{t+1}^f\|^2 \leq (1-\eta_{t+1}^f)^2 \mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2\sigma_f^2 + 4(1-\eta_{t+1}^f)^2 L_K^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 \\ + 8(1-\eta_{t+1}^f)^2 L_K^2 \beta_t^2 \mathbb{E}\|e_t^g\|^2 + 8(1-\eta_{t+1}^f)^2 L_K^2 \beta_t^2 \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2,\end{aligned}$$

for all $t \in \{0, \dots, T-1\}$, with L_K defined in the statement of Lemma C.2. Here the expectation is taken w.r.t the stochasticity of the algorithm.

Proof. From the definition of e_t^f we have

$$\begin{aligned}
&\mathbb{E}\|e_{t+1}^f\|^2 \qquad (32)\\
&= \mathbb{E}\|h_{t+1}^f - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}\|^2 \\
&\overset{(a)}{=} \mathbb{E}\big\|\eta_{t+1}^f \bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1-\eta_{t+1})\big(h_t^f + \bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t; \xi_{t+1})\big) \\
&\qquad\qquad - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}\big\|^2 \\
&\overset{(b)}{=} \mathbb{E}\big\|(1-\eta_{t+1}^f)e_t^f + \eta_{t+1}^f(\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) \\
&\qquad + (1-\eta_{t+1}^f)\big((\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) \\
&\qquad\qquad - (\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t)\big)\big\|^2 \\
&\overset{(c)}{=} (1-\eta_{t+1}^f)^2 \mathbb{E}\big\|e_t^f\big\|^2 + \mathbb{E}\big\|\eta_{t+1}^f(\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) \\
&\qquad + (1-\eta_{t+1}^f)\big((\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) \\
&\qquad\qquad - (\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t)\big)\big\|^2 \\
&\overset{(d)}{\leq} (1-\eta_{t+1}^f)^2 \mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2 \mathbb{E}\big\|\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}\big\|^2 \\
&\qquad + 2(1-\eta_t^f)^2 \mathbb{E}\big\|\big(\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}\big) \\
&\qquad\qquad - \big(\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t\big)\big\|^2 \\
&\overset{(e)}{\leq} (1-\eta_{t+1}^f)^2 \mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2 \sigma_f^2 \\
&\qquad + 2(1-\eta_t^f)^2 \mathbb{E}\big\|\big(\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}\big) \\
&\qquad\qquad - \big(\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t\big)\big\|^2 \qquad (33)
\end{aligned}$$

where equality (a) uses the definition of the recursive gradient estimator (14); (b) results from the definition $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t) - B_t$; (c) follows from the fact that conditioned on $\mathcal{F}_{t+1} = \sigma\{y_0, x_0, \dots, y_t, x_t, y_{t+1}, x_{t+1}\}$

$$\begin{aligned}
&\mathbb{E}\Big\langle e_t^f, (\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) \\
&\qquad - (1-\eta_{t+1}^f)\big((\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t)\big)\Big\rangle \\
&\mathbb{E}\Big\langle e_t^f, \underbrace{\mathbb{E}\big[(\bar{\nabla} f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}) - B_{t+1}) - (1-\eta_{t+1}^f)\big((\bar{\nabla} f(x_t, y_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t) - B_t)\big)|\mathcal{F}_{t+1}\big]}_{=0}\Big\rangle = 0,
\end{aligned}$$

which follows from the fact that the second term in the inner product above is zero mean as a consequence of Assumption 4-(i) and inequality (d) utilizes (24); and (e) results from Assumption 4-(i).

Next, we bound the last term of (33) above

$$
\begin{aligned}
&2(1-\eta^f_{t+1})^2\mathbb{E}\big\|\big(\bar{\nabla} f(x_{t+1},y_{t+1};\xi_{t+1})-\bar{\nabla} f(x_t,y_t;\xi_{t+1})\big)\\
&\qquad\qquad\qquad -\Big(\big(\bar{\nabla} f(x_{t+1},y_{t+1})+B_{t+1}\big)-\big(\bar{\nabla} f(x_t,y_t)+B_t\big)\Big)\big\|^2\\
&\overset{(a)}{\leq} 2(1-\eta^f_{t+1})^2\mathbb{E}\big\|\bar{\nabla} f(x_{t+1},y_{t+1};\xi_{t+1})-\bar{\nabla} f(x_t,y_t;\xi_{t+1})\|^2\\
&\overset{(b)}{\leq} 4(1-\eta^f_{t+1})^2\mathbb{E}\big\|\bar{\nabla} f(x_{t+1},y_{t+1};\xi_{t+1})-\bar{\nabla} f(x_t,y_{t+1};\xi_{t+1})\big\|^2\\
&\qquad\qquad\qquad + 4(1-\eta^f_{t+1})^2\mathbb{E}\big\|\bar{\nabla} f(x_t,y_{t+1};\xi_{t+1})-\bar{\nabla} f(x_t,y_t;\xi_{t+1})\big\|^2\\
&\overset{(c)}{\leq} 4(1-\eta^f_{t+1})^2L_K^2\mathbb{E}\|x_{t+1}-x_t\|^2+4(1-\eta^f_{t+1})^2L_K^2\mathbb{E}\|y_{t+1}-y_t\|^2\\
&\overset{(d)}{\leq} 4(1-\eta^f_{t+1})^2L_K^2\alpha_t^2\mathbb{E}\|h_t^f\|^2+4(1-\eta^f_{t+1})^2L_K^2\beta_t^2\mathbb{E}\|h_t^g\|^2,\\
&\overset{(e)}{\leq} 4(1-\eta^f_{t+1})^2L_K^2\alpha_t^2\mathbb{E}\|h_t^f\|^2+8(1-\eta^f_{t+1})^2L_K^2\beta_t^2\mathbb{E}\|e_t^g\|^2\\
&\qquad\qquad\qquad + 8(1-\eta^f_{t+1})^2L_K^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2, \qquad (34)
\end{aligned}
$$

where (a) follows from the mean variance inequality: For a random variable Z we have $\mathbb{E}\|Z-\mathbb{E}[Z]\|^2 \leq \mathbb{E}\|Z\|^2$ with Z defined as $Z := \bar{\nabla} f(x_{t+1},y_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_t;\xi_{t+1})$; (b) again uses (24); (c) follows from Lemma C.2; inequality (d) uses Steps 5 and 7 of Algorithm 1; finally, (e) utilizes (24) and the definition of e_t^g.

Finally, substituting (34) in (33), we get the statement of the lemma.

Therefore, the lemma is proved. □

D.4 Descent in the gradient estimation error of the inner function

We consider the descent on the gradient estimation error of the inner function.

Lemma D.4. *Define $e_t^g := h_t^g - \nabla_y g(x_t,y_t)$. Then the iterates generated from Algorithm 1 satisfy*

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 \leq &\Big((1-\eta^g_{t+1})^2+8(1-\eta^g_{t+1})^2L_g^2\beta_t^2\Big)\mathbb{E}\|e_t^g\|^2+2(\eta^g_{t+1})^2\sigma_g^2\\
&+4(1-\eta^g_{t+1})^2L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2+8(1-\eta^g_{t+1})^2L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2
\end{aligned}
$$

for all $t \in \{0,1,\cdots,T-1\}$, where the expectation is taken w.r.t. the stochasticity of the algorithm.

Proof. From the definition of e_t^g we have

$$
\begin{aligned}
&\mathbb{E}\|e_{t+1}^g\|^2 = \mathbb{E}\|h_{t+1}^g - \nabla_y g(x_{t+1}, y_{t+1})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) + (1-\eta_{t+1}^g)\big(h_t^g - \nabla_y g(x_t, y_t; \zeta_{t+1})\big) - \nabla_y g(x_{t+1}, y_{t+1})\|^2 \\
&\overset{(b)}{=} \mathbb{E}\big\|(1-\eta_{t+1}^g)e_t^g + \big(\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_{t+1}, y_{t+1})\big) \\
&\qquad\qquad\qquad\qquad - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t, y_t; \zeta_{t+1}) - \nabla_y g(x_t, y_t)\big)\big\|^2 \\
&\overset{(c)}{=} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 \\
&+ \mathbb{E}\|\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_{t+1}, y_{t+1}) - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t, y_t; \zeta_{t+1}) - \nabla_y g(x_t, y_t)\big)\|^2 \\
&\overset{(d)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 2(1-\eta_{t+1}^g)^2\mathbb{E}\|\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_t, y_t; \zeta_{t+1})\|^2 \\
&\overset{(e)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 \\
&\qquad + 4(1-\eta_{t+1}^g)^2\mathbb{E}\|\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_t, y_{t+1}; \zeta_{t+1})\|^2 \\
&\qquad\qquad + 4(1-\eta_{t+1}^g)^2\mathbb{E}\|\nabla_y g(x_t, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_t, y_t; \zeta_{t+1})\|^2 \\
&\overset{(f)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\mathbb{E}\|x_{t+1} - x_t\|^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\mathbb{E}\|y_{t+1} - y_t\|^2 \\
&\overset{(g)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2\mathbb{E}\|h_t^g\|^2 \\
&\overset{(h)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\
&\qquad\qquad + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 \\
&\leq \Big((1-\eta_{t+1}^g)^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2\Big)\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 4(1-\eta_{t+1}^g)^2 L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\
&\qquad\qquad\qquad + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2,
\end{aligned}
$$

where equality (a) uses the definition of hybrid gradient estimator (13); (b) uses the definition of e_t^g; (c) uses the fact that conditioned on $\mathcal{F}_{t+1} = \sigma\{y_0, x_0, \ldots, y_t, x_t, y_{t+1}, x_{t+1}\}$

$$
\begin{aligned}
&\mathbb{E}\big\langle e_t^g, \big(\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_{t+1}, y_{t+1})\big) - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t, y_t; \zeta_{t+1}) - \nabla_y g(x_t, y_t)\big)\big\rangle \\
&= \mathbb{E}\big\langle e_t^g, \underbrace{\mathbb{E}\big[\big(\nabla_y g(x_{t+1}, y_{t+1}, \zeta_{t+1}) - \nabla_y g(x_{t+1}, y_{t+1})\big) - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t, y_t; \zeta_{t+1}) - \nabla_y g(x_t, y_t)\big)|\mathcal{F}_{t+1}\big]}_{=0}\big\rangle \\
&\quad = 0.
\end{aligned}
$$

Inequality (d) results from the application of (24) and Assumption 4-(ii); (e) again uses (24); (f) utilizes Assumption 2; (g) follows from Steps 5 and 7 of Algorithm 1 and finally, (h) follows from the application of (24) and the definition of e_t^g.

Therefore, the lemma is proved. □

D.5 Descent in the potential function

Let us define the potential function as:

$$
V_t := \ell(x_t) + \frac{2L}{3\sqrt{2}L_y}\|y_t - y^*(x_t)\|^2 + \frac{1}{\bar{c}_{\eta_f}}\frac{\|e_t^f\|^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_g}}\frac{\|e_t^g\|^2}{\alpha_{t-1}} \tag{35}
$$

where we define

$$
\bar{c}_{\eta_f} := \max\left\{36L_K^2, \frac{4L_K^2 L_{\mu_g}(\mu_g + L_g)c_\beta^2}{L^2}\right\} \quad \text{and} \quad \bar{c}_{\eta_g} := \max\left\{36L_g^2, \frac{4L_g^2 L_{\mu_g}(\mu_g + L_g)c_\beta^2}{L^2}\right\}. \tag{36}
$$

with L_{μ_g} defined as $L_{\mu_g} := \frac{\mu_g L_g}{\mu_g + L_g}$.

Next, we quantify the expected descent in the potential function $\mathbb{E}[V_{t+1} - V_t]$.

Lemma D.5. *Consider V_t defined in (35). Suppose the parameters of Algorithm 1 are chosen as*

$$\alpha_t := \frac{1}{(w+t)^{1/3}},\ \beta_t := c_\beta \alpha_t,\ \eta_{t+1}^f := c_{\eta_f}\alpha_t^2,\ \textit{and}\ \eta_{t+1}^g := c_{\eta_g}\alpha_t^2\ \textit{for all}\ \ t \in \{0,1,\dots,T-1\}.$$

with

$$c_\beta := \frac{6\sqrt{2}L_y L}{L_{\mu_g}},\ c_{\eta_f} := \frac{1}{3L_f} + \bar{c}_{\eta_f}\ \textit{and}\ c_{\eta_g} := \frac{1}{3L_f} + 8L_g^2 c_\beta^2 + \left[\frac{8L^2}{L_{\mu_g}^2} + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\right]\bar{c}_{\eta_g},$$

where $L_{\mu_g} := \frac{\mu_g L_g}{\mu_g + L_g}$ and

$$\bar{c}_{\eta_f} = \max\left\{36L_K^2, \frac{4L_K^2 L_{\mu_g}(\mu_g+L_g)c_\beta^2}{L^2}\right\} \quad \textit{and} \quad \bar{c}_{\eta_g} = \max\left\{36L_g^2, \frac{4L_g^2 L_{\mu_g}(\mu_g+L_g)c_\beta^2}{L^2}\right\}.$$

and the parameters

$$\gamma_t := \frac{\beta_t L_{\mu_g}/2}{1-\beta_t L_{\mu_g}} \qquad \textit{and} \qquad \delta_t := \frac{\beta_t L_{\mu_g}}{1-2\beta_t L_{\mu_g}}.$$

Then the iterates generated by Algorithm 1 when the outer problem is non-convex satisfy:

$$\mathbb{E}[V_{t+1} - V_t] \le -\frac{\alpha_t}{2}\mathbb{E}\|\nabla \ell(x_t)\|^2 + 2\alpha_t\|B_t\|^2 + \frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2.$$

for all $t \in \{0,1,\dots,T-1\}$

Proof. We have from Lemma D.2

$$\begin{aligned}\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2 \le & \left[(1+\gamma_t)(1+\delta_t)\left(1 - 2\beta_t\frac{\mu_g L_g}{\mu_g + L_g}\right) - 1\right]\mathbb{E}\|y_t - y^*(x_t)\|^2 \\ & - (1+\gamma_t)(1+\delta_t)\left(\frac{2\beta_t}{\mu_g + L_g} - \beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t,y_t)\| \\ & + (1+\gamma_t)\left(1+\frac{1}{\delta_t}\right)\beta_t^2\mathbb{E}\|e_t^g\|^2 + \left(1+\frac{1}{\gamma_t}\right)L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2.\end{aligned} \tag{37}$$

Let us consider coefficient of the first term of (37) above, choosing γ_t and δ_t such that we have

$$(1+\gamma_t)(1+\delta_t)(1-2\beta_t L_{\mu_g}) = 1 - \frac{\beta_t L_{\mu_g}}{2} \tag{38}$$

where we define $L_{\mu_g} := \frac{\mu_g L_g}{\mu_g + L_g}$. First we choose γ_t such that we have

$$(1+\delta_t)(1-2\beta_t L_{\mu_g}) = 1 - \beta_t L_{\mu_g} \quad \Rightarrow \quad 1+\delta_t = \frac{1-\beta_t L_{\mu_g}}{1-2\beta_t L_{\mu_g}} \quad \Rightarrow \quad \delta_t = \frac{\beta_t L_{\mu_g}}{1-2\beta_t L_{\mu_g}}$$

Moreover, this implies that we have:

$$1 + \frac{1}{\delta_t} = 1 + \frac{1-2\beta_t L_{\mu_g}}{\beta_t L_{\mu_g}} \le \frac{1}{\beta_t L_{\mu_g}}.$$

Using the definition of δ_t in (38) we

$$(1+\gamma_t)(1-\beta_t L_{\mu_g}) = 1 - \frac{\beta_t L_{\mu_g}}{2} \quad \Rightarrow \quad 1+\gamma_t = \frac{1-\frac{\beta_t L_{\mu_g}}{2}}{1-\beta_t L_{\mu_g}} \quad \Rightarrow \quad \gamma_t = \frac{\beta_t L_{\mu_g}/2}{1-\beta_t L_{\mu_g}}$$

Moreover, this implies that we have:

$$1 + \frac{1}{\gamma_t} = 1 + \frac{1-\beta_t L_{\mu_g}}{\beta_t L_{\mu_g}/2} \le \frac{2}{\beta_t L_{\mu_g}}.$$

Substituting the above bounds in (37), we get

$$\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2 \leq -\frac{\beta_t L_{\mu_g}}{2}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \left(\frac{2\beta_t}{\mu_g + L_g} - \beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t, y_t)\| + \frac{2}{\beta_t L_{\mu_g}}\beta_t^2 \mathbb{E}\|e_t^g\|^2 + \frac{2}{\beta_t L_{\mu_g}} L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2.$$

Choosing $\beta_t \leq \frac{1}{\mu_g + L_g}$ we get

$$\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2 \leq -\frac{\beta_t L_{\mu_g}}{2}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{\beta_t}{\mu_g + L_g}\mathbb{E}\|\nabla_y g(x_t, y_t)\| + \frac{2}{\beta_t L_{\mu_g}}\beta_t^2 \mathbb{E}\|e_t^g\|^2 + \frac{2}{\beta_t L_{\mu_g}} L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2.$$

Using the definition of $\beta_t = c_\beta \alpha_t$ and multiplying both sides by $\frac{4L^2}{c_\beta L_{\mu_g}}$ we get

$$\frac{4L^2}{c_\beta L_{\mu_g}}\mathbb{E}\big[\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2\big] \leq -2\alpha_t L^2 \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{4L^2\alpha_t}{L_{\mu_g}(\mu_g + L_g)}\mathbb{E}\|\nabla_y g(x_t, y_t)\| + \frac{8L^2\alpha_t}{L_{\mu_g}^2}\mathbb{E}\|e_t^g\|^2 + \frac{8L_y^2 L^2 \alpha_t}{c_\beta^2 L_{\mu_g}^2}\mathbb{E}\|h_t^f\|^2.$$

Finally, choosing $c_\beta = \frac{6\sqrt{2}L_y L}{L_{\mu_g}}$ such that $\frac{8L_y^2 L^2}{c_\beta^2 L_{\mu_g}^2} = \frac{1}{9}$

$$\frac{2L}{3\sqrt{2}L_y}\mathbb{E}\big[\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2\big] \leq -2\alpha_t L^2 \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{4L^2\alpha_t}{L_{\mu_g}(\mu_g + L_g)}\mathbb{E}\|\nabla_y g(x_t, y_t)\| + \frac{8L^2\alpha_t}{L_{\mu_g}^2}\mathbb{E}\|e_t^g\|^2 + \frac{\alpha_t}{9}\mathbb{E}\|h_t^f\|^2. \tag{39}$$

Next, we have from Lemma D.3

$$\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_{t-1}} \leq \left[\frac{(1 - \eta_{t+1}^f)^2}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right]\mathbb{E}\|e_t^f\|^2 + \frac{2(\eta_{t+1}^f)^2}{\alpha_t}\sigma_f^2 + 4L_K^2\alpha_t\mathbb{E}\|h_t^f\|^2 + \frac{8L_K^2\beta_t^2}{\alpha_t}\mathbb{E}\|e_t^g\|^2 + \frac{8L_K^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2, \tag{40}$$

where we have utilized the fact that $0 < 1 - \eta_t < 1$ for all $t \in \{0, 1, \ldots, T-1\}$. Now we consider the coefficient of the first term on the right hand side of (40), we have

$$\frac{(1 - \eta_{t+1}^f)^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{1}{\alpha_t} - \frac{\eta_{t+1}^f}{\alpha_t} - \frac{1}{\alpha_{t-1}}. \tag{41}$$

Using the definition of α_t we have

$$\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} = (w+t)^{1/3} - (w+t-1)^{1/3}] \overset{(a)}{\leq} \frac{1}{3(w+t-1)^{2/3}} \overset{(b)}{\leq} \frac{1}{3(w/2+t)^{2/3}}$$
$$= \frac{2^{2/3}}{3(w+2t)^{2/3}} \leq \frac{2^{2/3}}{3(w+t)^{2/3}} \overset{(c)}{\leq} \frac{2^{2/3}}{3}\alpha_t^2 \overset{(d)}{\leq} \frac{\alpha_t}{3L_f},$$

where (a) follows from $(x+y)^{1/3} - x^{1/3} \leq y/(3x^{2/3})$; (b) results from the fact that we choose $w \geq 2$ hence $1 \leq w/2$; (c) results from the definition of α_t and (d) uses the fact that we choose $\alpha_t \leq 1/3L_f$. Substituting in (41) and using $\eta_{t+1}^f = c_{\eta_f}\alpha_t^2$, we get

$$\frac{(1 - \eta_{t+1}^f)^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f} - c_{\eta_f}\alpha_t \leq -\bar{c}_{\eta_f}\alpha_t,$$

which follows from the choice

$$c_{\eta_f} = \frac{1}{3L_f} + \bar{c}_{\eta_f} \quad \text{with} \quad \bar{c}_{\eta_f} = \max\left\{36L_K^2, \frac{4L_K^2 L_{\mu_g}(\mu_g + L_g)c_\beta^2}{L^2}\right\}.$$

Substiuting in (40)

$$\frac{1}{\bar{c}_{\eta_f}}\mathbb{E}\left[\frac{\|e_{t+1}^f\|^2}{\alpha_t} - \frac{\|e_{t+1}^f\|^2}{\alpha_{t-1}}\right] \le -\alpha_t\mathbb{E}\|e_t^f\|^2 + \frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2 + \frac{\alpha_t}{9}\mathbb{E}\|h_t^f\|^2 + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\alpha_t\mathbb{E}\|e_t^g\|^2 + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2, \tag{42}$$

Next, from Lemma D.4, we have

$$\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}} \le \left[\frac{(1-\eta_{t+1}^g)^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right]\mathbb{E}\|e_t^g\|^2 + \frac{2(\eta_{t+1}^g)^2}{\alpha_t}\sigma_g^2 + 4L_g^2\alpha_t\mathbb{E}\|h_t^f\|^2 + \frac{8L_g^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 \tag{43}$$

where we have utilized the fact that $0 < 1 - \eta_t^g \le 1$ for all $t \in \{0, 1, \dots, T-1\}$. Let us consider the coefficient of the first term on the right hand side of (43) we have

$$\begin{aligned}\frac{(1-\eta_{t+1}^g)^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} &\le \frac{(1-\eta_{t+1}^g)}{\alpha_t}\left(1 + 8L_g^2\beta_t^2\right) - \frac{1}{\alpha_{t-1}} \\ &= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} + \frac{8L_g^2\beta_t^2}{\alpha_t} - c_{\eta_g}\alpha_t(1 + 8L_g^2\beta_t^2),\end{aligned}$$

using the fact that from earlier we have $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \le \frac{\alpha_t}{3L_f}$ and the definition of $\beta_t = c_\beta\alpha_t$, we have

$$\frac{(1-\eta_{t+1}^g)^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \le \frac{\alpha_t}{3L_f} + 8L_g^2 c_\beta^2\alpha_t - c_{\eta_g}\alpha_t,$$

Next choosing c_{η_g} as

$$c_{\eta_g} = \frac{1}{3L_f} + 8L_g^2c_\beta^2 + \left[\frac{8L^2}{L_{\mu_g}^2} + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\right]\bar{c}_{\eta_g} \quad \text{with} \quad \bar{c}_{\eta_g} = \max\left\{36L_g^2, \frac{4L_g^2 L_{\mu_g}(\mu_g + L_g)c_\beta^2}{L^2}\right\}.$$

Therefore, we get

$$\frac{(1-\eta_{t+1}^g)^2 + 8(1-\eta_{t+1}^g)^2 L_g^2\beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \le -\left[\frac{8L^2}{L_{\mu_g}^2} + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\right]\bar{c}_{\eta_g}\alpha_t,$$

Finally, replacing in (43) we get

$$\frac{1}{\bar{c}_{\eta_g}}\mathbb{E}\left[\frac{\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\|e_t^g\|^2}{\alpha_{t-1}}\right] \le -\left[\frac{8L^2}{L_{\mu_g}^2} + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\right]\alpha_t\mathbb{E}\|e_t^g\|^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2 + \frac{\alpha_t}{9}\mathbb{E}\|h_t^f\|^2 + \frac{2L^2}{L_{\mu_g}(\mu_g + L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 \tag{44}$$

Finally, adding (39), (42), (44) and the result of Lemma D.1 with $\alpha_t \le 1/3L_f$, we get

$$\mathbb{E}[V_{t+1} - V_t] \le -\frac{\alpha_t}{2}\mathbb{E}\|\nabla\ell(x_t)\|^2 + 2\alpha_t\|B_t\|^2 + \frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2.$$

Therefore, we have the statement of the Lemma.

D.6 Proof of Theorem 3.2

Summing the result of Lemma D.5 for $t = 0$ to $T - 1$, dividing by T on both sides and using the definition $\eta_{t+1}^f := c_{\eta_f}\alpha_t^2$ and $\eta_{t+1}^g := c_{\eta_g}\alpha_t^2$ we get

$$\frac{\mathbb{E}[V_T - V_0]}{T} \le -\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\ell(x_t)\|^2 + \frac{2}{T}\sum_{t=0}^{T}\alpha_t\|B_t\|^2 + \frac{2c_{\eta_f}^2\sigma_f^2}{\bar{c}_{\eta_f}}\sum_{t=0}^{T-1}\alpha_t^3 + \frac{2c_{\eta_g}^2\sigma_g^2}{\bar{c}_{\eta_g}}\sum_{t=0}^{T-1}\alpha_t^3. \tag{45}$$

Next considering $\sum_{t=0}^{T-1}\alpha_t$ in the last two terms on the right hand side of (45), we have from the definition of α_t that

$$\sum_{t=0}^{T-1}\alpha_t^3 = \sum_{t=0}^{T-1}\frac{1}{w+t} \overset{(a)}{\le} \sum_{t=0}^{T-1}\frac{1}{1+t} \le \log(T+1)$$

where inequality (a) results from the fact that we choose $w \ge 1$. Substituting the above in (45) we get

$$\frac{\mathbb{E}[V_T - V_0]}{T} \le -\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\ell(x_t)\|^2 + \frac{2}{T}\sum_{t=0}^{T}\alpha_t\|B_t\|^2 + \frac{2c_{\eta_f}^2}{\bar{c}_{\eta_f}}\frac{\log(T+1)}{T}\sigma_f^2 + \frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\frac{\log(T+1)}{T}\sigma_g^2$$

Rearranging the terms we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\ell(x_t)\|^2 \le \frac{\mathbb{E}[V_0 - \ell^*]}{T} + \frac{2}{T}\sum_{t=0}^{T}\alpha_t\|B_t\|^2 + \frac{2c_{\eta_f}^2}{\bar{c}_{\eta_f}}\frac{\log(T+1)}{T}\sigma_f^2 + \frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\frac{\log(T+1)}{T}\sigma_g^2$$

Using the fact that α_t is decreasing in t we have $\alpha_T \le \alpha_t$ for all $t \in \{0, 1, \dots, T-1\}$ and multiplying by $2/\alpha_T$ on both sides we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\ell(x_t)\|^2 \le \frac{2\mathbb{E}[V_0 - \ell^*]}{\alpha_T T} + \frac{4}{\alpha_T T}\sum_{t=0}^{T}\alpha_t\|B_t\|^2 + \frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\frac{\log(T+1)}{\alpha_T T}\sigma_f^2 + \frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\frac{\log(T+1)}{\alpha_T T}\sigma_g^2$$

Finally, we have from the definition of the Potential function

$$\begin{aligned}\mathbb{E}[V_0] &:= \mathbb{E}\left[\ell(x_0) + \frac{2L}{3\sqrt{2}L_y}\|y_0 - y^*(x_0)\|^2 + \frac{1}{\bar{c}_{\eta_f}}\frac{\|e_0^f\|^2}{\alpha_{-1}} + \frac{1}{\bar{c}_{\eta_g}}\frac{\|e_0^g\|^2}{\alpha_{-1}}\right] \\ &\le \ell(x_0) + \frac{2L}{3\sqrt{2}L_y}\|y_0 - y^*(x_0)\|^2 + \frac{\sigma_f^2}{\bar{c}_{\eta_f}\alpha_{-1}} + \frac{\sigma_g^2}{\bar{c}_{\eta_g}\alpha_{-1}},\end{aligned}$$

which follows from the assumption and the definition of h_t^f and h_t^g. Therefore, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\ell(x_t)\|^2 \le{}& \frac{2(\ell(x_0) - \ell^*)}{\alpha_T T} + \frac{4L}{3\sqrt{2}L_y}\frac{\|y_0 - y^*(x_0)\|^2}{\alpha_T T} + \frac{2}{\bar{c}_{\eta_f}\alpha_{-1}}\frac{\sigma_f^2}{\alpha_T T} + \frac{2}{\bar{c}_{\eta_g}\alpha_{-1}}\frac{\sigma_g^2}{\alpha_T T} \\ &+ \frac{4}{\alpha_T T}\sum_{t=0}^{T}\alpha_t\|B_t\|^2 + \frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\frac{\log(T+1)}{\alpha_T T}\sigma_f^2 + \frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\frac{\log(T+1)}{\alpha_T T}\sigma_g^2\end{aligned}$$

Finally, we have from the definition of $\alpha_T := 1/(w+T)^{1/3}$ and $\alpha_1 = \alpha_0$, moreover using the fact that for the choice of $K = (L_g/\mu_g)\log(C_{g_{xy}}C_{f_y}T/\mu_g)$ stochastic Hessian samples of $\nabla_{yy}^2 g(x,y)$ we have $\|B_t\| = 1/T$, we get

$$\mathbb{E}\|\nabla\ell(x_a(T))\|^2 \le \mathcal{O}\left(\frac{\ell(x_0) - \ell^*}{T^{2/3}}\right) + \mathcal{O}\left(\frac{\|y_0 - y^*(x_0)\|^2}{T^{2/3}}\right) + \tilde{\mathcal{O}}\left(\frac{\sigma_f^2}{T^{2/3}}\right) + \tilde{\mathcal{O}}\left(\frac{\sigma_g^2}{T^{2/3}}\right).$$

Hence, the theorem is proved. □

E Proof of Theorem 3.3: strongly-convex outer objective

To prove Theorem 3.3, we utilize the descent results obtained for the proof of Theorem 3.2 in Appendix D. The proof follows similar structure as the proof of non-convex case. We first consider the descent achieved by the consecutive iterates generated by Algorithm 1 when the outer function is strongly-convex and smooth.

E.1 Descent in the function value

Lemma E.1. *For strongly-convex and smooth $\ell(\cdot)$, with e_t^f defined as: $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_{t+1}) - B_t$, the consecutive iterates of Algorithm 1 satisfy:*

$$\mathbb{E}[\ell(x_{t+1}) - \ell^*] \leq \mathbb{E}\Big[(1 - \alpha_t \mu_f)\big(\ell(x_t) - \ell^*\big) - \frac{\alpha_t}{2}(1 - \alpha_t L_f)\|h_t^f\|^2 + \alpha_t \|e_t^f\|^2 + 2\alpha_t L^2 \|y_t - y^*(x_t)\|^2 + 2\alpha_t \|B_t\|^2\Big],$$

for all $t \in \{0, 1, \ldots, T-1\}$, where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Note that from Lemma D.1 derived in Appendix D, we have

$$\mathbb{E}[\ell(x_{t+1})] \leq \mathbb{E}\Big[\ell(x_t) - \frac{\alpha_t}{2}\|\nabla \ell(x_t)\|^2 - \frac{\alpha_t}{2}(1 - \alpha_t L_f)\|h_t^f\|^2 + \alpha_t \|e_t^f\|^2 + 2\alpha_t L^2 \|y_t - y^*(x_t)\|^2 + 2\alpha_t \|B_t\|^2\Big]. \tag{46}$$

Now using the fact that for a strongly convex function we have:

$$\|\nabla \ell(x)\|^2 \geq 2\mu_f(\ell(x) - \ell^*) \quad \text{for all} \quad x \in \mathbb{R}^{d_{\text{up}}},$$

substituting in (46), subtracting ℓ^* from both sides and rearranging the terms yields the statement of the Lemma. □

E.2 Descent in the iterates of the lower level problem

Lemma E.2. *The iterates of the inner problem generated according to Algorithm 1, satisfy*

$$\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 + \gamma_t)\big(1 - 2\beta_t \mu_g + \beta_t^2 L_g^2\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 + (1 + \gamma_t)\beta_t^2 \sigma_g^2.$$

for all $t \in \{0, \ldots, T-1\}$ with some $\gamma_t > 0$, where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Consider the term $\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2$, we have

$$\begin{aligned}\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 &\overset{(a)}{\leq} (1 + \gamma_t)\mathbb{E}\|y_{t+1} - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right)\mathbb{E}\|y^*(x_{t+1}) - y^*(x_t)\|^2 \\ &\overset{(b)}{\leq} (1 + \gamma_t)\mathbb{E}\|y_t - \beta_t h_t^g - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \mathbb{E}\|x_{t+1} - x_t\|^2 \\ &\overset{(c)}{\leq} (1 + \gamma_t)\mathbb{E}\|y_t - \beta_t h_t^g - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 \end{aligned} \tag{47}$$

where (a) results from Young's inequality; (b) uses Step 5 of Algorithm 1 and Lipschitzness of $y^*(\cdot)$ given in Lemma 2.2; and (c) uses Step 7 of Algorithm 1.

Next, we consider the first term of (47) above:

$$\begin{aligned}\mathbb{E}\|y_t - \beta_t h_t^g - y^*(x_t)\|^2 &= \mathbb{E}\|y_t - y^*(x_t)\|^2 + \beta_t^2 \mathbb{E}\|h_t^g\|^2 - \beta_t \mathbb{E}\langle y_t - y^*(x_t), h_t^g\rangle \\ &\overset{(a)}{\leq} \mathbb{E}\|y_t - y^*(x_t)\|^2 + \beta_t^2 \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 + \beta_t^2 \mathbb{E}\|h_t^g - \nabla_y g(x_t, y_t)\|^2 \\ &\qquad - \beta_t \mathbb{E}\langle y_t - y^*(x_t), \nabla_y g(x_t, y_t)\rangle \\ &\overset{(b)}{\leq} (1 - 2\mu_g \beta_t + \beta_t^2 L_g^2)\mathbb{E}\|y_t - y^*(x_t)\|^2 + + \beta_t^2 \sigma_g^2 \end{aligned} \tag{48}$$

where (a) utilizes the fact that for $\eta_t^g = 1$ we have $\mathbb{E}[h_t^g|\mathcal{F}_t] = \nabla_y g(x_t, y_t)$ and (b) uses the fact that (1) $\nabla_y g(x, y^*(x)) = 0$ and the Lipschitzness of $\nabla_y g(x, \cdot)$ in Assumption 2-(ii); (2) Assumption 4-(ii); and (3) $g(x, y)$ is μ_g-strongly convex w.r.t. y, we therefore have

$$\big\langle \nabla g_y(x, y_1) - \nabla g_y(x, y_2), y_1 - y_2 \big\rangle \geq \mu_g \|y_1 - y_2\|^2,$$

using $y_1 = y_t$ and $y_2 = y^*(x_t)$ yields inequality (b). Finally, substituting (48) in (47) yields the statement of the lemma. □

E.3 Descent in the gradient estimation error

Lemma E.3. *Define $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t) - B_t$. Then the consecutive iterates of Algorithm 1 satisfy:*

$$\mathbb{E}\|e_{t+1}^f\|^2 \leq (1 - \eta_{t+1}^f)^2 \mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2 \sigma_f^2 + 4(1 - \eta_{t+1}^f)^2 L_K^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 \\ + 8(1 - \eta_{t+1}^f)^2 L_K^2 \beta_t^2 \sigma_g^2 + 8(1 - \eta_{t+1}^f)^2 L_K^2 L_g^2 \beta_t^2 \mathbb{E}\|y_t - y^*(x_t)\|^2,$$

for all $t \in \{0, \ldots, T-1\}$, with L_K defined in the statement of Lemma C.2. Here the expectation is taken w.r.t the stochasticity of the algorithm.

Proof. From the statement of Lemma D.3, we have

$$\mathbb{E}\|e_{t+1}^f\|^2 \leq (1 - \eta_{t+1}^f)^2 \mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2 \sigma_f^2 + 4(1 - \eta_{t+1}^f)^2 L_K^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 \\ + 8(1 - \eta_{t+1}^f)^2 L_K^2 \beta_t^2 \mathbb{E}\|e_t^g\|^2 + 8(1 - \eta_{t+1}^f)^2 L_K^2 \beta_t^2 \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2,$$

The proof follows by noticing the fact that for the gradient estimate h_t^g with $\eta_t^g = 1$, we have $\mathbb{E}\|e_t^g\|^2 \leq \sigma_g^2$ from Assumption 4-(ii) and the Lipschitzness of $\nabla_y g(x, \cdot)$ combined with the fact that $\nabla_y g(x, y^*(x)) = 0$. □

E.4 Descent in potential function

In this section, we define the potential function as:

$$\widehat{V}_t := (\ell(x_t) - \ell^*) + \|e_t^f\|^2 + \|y_t - y^*(x_t)\|^2, \tag{49}$$

which is different from that of (35). We next show that the potential function decreases with appropriate choice of parameters.

Lemma E.4. *With the potential function, $\widehat{V}_t$, defined in (49), with the choice of parameters*

$$\eta_{t+1}^f = (\mu_f + 1)\alpha_t, \ \beta_t = \hat{c}_\beta \alpha_t \text{ with } \hat{c}_\beta = \frac{8L_y^2 + 8L^2 + 2\mu_f}{\mu_g} \text{ and } \gamma_t = \frac{\mu_g \beta_t}{2(1 - \mu_g \beta_t)} \text{ for all } t \in \{0, 1, \ldots, T-1\},$$

with $\alpha_{-1} = \alpha_0$, moreover, we choose

$$\alpha_t \leq \left\{ \frac{1}{\mu_f + 1}, \frac{1}{2\mu_g \hat{c}_\beta}, \frac{\mu_g}{\hat{c}_\beta L_g^2}, \frac{1}{8L_K^2 + L_f}, \frac{L^2 + 2L_y^2}{4L_K^2 L_g^2 \hat{c}_\beta^2} \right\}. \tag{50}$$

Further, we choose

$$K = \frac{L_g}{2\mu_g} \log \left(\left(\frac{C_{g_{xy}} C_{f_y}}{\mu_g} \right)^2 T \right)$$

such that we have $\|B_t\|^2 \leq 1/T$, then we have

$$\mathbb{E}[\widehat{V}_{t+1}] \leq (1 - \mu_f \alpha_{t+1}) \mathbb{E}[\widehat{V}_t] + \frac{2\alpha_t}{T} + \big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2 L_K^2)\sigma_g^2 + 2(\mu_f + 1)^2 \sigma_f^2 \big] \alpha_t^2,$$

for all $t \in \{0, 1, \ldots, T-1\}$.

Proof. From Lemma E.3, we have

$$\mathbb{E}\|e_{t+1}^f\|^2 \leq (1 - \eta_{t+1}^f)\mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2 \sigma_f^2 + 4L_K^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 + 8L_K^2 \beta_t^2 \sigma_g^2 + 8L_K^2 L_g^2 \beta_t^2 \mathbb{E}\|y_t - y^*(x_t)\|^2, \tag{51}$$

which follows from $1-\eta_{t+1}^f \le 1$. With the choice of $\eta_t = (\mu_f+1)\alpha_t$ and $\beta_t = \hat{c}_\beta \alpha_t$ we get from (51):

$$\begin{aligned}\mathbb{E}\|e_{t+1}^f\|^2 \le (1-(\mu_f+1)\alpha_t)\mathbb{E}\|e_t^f\|^2 + 2(\mu_f+1)^2\alpha_t^2\sigma_f^2 + 4L_K^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\ + 8L_K^2\hat{c}_\beta^2\alpha_t^2\sigma_g^2 + 8L_K^2L_g^2\hat{c}_\beta^2\alpha_t^2\mathbb{E}\|y_t - y^*(x_t)\|^2,\end{aligned} \tag{52}$$

Next, we consider the descent in the iterates of inner problem. Again using Lemma E.2 we have

$$\begin{aligned}\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \le (1+\gamma_t)\big(1 - 2\beta_t\mu_g + \beta_t^2L_g^2\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 \\ + \left(1 + \frac{1}{\gamma_t}\right)L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 + (1+\gamma_t)\beta_t^2\sigma_g^2.\end{aligned} \tag{53}$$

Using the fact that $\beta_t \le \frac{\mu_g}{L_g^2}$, $\beta_t \le \frac{1}{2\mu_g}$ and from the choice of γ_t we have $1 + \frac{1}{\gamma_t} \le \frac{2}{\mu_g\beta_t}$ Substituting the γ_t, β_t and the upper bound on $1+\frac{1}{\gamma_t}$ in (53) above we get:

$$\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \le \left(1 - \frac{\hat{c}_\beta\mu_g\alpha_t}{2}\right)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \frac{2L_y^2\alpha_t}{\mu_g\hat{c}_\beta}\mathbb{E}\|h_t^f\|^2 + 2\hat{c}_\beta^2\alpha_t^2\sigma_g^2. \tag{54}$$

Next, replacing the choice of $\hat{c}_\beta$ in (54), we get:

$$\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \le \big(1 - [4L_y^2 + 4L^2 + \mu_f]\alpha_t\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \frac{\alpha_t}{4}\mathbb{E}\|h_t^f\|^2 + 2\hat{c}_\beta^2\alpha_t^2\sigma_g^2. \tag{55}$$

Finally, to construct the potential function defined in (49) we add (52) and (55) to the expression of Lemma E.1, we get

$$\begin{aligned}\mathbb{E}[\widehat{V}_{t+1}] \le (1-\mu_f\alpha_t)\mathbb{E}[\widehat{V}_{t+1}] - \left(\frac{\alpha}{2}(1-\alpha_tL_f) - \frac{\alpha_t}{4} - 4L_K^2\alpha_t^2\right)\mathbb{E}\|h_t^f\|^2 + 2\alpha_t\|B_t\|^2 \\ - \big(4L^2\alpha_t + 4L_y^2\alpha_t - 2L^2\alpha_t - 8L_K^2L_g^2\hat{c}_\beta^2\alpha_t^2\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 \\ + (2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\alpha_t^2\sigma_g^2 + 2(\mu_f+1)^2\alpha_t^2\sigma_f^2.\end{aligned}$$

Noting the fact that $\alpha_t \le \frac{1}{8L_K^2 + L_f}$ and $\alpha_t \le \frac{L^2 + 2L_y^2}{4L_K^2L_g^2\hat{c}_\beta^2}$ and choosing B_t such that we have $\|B_t\|^2 \le \frac{1}{T}$, we get

$$\mathbb{E}[\widehat{V}_{t+1}] \le (1-\mu_f\alpha_t)\mathbb{E}[\widehat{V}_t] + \frac{2\alpha_t}{T} + \big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]\alpha_t^2.$$

This concludes the proof of the lemma. □

E.5 Proof of Theorem 3.3

Next, we conclude the proof for the case of strongly-convex outer objective function case based on fixed step sizes and momentum parameters.

Proof. With fixed step sizes, i.e. $\alpha_t = \alpha$ for all $t \in \{0, 1, \ldots, T-1\}$, we have from the Lemma E.4

$$\mathbb{E}[\widehat{V}_{t+1}] \le (1-\mu_f\alpha)\mathbb{E}[\widehat{V}_t] + \frac{2\alpha}{T} + \big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]\alpha^2.$$

applying the above inequality recursively we get

$$\begin{aligned}\mathbb{E}[\widehat{V}_t] &\le (1-\mu_f\alpha)^t\mathbb{E}[\widehat{V}_0] + \frac{2\alpha}{T}\sum_{k=0}^{t-1}(1-\mu_f\alpha)^k + \big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]\alpha^2\sum_{k=0}^{t-1}(1-\mu_f\alpha)^k \\ &\overset{(a)}{\le} (1-\mu_f\alpha)^t\big((\ell(x_0) - \ell^*) + \mathbb{E}\|e_0^f\|^2 + \mathbb{E}\|y_0 - y^*(x_0)\|^2\big) + \frac{2\alpha}{T}\sum_{k=0}^{t-1}(1-\mu_f\alpha)^k \\ &\qquad + \big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]\alpha^2\sum_{k=0}^{t-1}(1-\mu_f\alpha)^k \\ &\overset{(b)}{\le} (1-\mu_f\alpha)^t\big\{(\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big\} + \frac{2}{\mu_fT} + \frac{(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2}{\mu_f}\alpha,\end{aligned} \tag{56}$$

where (a) follows from the definition of $\widehat{V}_t$ given in (49) and (b) utilizes the summation of a geometric progression.

This concludes the proof of the theorem. □

Sample complexity of SUSTAIN in the strongly convex setting Let us estimate the total number of iterations, T, needed to reach an ϵ-optimal solution. First, we select a constant step size such that

$$\alpha \leq \frac{\mu_f}{4\big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2 L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]}\epsilon \quad \implies \quad \frac{\big[(2\hat{c}_\beta^2 + 8\hat{c}_\beta^2 L_K^2)\sigma_g^2 + 2(\mu_f+1)^2\sigma_f^2\big]}{\mu_f}\alpha \leq \frac{\epsilon}{4}, \tag{57}$$

which controls the last term in (56). Secondly, to control the second term in (56), we observe that $T \geq \frac{8}{\mu_f \epsilon}$ implies $\frac{2}{\mu_f T} \leq \frac{\epsilon}{4}$. Finally, controlling the first term in (56) requires

$$\frac{\epsilon}{2} \geq (1-\mu_f\alpha)^T \big((\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big) \tag{58}$$

which means we require:

$$\begin{aligned}
(1-\mu_f\alpha)^T &\leq \frac{\epsilon}{2\big((\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big)} \\
\iff T\log(1-\mu_f\alpha) &\leq \log\left(\frac{\epsilon}{2\big((\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big)}\right) \\
\iff T &\geq \frac{\log\left(\frac{2\big((\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big)}{\epsilon}\right)}{-\log(1-\mu_f\alpha)} \\
\impliedby T &\overset{(a)}{\geq} \log\left(\frac{2\big((\ell(x_0) - \ell^*) + \sigma_f^2 + \|y_0 - y^*(x_0)\|^2\big)}{\epsilon}\right)\frac{1}{\mu_f\alpha},
\end{aligned} \tag{59}$$

where (a) is due to $\log x \leq x - 1$ for all $x > 0$. This along with (57) imply that we require at most $T = \tilde{\mathcal{O}}(\epsilon^{-1})$ iterations to reach an ϵ-optimal solution, i.e., $\mathbb{E}[\ell(x_t) - \ell^*] \leq \epsilon$. Finally, as each iteration takes a batch of $K = \mathcal{O}(\log(T))$ samples, the total sample complexity required to reach an ϵ-optimal solution is bounded as $T = \tilde{\mathcal{O}}(\epsilon^{-1})$. □