A lower bound and a near-optimal algorithm for bilevel empirical risk minimization

Mathieu Dagréou, Thomas Moreau, Samuel Vaiter, Pierre Ablin

A Lower Bound and a Near-Optimal Algorithm for Bilevel Empirical Risk Minimization

Mathieu Dagréou
Université Paris-Saclay
Inria, CEA
Palaiseau, France

Thomas Moreau
Université Paris-Saclay
Inria, CEA
Palaiseau, France

Samuel Vaiter
Université de Côte d'Azur
CNRS, LJAD
Nice, France

Pierre Ablin
Apple
Paris, France

Abstract

Bilevel optimization problems, which are problems where two optimization problems are nested, have more and more applications in machine learning. In many practical cases, the upper and the lower objectives correspond to empirical risk minimization problems and therefore have a sum structure. In this context, we propose a bilevel extension of the celebrated SARAH algorithm. We demonstrate that the algorithm requires $\mathcal{O}((n+m)^{\frac{1}{2}}\varepsilon^{-1})$ oracle calls to achieve ε-stationarity with $n+m$ the total number of samples, which improves over all previous bilevel algorithms. Moreover, we provide a lower bound on the number of oracle calls required to get an approximate stationary point of the objective function of the bilevel problem. This lower bound is attained by our algorithm, making it optimal in terms of sample complexity.

1 Introduction

In the last few years, bilevel optimization has become an essential tool for the machine learning community thanks to its numerous applications. Among them, we can cite hyperparameter selection (Bengio, 2000; Pedregosa, 2016; Franceschi et al., 2017; Lorraine et al., 2020), implicit deep learning (Bai et al., 2019), neural architecture search (Liu et al., 2019; Zhang et al., 2021), data augmentation (Li et al., 2020; Rommel et al., 2022) or meta-learning (Franceschi et al., 2018; Rajeswaran et al., 2019). Bilevel optimization consists in minimizing a function under the constraint that one variable minimizes another function. This can be formalized as follows

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} h(x) = F(z^*(x), x) \ , \\ &\text{subject to } z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z,x) \ . \end{aligned} \tag{1}$$

The function F is called the outer function and the function G is the inner function. Likewise, we refer to x as the outer variable and z as the inner variable. The function h is the value function and it can be minimized using gradient descent. To compute its gradient, we use implicit differentiation which yields

$$\nabla h(x) = \nabla_2 F(z^*(x), x) + \nabla^2_{21} G(z^*(x), x) v^*(x) \tag{2}$$

where $v^*(x)$ is the solution of a linear system

$$v^*(x) = -\left[\nabla^2_{11} G(z^*(x), x)\right]^{-1} \nabla_1 F(z^*(x), x) \ . \tag{3}$$

When we have exact access to $z^*(x)$, solving (1) boils down to a smooth nonconvex optimization problem which can be solved using solvers for single-level problems. However, computing exactly $z^*(x)$ and $v^*(x)$ is often too costly, and implicit differentiation-based algorithms rely on approximations of $z^*(x)$ and $v^*(x)$ rather than their exact value. Depending on the precision of the different approximations, we are not ensured that the approximate gradient used is a descent direction. Results by Pedregosa (2016) characterized the approximation quality for $z^*(x)$ and $v^*(x)$ required to ensure convergence, opening the door to various algorithms to solve bilevel optimization problems (Lorraine et al., 2020; Ramzi et al., 2022).

In many applications of interest, the functions F and G correspond to Empirical Risk Minimization (ERM), and as a consequence have a finite sum structure

$$F(z,x) = \frac{1}{m}\sum_{j=1}^{m} F_j(z,x), \quad G(z,x) = \frac{1}{n}\sum_{i=1}^{n} G_i(z,x) \ .$$

For instance, in hyperparameter selection, F is the validation loss which is an average on the validation

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

set and G is the training loss which is an average on the training set. In single-level optimization, the finite sum structure has been widely leveraged to produce fast first-order algorithms that provably converge faster than gradient descent. Among them, we can cite stochastic methods such as stochastic gradient descent (SGD) (Robbins and Monro, 1951; Bottou, 2010) and its variance-reduced variants such as SAGA (Defazio et al., 2014), STORM (Cutkosky and Orabona, 2019) or SPIDER/SARAH (Fang et al., 2018; Nguyen et al., 2017) that use only a handful of samples at a time to make progress. To get faster methods than full-batch approaches, it is natural to extend these methods to the bilevel setting. The main obstacle comes from the difficulty of obtaining stochastic approximations of $\nabla h(x)$ because of its structure (2) which involves a Hessian inversion. Several strategies have been proposed to overcome this obstacle, and some works demonstrate that stochastic implicit differentiation-based algorithms for solving (1) have the same complexity as single-level analogous algorithms. For instance, ALSET from Chen et al. (2021) and SOBA from Dagréou et al. (2022) have the same convergence rate as SGD for nonconvex single-level problems (Ghadimi and Lan, 2013; Bottou et al., 2018). Also, Dagréou et al. (2022) show that SABA, an adaptation of SAGA (Defazio et al., 2014), has an analogous sample complexity to its single-level counterparts for nonconvex problems (Reddi et al., 2016).

Yet, in classical single-level optimization, it is known that neither of these algorithms is optimal: the SARAH algorithm (Nguyen et al., 2017) achieves a better sample complexity of $\mathcal{O}(m^{\frac{1}{2}}\varepsilon^{-1})$ with m the number of samples. Furthermore, this algorithm is *near-optimal* (*i.e.* optimal up to constant factors) because the lower bound for single-level nonconvex optimization is $\Omega(m^{\frac{1}{2}}\varepsilon^{-1})$ as proved by Zhou and Gu (2019). It is natural to ask if we can extend these results to bilevel optimization.

Contributions In Section 2, we introduce SRBA, an adaptation of the SARAH algorithm to the bilevel setting. We then demonstrate in Section 3 that, similarly to the single-level setting, $\mathcal{O}\left((n+m)^{\frac{1}{2}}\varepsilon^{-1} \vee (n+m)\right)$ oracle calls are sufficient to reach an ε-stationary point. As shown in Table 1, it achieves the best-known complexity in the regime $n+m \lesssim \mathcal{O}(\varepsilon^{-2})$. In Section 4, we analyze the lower bounds for such problems. We show that we need at least $\Omega(m^{\frac{1}{2}}\varepsilon^{-1})$ oracle calls to reach an ε-stationary point (see Definition 3.1), hereby matching the previous upper-bound in the case where $n \asymp m$ and $\varepsilon \leq m^{-\frac{1}{2}}$. SRBA is thus near-optimal in that regime. Even though our main contribution is theoretical, we illustrate the numerical performances of the algorithm in Section 5.

Related work There are several strategies to solve (1) with a stochastic method. The first one is the value-function-based method which consists in recasting Problem 1 as a single-level constrained optimization problem as done with F^{2}SA (Kwon et al., 2023) or BOME (Ye et al., 2022). The second way is to use first-order methods on h with *approximate* gradients. The approximate gradient of h can be estimated using two approaches: iterative differentiation (ITD) and approximate implicit differentiation (AID). On the one hand, in ITD algorithms, the Jacobian of z^* is estimated by differentiating the different steps used to compute an approximation of z^*. On the other hand, AID algorithms leverage the implicit gradient given by (2) replacing z^* and v^* by some approximations z and v. In the class of ITD algorithms, Maclaurin et al. (2015) propose to approximate the Jacobian of the solution of the inner problem by differentiating through the iterations of SGD with momentum. The complexity of the hypergradient computation in ITD solvers is studied in Franceschi et al. (2017); Grazzi et al. (2020); Ablin et al. (2020). For AID algorithms, Ghadimi and Wang (2018); Chen et al. (2021); Ji et al. (2021) propose to perform several SGD steps in the inner problem and then use Neumann approximations to approximate $v^*(x)$ defined in (3). A method consisting of alternating steps in the inner and outer variables was proposed in Hong et al. (2023). These methods can be improved by using a warm start strategy for the inner problem (Ji et al., 2021; Chen et al., 2021) and for the linear system (Arbel and Mairal, 2022). Some works adapt variance reduction methods to like STORM (Cutkosky and Orabona, 2019; Khanduri et al., 2021; Yang et al., 2021) or SAGA (Defazio et al., 2014; Dagréou et al., 2022). We take a similar approach and extend the SARAH variance reduction method to the bilevel setting. Recent works propose to approximate the Jacobian of z^* by stochastic finite difference (Sow et al., 2022) or to use Bregman divergence-based methods (Huang et al., 2022).

In single-level optimization, the problem of finding complexity lower bound has been widely studied since the seminal work of Nemirovsky and Yudin (1983). On the one hand, Agarwal and Bottou (2015) provided a lower bound to minimize strongly convex and smooth finite sum with deterministic algorithms that have access to individual gradients. These results were extended to randomized algorithms for (strongly) convex finite sum objective by Woodworth and Srebro (2016). On the other hand, Carmon et al. (2020) provided a lower bound for minimizing nonconvex functions with deterministic and randomized algorithms. The nonconvex finite sum case is treated in Fang et al. (2018); Zhou and Gu (2019). In the bilevel case, Ji and Liang (2023) showed a lower bound for deterministic algorithms. However, this result is restricted to the case where the value function h is convex or strongly convex, which

	Sample complexity	Stochastic setting	F	G
StocBiO (Ji et al., 2021)	$\tilde{\mathcal{O}}(\varepsilon^{-2})$	General expectation	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$
AmIGO (Arbel and Mairal, 2022)	$\mathcal{O}(\varepsilon^{-2})$	General expectation	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$
MRBO (Yang et al., 2021)	$\tilde{\mathcal{O}}(\varepsilon^{-\frac{3}{2}})$	General expectation	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$
VRBO (Yang et al., 2021)	$\tilde{\mathcal{O}}(\varepsilon^{-\frac{3}{2}})$	General expectation	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$
SABA (Dagréou et al., 2022)	$\mathcal{O}((n+m)^{\frac{2}{3}}\varepsilon^{-1})$	Finite sum	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$
F^{2}SA (Kwon et al., 2023)	$\mathcal{O}(\varepsilon^{-\frac{7}{2}})$	General expectation	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{2,2}$
SRBA	$\mathcal{O}((n+m)^{\frac{1}{2}}\varepsilon^{-1})$	Finite sum	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$

Table 1: Comparison between the sample complexities and the Assumptions of some stochastic bilevel solvers. It corresponds to the number of calls to gradient, Hessian-vector products, and Jacobian-vector product sufficient to get an ε-stationary point. The tilde on the $\tilde{\mathcal{O}}$ hide a factor $\log(\varepsilon^{-1})$. "SC" means "strongly convex". $\mathcal{C}_L^{p,k}$ means p-times differentiable with Lipschitz kth order derivatives for $k \leq p$.

is not the case with most ML-related bilevel problems. Our results are instead in a nonconvex setting.

Notation The quantity $A_\bullet$ refers to A_z, A_v, or A_x, depending on the context. If $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is a twice differentiable function, we denote $\nabla_i f(z,x)$ its gradient w.r.t. its i^{th} variable. Its Hessian with respect to z is denoted $\nabla_{11}^2 f(z,x) \in \mathbb{R}^{p\times p}$ and its cross derivative matrix $\left(\frac{\partial^2 f}{\partial z_i \partial x_j}\right)_{\substack{i\in[p]\\ j\in[d]}}$ is denoted $\nabla_{12}^2 f(z,x) \in \mathbb{R}^{p\times d}$. We denote $\Pi_{\mathcal{C}}$ the projection on a closed convex set $\mathcal{C}$.

2 SRBA: a Near-Optimal Algorithm for Bilevel Empirical Risk Minimization

In this section, we introduce SRBA (Stochastic Recursive Bilevel Algorithm), a novel algorithm for bilevel empirical risk minimization which is provably near-optimal for this problem. This algorithm is inspired by the algorithms SPIDER (Fang et al., 2018) and SARAH (Nguyen et al., 2017, 2022) which are known for being near-optimal algorithms for nonconvex finite sum minimization problems. It relies on a recursive estimation of directions of interest, which is restarted periodically. Proofs are deferred to the appendix.

2.1 Assumptions

Before presenting our algorithm, we formulate regularity Assumptions on the functions F and G.

Assumption 2.1. For all $j \in [m]$, F_j is twice differentiable and L_0^F-Lipschitz continuous. Its gradient is L_1^F-Lipschitz continuous and its Hessian is L_2^F-Lipschitz continuous.

Assumption 2.2. For all $i \in [n]$, G_i is three times differentiable. Its first, second, and third order derivatives are respectively L_1^G-Lipschitz continuous, L_2^G-Lipschitz continuous, and L_3^G-Lipschitz continuous. For $x \in \mathbb{R}^d$, the function $G_i(\,.\,,x)$ is μ_G-strongly convex.

The strong convexity and the smoothness with respect to z hold for instance when we consider an ℓ^2-regularized logistic regression problem with non-separable data. These regularity assumptions up to first-order for F and second-order for G are standard in the stochastic bilevel literature (Arbel and Mairal, 2022; Ji et al., 2021; Yang et al., 2021). The second-order regularity for F and third-order regularity for G are necessary for the analysis of the dynamics of v, as is the case in Dagréou et al. (2022). As shown in Ghadimi and Wang (2018, Lemma 2.2), these assumptions imply the smoothness of h, which is a fundamental property to get a descent.

Proposition 2.3. *Under Assumptions 2.1 and 2.2, the function h is L^h smooth for some $L^h > 0$ which is precised in Appendix A.2.*

Another consequence of Assumptions 2.1 and 2.2 is the boundedness of the function v^*.

Proposition 2.4. *Assume that Assumptions 2.1 and 2.2 hold. Then, for $R = \frac{L_0^F}{\mu_G}$ it holds that for any $x \in \mathbb{R}^d$, we have $\|v^*(x)\| \leq R$.*

We denote Γ the closed ball centered in 0 with radius R and Π_Γ the projection onto Γ. For $(z,v,x) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^d$, we denote $\Pi(z,v,x) = (z, \Pi_\Gamma(v), x)$.

2.2 Hypergradient Approximation

The gradient of h given by (2) is intractable in practice because it requires the perfect knowledge of $z^*(x)$ and $v^*(x)$ which are usually costly to compute. As classically done in the stochastic bilevel literature (Ji et al., 2021; Arbel and Mairal, 2022; Li et al., 2022), $z^*(x)$ and $v^*(x)$ are replaced by approximate surrogate variables z and v. The variable z is typically the output of one or several steps of an optimization procedure applied to $G(\,.\,,x)$. The variable v can be computed by using Neumann approximations or doing some optimization steps on the quadratic $v \mapsto \frac{1}{2}v^\top \nabla_{11}^2 G(z,x)v + \nabla_1 F(z,x)^\top v$.

We consider the approximate hypergradient given by

$$D_x(z,v,x) = \nabla^2_{21} G(z,x)v + \nabla_2 F(z,x) \ .$$

The motivation behind this direction is that if we take $z = z^*(x)$ and $v = v^*(x)$, we recover the true gradient, that is $D_x(z^*(x), v^*(x), x) = \nabla h(x)$. Proposition 2.5 from (Dagréou et al., 2022, Lemma 3.4) controls the hypergradient approximation error by the distances between z and $z^*(x)$ and between v and $v^*(x)$.

Proposition 2.5. *Let $x \in \mathbb{R}^d$. Assume that F is differentiable and L_1^F smooth with bounded gradient, G is twice differentiable with Lipschitz gradient and Hessian and $G(\,.\,,x)$ is μ_G-strongly convex. Then there exists a constant L_x such that*

$$\|D_x(z,v,x) - \nabla h(x)\|^2 \leq L_x^2(\|z - z^*(x)\|^2 + \|v - v^*(x)\|^2).$$

Thus, it is natural to make z and v move towards their respective equilibrium values which are given by $z^*(x)$ and $v^*(x)$. As a consequence, we also introduce the directions D_z and D_x as follows

$$D_z(z,v,x) = \nabla_1 G(z,x) \ ,$$
$$D_v(z,v,x) = \nabla^2_{11} G(z,x)v + \nabla_1 F(z,x) \ .$$

The interest of considering the directions D_z and D_v is expressed in Proposition 2.6.

Proposition 2.6. *Assume that G is strongly convex with respect to its first variable. Then for any $x \in \mathbb{R}^d$, it holds $D_z(z^*(x), v^*(x), x) = 0$ and $D_v(z^*(x), v^*(x), x) = 0$.*

The directions D_z, D_v, and D_x can be written as sums over the samples. Hence, following these directions enables to adapt any classical algorithm suited for single-level finite sum minimization to bilevel finite sum minimization. In what follows, for two indices $i \in [n]$ and $j \in [m]$, we consider the sampled directions $D_{z,i,j}$, $D_{v,i,j}$ and $D_{x,i,j}$ defined by

$$D_{z,i,j}(z,v,x) = \nabla_1 G_i(z,x) \tag{4}$$
$$D_{v,i,j}(z,v,x) = \nabla^2_{11} G_i(z,x)v + \nabla_1 F_j(z,x) \tag{5}$$
$$D_{x,i,j}(z,v,x) = \nabla^2_{21} G_i(z,x)v + \nabla_2 F_j(z,x) \ . \tag{6}$$

When i and j are randomly sampled uniformly, these directions are unbiased estimators of the true directions D_z, D_v, and D_x. Yet, as in Nguyen et al. (2017), we use them to recursively build biased estimators of the directions that enable fast convergence.

2.3 SRBA: Stochastic Recursive Bilevel Algorithm

In Algorithm 1, we present SRBA, a combination of the idea of recursive gradient coming from (Fang et al., 2018; Nguyen et al., 2022) and the framework proposed in (Dagréou et al., 2022). It relies on a recursive estimation of each direction D_z, D_v, D_x which is updated following the same strategy as SARAH. Let us denote by ρ the step size of the update for the variables z and v, and γ the step size for the update of the variable x. We use the same step size for z and v because the problems of minimizing the inner function G and solving the linear system (3) have the same conditioning driven by $\nabla^2_{11} G$. For simplicity, we denote the joint variable $\mathbf{u} = (z, v, x)$ and the joint directions weighted by the step sizes $\mathbf{\Delta} = (\rho D_z, \rho D_v, \gamma D_x) = (\mathbf{\Delta}_z, \mathbf{\Delta}_v, \mathbf{\Delta}_x)$.

Algorithm 1 Stochastic Recursive Bilevel Algorithm

Input: initializations $z_0 \in \mathbb{R}^p$, $x_0 \in \mathbb{R}^d$, $v_0 \in \mathbb{R}^p$, number of iterations T and q, step sizes ρ and γ.
Set $\tilde{\mathbf{u}}^{\mathbf{0}} = (z_0, v_0, x_0)$
for $t = 0, \dots, T-1$ **do**
  Reset $\mathbf{\Delta}$: $\mathbf{\Delta}^{t,0} = (\rho D_z(\tilde{\mathbf{u}}^t), \rho D_v(\tilde{\mathbf{u}}^t), \gamma D_x(\tilde{\mathbf{u}}^t))$
  Update $\mathbf{u}$: $\mathbf{u}^{t,1} = \Pi(\tilde{\mathbf{u}}^t - \mathbf{\Delta}^{\mathbf{t},\mathbf{0}})$,
  for $k = 1, \dots, q-1$ **do**
    Draw $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$
    $\mathbf{\Delta}_z^{t,k} = \rho(D_{z,i,j}(\mathbf{u}^{t,k}) - D_{z,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_z^{t,k-1}$
    $\mathbf{\Delta}_v^{t,k} = \rho(D_{v,i,j}(\mathbf{u}^{t,k}) - D_{v,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_v^{t,k-1}$
    $\mathbf{\Delta}_x^{t,k} = \gamma(D_{x,i,j}(\mathbf{u}^{t,k}) - D_{x,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_x^{t,k-1}$
    Update $\mathbf{u}$: $\mathbf{u}^{t,k+1} = \Pi(\mathbf{u}^{t,k} - \mathbf{\Delta}^{t,k})$
  end for
  Set $\tilde{\mathbf{u}}^{t+1} = \mathbf{u}^{t+1,q}$
end for
Return $(\tilde{z}^T, \tilde{v}^T, \tilde{x}^T) = \tilde{\mathbf{u}}^T$

At iteration t, the estimate direction $\mathbf{\Delta}$ is initialized by computing full batch directions:

$$\mathbf{\Delta}^{t,0} = (\rho D_z(\tilde{\mathbf{u}}^t), \rho D_v(\tilde{\mathbf{u}}^t), \gamma D_x(\tilde{\mathbf{u}}^t))$$

and a first update is performed by moving from $\tilde{\mathbf{u}}^t$ in the direction $-\mathbf{\Delta}^{t,0}$. As done in Hu et al. (2022), we project the variable v onto Γ to leverage the boundedness property of v^*. Then, during the kth iteration of an inner loop of size $q-1$, two indices $i \in [n]$ and $j \in [m]$ are sampled and the estimate directions are updated according to Equations (7) to (9)

$$\mathbf{\Delta}_z^{t,k} = \rho(D_{z,i,j}(\mathbf{u}^{t,k}) - D_{z,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_z^{t,k-1} \tag{7}$$
$$\mathbf{\Delta}_v^{t,k} = \rho(D_{v,i,j}(\mathbf{u}^{t,k}) - D_{v,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_v^{t,k-1} \tag{8}$$
$$\mathbf{\Delta}_x^{t,k} = \gamma(D_{x,i,j}(\mathbf{u}^{t,k}) - D_{x,i,j}(\mathbf{u}^{t,k-1})) + \mathbf{\Delta}_x^{t,k-1} \tag{9}$$

where the sampled directions $D_{z,i,j}$, $D_{v,i,j}$ and $D_{x,i,j}$ are defined by Equations (4) to (6). Then the joint variable $\mathbf{u}$ is updated by

$$\mathbf{u}^{t,k+1} = \Pi(\mathbf{u}^{t,k} - \mathbf{\Delta}^{t,k}) \ . \tag{10}$$

Recall that the projection is only performed on the variable v. The other variables z and x remain un-

changed after the projection step. At the end of the inner procedure, we set $\tilde{\mathbf{u}}^{t+1} = \mathbf{u}^{t,q}$.

In Algorithm 1, the variables z, v, and x are updated simultaneously rather than alternatively. From a computational perspective, this enables sharing the common computations between the different oracles and doing the update of each variable in parallel. So there is no sub-procedure to approximate the solution of the inner problem and the solution of the linear system.

Note that in Yang et al. (2021), the authors propose VRBO, another adaptation of SPIDER/SARAH for bilevel problems. VRBO has a double loop structure where the inner variable is updated by several steps in an inner loop. In this inner loop, the estimates of the gradient of G and the gradient of h are also updated using SARAH's update rules. SRBA has a different structure. First, in SRBA, the inner variable z is updated only once between two updates of the outer variable instead of several times. Second, the solution of the linear system evolves following optimization steps whereas in VRBO a Neumann approximation is used. Moreover, in Yang et al. (2021), the algorithm VRBO is analyzed in the case where the functions F and G are general expectations but not in the specific case of empirical risk minimization, as done in Section 3. Finally, VRBO requires three more parameters than SRBA: the number of inner steps, the number of terms and the scaling parameter in the Neumann approximations.

3 Theoretical Analysis of SRBA

In this section we provide the theoretical analysis of Algorithm 1 leading to a final sample complexity in $\mathcal{O}\left((n+m)^{\frac{1}{2}}\varepsilon^{-1} \vee (n+m)\right)$. The detailed proofs of the results are deferred to the appendix. In Definition 3.1, we recall what is an ε-stationary point.

Definition 3.1. Let d a positive integer, $f : \mathbb{R}^d \to \mathbb{R}$ a differentiable function and $\varepsilon > 0$. We say that a point $x \in \mathbb{R}^d$ is an ε-stationary point of f if $\|\nabla f(x)\|^2 \leq \epsilon$. In a stochastic context, we call ε-stationary point a random variable x such that $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$.

In this paper, the theoretical complexity of the algorithms is given in terms of number of calls to oracle, that is to say, the number of times the quantity

$$[\nabla F_j(z,x), \nabla G_i(z,x), \nabla^2_{11} G_i(z,x)v, \nabla^2_{21} G_i(z,x)v] \tag{11}$$

is queried for $i \in [n]$, $j \in [m]$, $z \in \mathbb{R}^p$, $v \in \mathbb{R}^p$ and $x \in \mathbb{R}^d$. Note that in practice, although the second-order derivatives of the inner functions $\nabla^2_{11} G_i(z,x) \in \mathbb{R}^{p\times p}$ and $\nabla^2_{21} G_i(z,x) \in \mathbb{R}^{d\times p}$ are involved, they are never computed or stored explicitly. We rather work with Hessian-vector products $\nabla^2_{11} G_i(z,x)v \in \mathbb{R}^p$ and Jacobian-vector products $\nabla^2_{21} G_i(z,x)v \in \mathbb{R}^d$ which can be computed efficiently thanks to automatic differentiation with a computational cost similar to the cost of computing the gradients $\nabla_1 G_i(z,x)$ and $\nabla_2 G_i(z,x)$ Pearlmutter (1994). The cost of one query (11) is therefore of the same order of magnitude as that of computing one stochastic gradient.

3.1 Mean Squared Error of the Estimated Directions

A strength of our method is its simple expression of the estimation error of the directions coming from the bias-variance decomposition provided by Nguyen et al. (2017). Let us denote the estimate directions $D_z^{t,k} = \mathbf{\Delta}_z^{t,k}/\rho$, $D_v^{t,k} = \mathbf{\Delta}_v^{t,k}/\rho$ and $D_x^{t,k} = \mathbf{\Delta}_x^{t,k}/\gamma$. We also introduce the residuals

$$S_\bullet^{t,k} = \sum_{r=1}^{k} \mathbb{E}[\|D_\bullet(\mathbf{u}^{t,r}) - D_\bullet(\mathbf{u}^{t,r-1})\|^2],$$
$$\tilde{S}_\bullet^{t,k} = \sum_{r=1}^{k} \mathbb{E}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1}\|^2] \ .$$

We provide a link between the mean squared error $\mathbb{E}[\|D_\bullet^{t,k} - D_\bullet(\mathbf{u}^{t,k})\|^2]$ and the residuals.

Proposition 3.2 (MSE of the estimate directions)**.** *For any $t \geq 0$ and $k \in \{1, \dots, q-1\}$, the estimate $D_\bullet^{t,k}$ of the direction $D_\bullet(\mathbf{u}^{t,k})$ satisfies*

$$\mathbb{E}[\|D_\bullet^{t,k} - D_\bullet(\mathbf{u}^{t,k})\|^2] = \tilde{S}_\bullet^{t,k} - S_\bullet^{t,k} \ .$$

The above error has two components: the accumulation of the difference between two successive full batch directions and the accumulation of the difference between two successive estimate directions.

3.2 Fundamental Lemmas

We establish descent lemmas which are key ingredients to get the final convergence result. Lemma 3.3 characterizes the dynamic of $\mathbf{u}$ on the inner problem. To do so, we define the function ϕ_z as

$$\phi_z(z,x) = G(z,x) - G(z^*(x),x) \ .$$

In the bilevel literature, direct control on the distance to optimum $\delta_z^{t,k} \triangleq \mathbb{E}[\|z^{t,k} - z^*(x^{t,k})\|^2]$ is established. Here, the biased nature of the estimate direction $D_z^{t,k}$ makes it hard to upper bound appropriately the scalar product $\langle D_z(\mathbf{u}^{t,k}) - D_z^{t,k}, z^{t,k} - z^*(x^{t,k})\rangle$. Therefore, we rather consider $\phi_z^{t,k}$. By combining the smoothness property of ϕ_z and the bias-variance decomposition provided in Proposition 3.2, we can show some descent property on the sequence $\phi_z^{t,k}$ defined by $\phi_z^{t,k} = \mathbb{E}[\phi_z(z^{t,k}, x^{t,k})]$. Before stating Lemma 3.3,

let us define $\mathcal{G}_v^{t,k} = \frac{1}{\rho}\left(v^{t,k} - \Pi_\Gamma(v^{t,k} - \rho D_v^{t,k})\right)$ so that $v^{t,k+1} = v^{t,k} - \rho\mathcal{G}_v^{t,k}$. This is the actual update direction of v. If there were no projections, we would have $\mathcal{G}_v^{t,k} = D_v^{t,k}$. Hence, it acts as a surrogate of $D_v^{t,k}$ in our analysis. We also define

$$V_z^{t,k} = \mathbb{E}[\|D_z^{t,k}\|^2], \quad V_v^{t,k} = \mathbb{E}[\|\mathcal{G}_v^{t,k}\|^2], \quad V_x^{t,k} = \mathbb{E}[\|D_x^{t,k}\|^2]$$

the variances and their respective sums over the inner loop

$$\mathcal{V}_z^{t,k} = \sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2], \quad \mathcal{V}_v^{t,k} = \sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2], \quad \mathcal{V}_x^{t,k} = \sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \ .$$

Lemma 3.3 (Descent on the inner level)**.** *Assume that the step sizes ρ and γ verify $\gamma \leq C_z\rho$ for some positive constant C_z specified in the appendix. Then it holds*

$$\begin{aligned}\phi_z^{t,k+1} \leq \left(1 - \frac{\mu_G}{2}\rho\right)\phi_z^{t,k} - \frac{\rho}{2}\left(1 - \Lambda_z\rho\right)V_z^{t,k} \\ + \rho^3\beta_{zz}\mathcal{V}_z^{t,k} + \gamma^2\rho\beta_{zv}\mathcal{V}_v^{t,k} + \gamma^2\rho\beta_{zx}\mathcal{V}_x^{t,k} \\ + \frac{\Lambda_z}{2}\gamma^2 V_x^{t,k} + \frac{\gamma^2}{\rho}\overline{\beta}_{zx}\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2]\end{aligned} \tag{12}$$

for some positive constants $\Lambda_z, \beta_{zz}, \beta_{zx}$ and $\overline{\beta}_{zx}$ that are specified in the appendix.

In (12) we recover a decrease of $\phi_z^{t,k}$ by a factor $(1 - \rho\mu_G)$. But the outer variable's movement and the noise make appear $D_x(\mathbf{u}^{t,k})$ and the variance hindering the convergence of z towards $z^*(x)$.

For the variable v, the quantity we consider is

$$\phi_v(v, x) \ = \ \Psi(z^*(x), v, x) - \Psi(z^*(x), v^*(x), x)$$

where $\Psi(z, v, x)$ is defined as

$$\Psi(z, v, x) = \frac{1}{2}v^\top\nabla_{11}^2 G(z, x)v + \nabla_1 F(z, x)^\top v \ .$$

The intuition behind considering this quantity is that solving the linear system (3) is equivalent to minimizing over v the function $\Psi(z^*(x), v, x)$.

Lemma 3.4. *Assume that the step sizes ρ and γ verify $\rho \leq B_v$ and $\gamma \leq C_v\rho$ for some positive constants B_v and C_v specified in the appendix. Then it holds*

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1 - \frac{\rho\mu_G}{16}\right)\phi_v^{t,k} - \tilde{\beta}_{vv}\rho V_v^t + \rho^3\beta_{vz}\mathcal{V}_z^{t,k} \\ + 2\rho^3\beta_{vv}\mathcal{V}_v^{t,k} + \gamma^2\rho\beta_{vx}\mathcal{V}_x^{t,k} + \rho\alpha_{vz}\phi_z^{t,k} \\ + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] + \frac{\gamma^2}{\rho}\overline{\beta}_{vx}\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2]\end{aligned}$$

for some positive constants $\Lambda_v, \beta_{vz}, \beta_{vx}, \tilde{\beta}_{vv}$ and $\overline{\beta}_{vx}$ that are specified in the appendix.

Lemma 3.4 is similar to Lemma 3.3 with a term in $\phi_z^{t,k}$ taking into account the error of $z^*(x)$'s approximation. Its proof harnesses the generalization of Polyak-Łojasiewicz inequality for composite functions introduced in Karimi et al. (2016).

The following lemma is a consequence of the smoothness of h. Let us denote the expected values $h^{t,k} = \mathbb{E}[h(x^{t,k})]$ and expected gradient $g^{t,k} = \mathbb{E}[\|\nabla h(x^{t,k})\|^2]$.

Lemma 3.5. *There exist constants β_{hz}, β_{hv}, $\beta_{hx} > 0$ such that*

$$\begin{aligned}h^{t,k+1} \leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma\frac{2L_x^2}{\mu_G}(\phi_z^{t,k} + \phi_v^{t,k}) + \gamma\rho^2 \\ + \rho\alpha_{vz}\phi_z^{t,k} + \gamma\rho^2\beta_{hv}\mathcal{V}_v^{t,k} + \gamma^3\beta_{hx}\mathcal{V}_x^{t,k} \\ - \frac{\gamma}{2}\left(1 - L^h\gamma\right)V_x^{t,k} \ .\end{aligned}$$

This lemma shows that the control of the approximation error $\phi_\bullet$ (Lemma 3.3 and Lemma 3.4) and the sum of variances $\mathcal{V}_\bullet$ is crucial to get a decrease of $\mathbb{E}[h(x^{t,k})]$.

3.3 Complexity Analysis of SRBA

In Theorem 1, we provide the convergence rate of SRBA towards a stationary point.

Theorem 1. Assume that Assumptions 2.1 and 2.2 hold. Assume that the step sizes verify $\rho \leq \overline{\rho}$ and $\gamma \leq \min(\overline{\gamma}, \xi\rho)$ for some constants ξ, $\overline{\rho}$ and $\overline{\gamma}$ specified in appendix. Then it holds

$$\frac{1}{Tq}\sum_{t=0}^{T-1}\sum_{k=0}^{q-1}\mathbb{E}[\|\nabla h(x^{t,k})\|^2] = \mathcal{O}\left(\frac{1}{qT\gamma}\right)$$

where $\mathcal{O}$ hides regularity constants that are independent from n and m.

The proof combines classical proof techniques from the bilevel literature and elements from SARAH's analysis (Nguyen et al., 2017, 2022). We introduce the Lyapunov function $\mathcal{L}(\mathbf{u}^{t,k}) = h^{t,k} + \psi_z\phi_z^{t,k} + \psi_v\phi_v^{t,k}$ where ψ_z and ψ_v are non-negative constants chosen so that we have the inequality $\mathcal{L}(\mathbf{u}^{t,k+1}) \leq \mathcal{L}(\mathbf{u}^{t,k}) - \frac{\gamma}{4}g^{t,k}$. Summing and telescoping this inequality provides the result.

Note that increasing q allows a faster convergence in terms of iterations but makes each iteration more expensive since the number of oracle calls per iteration is $(2n + 3m) + 2 \times 5(q - 1)$. Thus, there is a trade-off between the convergence rate and the overall complexity. In Corollary 3.6, we state that the value of q that gives the best sample complexity is $\mathcal{O}(n + m)$.

Corollary 3.6. *Suppose that Assumptions 2.1 and 2.2 hold. If we take $\rho = \overline{\rho}(n + m)^{-\frac{1}{2}}$, $\gamma = \min(\overline{\gamma}, \xi\rho)(n + m)^{-\frac{1}{2}}$ and $q = n + m$, then*

$\mathcal{O}\left((n+m)^{\frac{1}{2}}\varepsilon^{-1} \vee (n+m)\right)$ *calls to oracles are sufficient to find an ε-stationary point with SRBA.*

This sample complexity is analogous to the sample complexity of SARAH in the nonconvex finite-sum setting. To the best of our knowledge, such a rate is the best known for bilevel empirical risk minimization problems in terms of dependency on the number of samples $n+m$ and the precision ε. This improves by a factor $(n+m)^{-\frac{1}{6}}$ the previous result which was achieved by SABA (Dagréou et al., 2022). As a comparison, VRBO (Yang et al., 2021) achieves a sample complexity in $\tilde{\mathcal{O}}(\varepsilon^{-\frac{3}{2}})$. Note that, for large value of $n+m$ we can have actually $(n+m)^{\frac{1}{2}}\varepsilon^{-1} \gtrsim \varepsilon^{-2}$. This means that, just like single-level SARAH, the complexity of SRBA can be beaten by others when the number of samples is too high with respect to the desired accuracy (actually if $n+m = \Omega(\varepsilon^{-2})$).

4 Lower Bound for Bilevel ERM

In this section, we derive a lower bound for bilevel empirical risk minimization problems. This shows that SRBA is a near-optimal algorithm for this class of problems.

Function and Algorithm Classes We define the function and algorithm classes we consider.

Definition 4.1. Let n, m two positive integers, L_1^F and μ_G two positive constants. The class of the smooth empirical risk minimization problems denoted by $\mathcal{C}^{L_1^F,\mu_G}$ is the set of pairs of real-valued function families $((F_j)_{1\leq j\leq m}, (G_i)_{1\leq i\leq n})$ defined on $\mathbb{R}^p \times \mathbb{R}^d$ such that for all $j \in [m]$, F_j is L_1^F smooth and for all $i \in [n]$, G_i is twice differentiable and μ_G-strongly convex.

Note that we consider a class of nonconvex bilevel problems. This class contains, the functions defining the bilevel formulation of the datacleaning task.

For the algorithmic class, we consider algorithms that use approximate implicit differentiation.

Definition 4.2. Given initial points z^0, v^0, x^0, a *linear bilevel algorithm* $\mathcal{A}$ is a measurable mapping such that for any $((F_j)_{1\leq j\leq m}, (G_i)_{1\leq i\leq n}) \in \mathcal{C}^{L_1^F,\mu_G}$, the output of $\mathcal{A}((F_j)_{1\leq j\leq m}, (G_i)_{1\leq i\leq n})$ is a sequence $\{(z^t, v^t, x^t, i_t, j_t)\}_{t\geq 0}$ of points (z^t, v^t, x^t) and random variables $i_t \in [n]$ and $j_t \in [m]$ such that for all $t \geq 0$

$$
\begin{aligned}
z^{t+1} &\in z^0 + \mathrm{Span}\{\nabla_1 G_{i_0}(z^0, x^0), \dots, \nabla_1 G_{i_t}(z^t, x^t)\} \\
v^{t+1} &\in v^0 + \mathrm{Span}\{\nabla_{11}^2 G_{i_0}(z^0, x^0)v^0 + \nabla_1 F_{j_0}(z^0, x^0), \\
&\qquad \dots, \nabla_{11}^2 G_{i_t}(z^t, x^t)v^t + \nabla_1 F_{j_t}(z^t, x^t)\} \\
x^{t+1} &\in x^0 + \mathrm{Span}\{\nabla_{21}^2 G_{i_0}(z^0, x^0)v^0 + \nabla_2 F_{j_0}(z^0, x^0), \\
&\qquad \dots, \nabla_{21}^2 G_{i_t}(z^t, x^t)v^t + \nabla_2 F_{j_t}(z^t, x^t)\}.
\end{aligned}
$$

This algorithm class includes popular stochastic bilevel first-order algorithms, such as AmIGO (Arbel and Mairal, 2022), FSLA (Li et al., 2022), SOBA, and SABA (Dagréou et al., 2022). Moreover, despite the projection step, SRBA is part of this algorithm class since the projection of a vector onto Γ is actually just a rescaling.

Main Theorem Problem (1) is actually a smooth nonconvex optimization problem. The lower complexity bound for nonconvex finite sum problem has been studied in Fang et al. (2018); Zhou and Gu (2019). In particular, they show that the number of gradient calls needed to get an ε-stationary point for a smooth nonconvex finite sum is at least $\mathcal{O}(m^{\frac{1}{2}}\varepsilon^{-1})$, where m is the number of terms in the finite sum.

Intuitively, we expect the lower complexity bound to solve (1) to be larger. Indeed, bilevel problems are harder than single-level problems because a bilevel problem involves the resolution of several subproblems to progress in its resolution. Theorem 2 formalizes this intuition by showing that the classical single-level lower bound is also a lower bound for bilevel problems.

Theorem 2. For any linear bilevel algorithm $\mathcal{A}$, and any L^F, n, Δ, ε, p such that $\varepsilon \leq (\Delta L^F m^{-1})/10^3$, there exists a dimension $d = \mathcal{O}(\Delta\varepsilon^{-1}m^{\frac{1}{2}}L^F)$, an element $((F_j)_{1\leq j\leq m}, (G_i)_{1\leq i\leq n}) \in \mathcal{C}^{L_1^F,\mu_G}$ such that the value function h defined as in (1) satisfies $h(x^0) - \inf_{x\in\mathbb{R}^d} h(x) \leq \Delta$ and in order to find $\hat{x} \in \mathbb{R}^d$ such that $\mathbb{E}[\|\nabla h(\hat{x})\|^2] \leq \varepsilon$, $\mathcal{A}$ needs at least $\Omega(m^{\frac{1}{2}}\varepsilon^{-1})$ calls to oracles of the form (11).

The proof is an adaptation of the proof of Zhou and Gu (2019, Theorem 4.7). We take as outer function F defined by $F(z, x) = \sum_{j=1}^m f(U^{(j)}z)$ where f is the "worst-case function" used by Carmon et al. (2021), $U = [U^{(j)}, \dots, U^{(m)}]^\top$ is an orthogonal matrix and $G(z, x) = \frac{1}{2}\|z - x\|^2$. We leverage the fact that $\|\nabla f(y)\|^2 > K$ as long as the two last coordinates of y are zero for some known constant K. Then we use the "zero chain property" to bound the number of indices j such the two last components of $U^{(j)}x^t$ are zero at a given iteration t, implying $\|\nabla h(x^t)\|^2 > \epsilon$ when t is smaller than $\mathcal{O}(m^{\frac{1}{2}}\varepsilon^{-1})$.

As a comparison to the existing lower bound for bilevel optimization in Ji and Liang (2023), we consider randomized algorithms and do not assume the value function h to be convex or strongly convex.

5 Numerical Experiments

Even though our contribution is mostly theoretical, we run several experiments to highlight to compare the proposed algorithm with state-of-the-art stochastic

bilevel solvers. We compare our method to AmIGO (Arbel and Mairal, 2022), F^{2}SA (Kwon et al., 2023), MRBO (Yang et al., 2021), VRBO (Yang et al., 2021), StocBiO (Ji et al., 2021) and SABA (Dagréou et al., 2022). They are run on a synthetic problem with quadratic functions and on a hyperparameter selection problem for ℓ^2-regularized logistic regression with the dataset IJCNN1[1]. A more detailed description of the experiments is available in Appendix C and an additional experiment the datacleaning task is available in Appendix D.

Experiments on quadratics To evaluate the performance of stochastic bilevel optimizers in a controlled setting, we perform a benchmark on quadratic loss functions described in Appendix C. Here F and G are quadratic jointly in (z, x), allowing us to choose freely the conditioning of F, G, and h. We take for the Hessian and cross derivative matrices of each sample, the empirical correlation of random vectors drawn with a prescribed covariance matrix. The generation process is detailed in Appendix C. In Figure 1, we report the norm of the gradient of the value function function with respect to time. Our first observation is that among all the methods, SRBA and SABA converge the fastest. These two solvers share two key ingredients: variance reduction and warm-starting. Variance reduction makes the variance of the gradient estimate go to zero without using decreasing step sizes. The warm-starting strategy in both the approximation of $z^*(x^t)$ and the approximation of $v^*(x^t)$ enables getting an estimator of $\nabla h(x^t)$ which is asymptotically unbiased, without requiring an increasing number of inner iterations or batch-size. Note that solvers using Neumann iterations (VRBO, MRBO, stocBiO) fail to converge because Neumann iterations provide a biased estimate of $v^*(x)$. Moreover, AmiIGO and stocBiO evolve slowly after some iterations because they require vanishing step sizes to converge. Finally, SRBA is faster than SABA, which is consistent with the theory.

Hyperparameter selection We also run an experiment on hyperparameter selection problem for ℓ^2-regularized logistic regression with the IJCNN1 dataset. SRBA shows good performances in the experiment, both in speed and accuracy. It is competitive with other state-of-the-art methods AmIGO and SABA, while going faster than Amigo and requiring less memory than SABA. VRBO –another extension of SARAH for bilevel problems– is slower in all problems. This is due to the burden of computing the approximate hypergradient at each inner iteration without updating the outer parameter. We can also notice that in the experiment on IJCNN1, the slowest method are method implementing Neumann approximations to approximate $v^*(x)$. Note that this last experiment does not include F^{2}SA because we find that on this problem, the norm of the iterates of F^{2}SA goes towards infinity.

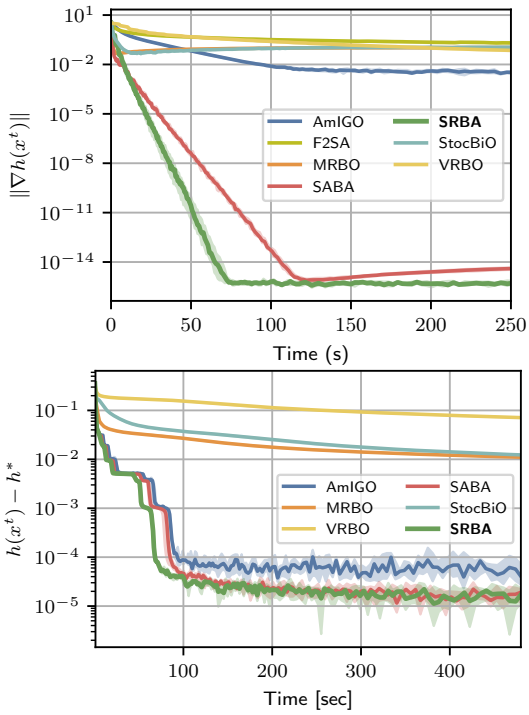

Figure 1: Comparison of the behavior of SRBA with other stochastic bilevel solvers. For each experiment, the solvers are run with 10 different seeds and the median performance over these seeds is reported. The shaded area corresponds to the performances between the 20% and the 80% percentiles. The performances are reported with respect to wall-clock time. **Top:** Experiments on quadratic functions. We report the gradient norm of the value function. **Bottom:** Hyperparameter selection with the IJCNN1 dataset.

6 Conclusion

In this paper, we have introduced SRBA, an algorithm for bilevel empirical risk minimization. We have demonstrated that the sample complexity of SRBA is $\mathcal{O}((n+m)^{\frac{1}{2}}\varepsilon^{-1})$ for any bilevel problem where the inner problem is strongly convex. Then, we have demonstrated that any bilevel empirical risk minimization algorithm has a sample complexity of at least $\mathcal{O}(m^{\frac{1}{2}}\varepsilon^{-1})$ on some problems where the inner problem is strongly convex. This demonstrates that SRBA is optimal, up to constant factors, and that bilevel ERM is as hard as single-level nonconvex ERM.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

References

Pierre Ablin, Gabriel Peyré, and Thomas Moreau. Super-efficiency of automatic differentiation for functions defined as a minimum. In *International Conference on Machine Learning (ICML)*, 2020.

Alekh Agarwal and Léon Bottou. A Lower Bound for the Optimization of Finite Sums. In *International Conference on Machine Learning (ICML)*, 2015.

Michael Arbel and Julien Mairal. Amortized Implicit Differentiation for Stochastic Bilevel Optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yoshua Bengio. Gradient-Based Optimization of Hyperparameters. *Neural Computation*, 12(8):1889–1900, 2000. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976600300015187.

Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In *International Conference on Computational Statistics (COMPSTAT)*, pages 177–186, 2010.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Reviews*, 60(2):223–311, 2018.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: First-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the Gap: Tighter Analysis of Alternating Stochastic Gradient Methods for Bilevel Problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path Integrated Differential Estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and Reverse Gradient-Based Hyperparameter Optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel Programming for Hyperparameter Optimization and Meta-Learning. In *International Conference on Machine Learning (ICML)*, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23 (4):2341–2368, 2013.

Saeed Ghadimi and Mengdi Wang. Approximation Methods for Bilevel Programming. *arXiv preprint arXiv:1802.02246*, 2018.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In Hal Daumé III and Aarti Singh, editors, *International Conference on Machine Learning (ICML)*, 2020.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A Two-Timescale Stochastic Algorithm Framework for Bilevel Optimization: Complexity Analysis and Application to Actor-Critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Quanqi Hu, Yongjian Zhong, and Tianbao Yang. Multi-block Min-max Bilevel Optimization with Applications in Multi-task Deep AUC Maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Feihu Huang, Junyi Li, Shangqian Gao, and Heng Huang. Enhanced Bilevel Optimization via Bregman Distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kaiyi Ji and Yingbin Liang. Lower Bounds and Accelerated Algorithms for Bilevel Optimization. *Journal of Machine Learning Research*, 24(22):1–56, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel Optimization: Convergence Analysis and Enhanced Design. In *International Conference on Machine Learning (ICML)*, 2021.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-\Lojasiewicz Condition. In *European Conference on Machine Learning (ECML)*, 2016.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A Near-Optimal Algorithm for Stochastic Bilevel Optimization via Double-Momentum. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. A Fully First-Order Method for Stochastic Bilevel Optimization. In *International Conference on Machine Leaning (ICML)*, 2023.

Junyi Li, Bin Gu, and Heng Huang. A Fully Single Loop Algorithm for Bilevel Optimization without Hessian Inverse. In *Proceedings of the Thirty-sixth AAAI Conference on Artificial Intelligence, AAAI'22*, 2022.

Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M. Robertson, and Yongxin Yang. DADA: Differentiable Automatic Data Augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable Architecture Search. In *International Conference on Learning Representations (ICLR)*, 2019.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing Millions of Hyperparameters by Implicit Differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*, pages 1540–1552, 2020.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based Hyperparameter Optimization through Reversible Learning. In *International Conference on Machine Learning (ICML)*, 2015.

Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Arkadiĭ Semenovich Nemirovsky and David Berkovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester ; New York, 1983. ISBN 978-0-471-10345-5.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Berlin Heidelberg, New York, NY, 2018. ISBN 978-3-319-91577-7.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. In *International Conference on Machine Learning (ICML)*, 2017.

Lam M. Nguyen, Marten van Dijk, Dzung T. Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R. Kalagnanam. Finite-sum smooth optimization with SARAH. *Computational Optimization and Applications*, 82(3):561–593, 2022. ISSN 0926-6003, 1573-2894. doi: 10.1007/s10589-022-00375-x.

Barak A. Pearlmutter. Fast Exact Multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1994.6.1.147.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning (ICML)*, 2016.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-Learning with Implicit Gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zaccharie Ramzi, Florian Mannel, Shaojie Bai, Jean-Luc Starck, Philippe Ciuciu, and Thomas Moreau. SHINE: SHaring the INverse Estimate from the forward pass for bi-level optimization and implicit models. In *International Conference on Learning Representations (ICLR)*, 2022.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alex Smola. Fast Incremental Method for Nonconvex Optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, IEEE, pages 1971–1977, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. CADDA: Class-wise Automatic Differentiable Data Augmentation for EEG Signals. In *International Conference on Learning Representations (ICLR)*, 2022.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the Convergence Theory for Hessian-Free Bilevel Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Blake E Woodworth and Nati Srebro. Tight Complexity Bounds for Optimizing Composite Objectives. In *Advances in Neural Information Systems Processing (NeurIPS)*, 2016.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably Faster Algorithms for Bilevel Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mao Ye, Bo Liu, Stephen Wright, Peter Stone, and Qiang Liu. BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Miao Zhang, Steven W. Su, Shirui Pan, Xiaojun Chang, Ehsan M Abbasnejad, and Reza Haffari. iDARTS: Differentiable Architecture Search with Stochastic Implicit Gradients. In *International Conference on Machine Learning (ICML)*, 2021.

Dongruo Zhou and Quanquan Gu. Lower Bounds for Smooth Nonconvex Finite-Sum Optimization. In *International Conference on Machine Learning (ICML)*, 2019.

Acknowledgements

SV acknowledges the support of the ANR GraVa ANR-18-CE40-0005. This work is supported by a public grant overseen by the French National Research Agency (ANR) through the program UDOPIA, project funded by the ANR-20-THIA-0013-01 and DATAIA convergence institute (ANR-17-CONV-0003).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes** [Section 2]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes** [Section 3]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes** [Appendix C]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes** [Section 2.1]
   (b) Complete proofs of all theoretical results. **Yes** [Appendix A and Appendix B]
   (c) Clear explanations of any assumptions. **Yes** [Section 2.1]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes** [Appendix C]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes** [Appendix C and Appendix D]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes** [Figure 1]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes** [Appendix C]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**
   (b) The license information of the assets, if applicable. **Not applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes**
   (d) Information about consent from data providers/curators. **Not applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Appendix A contains the necessary lemmas and proofs of Section 3. Appendix B contains the proof of the lower bound for stochastic bilevel optimization. Appendix C details the setting of the numerical experiments. Finally, Appendix D contains two more experiments on hyperparameter selection and datacleaning tasks.

A Convergence analysis of SRBA

A.1 Proof of Proposition 2.6

Proof. Let $x \in \mathbb{R}^d$. Since $G(\,.\,, x)$ is differentiable and $z^*(x)$ minimizes $G(\,.\,, x)$, the first order optimality condition ensures $\nabla_1 G(z^*(x), x) = 0 = D_z(z^*(x), v^*(x), x)$. Since G is strongly convex with respect to z, the Hessian $\nabla^2_{11} G(z^*(x), x)$ is invertible. As a consequence, the equation in v

$$D_v(z^*(x), v, x) = \nabla^2_{11} G(z^*(x), x)v + \nabla_1 F(z^*(x), x) = 0 \tag{13}$$

admits a unique solution given by $v^*(x)$. □

A.2 Smoothness constant of h

We can find in Ghadimi and Wang (2018, Lemma 2.2) the following value for the smoothness constant of h

$$L^h = L_1^F + \frac{2L_1^F L_2^G + (L_0^F)^2 L_2^G}{\mu_G} + \frac{L_{11}^G L_1^G L_0^F + L_1^G L_2^G L_0^F + (L_1^G)^2 L_1^F}{\mu_G^2} + \frac{(L_1^G)^2 L_2^G L_0^F}{\mu_G^3} \ .$$

A.3 Proof of Proposition 3.2

Proof. Let $t > 0$ and $k \in [q-1]$. For $k = 0$, we directly have $\mathbb{E}[\|D_\bullet^{t,k} - D_\bullet(\mathbf{u}^{t,k})\|^2] = 0$. For $k \geq 1$ and $r \in \{1, \dots, k\}$, the bias/variance decomposition of $D_\bullet^{t,r}$ reads

$$\begin{aligned}
\mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet(\mathbf{u}^{t,r})\|^2] &= \mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1} + D_\bullet(\mathbf{u}^{t,r-1}) - D_\bullet(\mathbf{u}^{t,r})\|^2] \\
&\quad + \|D_\bullet(\mathbf{u}^{t,r}) + D_\bullet(\mathbf{u}^{t,r-1}) - D_\bullet^{t,r-1} - D_\bullet(\mathbf{u}^{t,r})\|^2 \\
&= \mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1} - (D_\bullet(\mathbf{u}^{t,r-1}) - D_\bullet(\mathbf{u}^{t,r}))\|^2] \\
&\quad + \|D_\bullet^{t,r-1} - D_\bullet(\mathbf{u}^{t,r-1})\|^2
\end{aligned}$$

The term $\mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1} - (D_\bullet(\mathbf{u}^{t,r-1}) - D_\bullet(\mathbf{u}^{t,r}))\|^2]$ is the variance of $D_\bullet^{t,r} - D_\bullet^{t,r-1}$, and then can written as

$$\begin{aligned}
\mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1} - (D_\bullet(\mathbf{u}^{t,r-1}) - D_\bullet(\mathbf{u}^{t,r}))\|^2] &= \mathbb{E}_{t,r}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1}\|^2] \\
&\quad - \|D_\bullet(\mathbf{u}^{t,r}) - D_\bullet(\mathbf{u}^{t,r-1})\|^2
\end{aligned}$$

Plugging this in the previous inequality and taking the total expectation leads to

$$\begin{aligned}
\mathbb{E}[\|D_\bullet^{t,r} - D_\bullet(\mathbf{u}^{t,r})\|^2] &= \mathbb{E}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1}\|^2] - \mathbb{E}[\|D_\bullet(\mathbf{u}^{t,r}) - D_\bullet(\mathbf{u}^{t,r-1})\|^2] \\
&\quad + \mathbb{E}[\|D_\bullet^{t,r-1} - D_\bullet^{t,r-1}(\mathbf{u}^{t,r-1})\|^2]
\end{aligned}$$

Summing for $r \in \{1, \dots, k\}$ and telescoping gives the final result (taking into account that $D_\bullet^{t,0} = D_\bullet(\mathbf{u}^{t,0})$):

$$\mathbb{E}[\|D_\bullet^{t,k} - D_\bullet(\mathbf{u}^{t,k})\|^2] = \sum_{r=1}^{k} \mathbb{E}[\|D_\bullet^{t,r} - D_\bullet^{t,r-1}\|^2] - \sum_{r=1}^{k} \mathbb{E}[\|D_\bullet(\mathbf{u}^{t,r}) - D_\bullet(\mathbf{u}^{t,r-1})\|^2] \ .$$

□

A.4 Technical lemmas

Lemma A.1. *There exists constant L_{z^*} and L_{v^*} such that for any x_1, $x_2 \in \mathbb{R}^d$, we have*

$$\|z^*(x_1) - z^*(x_2)\| \leq L_{z^*}\|x_1 - x_2\| \quad \text{and} \quad \|v^*(x_1) - v^*(x_2)\| \leq L_{v^*}\|x_1 - x_2\|$$

Proof. The Jacobian of z^* reads $\mathrm{d}z^*(x) = [\nabla^2_{11}G(z^*(x), x)]^{-1}\nabla^2_{12}G(z^*(x), x)$. By μ_G-strong convexity and L_1^G-smoothness of G, we have $\|\mathrm{d}z^*(x)\| \leq \frac{L_1^G}{\mu_G}$ which implies that z^* is L_{z^*}-Lipschtiz with $L_{z^*} = \frac{L_1^G}{\mu_G}$.

For v^* we do the computation directly:

$$\begin{aligned}
\|v^*(x_1) - v^*(x_2)\| &= \|[\nabla^2_{11}G(z^*(x_1), x_1)]^{-1}\nabla_1 F(z^*(x_1), x_1) \\
&\quad - [\nabla^2_{11}G(z^*(x_2), x_2)]^{-1}\nabla_1 F(z^*(x_2), x_2)\| \\
&\leq \|[\nabla^2_{11}G(z^*(x_1), x_1)]^{-1}(\nabla_1 F(z^*(x_1), x_1) - \nabla_1 F(z^*(x_2), x_2))\| \\
&\quad + \|([\nabla^2_{11}G(z^*(x_1), x_1) - [\nabla^2_{11}G(z^*(x_2), x_2)]^{-1}]^{-1}\nabla_1 F(z^*(x_2), x_2)\| \\
&\leq \left(\frac{L_1^F}{\mu_G} + \frac{L_2^G L_0^F}{\mu_G^2}\right)\|(z^*(x_1), x_1) - (z^*(x_2), x_2)\| \\
&\leq \left(\frac{L_1^F}{\mu_G} + \frac{L_2^G L_0^F}{\mu_G^2}\right)(\|z^*(x_1) - z^*(x_2)\| + \|x_1 - x_2\|) \\
&\leq \left(1 + \frac{L_1^G}{\mu_G}\right)\left(\frac{L_1^F}{\mu_G} + \frac{L_2^G L_0^F}{\mu_G^2}\right)\|x_1 - x_2\|
\end{aligned}$$

Then taking $L_{v^*} = \left(1 + \frac{L_1^G}{\mu_G}\right)\left(\frac{L_1^F}{\mu_G} + \frac{L_2^G L_0^F}{\mu_G^2}\right)$ concludes the proof. □

Lemma A.2. *Let us consider the update directions $D_z^{t,k} = \mathbf{\Delta}_z^{t,k}/\rho$, $D_v^{t,k} = \mathbf{\Delta}_v^{t,k}/\rho$ and $D_x^{t,k} = \mathbf{\Delta}_x^{t,k}/\gamma$ where $\mathbf{\Delta}_z^{t,k}$, $\mathbf{\Delta}_v^{t,k}$ and $\mathbf{\Delta}_x^{t,k}$ verify Equations (7) to (9). Then it holds*

$$\begin{aligned}
\mathbb{E}[\|D_z^{t,k} - D_z(\mathbf{u}^{t,k})\|^2] &\leq \sum_{r=1}^{k} L_1^G(\rho^2\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\mathbb{E}[\|D_z^{t,r-1}\|^2]) \\
\mathbb{E}[\|D_v^{t,k} - D_v(\mathbf{u}^{t,k})\|^2] &\leq 4\rho^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\rho^2(L_1^G)^2\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\
&\quad + 4\gamma^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \\
\mathbb{E}[\|D_x^{t,k} - D_x(\mathbf{u}^{t,k})\|^2] &\leq 4\rho^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\rho^2(L_1^G)^2\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\
&\quad + 4\gamma^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \ .
\end{aligned}$$

Proof. **Direction D_z**

We start from Proposition 3.2.

$$\begin{aligned}
\mathbb{E}[\|D_z^{t,k} - D_z(\mathbf{u}^{t,k})\|^2] &= \mathbb{E}[\|D_z^{t,k} - \nabla_1 G(z^{t,k}, x^{t,k})\|^2] \\
&= \sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r} - D_z^{t,r-1}\|^2] - \sum_{r=1}^{k}\mathbb{E}[\|\nabla_1 G(z^{t,r}, x^{t,r}) - \nabla_1 G(z^{t,r-1}, x^{t,r-1})\|^2] \\
&\leq \sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r} - D_z^{t,r-1}\|^2] \\
&\leq \sum_{r=1}^{k} L_1^G(\rho^2\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\mathbb{E}[\|D_z^{t,r-1}\|^2])
\end{aligned}$$

where the last inequality comes from the smoothness of each G_i.

Direction D_v For D_v, the proof is almost the same. Proposition 3.2 gives us

$$\mathbb{E}[\|D_v^{t,k} - D_v(\mathbf{u}^{t,k})\|^2] \leq \sum_{r=1}^{k} \mathbb{E}[\|D_v^{t,r} - D_v^{t,r-1}\|^2] \ .$$

Then, using the boundedness of v and regularity of each G_i and F_j, we have

$$\begin{aligned}
\mathbb{E}[\|D_v^{t,r} - D_v^{t,r-1}\|^2] &\leq 2(\mathbb{E}[\|\nabla_{11}^2 G_i(z^{t,r}, x^{t,r})v^{t,r} - \nabla_{11}^2 G_i(z^{t,r-1}, x^{t,r-1})v^{t,r-1}\|^2] \\
&\quad + \mathbb{E}[\|\nabla_2 F_j(z^{t,r}, x^{t,r}) - \nabla_2 F_j(z^{t,r-1}, x^{t,r-1})\|^2]) \\
&\leq 4(\mathbb{E}[\|\nabla_{11}^2 G_i(z^{t,r}, x^{t,r})(v^{t,r} - v^{t,r-1})\|^2] \\
&\quad + \mathbb{E}[\|(\nabla_{11}^2 G_i(z^{t,r}, x^{t,r}) - \nabla_{11}^2 G_i(z^{t,r-1}, x^{t,r-1}))v^{t,r-1}\|^2] \\
&\quad + (L_1^F)^2(\gamma^2 \mathbb{E}[\|D_z^{t,r-1}\|] + \rho^2 \mathbb{E}[\|D_x^{t,r-1}\|^2])) \\
&\leq 4((L_1^G)^2 \rho^2 \mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\
&\quad + (L_2^G)^2 R^2(\rho^2 \mathbb{E}[\|D_z^{t,r-1}\|] + \gamma^2 \mathbb{E}[\|D_x^{t,r-1}\|^2]) \\
&\quad + (L_1^F)^2(\rho^2 \mathbb{E}[\|D_z^{t,r-1}\|] + \gamma^2 \mathbb{E}[\|D_x^{t,r-1}\|^2])) \\
&\leq 4\rho^2 \left((L_2^G R)^2 + (L_1^F)^2\right) \mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\rho^2 (L_1^G)^2 \mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\
&\quad + 4\gamma^2 \left((L_2^G R)^2 + (L_1^F)^2\right) \mathbb{E}[\|D_x^{t,r-1}\|^2] \ .
\end{aligned}$$

Direction D_x The proof is the same as the proof for D_v. □

A.5 Proof of Lemma 3.3

Let $\phi_z(z, x) = G(z, x) - G(z^*(x), x)$ the inner suboptimality gap. The proof of Lemma 3.3 is based on the smoothness of ϕ_z, which is the object of the following lemma.

Lemma A.3. *The function ϕ_z has Λ_z-Lipschitz continuous gradient on $\mathbb{R}^p \times \mathbb{R}^d$, for some constant Λ_z.*

Proof. For any $(z, x) \in \mathbb{R}^p \times \mathbb{R}^d$, we have

$$\nabla_1 \phi_z(z, x) = \nabla_1 G(z, x) \text{ and } \nabla_2 \phi_z(z, x) = \nabla_2 G(z, x) - \nabla_2 G(z^*(x), x) \ .$$

Let us consider $(z, x) \in \mathbb{R}^p \times \mathbb{R}^d$ and $(z', x') \in \mathbb{R}^p \times \mathbb{R}^d$. Since ∇G is L_1^G-Lipschitz continuous, we have directly

$$\|\nabla_1 \phi_z(z, x) - \nabla_1 \phi_z(z', x')\| \leq L_1^G \|(z, x) - (z', x')\| \ .$$

Moreover, we have

$$\begin{aligned}
\|\nabla_2 \phi_z(z, x) - \nabla_2 \phi_z(z', x')\| &\leq \|\nabla_2 G(z, x) - \nabla_2 G(z', x')\| \\
&\quad + \|\nabla_2 G(z^*(x), x) - \nabla_2 G(z^*(x'), x')\| \\
&\leq L_1^G \|(z, x) - (z', x')\| + L_1^G \|(z^*(x), x) - (z^*(x'), x')\| \\
&\leq L_1^G \|(z, x) - (z', x')\| + L_1^G (\|z^*(x) - z^*(x')\| + \|x - x'\|) \ .
\end{aligned}$$

From Lemma A.1, z^* is L_* Lipschitz continuous, so

$$\begin{aligned}
\|\nabla_2 \phi_z(z, x) - \nabla_2 \phi_z(z', x')\| &\leq L_1^G \|(z, x) - (z', x')\| + L_1^G (\|z^*(x) - z^*(x')\| + \|x - x'\|) \\
&\leq L_1^G \|(z, x) - (z', x')\| + L_1^G (L_* + 1)\|x - x'\| \\
&\leq L_1^G (L_{z^*} + 2)\|(z, x) - (z', x')\| \ .
\end{aligned}$$

As a consequence

$$\begin{aligned}
\|\nabla \phi_z(z, x) - \nabla \phi_z(z', x')\| &\leq \|\nabla_1 \phi_z(z, x) - \nabla_1 \phi_z(z', x')\| + \|\nabla_2 \phi_z(z, x) - \nabla_2 \phi_z(z', x')\| \\
&\leq L_1^G (L_{z^*} + 3)\|(z, x) - (z', x')\| \ .
\end{aligned}$$

Hence, ϕ_z is Λ_z smooth with $\Lambda_z = L_1^G (L_{z^*} + 3)$. □

We can now turn to the proof of Lemma 3.3.

Proof. The smoothness of ϕ_z provides us the following upper bound

$$\begin{aligned}\phi_z(z^{t,k+1}, x^{t,k+1}) \leq \phi_z(z^{t,k}, x^{t,k}) - \rho\langle D_z^{t,k}, \nabla_1 G(z^{t,k}, x^{t,k})\rangle + \frac{\Lambda_z}{2}\rho^2\|D_z^{t,k}\|^2 \\ - \gamma\langle D_x^{t,k}, \nabla_2 G(z^{t,k}, x^{t,k}) - \nabla_2 G(z^*(x^{t,k}), x^{t,k})\rangle + \frac{\Lambda_z}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{14}$$

Using the equality $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$, we get

$$\begin{aligned}-\langle D_z^{t,k}, \nabla_1 G(z^{t,k}, x^{t,k})\rangle + \frac{\Lambda_z}{2}\rho\|D_z^{t,k}\|^2 = \frac{1}{2}(\|D_z^{t,k} - \nabla_1 G(z^{t,k}, x^{t,k})\|^2 \\ - \|\nabla_1 G(z^{t,k}, x^{t,k})\|^2 - (1 - \Lambda_z\rho)\|D_z^{t,k}\|^2) \ .\end{aligned} \tag{15}$$

Plugging Equation (15) into Equation (14) and tacking the expectation conditionally to the past iterates yields

$$\begin{aligned}\mathbb{E}_{t,k}[\phi_z^{t,k+1}] \leq \phi_z^{t,k} + \frac{\rho}{2}\mathbb{E}_{t,k}[\|D_z^{t,k} - \nabla_1 G(z^{t,k}, x^{t,k})\|^2] \\ - \frac{\rho}{2}\|\nabla_1 G(z^{t,k}, x^{t,k})\|^2 - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}_{t,k}[\|D_z^{t,k}\|^2] \\ - \gamma\langle \mathbb{E}_{t,k}[D_x^{t,k}], \nabla_2 G(z^{t,k}, x^{t,k}) - \nabla_2 G(z^*(x^{t,k}), x^{t,k})\rangle + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}_{t,k}[\|D_x^{t,k}\|^2] \ .\end{aligned} \tag{16}$$

From Young inequality, we have for any $c > 0$

$$\begin{aligned}\langle \mathbb{E}_{t,k}[D_x^{t,k}], \nabla_2 G(z^{t,k}, x^{t,k}) - \nabla_2 G(z^*(x^{t,k}), x^{t,k})\rangle \leq \frac{1}{2c}\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2 \\ + \frac{c}{2}\|\nabla_2 G(z^{t,k}, x^{t,k}) - \nabla_2 G(z^*(x^{t,k}), x^{t,k})\|^2\end{aligned} \tag{17}$$

The smoothness of G and strong convexity give us

$$\|\nabla_2 G(z^{t,k}, x^{t,k}) - \nabla_2 G(z^*(x^{t,k}), x^{t,k})\|^2 \leq L_1^G\|z^{t,k} - z^*(x^{t,k})\|^2 \leq \frac{2L_1^G}{\mu_G}\phi_z(z^{t,k}, x^{t,k}) \tag{18}$$

Let us denote $L' = \frac{L_1^G}{\mu_G}$. Plugging Inequalities (17) and (18) into Equation (16) yields

$$\begin{aligned}\mathbb{E}_{t,k}[\phi_z(z^{t,k+1}, x^{t,k+1})] \leq (1 + cL'\gamma)\phi_z(z^{t,k+1}, x^{t,k+1}) - \frac{\rho}{2}\mathbb{E}_{t,k}[\|\nabla_1 G(z^{t,k}, x^{t,k})\|^2] \\ + \frac{\rho}{2}\mathbb{E}_{t,k}[\|D_z^{t,k} - \nabla_1 G(z^{t,k}, x^{t,k})\|^2] - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}_{t,k}[\|D_z^{t,k}\|^2] \\ + \frac{\gamma}{2c}\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2 + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}_{t,k}[\|D_x^{t,k}\|^2]\end{aligned} \tag{19}$$

From Lemma A.2, we have

$$\mathbb{E}[\|D_z^{t,k} - \nabla_1 G(z^{t,k}, x^{t,k})\|^2] \leq \sum_{r=1}^{k} L_1^G(\rho^2\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\mathbb{E}[\|D_z^{t,r-1}\|^2]) \ .$$

Taking the total expectation and plugging the previous inequality into Equation (19) yields

$$\begin{aligned}\phi_z^{t,k+1} \leq (1 + cL'\gamma)\phi^{t,k} + \frac{L_1^G}{2}\sum_{r=1}^{k}(\rho^3\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\rho\mathbb{E}[\|D_x^{t,r-1}\|^2]) \\ - \frac{\rho}{2}\mathbb{E}[\|\nabla_1 G(z^{t,k}, x^{t,k})\|^2] - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}[\|D_z^{t,k}\|^2] \\ + \frac{\gamma}{2c}\mathbb{E}[\|\mathbb{E}[D_x^{t,k}]\|^2] + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2]\end{aligned} \tag{20}$$

Since G is μ_G-strongly convex with respect to z, Polyak-Łojasiewicz inequality holds:

$$\|\nabla_1 G(z^{t,k}, x^{t,k})\|^2 \geq 2\mu_G \phi_z(z^{t,k}, x^{t,k})$$

As a consequence, Equation (20) becomes

$$\begin{aligned}\phi_z^{t,k+1} \leq (1 + cL'\gamma - \mu_G\rho)\,\phi^{t,k} + \frac{L_1^G}{2}\sum_{r=1}^{k}(\rho^3\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\rho\mathbb{E}[\|D_x^{t,r-1}\|^2]) \\ - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}[\|D_z^{t,k}\|^2] + \frac{\gamma}{2c}\mathbb{E}[\|\mathbb{E}[D_x^{t,k}]\|^2] + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2]\end{aligned}$$

Taking $c = \frac{\mu_G\rho}{2L'\gamma}$ yields

$$\begin{aligned}\phi_z^{t,k+1} \leq \left(1 - \frac{\mu_G}{2}\rho\right)\phi^{t,k} + \frac{L_1^G}{2}\sum_{r=1}^{k}(\rho^3\mathbb{E}[\|D_z^{t,r-1}\|^2] + \gamma^2\rho\mathbb{E}[\|D_x^{t,r-1}\|^2]) \\ - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}[\|D_z^{t,k}\|^2] + \frac{L'}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|\mathbb{E}[D_x^{t,k}]\|^2] + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2]\end{aligned}$$

For the term $\mathbb{E}[\|\mathbb{E}_{t,k}[D_z^{t,k}]\|^2]$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2] &= \mathbb{E}[\|D_x(z^{t,k}, v^{t,k}, x^{t,k}) - D_x(z^{t,k-1}, v^{t,k-1}, x^{t,k-1}) + D_x^{t,k-1}\|^2] \\ &= \mathbb{E}[\|D_x(z^{t,k}, v^{t,k}, x^{t,k}) - D_x(z^{t,k-1}, v^{t,k-1}, x^{t,k-1}) - \mathbb{E}[D_x^{t,k-1}]\|^2] \\ &\quad + \mathbb{E}[\|D_x^{t,k-1} - \mathbb{E}[D_x^{t,k-1}]\|^2] \\ &= \mathbb{E}[\|D_x(z^{t,k}, v^{t,k}, x^{t,k})\|^2] \\ &\quad + \mathbb{E}[\|D_x^{t,k-1} - D_x(z^{t,k-1}, v^{t,k-1}, x^{t,k-1})\|^2]\ .\end{aligned} \tag{21}$$

Using Lemma A.2, we get

$$\begin{aligned}\mathbb{E}[\|D_x^{t,k-1} - D_x(\mathbf{u}^{t,k-1})\|^2] \leq 4\rho^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k-1}\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + 4\rho^2(L_1^G)^2\sum_{r=1}^{k-1}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 4\gamma^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k-1}\mathbb{E}[\|D_x^{t,r-1}\|^2]\ .\end{aligned}$$

Putting all together yields

$$\begin{aligned}\phi_z^{t,k+1} \leq \left(1 - \frac{\mu_G}{2}\rho\right)\phi^{t,k} - \frac{\rho}{2}(1 - \Lambda_z\rho)\,\mathbb{E}[\|D_z^{t,k}\|^2] + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] \\ + \frac{L'}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x^{t,k}(\mathbf{u}^{t,k})\|^2] + 4(L_1^G)^2\frac{L'}{\mu_G}\gamma^2\rho\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + \rho\left[\rho^2\frac{L_1^G}{2} + \frac{4(L_2^G R)^2 L'}{\mu_G}\gamma^2 + \frac{4(L_1^F)^2 L'}{\mu_G}\gamma^2\right]\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + \gamma^2\left[\rho\frac{L_1^G}{2} + 4(L_2^G R)^2\frac{L'}{\mu_G}\frac{\gamma^2}{\rho} + 4(L_1^F)^2\frac{L'}{\mu_G}\frac{\gamma^2}{\rho}\right]\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2]\end{aligned} \tag{22}$$

By assumption, $\gamma \leq C_z\rho$, with $C_z = \sqrt{\frac{\mu_G L_1^G}{8L'((L_2^G R)^2 + (L_1^F)^2)}}$ therefore

$$\begin{aligned}\phi_z^{t,k+1} &\leq \left(1-\frac{\mu_G}{2}\rho\right)\phi_z^{t,k} - \frac{\rho}{2}\left(1-\Lambda_z\rho\right)\mathbb{E}[\|D_z^{t,k}\|^2] + \frac{\Lambda_z}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] \\ &\quad + \frac{L'}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x^{t,k}(\mathbf{u}^{t,k})\|^2] + \rho^3 L_1^G \sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ &\quad + 4(L_1^G)^2\frac{L'}{\mu_G}\gamma^2\rho\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] + \gamma^2\rho L_1^G\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \\ &\leq \left(1-\frac{\mu_G}{2}\rho\right)\phi^{t,k} - \frac{\rho}{2}\left(1-\Lambda_z\rho\right)V_z^{t,k} + \frac{\Lambda_z}{2}\gamma^2 V_x^{t,k} + \overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x^{t,k}(\mathbf{u}^{t,k})\|^2] \\ &\quad + \rho^3\beta_{zz}\mathcal{V}_z^{t,k} + \gamma^2\rho\beta_{zv}\mathcal{V}_v^{t,k} + \gamma^2\rho\beta_{zx}\mathcal{V}_x^{t,k}\end{aligned}$$

with $\beta_{zz} = L_1^G$, $\beta_{zv} = \frac{4(L_1^G)^2 L'}{\mu_G}$, $\beta_{zx} = L_1^G$ and $\overline{\beta}_{zx} = \frac{L'}{\mu_G}$. □

A.6 Proof of Lemma 3.4

Recall that we denote $\Psi(z,v,x) = \frac{1}{2}v^\top\nabla_{11}^2 G(z,x)v + \nabla_1 F(z,x)^\top v$ and $\phi_v(v,x) = \Psi(z^*(x),v,x) - \Psi(z^*(x),v^*(x),x)$. As for Lemma 3.3, the key property we need is the smoothness of ϕ_v. The derivatives of ϕ_v involve the third derivative of G. For a tensor $T \in \mathbb{R}^{p_1\times p_2\times p_3}$ and a vector $a\in\mathbb{R}^{p_3}$ we denote $(T|a)$ the matrix in $\mathbb{R}^{p_1\times p_2}$ defined by:

$$(T|a) = \left[\sum_{k=1}^{p_3} T_{i,j,k}a_k\right]_{\substack{1\leq i\leq p_1\\ 1\leq j\leq p_2}}.$$

Lemma A.4. *The function ϕ_v has Λ_v-Lipschitz continuous gradient on $\Gamma\times\mathbb{R}^d$, for some constant Λ_v.*

Proof. For any $(v,x)\in\Gamma\times\mathbb{R}^d$, we have

$$\nabla_1\phi_v(v,x) = D_v(z^*(x),v,x)$$

and

$$\begin{aligned}\nabla_2\phi_v(v,x) = (\mathrm{d}z^*(x))^\top &\left[\frac{1}{2}(\nabla_{111}^3 G(z^*(x),x)|v)v - \frac{1}{2}(\nabla_{111}^3 G(z^*(x),x)|v^*(x))v^*(x)\right. \\ &\left.+\nabla_{11}^2 F(z^*(x),x)v - \nabla_{11}^2 F(z^*(x),x)v^*(x)\right] \\ &+\left[\frac{1}{2}(\nabla_{211}^3 G(z^*(x),x)|v)v - \frac{1}{2}(\nabla_{211}^3 G(z^*(x),x)|v^*(x))v^*(x)\right. \\ &\left.+\nabla_{21}^2 F(z^*(x),x)v - \nabla_{21}^2 F(z^*(x),x)v^*(x)\right].\end{aligned}$$

Let us consider $(v,x)\in\Gamma\times\mathbb{R}^d$ and $(v',x')\in\Gamma\times\mathbb{R}^d$. We have

$$\begin{aligned}\|\nabla_1\phi_v(v,x) - \nabla_1\phi_v(v',x')\| &\leq \|\nabla_{11}^2 G(z^*(x),x)v - \nabla_{11}^2 G(z^*(x'),x')v'\| \\ &\quad + \|\nabla_1 F(z^*(x),x) - \nabla_1 F(z^*(x'),x')\|\end{aligned}$$

For the first term,

$$\begin{aligned}\|\nabla_{11}^2 G(z^*(x),x)v - \nabla_{11}^2 G(z^*(x'),x')v'\| &\leq \|\nabla_{11}^2 G(z^*(x),x)(v-v')\| \\ &\quad + \|(\nabla_{11}^2 G(z^*(x),x) - \nabla_{11}^2 G(z^*(x'),x'))v'\| \\ &\quad + \|\nabla_{11}^2 G(z^*(x'),x')(v-v')\| \\ &\leq 2L_1^G\|v-v'\| + L_2^G(L_{z^*}+1)\|v'\|\|x-x'\| \\ &\leq [2L_1^G + L_2^G(L_{z^*}+1)R]\|(v,x)-(v',x')\|\end{aligned}$$

For the second terms, we use the smoothness of F and the Lipschitz continuity of z^* (Lemma A.1):

$$\begin{aligned}\|\nabla_1 F(z^*(x),x) - \nabla_1 F(z^*(x'),x')\| &\leq L_1^F \|(z^*(x),x) - (z^*(x'),x')\| \\ &\leq L_1^F(\|z^*(x) - z^*(x')\| + \|x - x'\|) \\ &\leq L_1^F(L_{z^*}+1)\|x - x'\| \\ &\leq L_1^F(L_{z^*}+1)\|(x,v) - (x',v')\| \ .\end{aligned}$$

As a consequence

$$\|\nabla_1 \phi_v(v,x) - \nabla_1 \phi_v(v',x')\| \leq \Lambda_1 \|(v,x) - (v',x')\| \tag{23}$$

with

$$\Lambda_1 = L_1^F(L_{z^*}+1) + 2L_1^G + L_2^G(L_{z^*}+1)R \ . \tag{24}$$

To prove the Lipschitz continuity of $\nabla_2 \phi_v$, we remark that $\nabla^3_{111} G$, $\nabla^3_{211} G$ are Lipschitz and bounded by assumption. $(v \mapsto v)$ is Lipschitz and bounded on Γ. Also by Lemma A.1, z^* and v^* are Lipschitz and bounded. Finally, $\mathrm{d}z^*$ is bounded (Lemma A.1) and Lipschitz according to Chen et al. (2021)[Lemma 9]. As a consequence, $\nabla_2 \phi_v$ is Λ_2-Lpischitz for some constant $\Lambda_2 > 0$. Hence, $\nabla \phi_v$ is Λ_v-Lipschitz continuous with $\Lambda_v = \Lambda_1 + \Lambda_2$.

□

Lemma A.5. *Let $t > 0$. For $k \in [q-1]$, we have*

$$0 \leq -\left\langle \frac{1}{\rho}(v^{t,k+1} - v^{t,k}) + D_v^{t,k}, v^{t,k+1} - v^{t,k} \right\rangle$$

Proof. The function ι_Γ being convex (since Γ is convex), let us consider its sub-differential

$$\partial \iota_\gamma(v) = \{\eta \in \mathbb{R}^p, \forall v' \in \mathbb{R}^p, \iota_\Gamma(v') \geq \iota_\Gamma(v) + \langle \eta, v' - v \rangle\}$$

By definition

$$v^{t,k+1} = \arg\min_v (\iota_\Gamma(v) + \frac{1}{2\rho}\|v - (v^{t,k} - \rho D_v^{t,k})\|^2) \ .$$

Using Fermat's rule, we get

$$-\frac{1}{\rho}(v^{t,k+1} - v^{t,k}) - D_v^{t,k} \in \partial \iota_\Gamma(v^{t,k+1}) \ .$$

We can use the definition of the sub-differential with $\eta = -\frac{1}{\rho}(v^{t,k+1} - v^{t,k}) - D_v^{t,k}$ to get

$$\underbrace{\iota_\Gamma(v^{t,k+1})}_{=0} \leq \underbrace{\iota_\Gamma(v^{t,k})}_{=0} - \left\langle \frac{1}{\rho}(v^{t,k+1} - v^{t,k}) + D_v^{t,k}, v^{t,k+1} - v^{t,k} \right\rangle \ .$$

□

We can now turn to the proof of Lemma 3.4.

Proof. The smoothness of ϕ_v provides us the following upper bound

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \phi_v(v^{t,k}, x^{t,k}) &+ \langle \Pi_\Gamma(v^{t,k} - \rho D_v^{t,k}) - v^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}) \rangle \\ &+ \frac{\Lambda_v}{2}\rho^2 \|\Pi_\Gamma(v^{t,k} - \rho D_v^{t,k}) - v^{t,k}\|^2 \\ &- \gamma \langle D_x^{t,k}, \nabla_2 \phi_v(v^{t,k}, x^{t,k}) \rangle + \frac{\Lambda_v}{2}\gamma^2 \|D_x^{t,k}\|^2 \ .\end{aligned} \tag{25}$$

Let us denote $\Delta_\Pi^{t,k} = \Pi_\Gamma(v^{t,k} - \rho D_v^{t,k}) - \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}))$. Adding and subtracting

$$\begin{aligned}&\langle \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}) - v^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &\quad + \frac{\Lambda_v}{2}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}) - v^{t,k}\|^2\end{aligned}$$

yields

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \phi_v(v^{t,k}, x^{t,k}) &+ \langle \Delta_\Pi^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{\Lambda_v}{2}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^t), v^{t,k}, x^{t,k})) - v^{t,k}\|^2 \\ &+ \langle \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 + \Lambda_v\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{26}$$

Taking $\rho \leq \frac{1}{\Gamma_v}$ gives

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \phi_v(v^{t,k}, x^{t,k}) &+ \langle \Delta_\Pi^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{1}{2\rho}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^t), v^{t,k}, x^{t,k})) - v^{t,k}\|^2 \\ &+ \langle \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 + \Lambda_v\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{27}$$

Let ι_Γ the indicator function of the convex set Γ. Similarly to Karimi et al. (2016, Equation 13) we define for any $\alpha > 0$ and $v \in \mathbb{R}^p$

$$\mathcal{D}_{\iota_\Gamma}(v, x, \alpha) = -2\alpha \min_{v' \in \mathbb{R}^p}\left[\langle \nabla_1\phi_v(v, x), v' - v\rangle + \frac{\alpha}{2}\|v' - v\|^2 + \iota_\Gamma(v') - \iota_\Gamma(v)\right] \ .$$

Hence, for $v \in \Gamma$ and $x \in \mathbb{R}^d$, we have

$$\begin{aligned}-\frac{\rho}{2}\mathcal{D}_{\iota_\Gamma}\left(v, x, \frac{1}{\rho}\right) &= \langle \Pi_\Gamma(v - \rho D_v(z^*(x), v, x)) - v, D_v(z^*(x), v, x)\rangle \\ &\quad + \frac{1}{2\rho}\|\Pi_\Gamma(v - \rho D_v(z^*(x), v, x)) - v\|^2 \ .\end{aligned}$$

Therefore, Equation (27) can be written as

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \phi_v(v^{t,k}, x^{t,k}) &- \frac{\rho}{2}\mathcal{D}_{\iota_\Gamma}\left(v^{t,k}, x^{t,k}, \frac{1}{\rho}\right) \\ &+ \langle \Delta_\Pi^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 + \Lambda_v\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned}$$

By strong convexity of ϕ_v with respect top v and smoothness, we have $\mathcal{D}_{\iota_\Gamma}(v^{t,k}, x^{t,k}, \Lambda_v) \geq 2\mu_G\phi_v(v^{t,k}, x^{t,k})$. According to Karimi et al. (2016, Lemma 1), $\mathcal{D}_{\iota_\Gamma}(v^{t,k}, x^{t,k}, \bullet)$ is an increasing function. As a consequence, since $\Lambda_v \leq \frac{1}{\rho}$, we have $\mathcal{D}_{\iota_\Gamma}\left(v^{t,k}, x^{t,k}, \frac{1}{\rho}\right) \geq 2\mu_G\phi_v(v^{t,k}, x^{t,k})$. This leads to

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq (1 - \rho\mu_G)\phi_v(v^{t,k}, x^{t,k}) &+ \langle \Delta_\Pi^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle \\ &+ \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 + \Lambda_v\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{28}$$

The non-expansiveness of Π_Γ yields

$$\|\Delta_\Pi^{t,k}\| \leq \rho \|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\| \tag{29}$$

and

$$\begin{aligned}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}) - \underbrace{v^{t,k}}_{\in\Gamma}\| &= \|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - \Pi_\Gamma(v^{t,k})\| \\ &\leq \rho \|D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\| \ .\end{aligned} \tag{30}$$

Moreover, using Equation (29) and Young Inequality, we have for any $c > 0$

$$\begin{aligned}\langle \Delta_\Pi^{t,k}, D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\rangle &\leq \frac{c}{2}\|\Delta_\Pi\|^2 + \frac{1}{2c}\|D_v(z^*(x^{t,k}), v^{t,k}, x^{tk,})\|^2 \\ &\leq \frac{c\rho^2}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &\quad + \frac{1}{2c}\|D_v(z^*(x^{t,k}), v^{t,k}, x^{tk,}) - \underbrace{D_v(z^*(x^{t,k}), v^*(x^{t,k}), x^{tk,})}_{=0}\|^2 \\ &\leq \frac{c\rho^2}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &\quad + \frac{L_1^G}{\mu_G c}\phi_v(v^{t,k}, x^{t,k})\end{aligned} \tag{31}$$

Plugging Equation (31) into Equation (28) with $c = \frac{2L_1^G}{\mu_G^2\rho}$ yields

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \left(1 - \frac{\rho\mu_G}{2}\right)\phi_v(v^{t,k}, x^{t,k}) &+ \frac{L_1^G\rho}{\mu_G^2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &+ \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 + \Lambda_v\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{32}$$

Using Equation (29), Equation (30) and Young Inequality for $d > 0$ yields

$$\begin{aligned}\langle \Delta_\Pi^{t,k}, \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\rangle &\leq \frac{d}{2}\|\Delta_\Pi^{t,k}\|^2 \\ &\quad + \frac{1}{2d}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\|^2 \\ &\leq \frac{d\rho^2}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &\quad + \frac{\rho^2}{2d}\|D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &\leq \frac{d\rho^2}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &\quad + \frac{L_1^G\rho^2}{\mu_G d}\phi_v(v^{t,k}, x^{t,k}) \ .\end{aligned} \tag{33, 34}$$

Plugging Equation (34) into Equation (32) with $d = \frac{4L_1^G\Lambda_v)\rho}{\mu_G^2}$ gives

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq \left(1 - \frac{\rho\mu_G}{4}\right)\phi_v(v^{t,k}, x^{t,k}) &\\ + \left[\frac{L_1^G\rho}{\mu_G^2} + \frac{2L_1^G\Lambda_v^2\rho^3}{\mu_G^2}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 &\\ + \frac{\Lambda_v}{2}\|\Delta_\Pi^{t,k}\|^2 &\\ - \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .&\end{aligned} \tag{35}$$

Using once again (29), we get

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq &\left(1 - \frac{\rho\mu_G}{4}\right)\phi_v(v^{t,k}, x^{t,k}) \\ &+ \left[\frac{L_1^G\rho}{\mu_G^2} + \frac{2L_1^G\Lambda_v^2\rho^3}{\mu_G^2} + \frac{\Lambda_v\rho^2}{2}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{36}$$

By Lemma A.5, we have for any $\alpha > 0$

$$0 \leq -\alpha\left\langle \frac{1}{\rho}(v^{t,k+1} - v^{t,k}) + D_v^{t,k}, v^{t,k+1} - v^{t,k}\right\rangle \ .$$

By adding this to Equation (36), we get

$$\begin{aligned}\phi_v(v^{t,k+1}, x^{t,k+1}) \leq &\left(1 - \frac{\rho\mu_G}{4}\right)\phi_v(v^{t,k}, x^{t,k}) \\ &- \frac{\alpha}{\rho}\|v^{t,k+1} - v^{t,k}\|^2 - \alpha\langle D_v^{t,k}, v^{t,k+1} - v^{t,k}\rangle \\ &+ \left[\frac{L_1^G\rho}{\mu_G^2} + \frac{2L_1^G\Lambda_v^2\rho^3}{\mu_G^2} + \frac{\Lambda_v\rho^2}{2}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &- \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .\end{aligned} \tag{37}$$

We can control $-\left\langle D_v^{t,k}, v^{t,k+1} - v^{t,k}\right\rangle$ by Cauchy-Schwarz and Young for some $c, d, e, f > 0$

$$\begin{aligned}-\left\langle D_v^{t,k}, v^{t,k+1} - v^{t,k}\right\rangle = &-\left\langle D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}), \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}) - v^{t,k}\right\rangle \\ &- \left\langle D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}), \Delta_\Pi^{t,k}\right\rangle \\ &- \left\langle D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}), \Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x))) - v^{t,k}\right\rangle \\ &- \left\langle D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k}), \Delta_\Pi^{t,k}\right\rangle \\ \leq &\frac{c}{2}\|D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &+ \frac{1}{2c}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\|^2 \\ &+ \frac{d}{2}\|D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 + \frac{1}{2d}\|\Delta_\Pi^{t,k}\|^2 \\ &+ \frac{e}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &+ \frac{1}{2e}\|\Pi_\Gamma(v^{t,k} - \rho D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})) - v^{t,k}\|^2 \\ &+ \frac{f}{2}\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 + \frac{1}{2f}\|\Delta_\Pi^{t,k}\|^2 \\ \leq &\left(\frac{c+d}{2} + \rho^2\left(\frac{1}{2c} + \frac{1}{2e}\right)\right)\|D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ &+ \left(\frac{e+f}{2} + \rho^2\left(\frac{1}{2d} + \frac{1}{2f}\right)\right)\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\ \leq &\left(\frac{c+d}{2} + \rho^2\left(\frac{1}{2c} + \frac{1}{2e}\right)\right)\frac{2L_1^G}{\mu_G}\phi_v(v^{t,k}, x^{t,k}) \\ &+ \left(\frac{e+f}{2} + \rho^2\left(\frac{1}{2d} + \frac{1}{2f}\right)\right)\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2\end{aligned}$$

Let us take $c = d = e = f = \rho$. We get

$$-\left\langle D_v^{t,k}, v^{t,k+1} - v^{t,k}\right\rangle \leq \frac{4L_1^G}{\mu_G}\rho\phi_v(v^{t,k}, x^{t,k}) + 2\rho\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \ . \tag{38}$$

Then, by plugging the last Inequality in Equation (37) and setting $\alpha = \frac{\mu_G^2}{32L_1^G}$, we end up with

$$\begin{aligned}
\phi_v(v^{t,k+1}, x^{t,k+1}) &\leq \left(1 - \frac{\mu_G}{8}\rho\right)\phi_v(v^{t,k}, x^{t,k}) - \frac{\alpha}{\rho}\|v^{t,k+1} - v^{t,k}\|^2 \\
&\quad + \rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G} + \frac{\Lambda_v\rho}{2} + \frac{2L_1^G\Lambda_v^2\rho^2}{\mu_G^2}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\
&\quad - \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \\
&\leq \left(1 - \frac{\mu_G}{8}\rho\right)\phi_v(v^{t,k}, x^{t,k}) - \frac{\mu_G^2}{32L_1^G}\rho\|\mathcal{G}_v^{t,k}\|^2 \\
&\quad + \rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G} + \frac{\Lambda_v\rho}{2} + \frac{2L_1^G\Lambda_v^2\rho^2}{\mu_G^2}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\
&\quad - \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .
\end{aligned}$$

Since $\rho \leq B_v \triangleq \left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\min\left(\frac{2}{\Lambda_v}, \frac{\mu_G}{\sqrt{2L_1^G}\Lambda_v}\right)$ yields

$$\begin{aligned}
\phi_v(v^{t,k+1}, x^{t,k+1}) &\leq \left(1 - \frac{\mu_G}{8}\rho\right)\phi_v(v^{t,k}, x^{t,k}) - \frac{\mu_G^2}{32L_1^G}\rho\|\mathcal{G}_v^{t,k}\|^2 \\
&\quad + 3\rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\|D_v^{t,k} - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\
&\quad - \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \\
&\leq \left(1 - \frac{\mu_G}{8}\rho\right)\phi_v(v^{t,k}, x^{t,k}) - \frac{\mu_G^2}{32L_1^G}\rho\|\mathcal{G}_v^{t,k}\|^2 \\
&\quad + 6\rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\|D_v^{t,k} - D_v(\mathbf{u}^{t,k})\|^2 \\
&\quad + 6\rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\|D_v(\mathbf{u}^{t,k}) - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2 \\
&\quad - \gamma\langle D_x^{t,k}, \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\|D_x^{t,k}\|^2 \ .
\end{aligned} \tag{39}$$

Tacking the expectation conditionally to the past iterates yields

$$\begin{aligned}
\mathbb{E}_{t,k}[\phi_v(v^{t,k+1}, x^{t,k+1})] &\leq \left(1 - \frac{\mu_G}{8}\rho\right)\phi_v(v^{t,k}, x^{t,k}) - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\
&\quad + 6\rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\mathbb{E}_{t,k}[\|D_v^{t,k} - D_v(\mathbf{u}^{t,k})\|^2] \\
&\quad + 6\rho\left[\frac{L_1^G}{\mu_G^2} + \frac{\mu_G^2}{16L_1^G}\right]\mathbb{E}_{t,k}[\|D_v(\mathbf{u}^{t,k}) - D_v(z^*(x^{t,k}), v^{t,k}, x^{t,k})\|^2] \\
&\quad - \gamma\langle \mathbb{E}_{t,k}[D_x^{t,k}], \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}_{t,k}[\|D_x^{t,k}\|^2] \ .
\end{aligned} \tag{40}$$

From Young inequality, we have for any $c > 0$

$$\langle \mathbb{E}_{t,k}[D_x^{t,k}], \nabla_2\phi_v(v^{t,k}, x^{t,k})\rangle \leq c^{-1}\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2 + c\|\nabla_2\phi_v(v^{t,k}, x^{t,k})\|^2 \ . \tag{41}$$

Moreover, using the Lipschitz continuity of z^*, of $\nabla_{11}^2 G$ and ∇_F and the fact that v and v^* are bounded, we have

$$\|\nabla_2\phi_v(v, x)\| \leq \|\mathrm{d}z * (x)\|\left[\left\|\frac{1}{2}(\nabla_{111}^3 G(z^*(x), x)|v)v - \frac{1}{2}(\nabla_{111}^3 G(z^*(x), x)|v^*(x))v^*(x)\right\|\right.$$

$$
\begin{aligned}
&+\|\nabla^2_{11}F(z^*(x),x)v-\nabla^2_{11}F(z^*(x),x)v^*(x)\|\big]\\
&+\|\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v)v-\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v^*(x))v^*(x)\|\\
&+\|\nabla^2_{21}F(z^*(x),x)v-\nabla^2_{21}F(z^*(x),x)v^*(x)\|\\
\leq L_*&\left[\left\|\frac{1}{2}(\nabla^3_{111}G(z^*(x),x)|v-v^*(x))v-\frac{1}{2}(\nabla^3_{111}G(z^*(x),x)|v^*(x))(v-v^*(x))\right\|\right.\\
&\left.+L_2^F\|v-v^*(x)\|\right]\\
&+\left\|\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v-v^*(x))v-\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v^*(x))(v-v^*(x))\right\|\\
&+L_2^F\|v-v^*(x)\|\\
\leq L_*&\left[\left\|\frac{1}{2}(\nabla^3_{111}G(z^*(x),x)|v-v^*(x))v\right\|\right.\\
&\left.+\left\|\frac{1}{2}(\nabla^3_{111}G(z^*(x),x)|v^*(x))(v-v^*(x))\right\|+L_2^F\|v-v^*(x)\|\right]\\
&+\left\|\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v-v^*(x))v\right\|\\
&+\left\|\frac{1}{2}(\nabla^3_{211}G(z^*(x),x)|v^*(x))(v-v^*(x))\right\|+L_2^F\|v-v^*(x)\|\\
\leq L_*&\left[\frac{L_2^G}{2}(\|v\|+\|v^*(x)\|)\|v-v^*(x))\|+L_2^F\|v-v^*(x)\|\right]\\
&+\frac{L_2^G}{2}(\|v\|+\|v^*(x)\|)\|v-v^*(x))\|+L_2^F\|v-v^*(x)\|\\
\leq (L_*+1)&\left[L_2^GR+L_2^F\right]\|v-v^*(x)\|\ .
\end{aligned}
$$

On the other hand, we have by strong convexity

$$\|v-v^*(x)\|^2\leq\frac{2}{\mu_G}\phi_v(v,x)\ .$$

As a consequence, we have

$$\|\nabla_2\phi_v(v^{t,k},x^{t,k})\|^2\leq L''\phi_v(v^{t,k},x^{t,k}) \tag{42}$$

with $L''=\frac{2(L_*+1)^2\left[L_2^GR+L_2^F\right]^2}{\mu_G}$.

Plugging Inequalities (41) and (42) into (40) yields

$$
\begin{aligned}
\mathbb{E}_{t,k}[\phi_v(v^{t,k+1},x^{t,k+1})]\leq&\left(1-\frac{\mu_G}{8}\rho+cL''\gamma\right)\phi_v(v^{t,k},x^{t,k})-\frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2]\\
&+6\rho\left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\mathbb{E}_{t,k}[\|D_v^{t,k}-D_v(\mathbf{u}^{t,k})\|^2]\\
&+6\rho\left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\mathbb{E}_{t,k}[\|D_v(\mathbf{u}^{t,k})-D_v(z^*(x^{t,k}),v^{t,k},x^{t,k})\|^2]\\
&+\frac{\gamma}{c}\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2+\frac{\Lambda_v}{2}\gamma^2\mathbb{E}_{t,k}[\|D_x^{t,k}\|^2]\ .
\end{aligned}
$$

The Lipschitz continuity of $\nabla^2_{11}G$ and ∇_1F and the boundedness of v give us

$$
\begin{aligned}
\|D_v(\mathbf{u}^{t,k})-D_v(z^*(x^{t,k}),v^{t,k},x^{t,k})\|^2\leq&\big(\|\nabla^2_{11}G(z^{t,k},x^{t,k})v^{t,k}-\nabla^2_{11}G(z^*(x^{t,k}),x^{t,k})v^{t,k}\|\\
&+\|\nabla_1F(z^{t,k},x^{t,k})-\nabla_1F(z^*(x^{t,k}),x^{t,k})\|\big)^2\\
\leq&(L_2^GR+L_1^F)^2\|z^{t,k}-z^*(x^{t,k})\|^2\\
\leq&\frac{2(L_2^GR+L_1^F)^2}{\mu_G}\phi_z(z^{t,k},x^{t,k})\ .
\end{aligned}
$$

As a consequence

$$\begin{aligned}\mathbb{E}_{t,k}[\phi_v(v^{t,k+1},x^{t,k+1})] \leq \left(1-\frac{\mu_G}{8}\rho+cL''\gamma\right)\phi_v(v^{t,k},x^{t,k}) - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\ + 6\rho\left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\mathbb{E}_{t,k}[\|D_v^{t,k}-D_v(\mathbf{u}^{t,k})\|^2] \\ + 6\rho\left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{2(L_2^GR+L_1^F)^2}{\mu_G}\phi_z(z^{t,k},x^{t,k}) \\ + \frac{\gamma}{c}\|\mathbb{E}_{t,k}[D_x^{t,k}]\|^2 + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}_{t,k}[\|D_x^{t,k}\|^2]\ .\end{aligned} \tag{43}$$

From Lemma A.2, we have

$$\begin{aligned}\mathbb{E}[\|D_v^{t,k}-D_v(\mathbf{u}^{t,k})\|^2] \leq 4\rho^2\left((L_2^GR)^2+(L_1^F)^2\right)\sum_{r=1}^k\mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\rho^2(L_1^G)^2\sum_{r=1}^k\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 4\gamma^2\left((L_2^GR)^2+(L_1^F)^2\right)\sum_{r=1}^k\mathbb{E}[\|D_x^{t,r-1}\|^2]\end{aligned}$$

Taking the total expectation and plugging the previous inequality in Equation (43) yields

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1-\frac{\mu_G}{8}\rho+cL''\gamma\right)\phi_v^{t,k} - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\ + 24\rho^3\left((L_2^GR)^2+(L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + 24\rho^3(L_1^G)^2\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 24\rho\gamma^2\left((L_2^GR)^2+(L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|D_x^{t,r-1}\|^2] \\ + \left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{12(L_2^GR+L_1^F)^2}{\mu_G}\rho\phi_z^{t,k} \\ + \frac{\gamma}{c}\mathbb{E}[\|[\mathbb{E}_{t,k}D_x^{t,k}]\|^2] + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2]\ .\end{aligned}$$

Taking $c=\frac{\mu_G\rho}{16L''\gamma}$ yields

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1-\frac{\mu_G}{16}\rho\right)\phi_v^{t,k} - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\ + 24\rho^3\left((L_2^GR)^2+(L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + 24\rho^3(L_1^G)^2\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 24\rho\gamma^2\left((L_2^GR)^2+(L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^k\mathbb{E}[\|D_x^{t,r-1}\|^2] \\ + \left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{12(L_2^GR+L_1^F)^2}{\mu_G}\rho\phi_z^{t,k} \\ + \frac{16L''}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|[\mathbb{E}_{t,k}D_x^{t,k}]\|^2] + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2]\ .\end{aligned}$$

Combining Equation (21) and Lemma A.2 yields

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1-\frac{\mu_G}{16}\rho\right)\phi_v^{t,k} - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\ + 8\rho\left((L_2^G R)^2 + (L_1^F)^2\right)\left[3\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\rho^2 + \frac{8L''}{\mu_G}\gamma^2\right]\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + 8\rho(L_1^G)^2\left[3\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\rho^2 + \frac{8L''}{\mu_G}\gamma^2\right]\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 8\gamma^2\left((L_2^G R)^2 + (L_1^F)^2\right)\left[3\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\gamma + \frac{8L''}{\mu_G}\frac{\gamma^2}{\rho}\right]\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \\ + \left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{12(L_2^G R + L_1^F)^2}{\mu_G}\rho\phi_z^{t,k} \\ + \frac{16L''}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(\mathbf{u}^{tk,})\|^2] + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] \ .\end{aligned}$$

By assumption, $\gamma \leq C_v\rho$ with $C_v = \sqrt{\frac{\mu_G}{8L''}\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)}$, therefore

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1-\frac{\mu_G}{16}\rho\right)\phi_v^{t,k} - \frac{\mu_G^2}{32L_1^G}\rho\mathbb{E}_{t,k}[\|\mathcal{G}_v^{t,k}\|^2] \\ + 32\rho^3\left((L_2^G R)^2 + (L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] \\ + 32\rho^3(L_1^G)^2\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ + 32\gamma^2\rho\left((L_2^G R)^2 + (L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \\ + \left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{12(L_2^G R + L_1^F)^2}{\mu_G}\rho\phi_z^{t,k} \\ + \frac{16L''}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(\mathbf{u}^{tk,})\|^2] + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] \ .\end{aligned}$$

We get finally

$$\begin{aligned}\phi_v^{t,k+1} \leq \left(1-\frac{\rho\mu_G}{16}\right)\phi_v^{t,k} - \tilde{\beta}_{vv}\rho V_v^t + \rho^3\beta_{vz}\mathcal{V}_z^{t,k} + 2\rho^3\beta_{vv}\mathcal{V}_v^{t,k} + \gamma^2\rho\beta_{vx}\mathcal{V}_x^{t,k} \\ + \rho\alpha_{vz}\phi_z^{t,k} + \frac{\Lambda_v}{2}\gamma^2\mathbb{E}[\|D_x^{t,k}\|^2] + \frac{\gamma^2}{\rho}\overline{\beta}_{vx}\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2]\end{aligned}$$

with $\beta_{vz} = \beta_{vx} = 32\left((L_2^G R)^2 + (L_1^F)^2\right)\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)$, $\beta_{vv} = (L_1^G)^2\left(\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right)$, $\overline{\beta}_{vx} = \frac{16L''}{\mu_G}$, $\tilde{\beta}_{vv} = \frac{\mu_G^2}{32L_1^G}$ and $\alpha_{vz} = \left[\frac{L_1^G}{\mu_G^2}+\frac{\mu_G^2}{16L_1^G}\right]\frac{12(L_2^G R+L_1^F)^2}{\mu_G}$. □

A.7 Proof of Lemma 3.5

Proof. The smoothness of h (Proposition 2.3) gives us

$$h(x^{t,k+1}) \leq h(x^{t,k}) - \gamma\langle\nabla h(x^{t,k}), D_x^{t,k}\rangle + \gamma^2\frac{L^h}{2}\|D_x^{t,k}\|^2 \ .$$

Then, we use the identity $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$ to get

$$\begin{aligned} h(x^{t,k+1}) &\leq h(x^{t,k}) - \frac{\gamma}{2}\|\nabla h(x^{t,k})\|^2 - \frac{\gamma}{2}\|D_x^{t,k}\|^2 + \frac{\gamma}{2}\|\nabla h(x^{t,k}) - D_x^{t,k}\|^2 + \gamma^2\frac{L^h}{2}\|D_x^{t,k}\|^2 \\ &\leq h(x^{t,k}) - \frac{\gamma}{2}\|\nabla h(x^{t,k})\|^2 - \frac{\gamma}{2}\|D_x^{t,k}\|^2 + \gamma\|\nabla h(x^{t,k}) - D_x(\mathbf{u}^{t,k})\|^2 \\ &\quad + \gamma\|D_x(\mathbf{u}^{t,k}) - D_x^{t,k}\|^2 + \gamma^2\frac{L^h}{2}\|D_x^{t,k}\|^2 \ . \end{aligned}$$

Then taking the expectation gives and using Proposition 2.5 yields

$$\begin{aligned} h^{t,k+1} &\leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma\mathbb{E}[\|\nabla h(x^{t,k}) - D_x(\mathbf{u}^{t,k})\|^2] \\ &\quad + \gamma\mathbb{E}[\|D_x(\mathbf{u}^{t,k}) - D_x^{t,k}\|^2] - \frac{\gamma}{2}\left(1 - L^h\gamma\right)\mathbb{E}[\|D_x^{t,k}\|^2] \\ &\leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma L_x^2(\mathbb{E}[\|z^{t,k} - z^*(x^{t,k})\|^2] + \mathbb{E}[\|v^{t,k} - v^*(x^{t,k})\|^2]) \\ &\quad + \gamma\mathbb{E}[\|D_x(\mathbf{u}^{t,k}) - D_x^{t,k}\|^2] - \frac{\gamma}{2}\left(1 - L^h\gamma\right)\mathbb{E}[\|D_x^{t,k}\|^2] \ . \end{aligned}$$

The μ_G-strong convexity of $G(\,.\,, x)$ ensures that $\|z - z^*(x)\|^2 \leq \frac{2}{\mu_G}\phi_z(z,x)$ and $\|v - v^*(x)\|^2 \leq \frac{2}{\mu_G}\phi_v(v,x)$. As a consequence

$$\begin{aligned} h^{t,k+1} &\leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma\frac{2L_x^2}{\mu_G}(\phi_z^{t,k} + \phi_v^{t,k}) + \gamma\mathbb{E}[\|D_x(z^{t,k}, v^{t,k}, x^{t,k}) - D_x^{t,k}\|^2] \\ &\quad - \frac{\gamma}{2}\left(1 - L^h\gamma\right)\mathbb{E}[\|D_x^{t,k}\|^2] \ . \end{aligned}$$

From Lemma A.2, we have

$$\begin{aligned} \mathbb{E}[\|D_x^{t,k} - D_x(\mathbf{u}^{t,k})\|^2] &\leq 4\rho^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\rho^2(L_1^G)^2\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ &\quad + 4\gamma^2\left((L_2^G R)^2 + (L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] \ . \end{aligned}$$

As a consequence

$$\begin{aligned} h^{t,k+1} &\leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma\frac{2L_x^2}{\mu_G}(\phi_z^{t,k} + \phi_v^{t,k}) \\ &\quad + 4\gamma\rho^2\left((L_2^G R)^2 + 2(L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_z^{t,r-1}\|^2] + 4\gamma\rho^2(L_1^G)^2\sum_{r=1}^{k}\mathbb{E}[\|\mathcal{G}_v^{t,r-1}\|^2] \\ &\quad + 4\gamma^3\left((L_2^G R)^2 + 2(L_1^F)^2\right)\sum_{r=1}^{k}\mathbb{E}[\|D_x^{t,r-1}\|^2] - \frac{\gamma}{2}\left(1 - L^h\gamma\right)\mathbb{E}[\|D_x^{t,k}\|^2] \\ &\leq h^{t,k} - \frac{\gamma}{2}g^{t,k} + \gamma\frac{2L_x^2}{\mu_G}(\phi_z^{t,k} + \phi_v^{t,k}) + \gamma\rho^2\beta_{hz}\mathcal{V}_z^{t,k} + \gamma\rho^2\beta_{hv}\mathcal{V}_v^{t,k} \\ &\quad + \gamma^3\beta_{hx}\mathcal{V}_x^{t,k} - \frac{\gamma}{2}\left(1 - L^h\gamma\right)\mathbb{E}[\|D_x^{t,k}\|^2] \end{aligned}$$

with $\beta_{hz} = 4\left((L_2^G R)^2 + 2(L_1^F)^2\right)$, $\beta_{hv} = 4(L_1^G)^2$ and $\beta_{hx} = 4\left((L_2^G R)^2 + 2(L_1^F)^2\right)$. □

A.8 Proof of Theorem 1 and Corollary 3.6

The constants involved in Theorem 1 are

$$\psi_z = \frac{1}{16\overline{\beta}_{zx}}, \quad \psi_v = \min\left[\frac{1}{16\overline{\beta}_{vx}}, \frac{\alpha_{zv}\mu_G}{12}\psi_z\right]$$

$$\overline{\rho} = \min\left[\sqrt{\frac{\psi_z}{12q(\psi_z\beta_{zz}+\psi_v\beta_{zv})}}, \sqrt{\frac{1}{6\Lambda_z}}, \sqrt{\frac{1}{12q\beta_{vv}}}, sqrt\frac{\tilde{\beta}_{vv}}{3\Lambda_v}, B_v\right] \ ,$$

$$\xi = \min\left[C_z, C_v, 1, \frac{\psi_v\mu_G^2}{16L_x^2}, \sqrt{\frac{\mu_G}{8\overline{\beta}_{vx}}}, \frac{\psi_z\mu_G^2}{24L_x^2}, \sqrt{\frac{\mu_G}{12\overline{\beta}_{zx}}}\right] \ ,$$

$$\overline{\gamma} = \min\left[\sqrt{\frac{1}{12q(\psi_z\beta_{zx}+\psi_v\beta_{vx})}}, \sqrt{\frac{1}{12q\beta_{hx}}}, \frac{1}{6(L^h+\psi_z\Lambda_z+\psi_v\Lambda_v)}, \sqrt{\frac{\psi_v\tilde{\beta}_{vv}}{6q(\beta_{hv}+\psi_z\beta_{vz})}}, \sqrt{\frac{\psi_z}{12q\beta_{hz}}}\right] \ .$$

Proof. The proof is a classical Lyapunov analysis. Consider the following Lyapunov function $\mathcal{L}^{t,k} = h^{t,k} + \psi_z\phi_z^{t,k} + \psi_v\phi_v^{t,k}$ for some positive constants ψ_z and ψ_v. We use use Lemmas 3.3 to 3.5 to upper bound $\mathcal{L}^{t,k} - \mathcal{L}^{t,k+1}$.

We have

$$\begin{aligned}
\mathcal{L}^{t,k+1} - \mathcal{L}^{t,k} \leq & -\frac{\gamma}{2}g^{t,k} + (\psi_z\overline{\beta}_{zx} + \psi_v\overline{\beta}_{vx})\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2] \\
& + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_z\frac{\mu_G}{2}\rho + \psi_v\alpha_{zv}\rho\right)\phi_z^{t,k} + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_v\frac{\mu_G}{16}\rho\right)\phi_v^{t,k} \\
& + \left(\psi_z\frac{\Lambda_z}{2}\rho^2 - \psi_z\frac{1}{2}\rho\right)V_z^{t,k} - \psi_v\tilde{\beta}_{vv}\rho V_v^{t,k} \\
& + \left(\frac{L^h}{2}\gamma^2 + \psi_z\frac{\Lambda_z}{2}\gamma^2 + \psi_v\frac{\Lambda_v}{2}\gamma^2 - \frac{\gamma}{2}\right)V_x^{t,k} \\
& + \left(\beta_{hz}\rho\gamma^2 + \psi_z\beta_{zz}\rho^3 + \psi_v\beta_{zv}\rho^3\right)\mathcal{V}_z^{t,k} \\
& + \left(\beta_{hv}\rho\gamma^2 + \psi_z\beta_{vz}\gamma^2\rho + \psi_v\beta_{vv}\rho^3\right)\mathcal{V}_v^{t,k} \\
& + \left(\beta_{hx}\gamma^3 + \psi_z\beta_{zx}\gamma^2\rho + \psi_v\beta_{vx}\rho^3\right)\mathcal{V}_x^{t,k} \ .
\end{aligned} \tag{44}$$

We bound $\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2]$ crudely by using Proposition 2.5

$$\begin{aligned}
\mathbb{E}[\|D_x(\mathbf{u}^{t,k})\|^2] &\leq 2\mathbb{E}[\|\nabla h(x^{t,k})\|^2] + 2\mathbb{E}[\|D_x(\mathbf{u}^{t,k}) - \nabla h(x^{t,k})\|^2] \\
&\leq 2g^{t,k} + 2(\mathbb{E}[\|z^{t,k} - z^*(x^{t,k})\|^2] + \mathbb{E}[\|v^{t,k} - v^*(x^{t,k})\|^2]) \\
&\leq 2g^{t,k} + \frac{4}{\mu_G}(\phi_z^{t,k} + \phi_v^{t,k}) \ .
\end{aligned}$$

Summing in (44) for $k = 0, \dots, q-1$ yields

$$\begin{aligned}
\mathcal{L}^{t,q} - \mathcal{L}^{t,0} \leq & -\left(\frac{\gamma}{2} - 2\psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} - 2\psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1} g^{t,k} \\
& + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_z\frac{\mu_G}{2}\rho + \psi_v\alpha_{zv}\rho + \psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1}\phi_z^{t,k} \\
& + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_v\frac{\mu_G}{16}\rho + \psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1}\phi_v^{t,k} - \psi_z\tilde{\beta}_{vv}\rho\sum_{k=0}^{q-1}V_z^{t,k} \\
& + \left(\psi_v\frac{\Lambda_v}{2}\rho^2 - \psi_v\frac{1}{2}\rho\right)\sum_{k=0}^{q-1}V_v^{t,k} + \left(\frac{L^h}{2}\gamma^2 + \psi_z\frac{\Lambda_z}{2}\gamma^2 + \psi_v\frac{\Lambda_v}{2}\gamma^2 - \frac{\gamma}{2}\right)\sum_{k=0}^{q-1}V_x^{t,k} \\
& + \left(\beta_{hz}\rho\gamma^2 + \psi_z\beta_{zz}\rho^3 + \psi_v\beta_{zv}\rho^3\right)\sum_{k=0}^{q-1}\mathcal{V}_z^{t,k} \\
& + \left(\beta_{hv}\rho\gamma^2 + \psi_z\beta_{vz}\gamma^2\rho + \psi_v\beta_{vv}\rho^3\right)\sum_{k=0}^{q-1}\mathcal{V}_v^{t,k} \\
& + \left(\beta_{hx}\gamma^3 + \psi_z\beta_{zx}\gamma^2\rho + \psi_v\beta_{vx}\rho^3\right)\sum_{k=0}^{q-1}\mathcal{V}_x^{t,k} \ .
\end{aligned} \tag{45}$$

Since we have

$$\begin{aligned}
\sum_{k=0}^{q-1}\mathcal{V}_\bullet^{t,k} &= \sum_{k=0}^{q-1}\sum_{r=1}^{k}\mathbb{E}[\|D_\bullet^{t,r-1}\|^2] = \sum_{r=1}^{q-1}\sum_{k=r}^{q-1}\mathbb{E}[\|D_\bullet^{t,r-1}\|^2] \\
&= \sum_{r=1}^{q-1}(q-r)\mathbb{E}[\|D_\bullet^{t,r-1}\|^2] \leq q\sum_{k=1}^{q-1}\mathbb{E}[\|D_\bullet^{t,k-1}\|^2]
\end{aligned}$$

we get

$$\begin{aligned}
\mathcal{L}^{t,q} - \mathcal{L}^{t,0} \leq & -\left(\frac{\gamma}{2} - 2\psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} - 2\psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1} g^{t,k} \\
& + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_z\frac{\mu_G}{2}\rho + \psi_v\alpha_{zv}\rho + \psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1}\phi_z^{t,k} \\
& + \left(\frac{2L_x^2}{\mu_G}\gamma - \psi_v\frac{\mu_G}{2}\rho + \psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1}\phi_v^{t,k} \\
& + \left(\psi_z\frac{\Lambda_z}{2}\rho^2 - \psi_z\frac{1}{2}\rho + q\left(\beta_{hz}\rho\gamma^2 + \psi_z\beta_{zz}\rho^3 + \psi_v\beta_{zv}\rho^3\right)\right)\sum_{k=0}^{q-1}V_z^{t,k} \\
& + \left(\psi_v\frac{\Lambda_v}{2}\rho^2 - \psi_v\tilde{\beta}_{vv}\rho + q\left(\beta_{hv}\rho\gamma^2 + \psi_z\beta_{vz}\gamma^2\rho + \psi_v\beta_{vv}\rho^3\right)\right)\sum_{k=0}^{q-1}V_v^{t,k} \\
& + \left(\frac{L^h}{2}\gamma^2 + \psi_z\frac{\Lambda_z}{2}\gamma^2 + \psi_v\frac{\Lambda_v}{2}\gamma^2 - \frac{\gamma}{2} + q\left(\beta_{hx}\gamma^3 + \psi_z\beta_{zx}\gamma^2\rho + \psi_v\beta_{vx}\rho\gamma^2\right)\right)\sum_{k=0}^{q-1}V_x^{t,k} \ .
\end{aligned} \tag{46}$$

Since $\rho \leq \overline{\rho} \leq \min\left[\sqrt{\frac{\psi_z}{12q(\psi_z\beta_{zz}+\psi_v\beta_{zv})}}, \sqrt{\frac{1}{6\Lambda_z}}\right]$ and $\gamma \leq \overline{\gamma} \leq \sqrt{\frac{\psi_z}{12q\beta_{hz}}}$, we have

$$\psi_z\frac{\Lambda_z}{2}\rho^2 - \psi_z\frac{1}{2}\rho + q\left(\beta_{hz}\rho\gamma^2 + \psi_z\beta_{zz}\rho^3 + \psi_v\beta_{zv}\rho^3\right) < 0 \ . \tag{47}$$

Moreover, the conditions $\rho \leq \overline{\rho} \leq \min\left[\sqrt{\frac{\tilde{\beta}_{vv}}{6q\beta_{vv}}}, \sqrt{\frac{\tilde{\beta}_{vv}}{3\Lambda_v}}\right]$ and $\gamma \leq \overline{\gamma} \leq \sqrt{\frac{\psi_v \tilde{\beta}_{vv}}{6q(\beta_{hv}+\psi_z\beta_{vz})}}$, ensure that

$$\psi_v \frac{\Lambda_v}{2}\rho^2 - \psi_v\tilde{\beta}_{vv}\rho + q\left(\beta_{hv}\rho\gamma^2 + \psi_z\beta_{vz}\gamma^2\rho + \psi_v\beta_{vv}\rho^3\right) < 0 \ . \tag{48}$$

The conditions $\rho \leq \overline{\rho} \leq \sqrt{\frac{1}{12q(\psi_z\beta_{zx}+\psi_v\beta_{vx})}}$ and $\gamma \leq \overline{\gamma} \leq \min\left[\sqrt{\frac{1}{12q(\psi_z\beta_{zx}+\psi_v\beta_{vx})}}, \sqrt{\frac{1}{12q\beta_{hx}}}, \frac{1}{6(L^h+\psi_z\Lambda_z+\psi_v\Lambda_v)}\right]$ yield

$$\frac{L^h}{2}\gamma^2 + \psi_z\frac{\Lambda_z}{2}\gamma^2 + \psi_v\frac{\Lambda_v}{2}\gamma^2 - \frac{\gamma}{2} + q\left(\beta_{hx}\gamma^3 + \psi_z\beta_{zx}\gamma^2\rho + \psi_v\beta_{vx}\rho\gamma^2\right) < 0 \ . \tag{49}$$

The condition $\gamma \leq \xi\rho \leq \min\left[\frac{\psi_v\mu_G^2}{16L_x^2}, \sqrt{\frac{\mu_G}{8\overline{\beta}_{vx}}}\right]\rho$ ensures

$$\frac{2L_x^2}{\mu_G}\gamma - \psi_v\frac{\mu_G}{2}\rho + \psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho} \leq 0 \tag{50}$$

By definition, we have $\psi_v \leq \frac{\alpha_{zv}\mu_G}{12}\psi_z$ and by assumptions $\gamma \leq \xi\rho \leq \min\left[\frac{\psi_z\mu_G^2}{24L_x^2}, \sqrt{\frac{\mu_G}{12\overline{\beta}_{zx}}}\right]\rho$. As a consequence

$$\frac{2L_x^2}{\mu_G}\gamma - \psi_z\frac{\mu_G}{2}\rho + \psi_v\alpha_{zv}\rho + \psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} < 0 \ . \tag{51}$$

Plugging Inequalities (47) to (51) into Equation (46) gives

$$\mathcal{L}^{t,q} - \mathcal{L}^{t,0} \leq -\left(\frac{\gamma}{2} - 2\psi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} - 2\psi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\sum_{k=0}^{q-1} g^{t,k} \ .$$

Since $\psi_z = \frac{\rho}{16\overline{\beta}_{zx}}$ and $\psi_v \leq \frac{\rho}{16\overline{\beta}_{vx}}$ and $\frac{\gamma^2}{\rho} \leq \xi \leq 1$, we get

$$\underbrace{\mathcal{L}^{t,q} - \mathcal{L}^{t,0}}_{\mathcal{L}^{t+1,0}-\mathcal{L}^{t,0}} \leq -\frac{\gamma}{4}\sum_{k=0}^{q-1} g^{t,k} \ .$$

Summing, telescoping and dividing by Tq gives

$$\frac{1}{Tq}\sum_{t=0}^{T-1}\sum_{k=0}^{q-1} g^{t,k} \leq \frac{4}{Tq\gamma}\underbrace{\left(h^{0,0} - h^* + \psi_z\phi^{0,0} + \psi_v\phi^{0,0}]\right)}_{\Gamma^0} \ .$$

□

From Theorem 1 we deduce Corollary 3.6.

Proof. Let us take $\rho = \overline{\rho}(n+m)^{-\frac{1}{2}}$, $\gamma = \min(\xi\rho, \overline{\gamma})$ and $q = n+m$. Then Theorem 1 holds:

$$\frac{1}{Tq}\sum_{t=0}^{T-1}\sum_{k=0}^{q-1} g^{t,k} \leq \frac{4}{Tq\gamma}\Gamma^0 \ .$$

with $\Gamma_0 = \mathcal{O}(1)$. To get an ε-stationary solution, we set $T \geq \frac{4}{q\gamma}\Gamma^0\varepsilon^{-1} \vee 1 = \mathcal{O}\left(\frac{1}{q\gamma\varepsilon} \vee 1\right)$. One iteration has $\Theta(q) = \Theta(n+m)$ oracle complexity. As a consequence, the sample complexity to get an ε-stationary point is $\mathcal{O}\left((n+m)^{\frac{1}{2}}\epsilon^{-1} \vee (n+m)\right)$. □

B Lower bound for bilevel problems (proof of Theorem 2)

The proof of Theorem 2 is an adaptation of the proof of (Zhou and Gu, 2019, Theorem 4.7) from single-level to bilevel problems. We build the outer function from the worst-case instance of (Zhou and Gu, 2019, Theorem 4.7) and we add a bilevel component by using as inner function the function G defined by $G(z,x) = \frac{\mu_G}{2}\|z-x\|^2$. We start by introducing the different tools used in this proof.

B.1 Preliminary results

In what follows, we provide the building blocks of our worst-case instance. The proof uses the following quadratic function presented by (Nesterov, 2018).

Definition B.1. Let $d \in \mathbb{N}_{>0}$, $\xi \in [0\,,\,+\infty)$ and $\zeta \leq 1$. We define $\mathbf{Q}(.;\xi,d) : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathbf{Q}(x;\xi,d) = \frac{\xi}{2}(x_1-1)^2 + \frac{1}{2}\sum_{k=1}^{d-1}(x_{k+1}-x_k)^2 \ .$$

Proposition B.2 proposition comes directly from (Zhou and Gu, 2019, Proposition 3.5). The first part of the proposition gives us the regularity of $\mathbf{Q}$. In the second part shows that a function defined as $\mathbf{Q}(U\times\cdot;\xi,d) + \sum_{p=1}^q g(\langle u_p,\cdot\rangle)$ verifies the so-called "zero-chain property" Carmon et al. (2020): if $Ux \in \mathrm{Span}(u_1,\dots,u_k)$, then we gain a non zero coordinate by calling the gradient $\nabla\left[\mathbf{Q}\left(U\times\cdot;\xi,d\right) + \sum_{p=1}^q g(\langle u_p,\cdot\rangle)\right](x)$. In other words, that makes us progress in the problem resolution.

Proposition B.2. *For $d \in \mathbb{N}_{>0}$, $\xi \in [0\,,\,+\infty)$ and $\zeta \leq 1$. The following holds:*

1. *$\mathbf{Q}(\,\cdot\,;\xi,d)$ is convex and 4-smooth.*
2. *Let $q \in \mathbb{N}_{>0}$, $U = [u_1,\dots,u_d]^\top \in \mathbb{R}^{d\times q}$ such that $UU^\top = I$ and for $k \leq d$, $U^{(\leq k)} = [u_1,\dots,u_k,0,\dots,0]^\top \in \mathbb{R}^{d\times q}$. Let $g:\mathbb{R}\to\mathbb{R}$ differentiable such that $g'(0)=0$. Then for any $x\in\mathbb{R}^q$ such that $Ux = U^{(\leq k)}x$, then*

$$\nabla\left[\mathbf{Q}(U\times\cdot;\xi,d) + \sum_{p=1}^d g(\langle u_p,\cdot\rangle)\right](x) \in \mathrm{Span}(u_1,\dots,u_k,u_{k+1}) \ .$$

Proof. Let $x\in\mathbb{R}^q$ such that $Ux = U^{(\leq k)}x$. For $0 \geq k \leq d$, we denote

$$\mathbb{R}^{k,d} = \{v\in\mathbb{R}^d, v_{k+1} = \cdots = v_d = 0\} \ .$$

Let us write $\mathbf{Q}(x;\xi,d) = \frac{1}{2}x^\top Ax + b^\top x + c$ with

$$A = \begin{bmatrix} 1+\xi & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d\times d} \ ,$$

$b = \xi(1,0,\dots,0)^\top$ and $c = \frac{\xi}{2}(1,0,\dots,0)^\top$.

On the one hand it is known from (Nesterov, 2018, Lemma 2.5.1) that if $v\in\mathbb{R}^{k,d}$,

$$\nabla\mathbf{Q}(v;\xi,d) \in \mathbb{R}^{k+1,d}$$

As a consequence,

$$\nabla\mathbf{Q}(Ux;\xi,d) = \nabla\mathbf{Q}(\underbrace{U^{(\leq k)}x}_{\in\mathbb{R}^{k,d}};\xi,d) \in \mathbb{R}^{k+1,d}$$

and

$$\nabla[\mathbf{Q}(U\times\cdot;\xi,d)](x) = U^\top\nabla\mathbf{Q}(Ux;\xi,d) \in \mathrm{Span}(u_1,\dots,u_{k+1}) \ .$$

On the other hand,

$$\nabla\left[\sum_{p=1}^{d} g(\langle u_p, \cdot\rangle)\right](x) = \sum_{p=1}^{d} g'(\langle u_p, x\rangle)u_p = \sum_{p=1}^{k} g'(\langle u_p, x\rangle)u_p \in \mathrm{Span}(u_1, \dots, u_{k+1}) \ .$$

Thus

$$\nabla\left[\mathbf{Q}(U\times\cdot\,;\xi,d) + \sum_{p=1}^{d} g(\langle u_p, \cdot\rangle)\right](x) \in \mathrm{Span}(u_1, \dots, u_k, u_{k+1}) \ .$$

□

However, the function $\mathbf{Q}$ is convex. That is why we also use the function Γ introduced in Carmon et al. (2021). As explained in Carmon et al. (2021), this function is essential to lower bound the gradient of our worst wase instance.

Definition B.3. Let $d \in \mathbb{N}_{>0}$. We define $\Gamma(\cdot\,; d) : \mathbb{R}^{d+1} \to \mathbb{R}$ by

$$\Gamma(x; d) = 120 \sum_{k=1}^{d} \int_1^{x_k} \frac{t^2(t-1)}{1+t^2}\mathrm{d}t \ .$$

An important property of Γ shown in (Carmon et al., 2021, Lemma 2) is the smoothness of the function Γ.

Proposition B.4. *There exists a constant $c > 0$ such that $\Gamma(\cdot\,; d)$ is c-smooth.*

Now we introduce the function f_{nc} we use to build our worst-case instance. This function comes from (Zhou and Gu, 2019, Definition 3.5). It is the sum of the quadratic function defined by B.1 and the nonconvex component given by Definition B.3.

Definition B.5. For $\alpha > 0$ and $d \in \mathbb{N}_{>0}$, $f_{\mathrm{nc}}(\cdot\,; \alpha, d) : \mathbb{R}^{d+1} \to \mathbb{R}$ is defined a

$$f_{\mathrm{nc}}(x; \alpha, d) = \mathbf{Q}(x; \sqrt{\alpha}, d+1) + \alpha\Gamma(x) \ .$$

The essential properties of f_{nc} come from (Carmon et al., 2021, Lemmas 2, 3, 4). The first part provides the regularity properties of f_{nc}. The second part bounds the distance between $f_{\mathrm{nc}}(\cdot\,; \alpha, d)$ and the optimal value of the function. The third part will be key to the overall proof. In words, it states that as long $x \in \mathbb{R}^{d+1}$ has its two last components equal to zero, the norm of the gradient of $f_{\mathrm{nc}}(\cdot\,; \alpha, d)$ is higher than a constant controlled by α. As a consequence, if α is properly chosen, as soon as $x_d = x_{d+1} = 0$, we are ensured that $\|\nabla f_{\mathrm{nc}}(x; \alpha, d)\| \geq \epsilon$.

Proposition B.6. *For $\alpha \in [0\,,\,1]$, it holds*

1. $-\alpha c \preceq \nabla^2 f_{\mathrm{nc}} \preceq 4 + \alpha c$.
2. $f_{\mathrm{nc}}(0; \alpha, d) - \inf_x f_{\mathrm{nc}}(x; \alpha, d) \leq \frac{\sqrt{\alpha}}{2} + 10\alpha d$.
3. *For $x \in \mathbb{R}^{d+1}$ such that $x_d = x_{d+1} = 0$, $\|\nabla f_{\mathrm{nc}}(x; \alpha, d)\| \geq \frac{\alpha^{\frac{3}{4}}}{4}$.*

From now we denote

$$\mathcal{O}(a, b) = \{U \in \mathbb{R}^{a\times b}, UU^\top = I_a\} \ .$$

The following Lemma adapted from Zhou and Gu (2019) is fundamental for our lower bound proof.

Lemma B.7. *Let $d, m \in \mathbb{N}_{>0}$ and $U \in \mathcal{O}((d+1)m, (d+1)m)$. We denote $U = \begin{bmatrix} U^{(1)} \\ \vdots \\ U^{(m)} \end{bmatrix}$ with $U^{(i)} \in \mathcal{O}(d+1, (d+1)m)$. Let $\{h_j\}_{j\in[m]}$ with $h_j(x) = f_{\mathrm{nc}}(U^{(j)}x; \alpha, d)$ and $h = \frac{1}{m}\sum_{j=1}^m h_j$. Let $x \in \mathbb{R}^{(d+1)m}$ and $y^{(j)} = U^{(j)}x \in \mathbb{R}^{d+1}$. Let $\mathcal{I} = \{j \in [m], y_d^{(i)} = y_{d+1}^{(i)} = 0\}$. Then it holds*

$$\|\nabla h(x)\|^2 \geq \frac{\alpha^{\frac{3}{2}}|\mathcal{I}|}{16m^2} \ .$$

Proof. We have

$$\begin{aligned}
\|\nabla h(x)\|^2 &= \left\|\frac{1}{m}\sum_{j=1}^{d}\nabla h_j(x)\right\|^2 \\
&= \left\|\frac{1}{m}\sum_{j=1}^{m}(U^{(j)})^\top \nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)\right\|^2 \\
&= \frac{1}{m^2}\left\|\sum_{j=1}^{m}(U^{(j)})^\top \nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)\right\|^2 \\
&\quad + \frac{2}{m^2}\sum_{\substack{j,l=1\\ j\neq l}}^{m}\nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)^\top U^{(l)}(U^{(j)})^\top \nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d) \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\left\|(U^{(j)})^\top \nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)\right\|^2
\end{aligned}$$

where the last equality comes from the fact that for $j \neq l$, $U^{(l)}(U^{(j)})^\top = 0$ since $U \in \mathcal{O}((d+1)m,(d+1)m)$. Now, using the third part of Proposition B.6, we get

$$\begin{aligned}
\|\nabla h(x)\|^2 &\geq \frac{1}{m^2}\sum_{j\in\mathcal{I}}\left\|(U^{(j)})^\top \nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)\right\|^2 \\
&\geq \frac{1}{m^2}\sum_{j\in\mathcal{I}}\left\|\nabla f_{\mathrm{nc}}(U^{(j)}x;\alpha,d)\right\|^2 \\
&\geq \frac{\alpha^{\frac{3}{2}}|\mathcal{I}|}{16m^2} \ .
\end{aligned}$$

□

B.2 Main proof

Now we are ready to prove Theorem 2.

Proof. We consider $U \in \mathcal{O}((T+1)m,(T+1)m)$ and we denote

$$U = \begin{bmatrix} U^{(1)} \\ \vdots \\ U^{(m)} \end{bmatrix}$$

with $U^{(j)} = (u_1^{(j)},\dots,u_{T+1}^{(j)})^\top \in \mathcal{O}(T+1,(T+1)m)$.

For $j \in [m]$, we choose $\overline{F}_j : \mathbb{R}^{(T+1)m+(T+1)m} \to \mathbb{R}$ defined by

$$\overline{F}_j(z,x) = f_{\mathrm{nc}}(U^{(j)}z;\alpha,T)$$

and we set $\overline{F} = \frac{1}{m}\sum_{j=1}^{m}\overline{F}_j$. We also define for $i \in [n]$ $\overline{G}_i(z,x) = \frac{1}{2}\|z-x\|^2$, $\overline{G} = \frac{1}{n}\sum_{i=1}^{n}\overline{G}_i$, $\overline{z}^*(x) = \arg\min_z \overline{G}(z,x)$ and $\overline{h}(x) = \overline{F}(\overline{z}^*(x),x) = f_{\mathrm{nc}}(U^{(j)}x;\alpha,T)$. By Proposition B.6, $\overline{F}_j$ is $4+\frac{\alpha c}{m}$ smooth, and $\overline{G}_i$ is 1-smooth and 1-strongly convex.

We have

$$\overline{h}(0) - \inf_x \overline{h}(x) \leq \sqrt{\alpha} + 10\alpha T \ .$$

We finally consider $F_j(z,x) = \lambda_F \overline{F}_j(z/\beta, x/\beta)$, $G_i(z,x) = \lambda_G \overline{G}_i(z/\beta, x/\beta)$. As a consequence, we have $z^*(x) = \arg\min G = \overline{z}^*(x)$ and $h(x) = F(z^*(x), x)$. We also consider a *fixed* indices sequence (i_t, j_t). We set

$$\alpha = \min\left\{1, \frac{m}{c}\right\}, \quad \lambda_F = \frac{160 m\epsilon}{L_1^F \alpha^{3/2}}, \quad \beta = \sqrt{5\lambda_F / L_1^F}, \quad \lambda_G = \beta^2 \mu_G, \quad T = \frac{\Delta L_1^F}{1760 m\epsilon}\sqrt{\alpha} \ .$$

We can check that each F_j is L_1^F-smooth, and each G_i is μ_G-strongly convex. Assuming $\epsilon \leq \Delta L_1^F \alpha/(1760m)$ ensures that $h(0) - \inf_x h(x) \leq \Delta$ (we can check that $h(0) = \lambda_F \overline{h}(0)$ and $\inf h = \lambda_F \inf \overline{h}$).

Let us assume without loss of generality that the algorithm at initialization we have $z^0 = v^0 = x^0 = 0$ and consider (z^t, v^t, x^t) the output of an algorithm with the known sequence (i_t, j_t).

Given our inner function and the fact that $\nabla_2 F(z,x) = 0$ for any $(z,x) \in \mathbb{R}^{(m+1)d + (m+1)d}$, we have

$$z^{t+1} \in \mathrm{Span}(z^0 - x^0, \dots, z^t - x^t) \tag{52}$$
$$v^{t+1} \in \mathrm{Span}(v^0 + \nabla_1 F_{j_0}(z^0, x^0), \dots, v^t + \nabla_1 F_{j_t}(z^t, x^t)) \tag{53}$$
$$x^{t+1} \in \mathrm{Span}(v^0, \dots, v^t) \ . \tag{54}$$

Since $v^0 = 0$, we have by Equation (53) $v^1 \in \mathrm{Span}(\nabla_1 F_{j_0}(z^0, x^0))$ and by induction

$$v^{t+1} \ \in \ \mathrm{Span}(\nabla_1 F_{j_0}(z^0, x^0), \dots, \nabla_1 F_{j_t}(z^t, x^t)) \ .$$

Therefore, using Equation (54), we have

$$x^{t+1} \in \mathrm{Span}(\nabla_1 F_{j_0}(z^0, x^0), \dots, \nabla_1 F_{j_t}(z^t, x^t)) \ .$$

Since $z^0 = 0$, by Equation (52), $z^1 \in \mathrm{Span}(x^0)$ and by induction

$$z^{t+1} \in \mathrm{Span}(x^0, \dots, x^t) \ .$$

As a consequence,

$$x^t \in \mathrm{Span}(\nabla_1 F_{j_0}(\mathrm{Span}(x^0), x^0), \dots, \nabla_1 F_{j_t}(\mathrm{Span}((x^s)_{s\leq t}), x^t)) \ .$$

Let us denote $y^{(j,t)} = U^{(j)} x^t$. Since $x^0 = 0$, $y^{(j_0,0)} = 0$ and by the second part of Proposition B.2, $x^1 \in \mathrm{Span}(u_1^{(j_0)})$.

Now we assume that for all $s \leq t$ we have

$$x^s \ \in \ \mathrm{Span}(u_1^{(j_0)}, \dots, u_s^{(j_0)}, \dots, u_1^{(j_{s-1})}, \dots, u_s^{(j_{s-1})}) \ .$$

There exist scalars $\alpha_1, \dots, \alpha_r, \beta_{1,1}, \beta_{2,1}, \beta_{2,2}, \dots, \beta_{t,1}, \dots, \beta_{t,t}$ such that

$$x^{t+1} = \sum_{r=1}^{t} \alpha_r \nabla_1 F_{j_r}\left(\sum_{s=1}^{r} \beta_{r,s} x^s, x^r\right) \ .$$

Let $X^r = \sum_{s=1}^{r} \beta_{r,s} x^s$. For $r \in \{1, \dots, t\}$, we have by induction hypothesis

$$X^r \in \mathrm{Span}(u_1^{(j_0)}, \dots, u_r^{(j_0)}, \dots, u_1^{(j_{r-1})}, \dots, u_r^{(j_{r-1})}) \ .$$

By orthogonality, we have

$$\mathrm{Span}(u_1^{(j_0)}, \dots, u_r^{(j_0)}, \dots, u_1^{(j_{r-1})}, \dots, u_r^{(j_{r-1})}) \perp \mathrm{Span}(u_{r+1}^{(j_r)}, \dots, u_{T+1}^{(j_r)}) \ .$$

As a consequence

$$U^{(j_r)} X^r = (\langle u_1^{(j_r)}, X^r \rangle, \dots, \langle u_r^{(j_r)}, X^r \rangle, 0, \dots, 0) \ .$$

We can use Proposition B.2 to say

$$\nabla_1 F_{j_r}(X^r, x^r) \in \mathrm{Span}(u_1^{(j_r)}, \dots, u_{r+1}^{(j_r)}) \subset \mathrm{Span}(u_1^{(j_0)}, \dots, u_r^{(j_0)}, u_{r+1}^{(j_0)}, \dots, u_1^{(j_r)}, \dots, u_{r+1}^{(j_r)}) \ .$$

And we get finally

$$x^{t+1} \in \text{Span}(u_1^{(j_0)}, \dots, u_t^{(j_0)}, u_{t+1}^{(j_0)}, \dots, u_1^{(j_t)}, \dots, u_{t+1}^{(j_t)}) \ .$$

By induction, for any t, we have

$$x^t \in \text{Span}(\underbrace{u_1^{(j_0)}, \dots, u_t^{(j_0)}, \dots, u_1^{(j_t)}, \dots, u_t^{(j_t)}}_{\text{at most } mt \text{ vectors}})$$

and so

$$x^t \perp \text{Span}((u_1^{(j)}, \dots, u_{T+1}^{(j)})_{j \in [m] \setminus \{j_0, \dots, j_t\}}, (u_{t+1}^{(j)}, \dots, u_{T+1}^{(j)})_{j \in \{j_0, \dots, j_t\}}) \ .$$

As a consequence, for $t \leq \frac{m}{2}T$, let $\mathcal{I} = \{j, y_T^{(j,t)} = y_{T+1}^{(j,t)} = 0\}$ with $y^{(j,t)} = U^{(j)}x^t$. Since $t \leq \frac{m}{2}T$, we have $|\mathcal{I}| \leq \frac{m}{2}$ and by Lemma B.7, we have

$$\|\nabla h(x^t)\| \geq \epsilon \ .$$

If we define $T((x^t)_t, h) = \inf\{t \in \mathbb{N}, \|\nabla h(x^t)\|^2 \leq \epsilon\}$, we just showed that for the fixed sequence (i_t, j_t), we have

$$T((x^t)_t, h) \geq \frac{m}{2}T = \Omega(\sqrt{m}\epsilon^{-1}) \ .$$

The right-hand side being independent from the sequence (i_t, j_t), for $t \leq \frac{m}{2}T$, we have

$$\mathbb{E}[\|\nabla h(x^t)\|^2] > \epsilon$$

where the expectation is taken over the random choice of $i_0, \dots, i_{t-1}, j_0, \dots, j_{t-1}$. □

C Details on the experiments

We performed the experiments with the Python package Benchopt (Moreau et al., 2022)[2]. For each experiment, we use minibatches instead of single samples to estimate oracles because it is more efficient in practice. We use a batch size of 64 for the stochastic inner and outer oracles. All the experiments were performed on processors AMD EPYC 7742 (4 CPUs/experiment).

C.1 Benchmark on quadratics

For this benchmark, we consider

$$F(z,x) = \frac{1}{m}\sum_{j=1}^{m} F_j(z,x), \quad G(z,x) = \frac{1}{n}\sum_{i=1}^{n} G_i(z,x) \ .$$

The functions F_j and G_i are defined as

$$F_j(z,x) = \frac{1}{2}z^\top A_z^{F_j} z + \frac{1}{2}x^\top A_x^{F_j} + xB^{F_j}z + (d_z^{F_j})^\top z + (d_x^{F_j})^\top x$$
$$G_i(z,x) = \frac{1}{2}z^\top A_z^{G_i} z + \frac{1}{2}x^\top A_x^{G_i} + xB^{G_i}z + (d_z^{G_i})^\top z + (d_x^{G_i})^\top x$$

with $A_z^{F_j}, A_z^{G_i} \in \mathbb{R}^{p\times p}$, $A_x^{F_j}, A_x^{G_i} \in \mathbb{R}^{d\times d}$, $B^{F_j}, B^{G_i} \in \mathbb{R}^{d\times p}$, $d_z^{F_j}, d_z^{G_i} \in \mathbb{R}^p$ and $d_x^{F_j}, d_x^{G_i} \in \mathbb{R}^d$. The vectors $d_x^{F_j}$, $d_x^{G_i}$ are drawn randomly according to a normal distribution $\mathcal{N}(0, I_d)$. The vectors $d_z^{F_j}$, $d_z^{G_i}$ are drawn randomly according to a normal distribution $\mathcal{N}(0, I_p)$. For the Hessian matrices with respect to z, we generate $A_z^{G_i}$ so that $\frac{1}{n}\sum_{i=1}^n A_z^{G_i} = A$ for a symetric positive definite matrix A with spectrum in $[0.1, 1]$. To do so, we generate $x_i \sim \mathcal{N}(0, I_p)$ and set $A_z^{G_i} = \sqrt{A}x_i(\sqrt{A}x_i)^\top$. We proceed similarly for $A_z^{F_j}$, $A_z^{G_i}$, $A_x^{F_j}$. For B^{G_i}, we want $\frac{1}{n}\sum_{i=1}^n B^{G_i} = B$ for a prescribed matrix $B \in \mathbb{R}^{d\times p}$ such that $\|B\| = 0.1$. Let $B = U\Sigma V^\top$ the singular values decomposition of B. To get B^{G_i}, we generate $x_i \sim \mathcal{N}(0, I_p)$ and set $B^{G_i} = (V\Sigma x_i)(Ux_i)^\top$. We proceed similarly for B^{F_j}. In our experiment, we take $n = 32768$ and $m = 1024$. To select the parameters of the solvers, we

[2]The code of the benchmark is available at https://github.com/benchopt/benchmark_bilevel and the results are displayed in https://benchopt.github.io/results/benchmark_bilevel.html.

perform a grid search. More precisely, for each solver, we take the inner step size in the form of αt^{-a} where a is the theoretical decrease rate of each solver and α is chosen in $\{0.01, 0.1\}$. The outer step size is taken as $\frac{\alpha}{r}t^{-b}$ where b is the theoretical decrease rate and r is chosen in $\{0.1, 1, 10, 100\}$. For the two-loops algorithms (*i.e.* StocBiO, VRBO, AmIGO), the number of inner steps is set to 10 after a manual search. In the methods implementing Neumann approximations (MRBO, VRBO, StocBiO), the number of terms in the Neumann series is also set to 10 and the scaling parameter η is set to 0.5. To get the fastest convergence, we keep for each solver the set of parameters that give the best decrease of h on the 100 first epochs. The period of full batch computation of VRBO and SRBA q is parametrized as $q = a\frac{n+m}{b}$ where $b = 64$ is the batch size and a is chosen in $\{2^{-6}, 2^{-3}, 2^{-1}, 2^3, 2^6\}$. For F^{2}SA, we take $\lambda_0 = 1$ and $\delta_t = \alpha t^{-\frac{1}{7}}$ with α chosen in $\{0.01, 0.1, 1\}$.

C.2 Hyperparameter selection with IJCNN1

We solve a regularization selection problem for an ℓ^2-regularized logistic regression problem. Here, we assume that we have a regularization parameter per feature. We have $n_{\text{train}} = 49,990$ training samples $(d_i^{\text{train}}, y_i^{\text{train}})_{i\in[n_{\text{train}}]}$ and $n_{\text{val}} = 91,701$ validation samples $(d_i^{\text{val}}, y_i^{\text{val}})_{i\in[n_{\text{train}}]}$ coming from the IJCNN1[3] dataset. Mathematically, it boils down to solve Problem (1) with F and G given by

$$F(\theta, \lambda) = \frac{1}{n_{\text{val}}}\sum_{j=1}^{n_{\text{val}}} \varphi(y_j^{\text{val}}\langle d_j^{\text{val}}, \theta\rangle)$$

$$G(\theta, \lambda) = \frac{1}{n_{\text{train}}}\sum_{i=1}^{n_{\text{train}}} \varphi(y_i^{\text{train}}\langle d_i^{\text{train}}, \theta\rangle) + \frac{1}{2}\sum_{k=1}^{p} \mathrm{e}^{\lambda_k}\theta_k^2$$

where φ is the logistic loss defined by $\varphi(u) = \log(1 + \mathrm{e}^{-u})$. The inner and outer step sizes are set to 0.05.

To make our comparison, we select the parameters of each solver with an extensive grid search. More precisely, for each solver, we take the inner step size in the form of αt^{-a} where a is the theoretical decrease rate of each solver and α is chosen in $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}\}$. The outer step size is taken as $\frac{\alpha}{r}t^{-b}$ where b is the theoretical decrease rate and r is chosen in $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}\}$. For the two-loops algorithms (*i.e.* StocBiO, VRBO, AmIGO), the number of inner steps is set to 10 after a manual search. In the methods implementing Neumann approximations (MRBO, VRBO, StocBiO), the number of terms in the Neumann series is also set to 10 and the scaling parameter η is set to 0.5. To get the fastest convergence, we keep for each solver the set of parameters that give the best decrease of h on the 100 first epochs. The period of full batch computation of VRBO and SRBA q is parametrized as $q = a\frac{n+m}{b}$ where $b = 64$ is the batch size and a is chosen in $\{2^{-6}, 2^{-3}, 2^{-1}, 2^3, 2^6, 2^9\}$. For F^{2}SA, we take $\lambda_0 = 1$ and $\delta_t = \alpha t^{-\frac{1}{7}}$ with α chosen in $\{0.01, 0.1, 1\}$.

D Additional experiment: Datacleaning task

We run an additional experiment. For each experiment, the parameters of the solvers are chosen by an extensive grid search. Then we select the curve that gives the best validation accuracy for each solver and finally plot the corresponding test error on Figure D.1.

The third experiment is the datacleaning task. It aims to train a multiclass classifier while having some training samples with noisy labels. On the one hand we have $n_{\text{train}} = 20,000$ training labelled samples $(d_i^{\text{train}}, y_i^{\text{train}})_{i\in[n_{\text{train}}]}$ with potentially corrupted labels with probability p_c (in the experiments $p_c = 0.5$). On the other hand, we have a validation set $(d_j^{\text{val}}, y_j^{\text{val}})_{j\in[n_{\text{val}}]}$ of $n_{\text{val}} = 5,000$ samples where all the samples are clean. We also have $10,000$ clean test samples. The datacleaning problem consists in learning a classifier on all these samples by giving less weight to corrupted labels. It can be cast as a bilevel optimization problem like (1) where the function F and G are given by

$$F(\theta, \lambda) = \frac{1}{n_{\text{val}}}\sum_{j=1}^{n_{\text{val}}} \ell(\theta d_j^{\text{val}}, y_j^{\text{val}})$$

$$G(\theta, \lambda) = \frac{1}{n_{\text{train}}}\sum_{i=1}^{n_{\text{train}}} \sigma(\lambda_i)\ell(\theta d_i^{\text{train}}, y_i^{\text{train}}) + C_r\|\theta\|^2$$

[3] 1https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

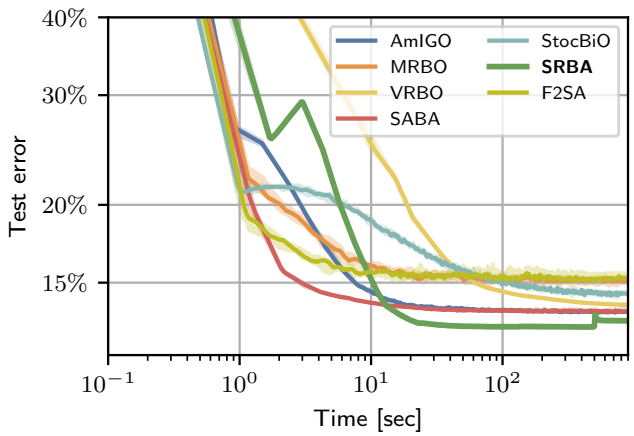

Figure D.1: Comparison of stochastic bilevel solvers. Each solver is run on 10 random seeds and the lines show the median performances. The shaded area corresponds to the performances between the 20% and the 80% percentiles. Test error on the datacleaning task with the MNIST dataset with a corruption rate 0.5.

where $\theta \in \mathbb{R}^{C\times p}$, $\lambda \in \mathbb{R}^{n_{\text{train}}}$, ℓ is the cross entropy loss and σ is the sigmoid function defined by $\sigma(\lambda) = \frac{1}{1+\mathrm{e}^{-\lambda}} \in (0\,,\,1]$.

We run this experiment on the MNIST dataset. We used $20,000$ training samples, $5,000$ validation samples, and $10,000$ test samples. The parameter C_r is set to 0.2 after a manual search to get the best performance. For the tuning of the step sizes of each method, we set $(\rho_t, \gamma_t) = (\alpha t^{-a}, \beta t^{-b})$ where (a, b) are the rate provided by the analysis of each method, α is chosen among 4 values between 10^{-3} and 10^{0} spaced on a logarithmic scale. The scaling parameter β is set to $\frac{\beta}{r}$ where r is chosen among 6 values between 10^{-5} and 10^{0} spaced on a logarithmic scale. The other parameters are chosen in the same way as the IJCNN1 experiments (see Appendix C.2).

We plot the test error on the Figure D.1 (right). On the one hand, SRBA reaches the best final value. On the other hand, in terms of speed, it is the second fastest after SABA. The other methods are slower and reach a worse final accuracy.