A Extensive comparison between stochastic methods for bilevel optimization

We provide here tables summarizing other methods in stochastic bilevel optimization. They are grouped between methods that are based on two nested loops and methods that use only one loop.

In the following tables, the inner iterations are referred with the variable k and the outer iterations are referred with the variable t (or T for the total number of iterations).

In the literature, there are three main ways to perform Hessian inversion. The HIA, first proposed in [19], and SHIA, proposed in [26], procedures used for Hessian inversion are precised in Algorithm 2 and 3. These methods are based on Neumann approximation of the inverse of a matrix. SGD for Hessian inversion refers to Stochastic Gradient Descent on $v \mapsto \frac{1}{2}\langle \nabla^2_{11} G(z,x)v, v\rangle - \langle \nabla_1 F(z,x), v\rangle$. The complexity refers to the number of call to the oracles to get an ϵ-stationary solution. In these complexities, the notation $\tilde{O}$ hide polynomial factors in $\log \epsilon^{-1}$.

Algorithm 2 Hessian Inverse Approximation (HIA)

Input: variables $z \in \mathbb{R}^p$, $x \in \mathbb{R}^d$, gradient $\nabla_1 F(z,x) \in \mathbb{R}^p$, maximum number of iterations b, a parameter η.
Set $v^0 = \nabla_1 F(z,x)$
Choose $p \in \{0, \dots, b-1\}$ randomly.
for $k = 1, \dots, p$ **do**
  Sample $i \in [n]$
  Update v : $v^{k+1} = (I - \eta \nabla^2_{11} G(z,x))v^k$
end for
Return: $b\eta v_{p+1}$

Algorithm 3 Summed Hessian Inverse Approximation (SHIA)

Input: variables $z \in \mathbb{R}^p$, $x \in \mathbb{R}^d$, gradient $\nabla_1 F(z,x) \in \mathbb{R}^p$, maximum number of iterations b, a parameter η.
Set $v^0 = \nabla_1 F(z,x)$
Set $s^0 = v^0$
for $k = 0 \dots, b-1$ **do**
  Sample $i \in [n]$
  Update v: $v^{k+1} = (I - \eta \nabla^2_{11} G(z,x))v^k$
  Update s: $s^{k+1} = s^k + v^{k+1}$
end for
Return: ηs^b

The momentum column refers to the use of STORM [12] momentum in the inner loop or the outer loop. This momentum can be applied to either the inner or the implicit gradient estimate. If we consider the current estimate $y^t = (z^t, v^t, x^t)$ and the previous estimate $y^{t-1} = (z^{t-1}, v^{t-1}, x^{t-1})$, and we apply STORM to the quantity $\phi(y^t)$ with the memory $\hat{\phi}^t$, the momentum update rule reads

$$\hat{\phi}^{(t+1)} = \eta \phi(y^t) + (1-\eta)(\hat{\phi}^t + \phi(y^t) - \phi(y^{t-1})) \ .$$

Note that this update requires to evaluate the quantity ϕ twice per iteration, once in y^t and once in y^{t-1}. The memory is need to store the previous estimates y^{t-1} as well as the running estimate of the gradient $\hat{\phi}$.

Method (Two-loops)	Hessian inversion	Inner loop	Momentum	LR in	LR out	Complexity
BSA [19]	HIA	SGD on inner	No	$O(k^{-1})$	$O(T^{-1/2})$	$O(\epsilon^{-3})$
stocBiO [26]	SHIA	SGD on inner	No	Constant	Constant	$\tilde{O}(\epsilon^{-2})$
VRBO [47]	SHIA	SPIDER on inner	Yes (SPIDER)	Constant	Constant	$\tilde{O}(\epsilon^{-3/2})$
AmIGO [2]	SGD	SGD on inner	No	Constant	Constant	$O(\epsilon^{-2})$
Method (One-loop)	**Hessian inversion**	**Inner step**	**Momentum**	**LR in**	**LR out**	**Complexity**
TTSA [24]	HIA	SGD	No	$O(T^{-2/5})$	$O(T^{-3/5})$	$\tilde{O}(\epsilon^{-5/2})$
SMB [23]	HIA	SGD with momentum	Yes	Constant	Constant	$\tilde{O}(\epsilon^{-4})$
MRBO [47]	SHIA	SGD with STORM	Yes (STORM)	$O(t^{-1/3})$	$O(t^{-1/3})$	$\tilde{O}(\epsilon^{-3/2})$
STABLE [10]	Direct	SGD	No	$O(T^{-1/2})$	$O(T^{-1/2})$	$O(\epsilon^{-2})$
SUSTAIN [28]	HIA	SGD with STORM	Yes (STORM)	$O(t^{-1/3})$	$O(t^{-1/3})$	$O(\epsilon^{-3/2})$
SVRB [22]	Direct + momentum	SGD with momentum	Yes	$O(t^{-1/3})$	$O(t^{-1/3})$	$\tilde{O}(\epsilon^{-3})$
SBFW [1]	HIA	SGD	No	$O(t^{-1/2})$	$O(T^{-3/4})$	$\tilde{O}(\epsilon^{-4})$
FSLA [30]	SGD with STORM	SGD with STORM	Yes (STORM)	$O(t^{-1/2})$	$O(T^{-1/2})$	$O(\epsilon^{-2})$
SOBA	SGD step	SGD	No	$O(t^{-1/2})$	$O(t^{-1/2})$	$O(\epsilon^{-1/2})$
SABA	SAGA step	SAGA	No	Constant	Constant	$\boldsymbol{O((n+m)^{2/3}\epsilon^{-1})}$

Table 1: Comparison of the stochastic bilevel optimization solvers in the literature.
The complexity represents the number of oracle calls necessary to attain an ϵ accurate stationary point.

B Details on experiments

We provide here additional informations on the experiments.

B.1 Generalities

All the experiments are performed with `Python`, using the package `Benchopt` [35] and Numba [29] for fast implementation of stochastic methods. For each problem, we use oracles for a function given function f that $(f(z,x), \nabla_1 f(z,x), \nabla^2_{11} f(z,x)v, \nabla^2_{21} f(z,x)v)$ avoiding duplicate computation of intermediate results for these quantities.

We find that using mini-batches instead of individual samples to compute the stochastic estimates allowed for much faster computations, thanks to hardware acceleration and vectorization of the computations. We use continuous batches to avoid random memory access that slow down the computations. Concretely, if i_b is the index of the current batch and B is the batch-size, the indices of the corresponding samples are those in the set $\{i_b \times B, \dots, (i_b+1) \times B - 1\}$. By doing so, the samples in a same batch are contiguous in memory, which facilitates the access. We use a batch-size of 64 in all experiments.

For the methods involving an inner loop (stocBiO, BSA, AmIGO), we perform 10 inner steps at each outer iteration as proposed in the papers which introduced these methods. For the approximate Hessian vector product, we perform 10 steps per outer iteration for each methods using HIA (BSA, TTSA, SUSTAIN), SHIA (MRBO, stocBiO) or SGD (AmIGO) for the inversion of the linear system.

For the step sizes, they all have the form $\rho^t = \alpha/t^a$ and $\gamma^t = \beta/t^b$. For the pair of exponents (a,b), we choose the theoretical one from the original papers, that is $(1/2, 1/2)$ for BSA and FSLA, $(1/3, 1/3)$ for MRBO and SUSTAIN, $(0,0)$ for SABA, AmIGO and stocBiO, $(2/5, 3/5)$ for TTSA and SOBA. For (α, β), we perform a grid search (the grid is precised in the subsection dedicated to each experiment) and we keep for each method, the pair (α, β) that gives the lowest value of h (for the hyperparameters) or the lowest test accuracy (for the data cleaning task) in median over 10 runs for each possible pair. When we use HIA or SHIA for the Hessian inversion, we set $\eta = \alpha$ since the Hessian inversion problem has the same conditioning as the inner optimization problem.

For the STORM's momentum parameter in MRBO and SUSTAIN, we take $0.5/t^{2/3}$.

For SABA, we have to maintain the estimate $S[\phi, w]^i_t = \phi_i(w_i^{t+1}) - \phi_i(w_i^t) + \frac{1}{n}\sum_{i'=1}^n \phi_{i'}(w_{i'}^t)$ of $\frac{1}{n}\sum_{i=1}^n \phi_i(y^t)$ (see Section 2.2 for the notations). The sum inside S is maintain by performing a rolling mean on the past gradients computed. More precisely, $A_t = \frac{1}{n}\sum_{i'=1}^n \phi_{i'}(w_{i'}^t)$. To get A_{t+1}, instead of computing the summing all the gradients stored, which has $O(n)$ computational complexity, we do $A_{t+1} = A_t + \frac{1}{n}(\phi_i(w_i^{t+1}) - \phi_i(w_i^t))$, which is equivalent mathematically but has $O(1)$ computational complexity.

B.2 Hyperparameter selection on a toy problem

The Figure 1 corresponds to the methods SABA et SOBA applied to an hyperparameter selection problem for a Ridge regression. We generate 1000 samples $x_1, \dots, x_{1000} \in \mathbb{R}^1 0$ for $\mathcal{N}(0, I_1 0)$. We generate a parameter $\beta \sim \mathcal{N}(0, I_{10})$ and do $y = (X \odot W)\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.01 I_{10})$ and the entries of W have the form $W_{i,j} = 1 + u_j v_{i,j}$ with $v_{i,j} \sim \mathcal{U}([0,1])$ and $u_j \sim \mathcal{U}([0,1])$ if $1 \leq j \leq 5$ or $u_j \sim \mathcal{U}([0,10])$ if $6 \leq j \leq 10$. Then we use 750 pairs $(x_i^{\text{train}}, y_i^{\text{train}})_{1 \leq i \leq 750}$ as training samples and the remaining pairs $(x_i^{\text{val}}, y_i^{\text{val}})_{1 \leq i \leq 250}$ as validation samples. Finally, we solve (1) with

$$F(\theta, \lambda) = \frac{1}{2n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} ((x_i^{\text{val}})^\top \theta - y_i^{\text{val}})^2$$

and

$$G(\theta, \lambda) = \frac{1}{2n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} ((x_i^{\text{train}})^\top \theta - y_i^{\text{train}})^2 + \frac{\lambda}{2}\|\theta\|^2$$

with $n_{\text{train}} = 750$ and $n_{\text{val}} = 250$.

B.3 Hyperparameters selection on IJCNN1

In this experiment, we select the parameters regularization for a multiregularized logistic regression model precised in Equations (12) and (13) where we have one hyperparameter per feature

$$F(\theta, \lambda) = \frac{1}{m} \sum_{i=1}^{m} \varphi(y_i^{\text{val}} \langle d_i^{\text{val}}, \theta \rangle) \text{ and} \tag{12}$$

$$G(\theta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \varphi(y_i^{\text{train}} \langle d_i^{\text{train}}, \theta \rangle) + \frac{1}{2} \theta^\top \mathbf{diag}(e^{\lambda_1}, \dots, e^{\lambda_p}) \theta \ . \tag{13}$$

Note that the parametrization in e^λ of the penalty instead of λ can be surprising at first glance, but it is classical in the bilevel optimization literature [38, 26, 21] because it avoids positivity constraints on λ. In order to choose the select proper parameters (α, β) for each algorithm, we perform a grid search. We search α in a set of 9 values between 2^{-5} and 2^3 spaced on a log scale. For β, we choose r in a set of 7 values between 10^{-2} and 10 spaced on a logarithmic scale and we set $\beta = \frac{\alpha}{r}$.

For this experiments, we use Just-In-Time (JIT) compilation thanks to the package Numba [29], to decrease the python overhead in the iteration loop.

To evaluate the value function h, we use L-BFGS [32] to solve compute $z^*(x^t)$ and then evaluate the function $h(x^t) = F(z^*(x^t), x^t)$.

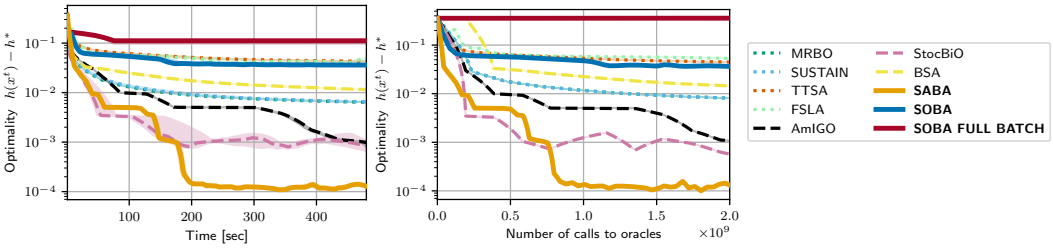


Figure B.1: Comparison of SOBA and SABA with other stochastic bilevel optimization methods in a problem of hyperparameter selection for ℓ^2 penalized logistic regression on IJCNN1 dataset. For each algorithm, we plot the median performance over 10 runs. In both plots, SABA achieves the best performance. The dashed lines are for one loop competitor methods, the dotted lines are for two loops methods and the solid lines are the proposed methods. **Left**: performance in running time, **Right**: performance in number of gradient/Hessian-vector products sampled.

B.4 Data hyper-cleaning

For the regularization parameter C_r, we choose $C_r = 0.2$ after a manual search in order to get the best final test accuracy.

In this experiment, the selection of the good pair (α, β) is also performed by grid search. The parameter α is picked in a set of 11 numbers between 10^{-3} and 100 spaced on a logarithmic scale. For β, we choose r in a set of 11 values between 10^{-5} and 1 spaced on a logarithmic scale and we set $\beta = \frac{\alpha}{r}$.

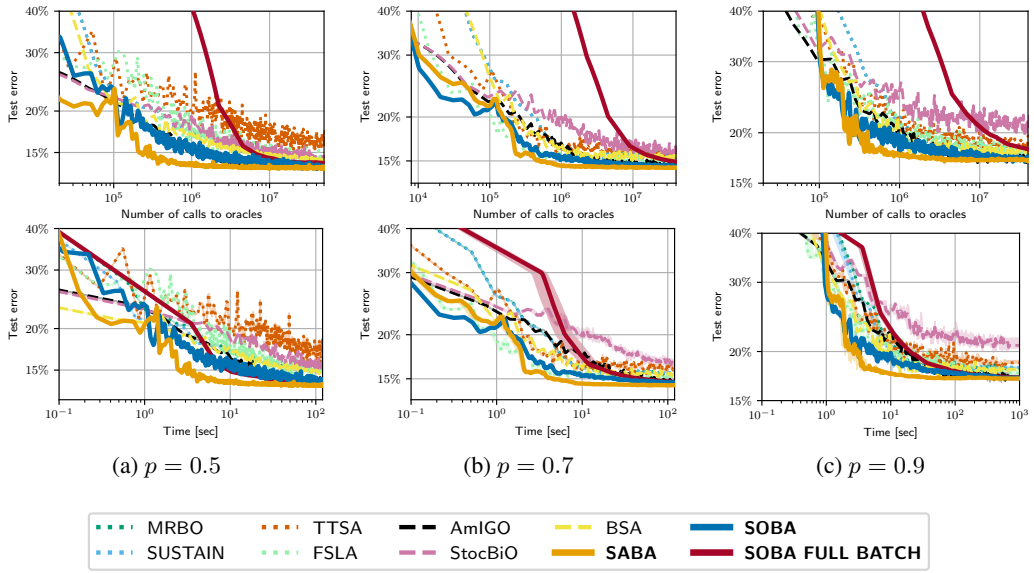
Note that in this case, we could not use JIT from Numba since at the moment of the experiment, the softmax function coming from Scipy was not compatible with Numba.

We report in Figure B.2 some additional convergence curves with different corruption probabilities $p \in \{0.5, 0.7, 0.9\}$ (the figure in the main text corresponds to $p = 0.5$). SABA is always the fastest algorithm to reach its final accuracy.

B.5 Additional experiment: Hyperparameter selection on the covtype dataset

We also perform an additional experiment which consists in selecting the best regularization parameter for a ℓ^2-regularized multinomial logistic regression problem on the covtype dataset[4]. This dataset

[4] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_covtype.html



(a) $p = 0.5$ (b) $p = 0.7$ (c) $p = 0.9$

Figure B.2: Datacleaning experiment, with different corruption probability (higher means that more data are contamined). **Top:** Performance with respect to the number of gradient/Hessian-vector product sampled, **Bottom:** Performance with respect to running time

contains $581,012$ samples with $p = 54$ features and there are $C = 7$ classes. We used $n = 371,847$ train samples, $m = 92,962$ samples and $n_{\text{test}} = 116,203$ test samples. We fit a multiclass logistic regression on this dataset, with one hyperparameter per class. This means that, if $(d_i^{\text{train}}, y_i^{\text{train}})_{i\in[n]}$ and $(d^{\text{val}}, y^{\text{val}})_{i\in[m]}$ are respectively the training samples and the validation samples, we solve the Problem (1) with

$$F(\theta, \lambda) = \frac{1}{m}\sum_{i=1}^{m} \ell(\theta d_i^{\text{val}}, y_i^{\text{val}}) \text{ and}$$

$$G(\theta, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta d_i^{\text{train}}, y_i^{\text{train}}) + \sum_{c=1}^{C} e^{\lambda_c} \sum_{i=1}^{p} \theta_{i,c}^2$$

where $\theta \in \mathbb{R}^{p\times C}$ and $\lambda \in \mathbb{R}^C$.

As for the other experiments, we performed and grid search over 63 pairs (α, β) to set the step sizes. The parameter α is chosen among values between 2^{-5} and 2^3 spaced in log scale. For β, we choose it in a set of values between 10^{-2} and 10 spaced in log scale. We used a batch size of 64. The experiment took 525 CPU hours.

We show in Figure B.3 the error on the test samples with respect to the running time and the number of gradients/Hessian-vector products sampled. We observe that SABA and SOBA achieve the best performances. The initial gap between the first and the second plot for SABA is due to the overhead of the initialization of the memory. This gap can be reduced by increasing the batch size.

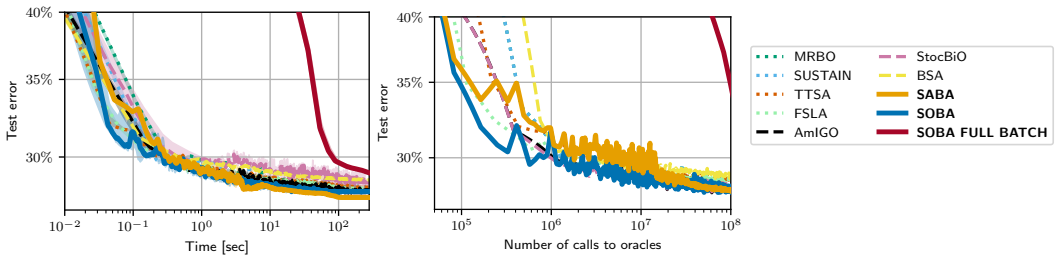


Figure B.3: Comparison of SOBA and SABA with other stochastic bilevel optimization methods in a problem of hyperparameter selection for ℓ^2 penalized multical logistic regression on covtype dataset. For each algorithm, we plot the median performance over 10 runs. The dashed lines are for one loop competitor methods, the dotted lines are for two loops methods and the solid lines are the proposed methods. **Left**: performance in running time, **Right**: performance in number of gradient/Hessian-vector products sampled.

C Proofs

C.1 Proof of Proposition 2.1

Proof. Let (z, v, x) a zero of (D_z, D_v, D_x). For D_z, this means that $\nabla_1 G(z, x) = 0$. Since $G(\cdot\,, x)$ is strongly convex, z is the minimizer of $G(\cdot\,, x)$, *i.e.* $z = z^*(x)$. The fact that (z, v, x) is a zero of D_v implies that $\nabla^2_{11} G(z, x)v = -\nabla_1 F(z, x)$. Replacing z by its value, we get $v = -\left[\nabla^2_{11} G(z^*(x), x)\right]^{-1} \nabla_1 F(z^*(x), x)$ which is $v^*(x)$ by definition. Putting all together and using the expression of $\nabla h(x)$ given by (2), we get

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

On the other hand, $D_x(z, v, x) = 0$ so $\nabla h(x) = 0$. □

C.2 Proof of Lemma 3.4

Proof. Let $(z, v, x) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^d$. Using the fact that $\nabla_1 G(z^*(x), x) = 0$ and the L_1^G-smoothness of $G(\cdot\,, x)$, we have

$$\boxed{\|D_z(z, v, x)\|^2 = \|\nabla_1 G(z, x) - \nabla_1 G(z^*(x), x)\|^2 \leq L_G^2 \|z - z^*(x)\|^2 \ .}$$

For D_v, since $\nabla^2_{11} G(z^*(x), x) v^*(x) = -\nabla_1 F(z^*(x), x)$, we write

$$\begin{aligned} \|D_v\| &= \|(\nabla^2_{11} G(z, x)v + \nabla_1 F(z, x)) - (\nabla^2_{11} G(z^*(x), x)v^*(x) + \nabla_1 F(z^*(x), x))\| & (14) \\ &\leq \|[\nabla^2_{11} G(z, x) - \nabla^2_{11} G(z^*(x), x)]v^*(x)\| + \|\nabla^2_{11} G(z, x)[v - v^*(x)]\| & (15) \\ &\quad + \|\nabla_1 F(z, x) - \nabla_1 F(z^*(x), x)\| \ . \end{aligned}$$

For the first term, we use the Lipschitz continuity of $\nabla^2_{11} G$:

$$\|[\nabla^2_{11} G(z, x) - \nabla^2_{11} G(z^*(x), x)]v^*(x)\| \leq L_2^G \|z - z^*(x)\| \|v^*(x)\| \ .$$

Then, since G in μ_G-strongly convex w.r.t. z, $\nabla_1 F(z^*(\cdot), \cdot)$ is bounded and $v^*(x) = -[\nabla^2_{11} G(z^*(x), x)]^{-1} \nabla_1 F(z^*(x), x)$, we have

$$\|[\nabla^2_{11} G(z, x) - \nabla^2_{11} G(z^*(x), x)]v^*(x)\| \leq \frac{L_2^G C_F}{\mu_G} \|z - z^*(x)\| \ . \tag{16}$$

For the second term, we use the L_1^G-smoothness of $G(\cdot\,, x)$ and for the third term, we use the L_1^F-smoothness of F and we finally get

$$\|D_v\| \leq \left(\frac{L_2^G C_F}{\mu_G} + L_1^F \right) \|z - z^*(x)\| + L_1^G \|v - v^*(x)\| \ . \tag{17}$$

Then, taking $L_v = \sqrt{2}\max\left(\frac{L_2^G C_F}{\mu_G} + L^F, L_1^G\right)$, we get

$$\boxed{\|D_v(z,v,x)\|^2 \leq L_v^2(\|z - z^*(x)\|^2 + \|v - v^*(x)\|^2) \ .} \tag{18}$$

For $D_x(z,v,x) - \nabla h(x)$ we start by writing

$$\|D_x(z,v,x) - \nabla h(x)\| \leq \|\nabla_2 F(z,x) - \nabla_2 F(z^*(x),x)\| + \|\nabla_{21}^2 G(z,x)v - \nabla_{21}^2 G(z^*(x),x)v^*(x)\| \tag{19}$$

$$\leq \|\nabla_2 F(z,x) - \nabla_2 F(z^*(x),x)\| + \|\nabla_{21}^2 G(z,x)\|\|v - v^*(x)\| \tag{20}$$
$$+ \|v^*(x)\|\|\nabla_{21}^2 G(z,x) - \nabla_{21}^2 G(z^*(x),x)\| \ .$$

We bound the first term using the fact that $\nabla_2 F$ is L_1^F-Lipschitz continuous. For the second term, the fact that $\nabla_{21}^2 G$ is bounded thanks to the Lipschitz continuity of $\nabla_1 G(z,\,\cdot\,)$. For the third term, we use that $\nabla_{21}^2 G(\,\cdot\,,x)$ is L_2^G-Lipschitz continuous and the same derivation as Equation (16). We finally get

$$\|D_x - \nabla h(x)\| \leq \left(L_1^F + \frac{C_F L_2^G}{\mu_G}\right)\|z - z^*(x)\| + L_1^G\|v - v^*(x)\| \ . \tag{21}$$

Taking $L_x = \sqrt{2}\max\left(L_1^F + \frac{C_F L_2^G}{\mu_G}, L_1^G\right)$ yields

$$\boxed{\|D_x(z,v,x) - \nabla h(x)\|^2 \leq L_x^2(\|z - z^*(x)\|^2 + \|v - v^*(x)\|^2) \ .} \tag{22}$$

□

C.3 Smoothness constant of h

From Ghadimi and Wang [19, Lemma 2.2], we get the Lemma 3.10 which states the L^h-smoothness of h with

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

C.4 Lemmas on the regularity of z^* and v^*

We start by showing the Lipschitz continuity of z^* and v^*.

Lemma C.1. *There exists a constant $L_* > 0$ such that for any $x_1, x_2 \in \mathbb{R}^d$ we have*

$$\|z^*(x_1) - z^*(x_2)\| \leq L_*\|x_1 - x_2\|, \quad \|v^*(x_1) - v^*(x_2)\| \leq L_*\|x_1 - x_2\| \ .$$

Proof. Let $x \in \mathbb{R}^d$. The Jacobian of z^* is given by $\mathrm{d}z^*(x) = -[\nabla_{11}^2 G(z^*(x),x)]^{-1}\nabla_{1,2}^2 G(z^*(x),x)$. Thanks to the μ_G-strong convexity of G and the fact that $\nabla_{21}^2 G$ is bounded, we have $\|\mathrm{d}z^*(x)\| \leq \frac{L_1^G}{\mu_G}$. Thus, z^* is Lipschitz continuous.

For $\|v^*(x_1) - v^*(x_2)\|$, we start from the definition v^*:

$$\|v^*(x_1) - v^*(x_2)\| = \|[\nabla_{11}^2 G(z^*(x_1),x_1)]^{-1}\nabla_1 F(z^*(x_1),x_1) - [\nabla_{11}^2 G(z^*(x_2),x_2)]^{-1}\nabla_1 F(z^*(x_2),x_2)\| \tag{23}$$

$$\leq \|([\nabla_{11}^2 G(z^*(x_1),x_1)]^{-1} - [\nabla_{11}^2 G(z^*(x_2),x_2)]^{-1}\nabla_1 F(z^*(x_1),x_1)\| \tag{24}$$
$$+ \|[\nabla_{11}^2 G(z^*(x_2),x_2)]^{-1}(\nabla_1 F(z^*(x_2),x_2) - \nabla_1 F(z^*(x_1),x_1))\| \ .$$

For the first term, we use that for any invertible matrix A and B we have $A^{-1} - b^{-1} = A^{-1}(B - A)B^{-1}$ to get

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

And then, since $\nabla_1 F(z^*(\,\cdot\,),\,\cdot\,)$ is bounded:

$$\left\|([\nabla^2_{11}G(z^*(x_1),x_1)]^{-1} - [\nabla^2_{11}G(z^*(x_2),x_2)]^{-1}\nabla_1 F(z^*(x_1),x_1)\right\| \leq \frac{C_F L_2^G}{\mu_G^2}\left[1 + \frac{L_1^G}{\mu_G}\right]\|x_1 - x_2\| \ .$$

For the second term, the strong convexity of $G(\,\cdot\,,x)$ and the fact that $\nabla_1 F$ is Lipschitz continuous lead to

$$\|[\nabla^2_{11}G(z^*(x_2),x_2)]^{-1}(\nabla_1 F(z^*(x_2),x_2) - \nabla_1 F(z^*(x_1),x_1))\| \leq \frac{1}{\mu_G}\|\nabla_1 F(z^*(x_2),x_2) - \nabla_1 F(z^*(x_1),x_1)\| \tag{25}$$

$$\leq \frac{L_1^F}{\mu_F}\|(z^*(x_1),x_1) - (z^*(x_2),x_2)\| \tag{26}$$

$$\leq \frac{L_1^F}{\mu_G}[\|z^*(x_1) - z^*(x_2)\| + \|x_1 - x_2\|] \tag{27}$$

$$\leq \frac{L_1^F}{\mu_G}\left[1 + \frac{L_1^G}{\mu_G}\right]\|x_1 - x_2\| \ . \tag{28}$$

Then we get

$$\|v^*(x_1) - v^*(x_2)\| \leq \left[\frac{C_F L_2^G}{\mu_G^2}\left[1 + \frac{L_1^G}{\mu_G}\right] + \frac{L_1^F}{\mu_G}\left[1 + \frac{L_1^G}{\mu_G}\right]\right]\|x_1 - x_2\| \ . \tag{29}$$

We conclude by setting

$$L_* = \max\left(\frac{L_1^G}{\mu_G}, \frac{C_F L_2^G}{\mu_G^2}\left[1 + \frac{L_1^G}{\mu_G}\right] + \frac{L_1^F}{\mu_G}\left[1 + \frac{L_1^G}{\mu_G}\right]\right) \ .$$

□

In what follows, we denote by $\mathbb{E}_t[\,\cdot\,]$ the expectation conditionally on z^t, v^t and x^t.

We have the smoothness property of z^* provided in [9, Lemma 2].

Lemma C.2. *Under the Assumptions 3.1, 3.2 and 3.3, the function $z^* : \mathbb{R}^d \to \mathbb{R}^p$ is L_{zx}-smooth with*

$$L_{zx} = \frac{L_2^G(1 + L_*)}{\mu_G} + \frac{L_1^G L_{11}^G(1 + L_*)}{\mu_G^2} \ . \tag{30}$$

We establish the same result for v^*. To this, we need more regularity on G and F.

Lemma C.3. *The function $v^* : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable and its differential is defined for any $x, \epsilon \in \mathbb{R}^d$ by:*

$$\begin{aligned} \mathrm{d}v^*(x).\epsilon = &[\nabla_1^2 G(z^*(x),x)]^{-1}[\nabla_{11}^2 F(z^*(x),x)\mathrm{d}z^*(x).\epsilon + \nabla_{12}^2 F(z^*(x),x).\epsilon] \\ &- [\nabla_1^2 G(z^*(x),x)]^{-1}[(\nabla_{111}^3 G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) + (\nabla_{112}^3 G(z^*(x),x)|\epsilon)] \\ &\times [\nabla_1^2 G(z^*(x),x)]^{-1}\nabla_1 F(z^*(x),x) \end{aligned} \tag{31}$$

where for any $z, \alpha \in \mathbb{R}^p$ and $x \in \mathbb{R}^d$, $(\nabla_{111}^3 G(z,x)|\alpha) \in \mathbb{R}^{p\times p}$ is defined by

$$(\nabla_{111}^3 G(z,x)|\alpha) = \left[\sum_{k=1}^p \frac{\partial^3 G}{\partial z_i \partial z_j \partial z_k}(z,x)\alpha_k\right]_{1\le i,j\le p}$$

and for any $\beta \in \mathbb{R}^d$, $(\nabla_{112}^3 G(z,x)|\beta) \in \mathbb{R}^{p\times p}$ is defined by

$$(\nabla_{112}^3 G(z,x)|\beta) = \left[\sum_{k=1}^p \frac{\partial^3 G}{\partial z_i \partial z_j \partial x_k}(z,x)\beta_k\right]_{1\le i,j\le p} .$$

Moreover, $\mathrm{d}v^*$ is L_{vx}-Lipschitz continuous.

Proof. Let $x, \epsilon \in \mathbb{R}^d$. Using the differentiability of $\nabla_{11}^2 G$, $\nabla_1 F$ and of the matrix inversion, we have

$$\begin{aligned} v^*(x+\epsilon) &= [\nabla_{11}^2 G(z^*(x+\epsilon), x+\epsilon)]^{-1}\nabla_1 F(z^*(x+\epsilon),\epsilon) \\ &= [\nabla_{11}^2 G(z^*(x),x) + (\nabla_{111}^3 G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) + (\nabla_{112}^3 G(z^*(x),x)|\epsilon) + o(\|\epsilon\|)]^{-1} \\ &\quad \times (\nabla_1 F(z^*(x),x) + \nabla_{11}^2 F(z^*(x),x)\mathrm{d}z^*(x).\epsilon + \nabla_{12}^2 F(z^*(x),x)\epsilon + o(\|\epsilon\|)) \\ &= \big\{[\nabla_{11}^2 G(z^*(x),x)]^{-1} \\ &\quad -[\nabla_{11}^2 G(z^*(x),x)]^{-1}[(\nabla_{111}^3 G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) + (\nabla_{112}^3 G(z^*(x),x)|\epsilon)] \\ &\quad \times[\nabla_{11}^2 G(z^*(x),x)]^{-1} + o(\|\epsilon\|)\big\} \\ &\quad \times (\nabla_1 F(z^*(x),x) + \nabla_{11}^2 F(z^*(x),x)\mathrm{d}z^*(x).\epsilon + \nabla_{12}^2 F(z^*(x),x)\epsilon + o(\|\epsilon\|)) \\ &= v^*(x) + [\nabla_1^2 G(z^*(x),x)]^{-1}[\nabla_{11}^2 F(z^*(x),x)\mathrm{d}z^*(x).\epsilon + \nabla_{12}^2 F(z^*(x),x).\epsilon] \\ &\quad - [\nabla_1^2 G(z^*(x),x)]^{-1}[(\nabla_{111}^3 G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) + (\nabla_{112}^3 G(z^*(x),x)|\epsilon)][\nabla_1^2 G(z^*(x),x)]^{-1} \\ &\quad \times \nabla_1 F(z^*(x),x) + o(\|\epsilon\|) \end{aligned}$$

that proves (31). Now, let $x, y, \epsilon \in \mathbb{R}^d$ with $\|\epsilon\| = 1$. Let us denote

$$A(x,\epsilon) = -[\nabla_1^2 G(z^*(x),x)]^{-1}[(\nabla_{111}^3 G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) + (\nabla_{112}^3 G(z^*(x),x)|\epsilon)][\nabla_1^2 G(z^*(x),x)]^{-1}$$

and

$$B(x,\epsilon) = \nabla_{11}^2 F(z^*(x),x)\mathrm{d}z^*(x).\epsilon + \nabla_{12}^2 F(z^*(x),x)$$

so that $\mathrm{d}v^*(x).\epsilon = [\nabla_{11}^2 G(z^*(x),x)]^{-1}B(x,\epsilon) + A(x,\epsilon)\nabla_1 F(z^*(x),x)$. We have

$$\begin{aligned} (\mathrm{d}v^*(x) - \mathrm{d}v^*(y)).\epsilon &= [\nabla_{11}^2 G(z^*(x),x)]^{-1}B(x,\epsilon) + A(x,\epsilon)\nabla_1 F(z^*(x),x) \\ &\quad - [\nabla_{11}^2 G(z^*(y),y)]^{-1}B(y,\epsilon) - A(y,\epsilon)\nabla_1 F(z^*(y),y) \end{aligned} \tag{32}$$

$$\begin{aligned} &= [\nabla_{11}^2 G(z^*(x),x)]^{-1}(B(x,\epsilon) - B(y,\epsilon)) \\ &\quad + ([\nabla_{11}^2 G(z^*(x),x)]^{-1} - [\nabla_{11}^2 G(z^*(y),y)]^{-1})B(y,\epsilon) \\ &\quad + A(x,\epsilon)(\nabla_1 F(z^*(x),x) - \nabla_1 F(z^*(y),y)) \\ &\quad + (A(x,\epsilon) - A(y,\epsilon))\nabla_1 F(z^*(y),y) \ . \end{aligned} \tag{33}$$

We can now bound each term using the regularity assumptions on G and F:

$$\|[\nabla^2_{11}G(z^*(x),x)]^{-1}(B(x,\epsilon)-B(y,\epsilon))\| \leq \frac{1}{\mu_G}(\|\nabla^2_{11}F(z^*(x),x)\mathrm{d}z^*(x)-\nabla^2_{11}F(z^*(y),y)\mathrm{d}z^*(y)\| \tag{34}$$

$$+\|\nabla^2_{12}F(z^*(x),x)-\nabla^2_{12}F(z^*(y),y)\|)$$

$$\leq \frac{1}{\mu_G}(\|\nabla^2_{11}F(z^*(x),x)-\nabla^2_{11}F(z^*(y),y)\|\|\mathrm{d}z^*(x)\| \tag{35}$$

$$+\|\mathrm{d}z^*(x)-\mathrm{d}z^*(y)\|\|\nabla^2_{11}F(z^*(y),y)\|$$

$$+L_2^F(\|z^*(x)-z^*(y)\|+\|x-y\|)$$

$$\leq \frac{1}{\mu_G}(L_2^F L_*(1+L_*)+L_{zx}L_1^F+L_2^F(1+L_*))\|x-y\| \tag{36}$$

$$\tag{37}$$

For the second term:

$$\|([\nabla^2_{11}G(z^*(x),x)]^{-1}-[\nabla^2_{11}G(z^*(y),y)]^{-1})B(y,\epsilon)\| \leq \frac{1}{\mu_G^2}\|\nabla^2_{11}G(z^*(x),x)-\nabla^2_{11}G(z^*(y),y)\|\|B(y,\epsilon)\| \tag{38}$$

$$\leq \frac{1}{\mu_G^2}\|\nabla^2_{11}G(z^*(x),x)-\nabla^2_{11}G(z^*(y),y)\| \tag{39}$$

$$\times(\|\nabla^2_{11}F(z^*(x),x)\|\|\mathrm{d}z^*(x)\|+\|\nabla^2_{12}F(z^*(x),x)\|)$$

$$\leq \frac{(L_2^G+L_1^F)(L_*+1)}{\mu_G^2}\|x-y\| \tag{40}$$

For the third term, we have:

$$\|A(x,\epsilon)(\nabla_1 F(z^*(x),x)-\nabla_1 F(z^*(y),y))\| \leq \frac{L_1^F(1+L^*)}{\mu_G^2}\|(\nabla^3_{111}G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon) \tag{41}$$

$$+\quad(\nabla^3_{112}G(z^*(x),x)|\epsilon)\|\|x-y\|$$

$$\leq \frac{(L_1^F+L_2^G)(1+L^*)}{\mu_G^2}\|x-y\| \tag{42}$$

And finally, for the forth term:

$$\|(A(x,\epsilon)-A(y,\epsilon))\nabla_1 F(z^*(y),y)\| \leq C^F\{\|[\nabla^2_{11}G(z^*(x),x)]^{-1}\| \tag{43}$$

$$\times\|(\nabla^3_{111}G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon)+(\nabla^3_{112}G(z^*(x),x)|\epsilon)\|$$

$$\times\|[\nabla^2_{11}G(z^*(x),x)]^{-1}-[\nabla^2_{11}G(z^*(y),y)]^{-1}\|$$

$$+\|[\nabla^2_{11}G(z^*(x),x)]^{-1}-[\nabla^2_{11}G(z^*(y),y)]^{-1}\|$$

$$\times\|(\nabla^3_{111}G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon)+(\nabla^3_{112}G(z^*(x),x)|\epsilon)\|$$

$$\times\|[\nabla^2_{11}G(z^*(y),y)]^{-1}\|$$

$$+\|[\nabla^2_{11}G(z^*(y),y)]^{-1}\|^2$$

$$\times(\|(\nabla^3_{111}G(z^*(x),x)|\mathrm{d}z^*(x).\epsilon)-(\nabla^3_{111}G(z^*(y),y)|\mathrm{d}z^*(y).\epsilon)\|$$

$$\|(\nabla^3_{112}G(z^*(x),x)|\epsilon)-(\nabla^3_{112}G(z^*(y),y)|\epsilon)\|)\} \tag{44}$$

$$\leq C^F\left\{2\frac{2L_2^G(1+L^*)}{\mu_G^3}+\frac{L_3^G(1+L^*)}{\mu_G^2}\right\}\|x-y\|$$

Thus v^* is L_{vx}-smooth with

$$L_{vx}=\frac{L_2^F L_*(1+L_*)+L_{zx}L_1^F+L_2^F(1+L_*)}{\mu_G}+2\frac{(L_2^G+L_1^F)(L_*+1)}{\mu_G^2}+\frac{C_F L_3^G(1+L^*)}{\mu_G^2}+4\frac{C_F L_2^G(1+L^*)}{\mu_G^3}\quad.$$

□

C.5 Proof of Lemma 3.9

We now provide the proof of Lemma 3.9.

Proof. **Inequality for δ_z.**
We start by expanding the square:

$$\begin{aligned}\|z^{t+1} - z^*(x^{t+1})\|^2 = \|z^{t+1} - z^*(x^t)\|^2 + \|z^*(x^{t+1}) - z^*(x^t)\|^2 \\ - 2\langle z^{t+1} - z^*(x^t), z^*(x^{t+1}) - z^*(x^t)\rangle\end{aligned} \tag{45}$$

We study each member, using the unbiasedness of D_z^t and the μ_G-strong convexity of $G(\,\cdot\,, x^t)$:

$$\mathbb{E}_t[\|z^{t+1} - z^*(x^t)\|^2] = \mathbb{E}_t[\|z^t - z^*(x^t)\|^2] - 2\rho\mathbb{E}_t[\langle D_z^t, z^t - z^*(x^t)\rangle] + \rho^2\mathbb{E}_t[\|D_z^t\|^2] \tag{46}$$

$$= \|z^t - z^*(x^t)\|^2 - 2\rho\langle \nabla_1 G(z^t, x^t), z^t - z^*(x^t)\rangle + \rho^2\mathbb{E}_t[\|D_z^t\|^2] \tag{47}$$

$$\leq (1 - \rho\mu_G)\|z^t - z^*(x^t)\|^2 + \rho^2\mathbb{E}_t[\|D_z^t\|^2] \ . \tag{48}$$

Taking the total expectation yields

$$\mathbb{E}[\|z^{t+1} - z^*(x^t)\|^2] \leq (1 - \rho\mu_G)\delta_z^t + \rho^2 V_z^t \ . \tag{49}$$

The second member is bounded using Lipschitz continuity of z^*:

$$\mathbb{E}[\|z^*(x^{t+1}) - z^*(x^t)\|^2] \leq L_*^2\mathbb{E}[\|x^{t+1} - x^t\|^2] = L_*^2\gamma^2 V_x^t \ .$$

For the remaining scalar product, we have

$$-2\langle z^{t+1} - z^*(x^t), z^*(x^{t+1}) - z^*(x^t)\rangle = -2[\langle z^t - z^*(x^t), z^*(x^{t+1}) - z^*(x^t)\rangle - \rho\langle D_z^t, z^*(x^{t+1}) - z^*(x^t)\rangle] \ . \tag{50}$$

The second term can be bounded using Cauchy-Schwarz inequality, the Lipschitz-continuity of z^* and Young inequality:

$$\mathbb{E}[\rho\langle D_z^t, z^*(x^{t+1}) - z^*(x^t)\rangle] \leq \mathbb{E}[\rho\|D_z^t\|\|z^*(x^{t+1}) - z^*(x^t)\|] \tag{51}$$

$$\leq \rho L_*\mathbb{E}[\|D_z^t\|\|x^{t+1} - x^t\|] \tag{52}$$

$$\leq \frac{\rho^2}{2}V_z^t + \frac{L_*^2}{2}\|x^{t+1} - x^t\|^2 \tag{53}$$

$$\leq \frac{\rho^2}{2}V_z^t + L_*^2\frac{\gamma^2}{2}V_x^t \ . \tag{54}$$

For $-2\langle z^t - z^*(x^t), z^*(x^{t+1}) - z^*(x^t)\rangle$, we follow the proof of [9] which consists in making appear the "unbiased part of $z^*(x^{t+1} - z^*(x^t)$ by a linear approximation. More precisely, we have

$$\begin{aligned}\langle z^t - z^*(x^t), z^*(x^{t+1}) - z^*(x^t)\rangle = \underbrace{\langle z^t - z^*(x^t), \mathrm{d}z^*(x^t)(x^{t+1} - x^t)\rangle}_{A} \\ \underbrace{\langle z^t - z^*(x^t), z^*(x^{t+1}) - z^*(x^t) - \mathrm{d}z^*(x^t)(x^{t+1} - x^t)\rangle}_{B} \ .\end{aligned} \tag{55}$$

For A, we use the unbiasedness of D_x^t, Cauchy-Schwarz inequality, the Lipschitz continuity of z^* (Lemma C.1) and the identity $ab \leq \eta a^2 + \frac{b^2}{\eta}$ for any $\eta > 0$:

$$-2\mathbb{E}[A] = -2\gamma\mathbb{E}[\langle z^t - z^*(x^t), \mathrm{d}z^*(x^t)D_x^t\rangle] \tag{56}$$

$$= -2\gamma\mathbb{E}[\langle z^t - z^*(x^t), \mathrm{d}z^*(x^t)\mathbb{E}_t[D_x^t]\rangle] \tag{57}$$

$$= -2\gamma\mathbb{E}[\langle z^t - z^*(x^t), \mathrm{d}z^*(x^t)D_x(z^t, v^t, x^t)\rangle] \tag{58}$$

$$\leq 2\gamma\mathbb{E}[\|z^t - z^*(x^t)\|\|\mathrm{d}z^*(x^t)D_x(z^t, v^t, x^t)\|] \tag{59}$$

$$\leq 2L_*\gamma\mathbb{E}[\|z^t - z^*(x^t)\|\|D_x(z^t, v^t, x^t)\|] \tag{60}$$

$$\leq 2\eta\delta_z^t + \frac{2L_*^2}{\eta}\gamma^2\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{61}$$

We take $\eta = \frac{\rho\mu_G}{4}$ and we get

$$-2\mathbb{E}[A] \leq \frac{\rho\mu_G}{2}\delta_z^t + \frac{8L_*^2}{\mu_G}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{62}$$

For B, we use Cauchy-Schwarz inequality, the smoothness of z^* (Lemma C.2), Young inequality and the boundedness of $\mathbb{E}_t[\|D_x^t\|^2]$ to get

$$\begin{aligned}
-2\mathbb{E}[B] &\leq 2\mathbb{E}[\|z^t - z^*(x^t)\|\|z^*(x^{t+1}) - z^*(x^t) - \mathrm{d}z^*(x^t)(x^{t+1} - x^t)\|] && (63)\\
&\leq L_{zx}\mathbb{E}[\|z^t - z^*(x^t)\|\|x^{t+1} - x^t\|^2] && (64)\\
&\leq L_{zx}\nu\mathbb{E}[\|z^t - z^*(x^t)\|^2\|x^{t+1} - x^t\|^2] + \frac{L_{zx}}{\nu}\mathbb{E}[\|x^{t+1} - x^t\|^2] && (65)\\
&\leq L_{zx}\nu\gamma^2\mathbb{E}[\|z^t - z^*(x^t)\|^2\mathbb{E}_t[\|D_x^t\|^2]] + \frac{L_{zx}\gamma^2}{\nu}V_x^t && (66)\\
&\leq L_{zx}B_x^2\nu\gamma^2\delta_z^t + \frac{L_{zx}\gamma^2}{\nu}V_x^t \ . && (67)
\end{aligned}$$

We take $\nu = \frac{L_{zx}}{L_*^2}$ and we get

$$-2\mathbb{E}[B] \leq \frac{L_{zx}^2 B_x^2\gamma^2}{L_*^2}\delta_z^t + L_*^2\gamma^2 V_x^t \tag{68}$$

Now, using $\gamma^2 \leq \frac{\rho\mu_G L_*^2}{B_x^2 L_{zx}^2}$, we end up with

$$\boxed{\delta_z^{t+1} \leq (1 - \frac{\rho\mu_G}{4})\delta_z^t + 2\rho^2 V_z^t + \beta_{zx}\gamma^2 V_x^t + \overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ ,} \tag{69}$$

with $\beta_{zx} = 3L_*^2$ and $\overline{\beta}_{zx} = \frac{8L_*^2}{\mu_G}$.

Inequality for δ_v. We proceed in a similar way for v:

$$\delta_v^{t+1} \leq \mathbb{E}[\|v^{t+1} - v^*(x^t)\|^2] + \mathbb{E}[\|v^*(x^{t+1}) - v^*(x^t)\|^2] - 2\mathbb{E}[\langle v^{t+1} - v^*(x^t), v^*(x^{t+1}) - v^*(x^t)\rangle] \ . \tag{70}$$

For the first term, we have

$$\mathbb{E}_t[\|v^{t+1} - v^*(x^t)\|^2] = \|v^t - v^*(x^t)\|^2 - 2\rho\langle D_v(z^t, v^t, x^t), v^t - v^*(x^t)\rangle + \rho^2\mathbb{E}_t[\|D_v^t\|^2] \tag{71}$$

Now, using that $D_v(z^*(x^t), v^*(x^t), x^t) = 0$:

$$\begin{aligned}
\langle D_v(z^t, v^t, x^t), v^t - v^*(x^t)\rangle &= \langle D_v(z^t, v^t, x^t) - D_v(z^*(x^t), v^*(x^t), x^t), v^t - v^*(x^t)\rangle && (72)\\
&= \langle \nabla_{11}^2 G(z^t, x^t)(v^t - v^*(x^t)), v^t - v^*(x^t)\rangle && (73)\\
&\quad + \langle (\nabla_{11}^2 G(z^t, x^t) - \nabla_{11}^2 G(z^*(x^t), x^t))v^*(x^t), v^t - v^*(x^t)\rangle\\
&\quad + \langle (\nabla_1 F(z^t, x^t) - \nabla_1 F(z^*(x^t), x^t)), v^t - v^*(x^t)\rangle\\
&\geq \mu_G\|v^t - v^*(x^t)\|^2 - \frac{L_2^G C_F}{\mu_G}\|z^t - z^*(x^t)\|\|v^t - v^*(x^t)\| && (74)\\
&\quad - L_1^F\|z^t - z^*(x^t)\|\|v^t - v^*(x^t)\|\\
&\geq \mu_G\|v^t - v^*(x^t)\|^2 - \omega\|z^t - z^*(x^t)\|\|v^t - v^*(x^t)\| && (75)
\end{aligned}$$

where $\omega = L_1^F + \frac{L_2^G C_F}{\mu_G}$. We then use $\omega\|z^t - z^*(x^t)\|\|v^t - v^*(x^t)\| \leq \frac{1}{2}c\|v^t - v^*(x^t)\|^2 + \frac{\omega^2}{2c}\|z^t - z^*(x^t)\|^2$ with $c = \mu_G$ to get

$$-\langle D_v(z^t, v^t, x^t), v^t - v^*(x^t)\rangle \leq -\frac{1}{2}\mu_G\delta_v^t + \frac{\omega^2}{2\mu_G}\delta_z^t \ .$$

We get the overall inequality by taking the total expectation

$$\mathbb{E}[\|v^{t+1} - v^*(x^t)\|^2] \leq \left(1 - \frac{\rho\mu_G}{2}\right)\delta_v^t + \rho\frac{\omega^2}{2\mu_G}\delta_z^t + \rho^2 V_v^t \ .$$

We also use Lipschitz on v^* to bound the other term

$$\mathbb{E}[\|v^*(x^{t+1}) - v^*(x^t)\|^2] \leq L_*^2\gamma^2 V_x^t \ .$$

As previously, the scalar product is bounded by:

$$-\mathbb{E}[\langle v^{t+1} - v^*(x^t), v^*(x^{t+1}) - v^*(x^t)\rangle] = -\mathbb{E}[\langle v^t - v^*(x^t), v^*(x^{t+1}) - v^*(x^t)\rangle] - \rho\mathbb{E}[\langle D_v^t, v^*(x^{t+1}) - v^*(x^t)\rangle] \tag{76}$$

$$\leq \mathbb{E}[\langle z^t - z^*(x^t), v^*(x^{t+1}) - v^*(x^t)\rangle] + \frac{\rho^2}{2}V_v^t + L_*^2\frac{\gamma^2}{2}V_x^t \tag{77}$$

We do similar manipulations pour v^*, thanks to Lemma C.3. We have as for z from Lemma C.1 for any $\eta > 0$:

$$-\mathbb{E}[\langle v^t - v^*(x^t), \mathrm{d}v^*(x^t)(x^{t+1} - x^t)\rangle] \leq \eta\delta_v^t + \frac{L_*^2\gamma^2}{\eta}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{78}$$

We take $\eta = \frac{\rho\mu_G}{8}$ and we get

$$-\mathbb{E}[\langle v^t - v^*(x^t), \mathrm{d}v^*(x^t)(x^{t+1} - x^t)\rangle] \leq \frac{\rho\mu_G}{8}\delta_v^t + \frac{8L_*^2\gamma^2}{\mu_G\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{79}$$

$$\tag{80}$$

Then smoothness of v^* for any $\eta > 0$ gives us

$$-\mathbb{E}[\langle v^t - v^*(x^t), v^*(x^{t+1}) - v^*(x^t) - \mathrm{d}v^*(x^t)(x^{t+1} - x^t)\rangle] \leq \frac{L_{vx}B_x^2\nu}{2}\gamma^2\delta_v^t + \frac{L_{vx}}{2\nu}\gamma^2 V_x^t \ . \tag{81}$$

With $\nu = \frac{L_{vx}}{L_*^2}$ we get

$$-\mathbb{E}[\langle v^t - v^*(x^t), v^*(x^{t+1}) - v^*(x^t) - \mathrm{d}v^*(x^t)(x^{t+1} - x^t)\rangle] \leq \frac{L_{vx}^2B_x^2}{2L_*^2}\gamma^2\delta_v^t + \frac{L_*^2}{2}\gamma^2 V_x^t \ . \tag{82}$$

With the assumption $\gamma^2 \leqslant \frac{\rho\mu_G L_*^2}{8L_{vx}^2B_x^2}$, we get

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{2} + \frac{\rho\mu_G}{4} + \frac{L_{vx}^2B_x}{L_*^2}\right)\delta_v^t + \rho\beta_{vz}\delta_z^t + 2\rho^2V_z^t + 3L_*^2\gamma^2V_x^t + \frac{16L_*^2\gamma^2}{\mu_G\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{83}$$

$$\leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_v^t + \rho\beta_{vz}\delta_z^t + 2\rho^2V_z^t + 3L_*^2\gamma^2V_x^t + \frac{16L_*^2\gamma^2}{\mu_G\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{84}$$

And finally we have

$$\boxed{\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_v^t + \rho\beta_{vz}\delta_z^t + 2\rho^2V_z^t + \beta_{vx}\gamma^2V_x^t + \overline{\beta}_{vx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2]} \tag{85}$$

with $\beta_{vz} = \frac{\omega^2}{2\mu_G}$, $\beta_{vx} = 3L_*^2$ and $\overline{\beta}_{vx} = \frac{16L_*^2\gamma^2}{\mu_G}$. □

C.6 Proof of Lemma 3.10

Proof. We use smoothness of h to get

$$\mathbb{E}_t[h(x^{t+1})] \leq h(x^t) - \gamma\langle D_x(z^t, v^t, x^t), \nabla h(x^t)\rangle + \frac{L^h}{2}\gamma^2\mathbb{E}_t[\|D_x^t\|^2] \tag{86}$$

$$\leq h(x^t) - \frac{\gamma}{2}(\|\nabla h(x^t)\|^2 + \|D_x(z^t, v^t, x^t)\|^2 - \|\nabla h(x^t) - D_x(z^t, v^t, x^t)\|^2) + \frac{L^h}{2}\gamma^2\mathbb{E}_t[\|D_x^t\|^2] \tag{87}$$

where the last inequality comes from the identity $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|)^2$. We take the total expectation and use the previous Lemma 3.4 to get

$$\boxed{h^{t+1} \leq h^t - \frac{\gamma}{2}g^t - \frac{\gamma}{2}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] + \frac{\gamma L_x^2}{2}(\delta_z^t + \delta_v^t) + \frac{L^h}{2}\gamma^2V_x^t} \tag{88}$$

□

C.7 Proof of Theorem 1

This section is devoted to the proof of Theorem 1 that we recall here.

Theorem 1 (Convergence of SOBA, fixed step size)**.** Fix an iteration $T > 1$ and assume that Assumptions 3.1 to 3.7 hold. We consider fixed steps $\rho^t = \frac{\overline{\rho}}{\sqrt{T}}$ and $\gamma^t = \xi\rho^t$ with $\overline{\rho}$ and ξ precised in the appendix. Let $(x^t)_{t\geq 1}$ the sequence of outer iterates for SOBA. Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla h(x^t)\|^2] = O(T^{-\frac{1}{2}}) \ .$$

The values of the differents constants are

$$\phi'_z = \frac{1}{8\overline{\beta}_{zx}}, \quad \phi'_v = \min\left(\frac{1}{8\overline{\beta}_{vx}}, \frac{\mu_G\phi'_z}{32\beta_{vz}}\right) \ , \quad \overline{\rho} = \min\left(\frac{16}{\mu_G}, \frac{\mu_G}{16L_z^2B_z^2}, \frac{\mu_G}{32L_v^2B_v^2}, \frac{\beta_{vz}}{L_v^2B_v^2}\right) \ ,$$
$$\text{and} \quad \xi^2 = \frac{\mu_G}{4}\min\left[\min\left(\frac{1}{L_{zx}^2}, \frac{1}{L_{vx}^2}\right)\frac{L_*^2}{B_x^2\overline{\rho}}, \min\left(\phi'_v, \phi'_z\right)\frac{1}{2L_x^2}\right] \ .$$

Before, one has to adapt our descent lemmas to the case of SOBA.

Lemma C.4. *Assume that the step sizes ρ and γ verify $\rho \leq \min\left(\frac{\mu_G}{16L_z^2B_z^2}, \frac{\mu_G}{32L_v^2B_v^2}, \frac{\beta_{vz}}{L_v^2B_v^2}\right)$ and $\gamma^2 \leq \min\left(\frac{\rho\mu_GL_*^2}{4B_x^2L_{zx}^2}, \frac{\rho\mu_GL_*^2}{8B_x^2L_{vx}^2}\right)$. Then it holds*

$$\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_z^t + 2\rho^2B_z^2 + \beta_{zx}\gamma^2B_x^2 + \overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{89}$$
$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{16}\right)\delta_v^t + 2\beta_{vz}\rho\delta_z^t + 2\rho^2B_v^2 + \beta_{vx}\gamma^2B_x^2 + \overline{\beta}_{vx}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{90}$$

Proof. From Assumption 3.6 and Lemma 3.4, we have

$$V_z^t \leq B_z^2(1 + D_z(z^t, v^t, x^t)) \leq B_z^2(1 + L_z^2\delta_z^t) \ .$$

Plugging this into Equation (69) and using $V_x^t \leq B_x^2$ yields

$$\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{4} + 2L_z^2B_z^2\rho^2\right)\delta_z^t + 2\rho^2B_z^2 + \beta_{zx}\gamma^2B_x^2 + \overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{91}$$

Since by assumption $\rho \leq \frac{\mu_G}{16L_z^2B_z^2}$, we have

$$\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_z^t + 2\rho^2B_z^2 + \beta_{zx}\gamma^2B_x^2 + \overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{92}$$

For δ_v^t, Assumption 3.3 and Lemma 3.4 provide us

$$V_v^t \leq B_v^2(1 + L_v^2(\delta_z^t + \delta_v^t)) \ .$$

Since the assumptions of Lemma 3.9 are verified, we can plug the previous inequality into Equation (85) to get

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8} + 2L_v^2B_v^2\rho^2\right)\delta_v^t + (\beta_{vz}\rho + 2L_v^2\rho^2B_v^2)\delta_z^t + 2\rho^2B_v^2 + \beta_{vx}\gamma^2B_x^2 + \overline{\beta}_{vx}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{93}$$

which can be simplified using $\rho \leq \min\left(\frac{\mu_G}{32L_v^2B_v^2}, \frac{\beta_{vz}}{L_v^2B_v^2}\right)$ to get finally

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{16}\right)\delta_v^t + 2\beta_{vz}\rho\delta_z^t + 2\rho^2B_v^2 + \beta_{vx}\gamma^2B_x^2 + \overline{\beta}_{vx}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ . \tag{94}$$

□

We can now prove Theorem 1.

Proof. Consider the Lyapunov function $\mathcal{L}^t = h^t + \phi_z \delta_z^t + \phi_v \delta_v^t$. Using the Equations (88), (69) and (85), we can bound $\mathcal{L}^{t+1} - \mathcal{L}^t$:

$$\begin{aligned}\mathcal{L}^{t+1} - \mathcal{L}^t \leq & -\frac{\gamma}{2} g^t - \left(\frac{\gamma}{2} - \phi_z \overline{\beta}_{zx} \frac{\gamma^2}{\rho} - \phi_v \overline{\beta}_{vx} \frac{\gamma^2}{\rho}\right) \mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \\ & - \left(\phi_z \frac{\mu_G}{8} \rho - \frac{L_x^2}{2} \gamma - 2\phi_v \beta_{vz} \rho\right) \delta_z^t \\ & - \left(\phi_v \frac{\mu_G}{16} \rho - \frac{L_x^2}{2} \gamma\right) \delta_v^t \\ & + \left(\frac{L^h}{2} + \phi_z \beta_{zx} + \phi_v \beta_{vx}\right) B_x^2 \gamma^2 \\ & + 2(\phi_z B_z^2 + \phi_v B_v^2) \rho^2 \ .\end{aligned} \tag{95}$$

Let $\phi_z' = \phi_z \frac{\gamma}{\rho}$ and $\phi_v' = \phi_v \frac{\gamma}{\rho}$, so that:

$$\begin{aligned}\mathcal{L}^{t+1} - \mathcal{L}^t \leq & -\frac{\gamma}{2} g^t - \left(\frac{\gamma}{2} - \phi_z' \overline{\beta}_{zx} \gamma - \phi_v' \overline{\beta}_{vx} \gamma\right) \mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \\ & - \left(\phi_z' \frac{\mu_G}{8} \frac{\rho^2}{\gamma} - \frac{L_x^2}{2} \gamma - 2\phi_v' \beta_{vz} \frac{\rho^2}{\gamma}\right) \delta_z^t \\ & - \left(\phi_v' \frac{\mu_G}{16} \frac{\rho^2}{\gamma} - \frac{L_x^2}{2} \gamma\right) \delta_v^t \\ & + \left(\frac{L^h}{2} + \phi_z' \beta_{zx} \frac{\rho}{\gamma} + \phi_v \beta_{vx} \frac{\rho}{\gamma}\right) B_x^2 \gamma^2 \\ & + 2\left(\phi_z' B_z^2 \frac{\rho}{\gamma} + \phi_v' B_v^2 \frac{\rho}{\gamma}\right) \rho^2 \ .\end{aligned} \tag{96}$$

In order to get a decrease, ϕ_z', ϕ_v', ρ and γ must verify

$$\begin{cases} \phi_z' \overline{\beta}_{zx} + \phi_v' \overline{\beta}_{vx} \leq \frac{1}{2} \\ \frac{L_x^2}{2} \gamma + 2\phi_v' \beta_{vz} \frac{\rho^2}{\gamma} \leq \phi_z' \frac{\mu_G}{8} \frac{\rho^2}{\gamma} \\ \frac{L_x^2}{2} \gamma \leq \phi_v' \frac{\mu_G}{16} \frac{\rho^2}{\gamma} \end{cases} \tag{97}$$

Let us take $\phi_z' = \frac{1}{8\overline{\beta}_{zx}}$ and $\phi_v' = \min\left(\frac{1}{8\overline{\beta}_{vx}}, \frac{\mu_G \phi_z'}{32\beta_{vz}}\right)$. We have

$$\phi_z' \overline{\beta}_{zx} + \phi_v' \overline{\beta}_{vx} \leq \frac{1}{4} < \frac{1}{2}$$

and

$$\frac{L_x^2}{2} \gamma + 2\phi_v' \beta_{vz} \frac{\rho^2}{\gamma} \leq \frac{L_x^2}{2} \gamma + \phi_z' \frac{\mu_G}{16} \frac{\rho^2}{\gamma} \ .$$

If we impose $\frac{L_x^2}{2} \gamma + \phi_z' \frac{\mu_G}{16} \frac{\rho^2}{\gamma} \leq \phi_z' \frac{\mu_G}{8} \frac{\rho^2}{\gamma}$, this combined with the third condition in Equation (97) gives the condition $\frac{L_x^2}{2} \gamma^2 \leq \min(\phi_v', \phi_z') \frac{\mu_G}{16} \rho^2$. We also have the conditions coming from the assumptions of C.4, that is

$$\rho \leq \overline{\rho} = \min\left(\frac{16}{\mu_G}, \frac{\mu_G}{16 L_z^2 B_z^2}, \frac{\mu_G}{32 L_v^2 B_v^2}, \frac{\beta_{vz}}{L_v^2}\right) \tag{98}$$

and $\gamma^2 \leq \min\left(\frac{1}{L_{zx}^2}, \frac{1}{L_{vx}^2}\right) \frac{\mu_G L_*^2}{4 B_x^2} \rho$. Let us take $\rho = \frac{\overline{\rho}}{\sqrt{T}}$ with $\gamma = \xi \rho$ where ξ is defined as

$$\xi^2 \triangleq \frac{\mu_G}{4} \min\left[\min\left(\frac{1}{L_{zx}^2}, \frac{1}{L_{vx}^2}\right) \frac{L_*^2}{B_x^2 \overline{\rho}}, \min(\phi_v', \phi_z') \frac{1}{2L_x^2}\right] \ . \tag{99}$$

From now, we have

$$\mathcal{L}^{t+1} - \mathcal{L}^t \leq -\frac{\gamma}{2} g^t + \frac{L^h}{2} B_x^2 \gamma^2 + (\phi_z' \beta_{zx} + \phi_v' \beta_{vx}) B_x^2 \rho \gamma + 2\left(\phi_z' B_z^2 + \phi_v' B_v^2\right) \frac{\rho^3}{\gamma} \ . \tag{100}$$

Summing and telescoping yields

$$\frac{1}{T}\sum_{t=1}^{T} g^t \leq \frac{2\mathcal{L}^1}{T\gamma} + L^h B_x^2 \gamma + 2\left(\phi_z' \beta_{zx} + 2\phi_v' \beta_{vx}\right) B_x^2 \rho + 4\left(\phi_z' B_z^2 + \phi_v' B_v^2\right) \frac{\rho^3}{\gamma^2} \tag{101}$$

$$\leq \frac{2\mathcal{L}^1}{\sqrt{T}\xi\overline{\rho}} + L^h B_x^2 \frac{\xi\alpha}{\sqrt{T}} + \left(\phi_z' \beta_{zx} + 2\phi_v' \beta_{vx}\right) B_x^2 \frac{\alpha}{\sqrt{T}} + 4\left(\phi_z' B_z^2 + \phi_v' B_v^2\right) \frac{\alpha}{\xi^2\sqrt{T}} \tag{102}$$

$$\tag{103}$$

and so

$$\boxed{\frac{1}{T}\sum_{t=1}^{T} g^t = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) .}$$

□

C.8 Proof of Theorem 2

Proof. In the decreasing step size case, we take $\rho^t = \overline{\rho}\sqrt{t}$ and $\gamma^t = \xi\rho^t$ where $\overline{\rho}$ is defined in Equation (98) and ξ is defined in Equation (99). We recall the integral majorization:

$$\sum_{t=1}^{T} t^{-1} \leq 1 + \int_1^T t^{-1} \mathrm{d}t = 1 + \log(T) .$$

With such definition of ρ^t and γ^t, Equation (100) is still valid for any $t \geqslant 1$. The only difference is that the step sizes decrease with t. Hence, by summing and rearranging in Equation (100), we get

$$\sum_{t=1}^{T} \gamma^t g^t \leq 2\mathcal{L}^1 + \left(L^h + 2\left(\phi_z' \beta_{zx} + 2\phi_v' \beta_{vx}\right) B_x^2 \frac{1}{\xi} + 4\left(\phi_z' B_z^2 + \phi_v' B_v^2\right) \frac{1}{\xi^3}\right) \sum_{t=1}^{T} (\gamma^t)^2 \tag{104}$$

The left-hand-side in Equation (104) can be lower bounded by

$$\sum_{t=1}^{T} \gamma^t g^t \geq \left(\inf_{t\in[T]} g^t\right) \xi\overline{\rho} \sum_{t=1}^{T} t^{-\frac{1}{2}} \geq \left(\inf_{t\in[T]} g^t\right) \xi\overline{\rho} T^{\frac{1}{2}} . \tag{105}$$

Also we have

$$\sum_{t=1}^{T} (\gamma^t)^2 = \xi^2\overline{\rho}^2 \sum_{t=1}^{T} t^{-1} \leq \xi^2\overline{\rho}^2 (1 + \log(T)) . \tag{106}$$

Plugging Equations (105) and (106) into Equation (104) and rearranging give

$$\inf_{t\in[T]} g^t \leq \frac{2\mathcal{L}^1}{\xi\overline{\rho}\sqrt{T}} + \xi\overline{\rho}\left(L^h + 2\left(\phi_z' \beta_{zx} + 2\phi_v' \beta_{vx}\right) B_x^2 \frac{1}{\xi} + 4\left(\phi_z' B_z^2 + \phi_v' B_v^2\right) \frac{1}{\xi^3}\right) \frac{1 + \log(T)}{\sqrt{T}} \tag{107}$$

that is to say

$$\boxed{\inf_{t\in[T]} g^t = \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{\log(T)}{\sqrt{T}}\right) .} \tag{108}$$

□

C.9 Proof of Theorem 3

In this section, we prove Theorem 3 that we recall here

Theorem 3 (Convergence of SABA, smooth case)**.** Assume that Assumptions 3.1 to 3.3 and 3.7 to 3.8 hold. We suppose $\rho = \rho' N^{-\frac{2}{3}}$ and $\gamma = \xi\rho$, where ρ' and ξ depend only on F and G and are specified in appendix. Let x^t the iterates of SABA. Then,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla h(x^t)\|^2] = O\left(N^{\frac{2}{3}} T^{-1}\right) .$$

The constants ρ' and ξ are given by

$$\rho' = \min\left(\sqrt{\frac{K_1}{K_5}}, \left(\frac{K_2}{K_5}\right)^{\frac{2}{5}}, \left(\frac{K_3}{K_5}\right)^{\frac{5}{7}}, \left(\frac{K_4}{K_5}\right)^{\frac{1}{3}}, \frac{\mu_G}{64L_z^2}, \frac{\overline{\beta}_{zx}}{2\beta_{zx}}, \frac{\mu_G}{128(L_v^2+L_v'')}, \frac{\beta_{vz}}{8(L_v^2+L_v'')}, \frac{\overline{\beta}_{vx}}{2\beta_{vx}}\right) \tag{109}$$

and

$$\xi = \min(K_1, K_2(\rho')^{-\frac{1}{2}}, K_3(\rho')^{-\frac{3}{2}}, K_4(\rho')^{-1}) \tag{110}$$

where

$$\phi_z'' = \frac{1}{32\overline{\beta}_{zx}}, \quad \phi_v'' = \min\left(\frac{1}{32\overline{\beta}_{vx}}, \phi_z'' \frac{\mu_G}{128\beta_{vz}}\right) \ ,$$

$$K_1 = \min\left(\sqrt{\frac{\phi_z''\mu_G}{32L_x^2}}, \sqrt{\frac{\phi_v''\mu_G}{48L_x^2}}, \sqrt{\frac{L_z'}{2L_x'\beta_{zx}}}, \sqrt{\frac{L_v'}{2L_x'\beta_{vx}}}\right) \ ,$$

$$K_2 = \min\left(\sqrt{\frac{\mu_G}{64\beta_{zx}L_x''}}, \sqrt{\frac{\mu_G}{128\beta_{vx}L_x''}}, \sqrt{\frac{\beta_{vz}}{4L_x''\beta_{vx}}}\right) \ ,$$

$$K_3 = \sqrt{\frac{\phi_v''\mu_G}{384\phi_z''L_x''}} \ , \quad K_4 = \min\left(\frac{1}{4L^h}, \frac{L_x^2}{2L^hL_x''}, \sqrt{\frac{\Gamma'}{6L^hL_x'}}, \frac{1}{8P'}, \frac{\phi_z''\mu_G}{32\beta_{sz}'}, \frac{\phi_v''\mu_G}{48\beta_{sv}'}\right)$$

and

$$K_5 = \frac{15(\phi_z''L_z' + \phi_v''L_v')}{\Gamma'} \ .$$

C.9.1 Control of distance from memory to iterates

We can view our method has having two "parallel" memories for each variable (z_i^t, v_i^t, x_i^t) for $i \in 1[n]$ corresponding to calls in G and $(z_j'^t, v_j'^t, x_j'^t)$ for $j \in [m]$ corresponding to calls to F. At each iteration, we sample i at random uniformly and do $(z_i^{t+1}, v_i^{t+1}, x_i^{t+1}) = (z^t, v^t, x^t)$ and $(z_{i'}^{t+1}, v_{i'}^{t+1}, x_{i'}^{t+1}) = (z_{i'}^t, v_{i'}^t, x_{i'}^t)$ for $i' \neq i$, and similarly for the other memory.

In what follows, we focus on controlling the error between the iterates and the memories. We define to make things simpler

$$E_z^t = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|z^t - z_i^t\|^2] \ , \quad E_v^t = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|v^t - v_i^t\|^2] \ , \quad E_x^t = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x^t - x_i^t\|^2] \ ,$$

and similarly $E_x'^t, E_v'^t$ and $E_x'^t$.

Lemma C.5. *We have the following inequalities:*

$$E_z^{t+1} \leq \left(1 - \frac{1}{2n}\right) E_z^t + \rho^2 \mathbb{E}\|D_z^t\|^2 + 2n\rho^2 \mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] \ ,$$

$$E_v^{t+1} \leq \left(1 - \frac{1}{2n}\right) E_v^t + \rho^2 \mathbb{E}\|D_v^t\|^2 + 2n\rho^2 \mathbb{E}[\|D_v(z^t, v^t, x^t)\|^2] \ ,$$

$$E_x^{t+1} \leq \left(1 - \frac{1}{2n}\right) E_x^t + \gamma^2 \mathbb{E}\|D_x^t\|^2 + 2n\gamma^2 \mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ ,$$

$$E_z'^{t+1} \leq \left(1 - \frac{1}{2m}\right) E_z^t + \rho^2 \mathbb{E}\|D_z^t\|^2 + 2m\rho^2 \mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] \ ,$$

$$E_v'^{t+1} \leq \left(1 - \frac{1}{2m}\right) E_v^t + \rho^2 \mathbb{E}\|D_v^t\|^2 + 2m\rho^2 \mathbb{E}[\|D_v(z^t, v^t, x^t)\|^2] \ ,$$

and

$$E_x'^{t+1} \leq \left(1 - \frac{1}{2m}\right) E_x^t + \gamma^2 \mathbb{E}\|D_x^t\|^2 + 2m\gamma^2 \mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ .$$

Proof. We provide the detailed proof for E_z^t. The approach for the five others is similar.
Let $i \in [n]$. Taking the expectation of $\|z^{t+1} - z_i^{t+1}\|^2$ conditionaly to z^t, v^t, x^t yields

$$\mathbb{E}_t[\|z^{t+1} - z_i^{t+1}\|^2] = \frac{1}{n}\mathbb{E}_t[\|z^{t+1} - z^t\|^2] + \frac{n-1}{n}\mathbb{E}_t[\|z^{t+1} - z_i^t\|^2] \ .$$

Then, using the fact that $\mathbb{E}_t[D_z^t(z^t, v^t, x^t)] = D_z(z^t, v^t, x^t)$, we have

$$\mathbb{E}_t[\|z^{t+1} - z_i^t\|^2] = \mathbb{E}_t[\|z^{t+1} - z^t\|^2] + \|z^t - z_i^t\|^2 - 2\rho\langle D_z(z^t, v^t, x^t), z^t - z_i^t\rangle \ . \tag{111}$$

We then upper-bound crudely the scalar product by Cauchy-Schwarz and Young inequalities with parameter β:

$$\mathbb{E}_t[\|z^{t+1} - z_i^t\|^2] \leq \mathbb{E}_t[\|z^{t+1} - z^t\|^2] + \rho\beta^{-1}\|D_z(z^t, v^t, x^t)\|^2 + (1+\rho\beta)\|z^t - z_i^t\|^2$$

As a consequence, by taking the total expectation and summing for all $i \in [n]$, we find

$$E_z^{t+1} \leq \rho^2\mathbb{E}[\|D_z^t\|^2] + \rho\beta^{-1}\left(1 - \frac{1}{n}\right)\mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] + (1+\rho\beta)\left(1 - \frac{1}{n}\right)E_z^t \ .$$

Finally, we take $\beta = \frac{1}{2n\rho}$ to obtain

$$\boxed{E_z^{t+1} \leq \left(1 - \frac{1}{2n}\right)E_z^t + \rho^2\mathbb{E}\|D_z^t(z^t, v^t, x^t)\|^2 + 2n\rho^2\mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] \ .} \tag{112}$$

□

C.9.2 Bounds on the variances

The following lemma gives us upper-bounds for $\mathbb{E}[\|D_z^t(z^t, v^t, x^t)\|^2]$, $\mathbb{E}[\|D_v^t(z^t, v^t, x^t)\|^2]$, and $\mathbb{E}[\|D_x^t(z^t, v^t, x^t)\|^2]$.

Lemma C.6. *For SABA, there are constants $L_z', L_v', L_x' > 0$ such that*

$$\mathbb{E}[\|D_z^t(z^t, v^t, x^t)\|]^2 \leq 2\mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] + 2L_z'(E_z^t + E_x^t) \ ,$$

$$\mathbb{E}[\|D_v^t(z^t, v^t, x^t)\|^2] \leq 2\mathbb{E}[\|D_v(z^t, v^t, x^t)\|^2] + 2L_v'(E_z^t + E_x^t + E_v^t + E_z'^t + E_x'^t) + 2L_v''(\delta_z^t + \delta_v^t)$$

and

$$\mathbb{E}[\|D_x^t(z^t, v^t, x^t)\|^2] \leq 2\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] + 2L_x'(E_z^t + E_x^t + E_v^t + E_z'^t + E_x'^t) + 2L_x''(\delta_z^t + \delta_v^t) \ .$$

Proof. For SABA, if we consider i sampled from $[n]$ at iteration t, we have

$$D_z^t = \nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t) + \frac{1}{n}\sum_{i'=1}^n \nabla_1 G_{i'}(z_{i'}^t, x_{i'}^t) \ .$$

Hence we get

$$\begin{aligned}\mathbb{E}_t[\|D_z^t(z^t, v^t, x^t)\|^2] &= \mathbb{E}_t[\|\nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t) + \frac{1}{n}\sum_{i'=1}^N \nabla_1 G_{i'}(z_{i'}^t, x_{i'}^t) \\ &\qquad - \nabla_1 G(z^t, x^t) + \nabla_1 G(z^t, x^t)\|^2] \\ &\leq 2\|\nabla_1 G(z^t, x^t)\|^2 + 2\mathbb{E}_t[\|\nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t) \\ &\qquad + \frac{1}{n}\sum_{i'=1}^N \nabla_1 G_{i'}(z_{i'}^t, x_{i'}^t) - \nabla_1 G(z^t, x^t)\|^2] \ .\end{aligned} \tag{113}$$

The second term is the variance of $\nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t)$, which is therefore upper-bounded by

$$\mathbb{E}_t[\|[\nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t)\|^2] = \frac{1}{n}\sum_{i=1}^{n} \|[\nabla_1 G_i(z^t, x^t) - \nabla_1 G_i(z_i^t, x_i^t)\|^2$$
$$\leq \frac{L_z'}{n}\sum_{i=1}^{n}(\|z^t - z_i^t\|^2 + \|x^t - x_i^t\|^2) \tag{114}$$

where the inequality comes from the Lipschitz continuity of each $\nabla_1 G_i$ with $L_z' = \max_{i\in[n]} L_1^{G_i}$. Then, by plugging (114) into (113) and taking the total expectation, we get

$$\boxed{\mathbb{E}[\|D_z^t(z^t, v^t, x^t)\|]^2 \leq 2\mathbb{E}[\|D_z(z^t, v^t, x^t)\|^2] + 2L_z'(E_z^t + E_x^t)\ .} \tag{115}$$

Things are quite similar for the other variables, albeit a bit more difficult.
In v, it holds

$$\mathbb{E}_t[\|D_v^t(z^t, v^t, x^t)\|^2] = \mathbb{E}_t[\|\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t) + \frac{1}{m}\sum_{j'=1}^{m}\nabla_1 F_{j'}(z_{j'}'^t, x_{j'}'^t) \tag{116}$$
$$+ \nabla_{11}^2 G_i(z^t, x^t)v^t - \nabla_{11}^2 G_i(z_i^t, x_i^t)v_i^t + \frac{1}{n}\sum_{i'=1}^{n}\nabla_1^2 G_{i'}(z_{i'}^t, x_{i'}^t)v_{i'^t}$$
$$- D_v(z^t, v^t, x^t) + D_v(z^t, v^t, x^t)\|^2]$$
$$\leq 2[\|D_v(z^t, v^t, x^t)\|^2 \tag{117}$$
$$+ 2\mathbb{E}_t[\|\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t) + \frac{1}{m}\sum_{j'=1}^{m}\nabla_1 F_{j'}(z_{j'}'^t, x_{j'}'^t)$$
$$+ \nabla_{11}^2 G_i(z^t, x^t)v^t - \nabla_{11}^2 G_i(z_i^t, x_i^t)v_i^t + \frac{1}{n}\sum_{i'=1}^{n}\nabla_1 G_{i'}^2(z_{i'}^t, x_{i'}^t)v_{i'}^t$$
$$- D_v(z^t, v^t, x^t)\|^2] \tag{118}$$

Here, we see that we need to control the variance of $\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t) + \nabla_{11}^2 G_i(z^t, x^t)v^t - \nabla_{11}^2 G_i(z_i^t, x_i^t)v_i^t$. Since i and j are independent, this is a sum of two independent random variables, hence its variance is the sum of the variances, which is upper-bounded by

$$\mathbb{E}_t[\|\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t)\|^2] + \mathbb{E}_t[\|\nabla_{11}^2 G_i(z^t, x^t)v^t - \nabla_{11}^2 G_i(z_i^t, x_i^t)v_i^t\|^2]\ .$$

For $\mathbb{E}_t[\|\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t)\|^2]$ we use the lipschitz continuity of the $\nabla_1 F_j$:

$$\mathbb{E}_t[\|\nabla_1 F_j(z^t, x^t) - \nabla_1 F_j(z_j'^t, x_j'^t)\|^2] \leq \left[\max_{j\in[m]} L_1^{F_j}\right]\mathbb{E}_t[\|z^t - z_j^t\|^2 + \|x^t - x_j^t\|^2] \tag{119}$$
$$\leq \left[\max_{j\in[m]} L_1^{F_j}\right]\frac{1}{m}\sum_{j=1}^{m}(\|z^t - z_j^t\|^2 + \|x^t - x_j^t\|^2)\ . \tag{120}$$

The control of $\mathbb{E}_t[\|\nabla^2_{11}G_i(z^t,x^t)v^t-\nabla^2_{11}G_i(z_i^t,x_i^t)v_i^t\|^2]$ is a bit harder without assuming the boundness of v beforehand. But, we can bypass the difficulty by introducing $\nabla^2_{11}G_i(z^*(x^t),x^t)v^*(x^t)$:

$$\begin{aligned}\mathbb{E}_t[\|\nabla^2_{11}G_i(z^t,x^t)v^t-\nabla^2_{11}G_i(z_i^t,x_i^t)v_i^t\|^2] &\leq 4\{\mathbb{E}_t[\|\nabla^2_{11}G_i(z^t,x^t)(v^t-v^*(x^t))\|^2] \qquad (121)\\ &\quad+\mathbb{E}_t[\|(\nabla^2_{11}G_i(z^t,x^t)-\nabla^2_{11}G_i(z^*(x^t),x^t))v^*(x^t)\|^2]\\ &\quad+\mathbb{E}_t[\|(\nabla^2_{11}G_i(z^*(x^t),x^t)-\nabla^2_{11}G_i(z_i^t,x_i^t))v^*(x^t)\|^2]\\ &\quad+\mathbb{E}_t[\|\nabla^2_{11}G_i(z_i^t,x_i^t)(v^*(x^t)-v_i^t)\|^2]\}\\ &\leq 4((\max_{i\in[n]}L_1^{G_i})\|v^t-v^*(x^t)\|^2+(\max_{i\in[n]}L_2^{G_i})\frac{C^F}{\mu_G}\|z^t-z^*(x^t)\|^2 \qquad (122)\\ &\quad+(\max_{i\in[n]}L_2^{G_i})\frac{C^F}{\mu_G}(\|x^t-x_i^t\|^2+2(\|z^t-z^*(x^t)\|^2+\|z^t-z_i^t\|^2))\\ &\quad+(\max_{i\in[n]}L_1^{G_i})(\|x^t-x_i^t\|^2+2(\|v^t-v^*(x^t)\|^2+\|v^t-v_i^t\|^2))\end{aligned}$$

Let $L'_v = 4\max\left(2\max_{i\in[n]}L_1^{G_i}, 2\max_{i\in[n]}L_2^{G_i}\frac{C^F}{\mu_G}, \max_{j\in[m]}L_1^{F_j}\right)$ and $L''_v = 4\max\left(3\max_{i\in[n]}L_1^{G_i}), 3\max_{i\in[n]}L_2^{G_i})\frac{C^F}{\mu_G}\right)$. Taking the total expectation and putting all together yields

$$\boxed{\mathbb{E}[\|D_v^t(z^t,v^t,x^t)\|^2]\leq 2\mathbb{E}[\|D_v(z^t,v^t,x^t)\|^2]+2L'_v(E_z^t+E_x^t+E_v^t+E_z'^t+E_x'^t)+2L''_v(\delta_z^t+\delta_v^t)\ .} \qquad (123)$$

In x we have similarly

$$\boxed{\mathbb{E}[\|D_x^t(z^t,v^t,x^t)\|^2]\leq 2\mathbb{E}[\|D_x(z^t,v^t,x^t)\|^2]+2L'_x(E_z^t+E_x^t+E_v^t+E_z'^t+E_x'^t)+2L''_x(\delta_z^t+\delta_v^t)\ .} \qquad (124)$$

□

We now form $S^t=E_z^t+E_x^t+E_v^t+E_z'^t+E_v'^t+E_x'^t$, and letting $\Gamma=\min(\frac{1}{m},\frac{1}{n})$. Note that by definition, each quantity E_z^t is smaller than S^t.

We will therefore use the cruder bounds on $\mathbb{E}[\|D_z^t\|^2]$, $\mathbb{E}[\|D_v^t\|^2]$ and $\mathbb{E}[\|D_x^t\|^2]$ as follows thanks to Lemma 3.4 and Lemma C.6

$$\mathbb{E}[\|D_z^t(z^t,v^t,x^t)\|^2]\leq 2L_z^2\delta_z^t+2L'_zS^t\ , \qquad (125)$$

$$\mathbb{E}[\|D_v^t(z^t,v^t,x^t)\|^2]\leq 2(L_v^2+L''_v)(\delta_z^t+\delta_v^t)+2L'_vS^t \qquad (126)$$

and

$$\mathbb{E}[\|D_x^t(z^t,v^t,x^t)\|^2]\leq 2\mathbb{E}[\|D_x\|^2]+2L'_xS^t+2L''_x(\delta_z^t+\delta_v^t)\ . \qquad (127)$$

We have the following lemma

Lemma C.7. *If* $4\rho^2(L'_z+L'_v)+4\gamma^2L'_x\leq\frac{\Gamma}{2}$ *and* $4L''_x\gamma^2\leq\rho^2(L_v^2+4L''_v)$*, it holds*

$$S^{t+1}\leq\left(1-\frac{\Gamma}{2}\right)S^t+\beta_{sz}\rho^2\delta_z^t+\beta_{sv}\rho^2\delta_v^t+P\gamma^2\mathbb{E}[\|D_x\|^2]$$

for some $L_s,\beta_{sz},P>0$.

Proof. It holds following eq. (112) (and omitting the dependencies in (z^t,v^t,x^t) in the direction for simplicity)

$$\begin{aligned}S^{t+1}\leq(1-\Gamma)\,S^t+\mathbb{E}\,\big[&2\rho^2(\|D_z^t\|^2+\|D_v^t\|^2)+2\gamma^2\|D_x^t\|^2\\ &+2(m+n)[\rho^2(\|D_z\|^2+\|D_v\|^2)+\gamma^2\|D_x\|^2]\big]\ .\end{aligned}$$

Using the previous bounds (115), (123) and (124), we get

$$\begin{aligned}S^{t+1}\leq\big(1-\Gamma+4\rho^2(L'_z+L'_v)+4\gamma^2L'_x\big)\,S^t+(2(m+n)+4)\mathbb{E}[\rho^2(\|D_z\|^2+\|D_v\|^2)\\ +\gamma^2\|D_x\|^2]+4L''_v\rho^2(\delta_z^t+\delta_v^t)+4L''_x\gamma^2(\delta_z^t+\delta_v^t)\ .\end{aligned}$$

Next, using $4\rho^2(L'_z+L'_v)+4\gamma^2L'_x \leq \frac{\Gamma}{2}$ and letting $P=(2(m+n)+4)$ we get

$$S^{t+1} \leq \left(1-\frac{\Gamma}{2}\right)S^t + P\mathbb{E}[\rho^2(\|D_z\|^2+\|D_v\|^2)+\gamma^2\|D_x\|^2] + +4L''_v\rho^2(\delta_z^t+\delta_v^t)+4L''_x\gamma^2(\delta_z^t+\delta_v^t) \ .$$

To finish, we use Lemma 3.4 to get

$$S^{t+1} \leq \left(1-\frac{\Gamma}{2}\right)S^t + P[\rho^2((L_z^2+L_v^2)\delta_z^t+L_v^2\delta_v^t)+(4L''_v\rho^2+4L''_x\gamma^2)(\delta_z^t+\delta_v^t)+\gamma^2\mathbb{E}[\|D_x\|^2]] \ .$$

Then, using that $4L''_x\gamma^2 \leq \rho^2(L_v^2+4L''_v)$, we get the bound, letting $L_{sz}=L_z^2+L_v^2+4L''_v$ and $L_{sv}=L_v^2+4L''_v$:

$$\boxed{S^{t+1} \leq \left(1-\frac{\Gamma}{2}\right)S^t + \beta_{sz}\rho^2\delta_z^t + \beta_{sv}\rho^2\delta_v^t + P\gamma^2\mathbb{E}[\|D_x\|^2]}$$

with $\beta_{sz}=2PL_{sz}$, $\beta_{sv}=2PL_{sv}$ □

C.9.3 Putting it all together

Recall that we denote $g^t=\mathbb{E}[\|\nabla h(x^t)\|^2]$ and $h^t=\mathbb{E}[h(x^t)]$. In the following lemma, we adapt Lemma 3.9 and Lemma 3.10 to the SABA algorithm.

Lemma C.8. *If*

$$\rho \leq \min\left(\frac{\mu_G}{64L_z^2}, \frac{\overline{\beta}_{zx}}{2\beta_{zx}}, \frac{\mu_G}{128(L_v^2+L''_v)}, \frac{\beta_{vz}}{8(L_v^2+L''_v)}, \frac{\overline{\beta}_{vx}}{2\beta_{vx}}\right)$$

and

$$\gamma \leq \min\left(\sqrt{\frac{\rho\mu_G}{64\beta_{zx}L''_x}}, \sqrt{\frac{L'_z}{2L'_x\beta_{zx}}}\rho, \sqrt{\frac{\rho\mu_G}{128\beta_{vx}L''_x}}, \sqrt{\frac{\rho\beta_{vz}}{4L''_x\beta_{vx}}}, \sqrt{\frac{L'_v}{2L'_x\beta_{vx}}}\rho, \frac{1}{4L^h}, \frac{L_x^2}{2L^hL''_x}\right)$$

then it holds

$$\delta_z^{t+1} \leq \left(1-\frac{\rho\mu_G}{8}\right)\delta_z^t + 2L''_x\beta_{zx}\gamma^2\delta_v^t + 5L'_z\rho^2S^t + 2\overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t,v^t,x^t)\|^2] \ , \tag{128}$$

$$\delta_v^{t+1} \leq \left(1-\frac{\rho\mu_G}{16}\right)\delta_v^t + 3\beta_{vz}\rho\delta_z^t + 5L'_v\rho^2S^t + 2\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t,v^t,x^t)\|^2] \tag{129}$$

and

$$h^{t+1} \leq h^t - \frac{\gamma}{2}g^t - \frac{\gamma}{4}\mathbb{E}[\|D_x(z^t,v^t,x^t)\|^2] + L_x^2\gamma(\delta_z^t+\delta_v^t) + L^hL'_x\gamma^2S^t \ . \tag{130}$$

Proof. We start from Lemma 3.9 and plug the bounds of Equations (125) and (126).

$$\begin{aligned}\delta_z^{t+1} \leq & \left(1-\frac{\rho\mu_G}{4}+4L_z^2\rho^2+4\beta_{zx}L''_x\gamma^2\right)\delta_z^t + 2L''_x\beta_{zx}\gamma^2\delta_v^t \\ & + (4L'_z\rho^2+2L'_x\beta_{zx}\gamma^2)S^t + \left(2\beta_{zx}\gamma^2+\overline{\beta}_{zx}\frac{\gamma^2}{\rho}\right)\mathbb{E}[\|D_x(z^t,v^t,x^t)\|^2]\end{aligned} \tag{131}$$

Since $\rho \leq \frac{\mu_G}{64L_z^2}$ and $\gamma^2 \leq \frac{\rho\mu_G}{64\beta_{zx}L''_x}$, we have

$$-\frac{\rho\mu_G}{4}+4L_z^2\rho^2+4\beta_{zx}L''_x\gamma^2 \leq -\frac{\rho\mu_G}{8} \ . \tag{132}$$

The condition $\gamma^2 \leq \frac{L'_z}{2L'_x\beta_{zx}}\rho^2$ gives us

$$4L'_z\rho^2+2L'_x\beta_{zx}\gamma^2 \leq 5L'_z\rho^2 \ . \tag{133}$$

With $\rho \leq \frac{\overline{\beta}_{zx}}{2\beta_{zx}}$, we get

$$2\beta_{zx}\gamma^2+\overline{\beta}_{zx}\frac{\gamma^2}{\rho} \leq 2\overline{\beta}_{zx}\frac{\gamma^2}{\rho} \ . \tag{134}$$

We can plug Equations (132), (133) and (134) into Equation (131) and we end up with

$$\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_z^t + 2L_x''\beta_{zx}\gamma^2\delta_v^t + 5L_z'\rho^2 S^t + 2\overline{\beta}_{zx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ .$$

The proof for δ_v^t is quite similar. From Lemma 3.9, Equations (126) and (127).

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{8}\right)\delta_v^t + \beta_{vz}\rho\delta_z^t + 2\rho^2 V_v^t + \beta_{vx}\gamma^2 V_x^t + \overline{\beta}_{vx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{135}$$

$$\leq \left(1 - \frac{\rho\mu_G}{8} + 4(L_v^2 + L_v'')\rho^2 + 4L_x''\beta_{vx}\gamma^2\right)\delta_v^t + (4(L_v^2 + L_v'')\rho^2 + 2L_x''\beta_{vx}\gamma^2 + \beta_{vz}\rho)\delta_z^t + \tag{136}$$

$$+ \left(4L_v'\rho^2 + 2L_x'\beta_{vx}\gamma^2\right)S^t + \left(2\beta_{vx}\gamma^2 + \overline{\beta}_{vx}\frac{\gamma^2}{\rho}\right)\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ .$$

Using $\rho \leq \frac{\mu_G}{128(L_v^2 + L_v'')}$ and $\gamma^2 \leq \frac{\rho\mu_G}{128L_x''\beta_{vx}}$, we get

$$-\frac{\rho\mu_G}{8} + 4(L_v^2 + L_v'')\rho^2 + 4L_x''\beta_{vx}\gamma^2 \leq -\frac{\rho\mu_G}{16} \ . \tag{137}$$

With $\gamma^2 \leq \frac{\rho\beta_{vz}}{4L_x''\beta_{vx}}$ and $\rho \leq \frac{\beta_{vz}}{8(L_v^2 + L_v'')}$, we have

$$4(L_v^2 + L_v'')\rho^2 + 2L_x''\beta_{vx}\gamma^2 + \beta_{vz}\rho \leq 3\beta_{vz}\rho \ . \tag{138}$$

The condition $\gamma^2 \leq \frac{L_v'}{2L_x'\beta_{vx}}\rho^2$ yields

$$4L_v'\rho^2 + 2L_x'\beta_{vx}\gamma^2 \leq 5L_v'\rho^2 \ . \tag{139}$$

With $\rho \leq \frac{\overline{\beta}_{vx}}{2\beta_{vx}}$ we get

$$2\beta_{vx}\gamma^2 + \overline{\beta}_{zx}\frac{\gamma^2}{\rho} \leq 2\overline{\beta}_{vx}\frac{\gamma^2}{\rho} \ . \tag{140}$$

As a consequence of Equations (135), (137), (138), (139) and (140), we have

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{16}\right)\delta_v^t + 3\beta_{vz}\rho\delta_z^t + 5L_v'\rho^2 S^t + 2\overline{\beta}_{vx}\frac{\gamma^2}{\rho}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \ .$$

For the inequality on h^t, we start from Equations (88) and (127)

$$h^{t+1} \leq h^t - \frac{\gamma}{2}g^t - \left(\frac{\gamma}{2} - L^h\gamma^2\right)\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \tag{141}$$

$$+ \left(\frac{L_x^2}{2}\gamma + L^h L_x''\gamma^2\right)(\delta_z^t + \delta_v^t) + L^h L_x'\gamma^2 S^t \ .$$

Assuming $\gamma \leq \min\left(\frac{1}{4L^h}, \frac{L_x^2}{2L^h L_x''}\right)$ leads

$$h^{t+1} \leq h^t - \frac{\gamma}{2}g^t - \frac{\gamma}{4}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] + L_x^2\gamma(\delta_z^t + \delta_v^t) + L^h L_x'\gamma^2 S^t \ . \tag{142}$$

□

We are now ready to prove Theorem 3.

Proof. We consider the Lyapunov function

$$\mathcal{L}^t = h^t + \phi_s S^t + \phi_z\delta_z^t + \phi_v\delta_v^t \tag{143}$$

for some constants ϕ_s, ϕ_z and ϕ_v.

We have

$$\begin{aligned}\mathcal{L}^{t+1} - \mathcal{L}^t \leq & -\frac{\gamma}{2} g^t - \left(\frac{\gamma}{4} - 2\phi_z \overline{\beta}_{zx} \frac{\gamma^2}{\rho} - 2\phi_v \overline{\beta}_{vx} \frac{\gamma^2}{\rho} - \phi_s P \gamma^2\right) \mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \\ & - \left(\phi_z \frac{\mu_G}{8} \rho - L_x^2 \gamma - 8\phi_v \beta_{vz} \rho - \phi_s \beta_{sz} \rho^2\right) \delta_z^t \\ & - \left(\phi_v \frac{\mu_G}{16} \rho - L_x^2 \gamma - 2\phi_z L_x'' \gamma^2 - \phi_s \beta_{sv} \rho^2\right) \delta_v^t \\ & - \left(\phi_s \frac{\Gamma}{2} - 5\phi_z L_z' \rho^2 - 5\phi_v L_v' \rho^2 - L^h L_x' \gamma^2\right) S^t \ .\end{aligned}$$

To get a decrease, ϕ_z, ϕ_v and ϕ_s, ρ and γ must be such that:

$$\begin{aligned} 2\phi_z \overline{\beta}_{zx} \frac{\gamma^2}{\rho} + 2\phi_v \overline{\beta}_{vx} \frac{\gamma^2}{\rho} + \phi_s P \gamma^2 &\leq \frac{\gamma}{4} \\ L_x^2 \gamma + 8\phi_v \beta_{vz} \rho + \phi_s \beta_{sz} \rho^2 &\leq \phi_z \frac{\mu_G}{8} \rho \\ L_x^2 \gamma + 8\phi_z L_x'' \gamma^2 + \phi_s \beta_{sv} \rho^2 &\leq \phi_v \frac{\mu_G}{16} \rho \\ 5\phi_z L_z' \rho^2 + 5\phi_v L_v' \rho^2 + L^h L_x' \gamma^2 &\leq \phi_s \frac{\Gamma}{2} \ . \end{aligned}$$

In order to take into account the scaling of the quantities with respect to $N = n + m$, we take $\rho = \rho' N^{n_\rho}$, $\gamma = \gamma' N^{n_\gamma}$, $\phi_z = \phi_z' N^{n_z}$, $\phi_v = \phi_v' N^{n_v}$ and $\phi_s = \phi_s' N^{n_s}$. Since $\Gamma = \mathcal{O}(N^{-1})$, $P = \mathcal{O}(N)$, $\beta_{sz} = \mathcal{O}(N)$ and $\beta_{sv} = \mathcal{O}(N)$, we also define $\Gamma' = \Gamma N$, $P' = PN^{-1}$, $\beta_{sz}' = \beta_{sz} N^{-1}$ and $\beta_{sv}' N^{-1}$. Now, the previous Equations read (after slight simplifications):

$$\begin{aligned} (2\phi_z' \overline{\beta}_{zx} + 2\phi_v' \overline{\beta}_{vx}) \frac{\gamma'}{\rho'} N^{n_z + n_\gamma - n_\rho} + \phi_s' P' \gamma' N^{n_s + n_\gamma + 1} &\leq \frac{1}{4} \\ L_x^2 \gamma' N^{n_\gamma} + 8\phi_v' \beta_{vz} \rho' N^{n_v + n_\rho} + \phi_s' \beta_{sz}' (\rho')^2 N^{2n_\rho + n_s + 1} &\leq \phi_z' \frac{\mu_G}{8} \rho' N^{n_z + n_\rho} \\ L_x^2 \gamma' N^{n_\gamma} + 8\phi_z' L_x'' (\gamma')^2 N^{2n_\gamma + n_z} + \phi_s' \beta_{sv}' (\rho')^2 N^{n_s + 2n_\rho + 1} &\leq \phi_v' \frac{\mu_G}{16} \rho' N^{n_v + n_\rho} \\ 5\phi_z' L_z' (\rho')^2 N^{n_z + 2n_\rho} + 5\phi_v' L_v' (\rho')^2 N^{2n_\rho + n_v} + L^h L_x' (\gamma')^2 N^{n_\gamma} &\leq \phi_s \frac{\Gamma'}{2} N^{n_s - 1} \ . \end{aligned}$$

In order to ensure that the exponents on N are lower in the left-hand-side than those on the right-hand-side, we take $n_z = n_v = 0$, $n_\rho = n_\gamma = -\frac{2}{3}$ and $n_s = -\frac{1}{3}$. The Equations become

$$\begin{aligned} (2\phi_z' \overline{\beta}_{zx} + 2\phi_v' \overline{\beta}_{vx}) \frac{\gamma'}{\rho'} + \phi_s' P' \gamma' &\leq \frac{1}{4} \\ L_x^2 \gamma' N^{-2/3} + 8\phi_v' \beta_{vz} \rho' N^{-2/3} + \phi_s' \beta_{sz}' (\rho')^2 N^{-2/3} &\leq \phi_z' \frac{\mu_G}{8} \rho' N^{-2/3} \\ L_x^2 \gamma' N^{-2/3} + 8\phi_z' L_x'' (\gamma')^2 N^{-4/3} + \phi_s' \beta_{sv}' (\rho')^2 N^{-2/3} &\leq \phi_v' \frac{\mu_G}{16} \rho' N^{-2/3} \\ 5\phi_z' L_z' (\rho')^2 N^{-4/3} + 5\phi_v' L_v' (\rho')^2 N^{-4/3} + L^h L_x' (\gamma')^2 N^{-4/3} &\leq \phi_s' \frac{\Gamma'}{2} N^{-4/3} \ . \end{aligned}$$

We can replace the penultimate equation by the stronger

$$L_x^2 \gamma' N^{-2/3} + 8\phi_z' L_x'' (\gamma')^2 N^{-2/3} + \phi_s' \beta_{sv}' (\rho')^2 N^{-2/3} \leq \phi_v' \frac{\mu_G}{16} \rho' N^{-2/3}$$

so that we can simplify all the equations by dropping the dependencies in N:

$$\begin{aligned} (2\phi_z' \overline{\beta}_{zx} + 2\phi_v' \overline{\beta}_{vx}) \frac{\gamma'}{\rho'} + \phi_s' P' \gamma' &\leq \frac{1}{4} \\ L_x^2 \gamma' + 8\phi_v' \beta_{vz} \rho' + \phi_s' \beta_{sz}' (\rho')^2 &\leq \phi_z' \frac{\mu_G}{8} \rho' \\ L_x^2 \gamma' + 8\phi_z' L_x'' (\gamma')^2 + \phi_s' \beta_{sv}' (\rho')^2 &\leq \phi_v' \frac{\mu_G}{16} \rho' \\ 5\phi_z' L_z' (\rho')^2 + 5\phi_v' L_v' (\rho')^2 + L^h L_x' (\gamma')^2 &\leq \phi_s' \frac{\Gamma'}{2} \ . \end{aligned}$$

Let us take $\phi'_s = 1$, $\phi'_z = \phi''_z \frac{\rho'}{\gamma'}$ and $\phi'_v = \phi''_v \frac{\rho'}{\gamma'}$ with $\phi''_z = \frac{1}{32\bar{\beta}_{zx}}$ and $\phi''_v = \min\left(\frac{1}{32\bar{\beta}_{vx}}, \phi''_z \frac{\mu_G}{128\beta_{vz}}\right)$. The equations become

$$\begin{aligned}
P'\gamma' &\leq \frac{1}{8} \\
L_x^2\gamma' + \beta'_{sz}(\rho')^2 &\leq \phi''_z \frac{\mu_G}{16}\frac{(\rho')^2}{\gamma'} \\
L_x^2\gamma' + 8\phi''_z L''_x \gamma'\rho' + \beta'_{sv}(\rho')^2 &\leq \phi''_v \frac{\mu_G}{16}\frac{(\rho')^2}{\gamma'} \\
5\phi''_z L'_z \frac{(\rho')^3}{\gamma'} + 5\phi''_v L'_v \frac{(\rho')^3}{\gamma'} + L^h L'_x (\gamma')^2 &\leq \frac{\Gamma'}{2} \ .
\end{aligned}$$

The condition $\gamma' \leq \frac{1}{8P'}$ ensures that the first equation is verified. With $\gamma' \leq \min\left(\sqrt{\frac{\phi''_z \mu_G}{32L_x^2}}\rho', \frac{\phi''_z \mu_G}{32\beta'_{sz}}\right)$, the second equations is verified. With $\gamma' \leq \min\left(\sqrt{\frac{\phi''_v \mu_G}{48L_x^2}}\rho', \frac{\phi''_v \mu_G}{48\beta'_{sv}}, \sqrt{\frac{\phi''_v \mu_G}{384\phi''_z L''_x \rho'}}\right)$, the third is verified. With $\gamma' \leq \sqrt{\frac{\Gamma'}{6L^h L'_x}}$, the last can be simplified:

$$(5\phi''_z L'_z + 5\phi''_v L'_v)(\rho')^3 \leq \frac{\Gamma'}{3}\gamma' \ .$$

Let us write $\gamma' = \xi\rho'$. If we want that equation does no contradict the previous upper bound on γ' involving ρ' and the conditions of Lemma C.8, that is

$$\begin{aligned}
\gamma' &\leq \underbrace{\min\left(\sqrt{\frac{\phi''_z \mu_G}{32L_x^2}}, \sqrt{\frac{\phi''_v \mu_G}{48L_x^2}}, \sqrt{\frac{L'_z}{2L'_x \beta_{zx}}}, \sqrt{\frac{L'_v}{2L'_x \beta_{vx}}}\right)}_{K_1} \rho' \\
\gamma' &\leq \underbrace{\min\left(\sqrt{\frac{\mu_G}{64\beta_{zx}L''_x}}, \sqrt{\frac{\mu_G}{128\beta_{vx}L''_x}}, \sqrt{\frac{\beta_{vz}}{4L''_x\beta_{vx}}}\right)}_{K_2} \sqrt{\rho'} \\
\gamma' &\leq \underbrace{\sqrt{\frac{\phi''_v \mu_G}{384\phi''_z L''_x}}}_{K_3} \frac{1}{\sqrt{\rho'}} \\
\gamma' &\leq \underbrace{\min\left(\frac{1}{4L^h}, \frac{L_x^2}{2L^h L''_x}, \sqrt{\frac{\Gamma'}{6L^h L'_x}}, \frac{1}{8P'}, \frac{\phi''_z \mu_G}{32\beta'_{sz}}, \frac{\phi''_v \mu_G}{48\beta'_{sv}}\right)}_{K_4} \\
\gamma' &\geq \underbrace{\frac{15(\phi''_z L'_z + \phi''_v L'_v)}{\Gamma'}}_{K_5} \rho^3
\end{aligned}$$

ξ must verify

$$\begin{aligned}
\xi &\leq K_1 \\
\xi &\leq K_2(\rho')^{-\frac{1}{2}} \\
\xi &\leq K_3(\rho')^{-\frac{3}{2}} \\
\xi &\leq K_4(\rho')^{-1} \\
\xi &\geq K_5(\rho')^2
\end{aligned}$$

which is possible if ρ' satisfies

$$\rho' \leq \min\left(\sqrt{\frac{K_1}{K_5}}, \left(\frac{K_2}{K_5}\right)^{-\frac{3}{2}}, \left(\frac{K_3}{K_5}\right)^{-\frac{5}{2}}, \left(\frac{K_4}{K_5}\right)^{-2}\right) \ .$$

Let us take

$$\rho' = \min\left(\sqrt{\frac{K_1}{K_5}}, \left(\frac{K_2}{K_5}\right)^{-\frac{3}{2}}, \left(\frac{K_3}{K_5}\right)^{-\frac{5}{2}}, \left(\frac{K_4}{K_5}\right)^{-2}, \frac{\mu_G}{64L_z^2}, \frac{\overline{\beta}_{zx}}{2\beta_{zx}}, \frac{\mu_G}{128(L_v^2+L_v'')}, \frac{\beta_{vz}}{8(L_v^2+L_v'')}, \frac{\overline{\beta}_{vx}}{2\beta_{vx}}\right) \tag{144}$$

and

$$\xi = \min(K_1, K_2(\rho')^{-\frac{1}{2}}, K_3(\rho')^{-\frac{3}{2}}, K_4(\rho')^{-1}) \ . \tag{145}$$

Finally, we have

$$\mathcal{L}^{t+1} - \mathcal{L}^t \leq -\frac{\gamma}{2} g^t$$

and therefore, summing and telescoping yields

$$\frac{1}{T}\sum_{t=1}^{T} g^t \leq \frac{\mathcal{L}^1}{\gamma T} = \frac{\mathcal{L}^0 N^{\frac{2}{3}}}{T} \ .$$

Since with respect to N we have

$$\mathcal{L}^0 = h^0 + \phi_z \delta_z^0 + \phi_v \delta_v^0 + \phi_s S^0 = \mathcal{O}(N^{-1} + 1 + 1 + N^{-\frac{1}{3}}) = \mathcal{O}(1) \ ,$$

we end up with

$$\boxed{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla h(x^t)\|^2] = \mathcal{O}\left(\frac{N^{\frac{2}{3}}}{T}\right) \ .}$$

□

C.10 Proof of Theorem 4

We are now going to prove Theorem 4 that we recall here:

Theorem 4 (Convergence of SABA, PL case)**.** Assume that h satisfies the PL inequality and that Assumptions 3.1 to 3.3 and 3.7 to 3.8 hold. We suppose $\rho = \rho' N^{-\frac{2}{3}}$ and $\gamma = \xi \rho' N^{-1}$, where ρ' and ξ depend only on F and G and are specified in appendix. Let x^t the iterates of SABA and $c' \triangleq \min\left(\mu_h, \frac{1}{16P'}\right)$ with P' specified in the appendix. Then,

$$\mathbb{E}[h^T] - h^* = (1 - c'\gamma)^T (h^0 - h^* + C^0)$$

where C^0 is a constant specified in appendix that depends on the initialization of z, v, x and memory.

Here, we have

$$\rho' = \min\left(\sqrt{\frac{K_1'}{K_5}}, \left(\frac{K_2}{K_5}\right)^{\frac{2}{5}}, \left(\frac{K_3}{K_5}\right)^{\frac{2}{7}}, \left(\frac{K_4'}{K_5}\right)^{\frac{1}{3}}, \frac{\mu_G}{64L_z^2}, \frac{\overline{\beta}_{zx}}{2\beta_{zx}}, \frac{\mu_G}{128(L_v^2+L_v'')}, \frac{\beta_{vz}}{8(L_v^2+L_v'')}, \frac{\overline{\beta}_{vx}}{2\beta_{vx}}\right) \ ,$$

and

$$\xi = \min(K_1', K_2(\rho')^{-\frac{1}{2}}, K_3(\rho')^{-\frac{3}{2}}, K_4'(\rho')^{-1}) \ .$$

where $P' = PN^{-1}$, $\Gamma' = \Gamma N$,

$$\phi_z'' = \frac{1}{32\overline{\beta}_{zx}} \ , \phi_v'' = \min\left(\frac{1}{32\overline{\beta}_{vx}}, \phi_z'' \frac{\mu_G}{128\beta_{vz}}\right) \ ,$$

$$K_1' = \min\left(\frac{\mu_G}{64c'}, \sqrt{\frac{\phi_z''\mu_G}{48L_x^2}}, \sqrt{\frac{\phi_v''\mu_G}{64L_x^2}}, \sqrt{\frac{L_z'}{2L_x'\beta_{zx}}}, \sqrt{\frac{L_v'}{2L_x'\beta_{vx}}}\right) \ ,$$

$$K_2 = \min\left(\sqrt{\frac{\mu_G}{64\beta_{zx}L_x''}}, \sqrt{\frac{\mu_G}{128\beta_{vx}L_x''}}, \sqrt{\frac{\beta_{vz}}{4L_x''\beta_{vx}}}\right) \ ,$$

$$K_3 = \sqrt{\frac{\phi_v''\mu_G}{512\phi_z''L_x''}} \ , \quad K_4' = \min\left(\frac{\Gamma'}{6c'}, \frac{1}{4L^h}, \frac{L_x^2}{2L^h L_x''}, \sqrt{\frac{\Gamma'}{6L^h L_x'}}, \frac{1}{18P'}, \frac{\phi_z''\mu_G}{48\beta_{sz}'}, \frac{\phi_v''\mu_G}{64\beta_{sv}'}\right)$$

and

$$K_5 = \frac{20(\phi_z''L_z' + \phi_v''L_v')}{\Gamma'} \ .$$

Proof. For simplicity, we assume that $h^* = 0$ and so for any $x \in \mathbb{R}^d$ the PL inequality reads:

$$\frac{1}{2}\|\nabla h(x)\|^2 \geq \mu_h h(x) \ . \tag{146}$$

Then, eq. (130) gives

$$h^{t+1} \leq \left(1 - \frac{\gamma\mu_h}{2}\right) h^t - \frac{\gamma}{4}\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] + \gamma L_x^2(\delta_z^t + \delta_v^t) + L^h L_x' \gamma^2 S^t \ .$$

We take $\mathcal{L}^t$ the Lyapunov function given in Equation (143). We find

$$\begin{aligned}
\mathcal{L}^{t+1} - \mathcal{L}^t \leq & -\gamma\mu_h h^t - \left(\frac{\gamma}{4} - 2\phi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} - 2\phi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho} - \phi_s P\gamma^2\right)\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \\
& - \left(\phi_z\frac{\mu_G}{8}\rho - L_x^2\gamma - 8\phi_v\beta_{vz}\rho - \phi_s\beta_{sz}\rho^2\right)\delta_z^t \\
& - \left(\phi_v\frac{\mu_G}{16}\rho - L_x^2\gamma - 2\phi_z L_x''\gamma^2 - \phi_s\beta_{sv}\rho^2\right)\delta_v^t \\
& - \left(\phi_s\frac{\Gamma}{2} - 5\phi_z L_z'\rho^2 - 5\phi_v L_v'\rho^2 - L^h L_x'\gamma^2\right) S^t \ .
\end{aligned}$$

We now try to find linear convergence, hence we add to this $c\mathcal{L}^t$ to get

$$\begin{aligned}
\mathcal{L}^{t+1} - (1-c)\mathcal{L}^t \leq & -(\gamma\mu_h - c)h^t - \left(\frac{\gamma}{4} - 2\phi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} - 2\phi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho} - \phi_s P\gamma^2 - c\right)\mathbb{E}[\|D_x(z^t, v^t, x^t)\|^2] \\
& - \left(\phi_z\frac{\mu_G}{8}\rho - L_x^2\gamma - 8\phi_v\beta_{vz}\rho - \phi_s\beta_{sz}\rho^2 - c\phi_z\right)\delta_z^t \\
& - \left(\phi_v\frac{\mu_G}{16}\rho - L_x^2\gamma - 2\phi_z L_x''\gamma^2 - \phi_s\beta_{sv}\rho^2 - c\phi_v\right)\delta_v^t \\
& - \left(\phi_s\frac{\Gamma}{2} - 5\phi_z L_z'\rho^2 - 5\phi_v L_v'\rho^2 - L^h L_x'\gamma^2 - c\phi_S\right) S^t \ .
\end{aligned}$$

Hence, the set of inequations for decrease becomes

$$\begin{aligned}
c &\leq \gamma\mu_h \\
2\phi_z\overline{\beta}_{zx}\frac{\gamma^2}{\rho} + 2\phi_v\overline{\beta}_{vx}\frac{\gamma^2}{\rho} + \phi_s P\gamma^2 + c &\leq \frac{\gamma}{4} \\
L_x^2\gamma + 8\phi_v\beta_{vz}\rho + \phi_s\beta_{sz}\rho^2 + \phi_z c &\leq \phi_z\frac{\mu_G}{8}\rho \\
L_x^2\gamma + 8\phi_z L_x''\gamma^2 + \phi_s\beta_{sv}\rho^2 + \phi_v c &\leq \phi_v\frac{\mu_G}{16}\rho \\
5\phi_z L_z'\rho^2 + 5\phi_v L_v'\rho^2 + L^h L_x'\gamma^2 + \phi_s c &\leq \phi_s\frac{\Gamma}{2} \ .
\end{aligned}$$

We see that it is more convenient to write $c = \gamma c'$. As previously, we write $\gamma = \gamma' N^{n_\gamma}$, $\rho = \rho' N^{n_\rho}$, $\phi_z = \phi_z' N^{n_z}$, $\phi_v = \phi_v' N^{n_v}$, $\phi_s = \phi_s' N^{n_s}$, $P = P'N$, $\Gamma = \Gamma' N^{-1}$, $\beta_{sx} = \beta_{sx}' N$ and $\beta_{sv} = \beta_{sv}' N$. The equations read:

$$\begin{aligned}
c' &\leq \mu_h \\
2\phi_z'\overline{\beta}_{zx}\frac{\gamma'}{\rho'}N^{n_z+n_\gamma-n_\rho} + 2\phi_v'\overline{\beta}_{vx}\frac{\gamma'}{\rho'}N^{n_v+n_\gamma-n_\rho} + \phi_s' P'\gamma' N^{n_s+1+n_\gamma} + c' &\leq \frac{1}{4} \\
L_x^2\gamma' N^{n_\gamma} + 8\phi_v'\beta_{vz}\rho' N^{n_v+n_\rho} + \phi_s'\beta_{sz}'(\rho')^2 N^{n_s+2n_\rho+1} + \phi_z' c'\gamma' N^{n_z+n_\gamma} &\leq \phi_z'\frac{\mu_G}{8}\rho' N^{n_\rho+n_z} \\
L_x^2\gamma' N^{n_\gamma} + 8\phi_z' L_x''(\gamma')^2 N^{n_z+2n_\gamma} + \phi_s'\beta_{sv}'(\rho')^2 N^{n_s+1+2n_\rho} + \phi_v' c'\gamma' N^{n_v+n_\gamma} &\leq \phi_v'\frac{\mu_G}{16}\rho' N^{n_v+n_\rho} \\
5\phi_z' L_z'(\rho')^2 N^{n_z+2n_\rho} + 5\phi_v' L_v'(\rho')^2 N^{n_v+2n_\rho} + L^h L_x'(\gamma')^2 N^{2n_\gamma} + \phi_s' c'\gamma' N^{n_s+n_\gamma} &\leq \phi_s'\frac{\Gamma'}{2}N^{n_s-1} \ .
\end{aligned}$$

In order to ensure that the exponents on N are lower in the left-hand-side than those on the right-hand-side, we take $n_z = n_v = 0$, $n_\rho = -\frac{2}{3}$, $n_\gamma = -1$ and $n_s = -\frac{1}{3}$. The Equations become

$$\begin{aligned}
c' &\leq \mu_h \\
2\phi'_z\overline{\beta}_{zx}\frac{\gamma'}{\rho'}N^{-\frac{1}{3}} + 2\phi'_v\overline{\beta}_{vx}\frac{\gamma'}{\rho'}N^{-\frac{1}{3}} + \phi'_s P'\gamma' N^{-\frac{1}{3}} + c' &\leq \frac{1}{4} \\
L_x^2\gamma' N^{-1} + 8\phi'_v\beta_{vz}\rho' N^{-\frac{2}{3}} + \phi'_s\beta'_{sz}(\rho')^2 N^{-\frac{2}{3}} + \phi'_z c'\gamma' N^{-1} &\leq \phi'_z\frac{\mu_G}{8}\rho' N^{-\frac{2}{3}} \\
L_x^2\gamma' N^{-1} + 8\phi'_z L''_x(\gamma')^2 N^{-2} + \phi'_s\beta'_{sv}(\rho')^2 N^{-\frac{2}{3}} + \phi'_v c'\gamma' N^{-1} &\leq \phi'_v\frac{\mu_G}{16}\rho' N^{-\frac{2}{3}} \\
5\phi'_z L'_z(\rho')^2 N^{-\frac{4}{3}} + 5\phi'_v L'_v(\rho')^2 N^{-2} + L^h L'_x(\gamma')^2 N^{-2} + \phi'_s c'\gamma' N^{-\frac{4}{3}} &\leq \phi'_s\frac{\Gamma'}{2}N^{-\frac{4}{3}} \ .
\end{aligned}$$

Now we have to find ρ', γ', ϕ'_z, ϕ'_v and ϕ'_s that verifies the following conditions (which are a bit stronger than thoose in the previous Equations):

$$\begin{aligned}
c' &\leq \mu_h \\
2\phi'_z\overline{\beta}_{zx}\frac{\gamma'}{\rho'} + 2\phi'_v\overline{\beta}_{vx}\frac{\gamma'}{\rho'} + \phi'_s P'\gamma' + c' &\leq \frac{1}{4} \\
L_x^2\gamma' + 8\phi'_v\beta_{vz}\rho' + \phi'_s\beta'_{sz}(\rho')^2 + \phi'_z c'\gamma' &\leq \phi'_z\frac{\mu_G}{8}\rho' \\
L_x^2\gamma' + 8\phi'_z L''_x(\gamma')^2 + \phi'_s\beta'_{sv}(\rho')^2 + \phi'_v c'\gamma' &\leq \phi'_v\frac{\mu_G}{16}\rho' \\
5\phi'_z L'_z(\rho')^2 + 5\phi'_v L'_v(\rho')^2 + L^h L'_x(\gamma')^2 + \phi'_s c'\gamma' &\leq \phi'_s\frac{\Gamma'}{2} \ .
\end{aligned}$$

As previously, we take $\phi'_s = 1$ and we denote $\phi'_z = \phi''_z\frac{\rho'}{\gamma'}$ with $\phi''_z = \frac{1}{32\overline{\beta}_{zx}}$ and $\phi'_z = \phi''_z\frac{\rho'}{\gamma'}$ with $\phi''_v = \min\left(\frac{1}{32\overline{\beta}_{vx}}, \phi''_z\frac{\mu_G}{128\beta_{vz}}\right)$, the equations become

$$\begin{aligned}
c' &\leq \mu_h \\
P'\gamma' + c' &\leq \frac{1}{8} \\
L_x^2(\gamma')^2 + \beta'_{sz}(\rho')^2\gamma' + \phi''_z c'\rho'\gamma' &\leq \phi''_z\frac{\mu_G}{16}(\rho')^2 \\
L_x^2(\gamma')^2 + 8\phi''_z L''_x\rho'(\gamma')^2 + \beta'_{sv}(\rho')^2\gamma' + \phi''_v c'\rho'\gamma' &\leq \phi''_v\frac{\mu_G}{16}(\rho')^2 \\
5\phi''_z L'_z(\rho')^3 + 5\phi''_v L'_v(\rho')^3 + L^h L'_x(\gamma')^3 + c'(\gamma')^2 &\leq \frac{\Gamma'}{2}\gamma' \ .
\end{aligned}$$

Since $c' \leq \frac{1}{16}$ and $\gamma' \leq \frac{1}{16P'}$, the second equation is verified. With $\gamma' \leq \min\left(\sqrt{\frac{\phi''_z\mu_G}{48L_x^2}}\rho', \frac{\phi''_z\mu_G}{48\beta_{sv}}\right)$ and $c' \leq \frac{\mu_G\rho'}{48\gamma'}$ the third is verified. The conditions $\gamma' \leq \min\left(\sqrt{\frac{\phi''_v\mu_G}{64L_x^2}}\rho', \sqrt{\frac{\phi''_v\mu_G}{512\phi''_z L''_x\rho'}}, \frac{\phi''_v\mu_G}{64\beta'_{sv}}\right)$ and $c' \leq \frac{\mu_G\rho'}{64\gamma'}$ ensure that the forth is verified. With $\gamma' \leq \sqrt{\frac{\Gamma'}{8L^h L'_x}}$ and $c' \leq \frac{\Gamma'}{8\gamma'}$, the fifth is simplified in

$$5\phi''_z L'_z(\rho')^3 + 5\phi''_v L'_v(\rho')^3 \leq \frac{\Gamma'}{4}\gamma' \ .$$

As in the proof of Theorem 3, let us denote $\gamma' = \xi\rho'$. To verify this equation and the previous bounds on γ' and c', we need

$$\gamma' \leq \underbrace{\min\left(\sqrt{\frac{\phi_z''\mu_G}{48L_x^2}}, \sqrt{\frac{\phi_v''\mu_G}{64L_x^2}}, \sqrt{\frac{L_z'}{2L_x'\beta_{zx}}}, \sqrt{\frac{L_v'}{2L_x'\beta_{zx}}}\right)}_{K_1}\rho' \ ,$$

$$\gamma' \leq \underbrace{\min\left(\sqrt{\frac{\mu_G}{64\beta_{zx}L_x''}}, \sqrt{\frac{\mu_G}{128\beta_{vx}L_x''}}, \sqrt{\frac{\beta_{vz}}{4L_x''\beta_{vx}}}\right)}_{K_2}\sqrt{\rho'} \ ,$$

$$\gamma' \leq \underbrace{\sqrt{\frac{\phi_v''\mu_G}{512\phi_z''L_x''}}}_{K_3}\frac{1}{\sqrt{\rho'}} \ ,$$

$$\gamma' \leq \underbrace{\min\left(\frac{1}{4L^h}, \frac{L_x^2}{2L^hL_x''}, \frac{\phi_z''\mu_G}{48\beta_{sv}}, \frac{\phi_v''\mu_G}{64\beta_{sv}'}, \frac{1}{16P'}, \sqrt{\frac{\Gamma'}{8L^hL_x'}}\right)}_{K_4}$$

$$\gamma' \geq \underbrace{\frac{20(\phi_z''L_z' + \phi_v''L_v')}{20}}_{K_5}(\rho')^3 \ ,$$

$$c' \leq \underbrace{\min\left(\mu_h, \frac{1}{16}, \frac{1}{16P'}\right)}_{K_6} \ ,$$

$$c' \leq \underbrace{\frac{\mu_G}{64}}_{K_7}\frac{1}{\xi} \ ,$$

$$c' \leq \underbrace{\frac{\Gamma'}{8}}_{K_8}\frac{1}{\gamma'} \ .$$

So, ξ, ρ' and c' must verify

$$\xi \leq \underbrace{\min\left(K_1, \frac{K_7}{c'}\right)}_{K_1'} \ ,$$

$$\xi \leq K_2(\rho')^{-\frac{1}{2}} \ ,$$

$$\xi \leq K_3(\rho')^{-\frac{3}{2}} \ ,$$

$$\xi \leq \underbrace{\min\left(K_4, \frac{K_8}{c'}\right)}_{K_4'}(\rho')^{-1}$$

$$\xi \geq K_5(\rho')^2 \ ,$$

$$c' \leq \underbrace{\min\left(\mu_h, \frac{1}{16}, \frac{1}{16P'}\right)}_{K_6} \ ,$$

which is possible if

$$\rho' \leq \min\left(\sqrt{\frac{K_1'}{K_5}}, \left(\frac{K_2}{K_5}\right)^{\frac{2}{5}}, \left(\frac{K_3}{K_5}\right)^{\frac{2}{7}}, \left(\frac{K_4'}{K_5}\right)^{\frac{1}{3}}\right) \ .$$

So let us take $c' = \min\left(\mu_h, \frac{1}{16}, \frac{1}{16P'}\right) = \min\left(\mu_h, \frac{1}{16P'}\right)$,

$$\rho' = \min\left(\sqrt{\frac{K_1'}{K_5}}, \left(\frac{K_2}{K_5}\right)^{\frac{2}{5}}, \left(\frac{K_3}{K_5}\right)^{\frac{2}{7}}, \left(\frac{K_4'}{K_5}\right)^{\frac{1}{3}}, \frac{\mu_G}{64L_z^2}, \frac{\overline{\beta}_{zx}}{2\beta_{zx}}, \frac{\mu_G}{128(L_v^2+L_v'')}, \frac{\beta_{vz}}{8(L_v^2+L_v'')}, \frac{\overline{\beta}_{vx}}{2\beta_{vx}}\right)$$

and

$$\xi = \min(K_1, K_2(\rho')^{-\frac{1}{2}}, K_3(\rho')^{-\frac{3}{2}}, K_4(\rho')^{-1}) \ .$$

We have

$$\mathcal{L}^{t+1} \leq (1-c)\mathcal{L}^t$$

therefore, unrolling yields

$$\boxed{h^t - h^* \leq \mathcal{L}^t \leq (1-c'\gamma)^t \mathcal{L}^0.}$$

□

D Convergence rates with weaker regularity assumptions

To get our rates, we need stronger assumptions than in the stochastic bilevel optimization literature [19, 24, 26, 2]. In this section, we shortly present the convergence rates we can expect if we replace Assumptions 3.1 and 3.2 by Assumptions D.1 and D.2.

Assumption D.1. The function F is differentiable. The gradient ∇F is Lipschitz continuous in (z, x) with Lipschitz constants L_1^F.

Assumption D.2. The function G is twice continuously differentiable on $\mathbb{R}^p \times \mathbb{R}^d$. For any $x \in \mathbb{R}^d$, $G(\,\cdot\,, x)$ is μ_G-strongly convex. The derivatives ∇G are $\nabla^2 G$ are Lipschitz continuous in (z, x) with respective Lipschitz constants L_1^G and L_2^G.

With these assumptions, we are not ensured that v^* is smooth, and so the descent lemmas take the form of Lemma D.3.

Lemma D.3. *Assume that $\rho \leq \frac{2}{\mu_G}$. We have:*

$$\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{2}\right)\delta_z^t + 2\rho^2 V_z^t + 4\frac{L_*^2}{\mu_G}\frac{\gamma^2}{\rho}V_x^t$$

$$\delta_v^{t+1} \leq \left(1 - \frac{\rho\mu_G}{4}\right)\delta_v^t + \rho\beta_{vz}\delta_z^t + 2\rho^2 V_v^t + 8\frac{L_*^2}{\mu_G}\frac{\gamma^2}{\rho}V_x^t$$

where L_* is the maximum between the Lipschitz constants of z^* and v^* (see Lemma C.1) and $\beta_{vz} = \frac{1}{\mu_G^3}(L^F\mu_G + L_2^G)^2$.

Proof. **Inequality for δ_z.**

Instead of expanding the square as done in the proof of Lemma 3.9 in Equation (45), we use Young's inequality for some $a > 0$

$$\delta_z^{t+1} \leq (1+a)\mathbb{E}[\|z^{t+1} - z^*(x^t)\|^2] + (1+a^{-1})\mathbb{E}[\|z^*(x^{t+1}) - z^*(x^t)\|^2] \ . \tag{147}$$

Treating $\mathbb{E}[\|z^{t+1} - z^*(x^t)\|^2]$ and $\mathbb{E}[\|z^*(x^{t+1}) - z^*(x^t)\|^2]$ as done in the proof of Lemma 3.9 leads to

$$\delta_z^{t+1} \leq (1+a)\left[(1-\rho\mu_G)\delta_z^t + \rho^2 V_z^t\right] + (1+a^{-1})L_*^2\gamma^2 V_x^t \tag{148}$$

In order to keep a decrease in δ_z, we might want to use $a = \frac{1}{2}\rho\mu_G$, which gives the bound

$$\boxed{\delta_z^{t+1} \leq \left(1 - \frac{\rho\mu_G}{2}\right)\delta_z^t + 2\rho^2 V_z^t + \beta_{zx}\frac{\gamma^2}{\rho}V_x^t} \tag{149}$$

with $\beta_{zx} = 4\frac{L_*^2}{\mu_G}$. Indeed, this gives $(1+\frac{1}{2}\rho\mu_G)(1-\rho\mu_G) \leq 1 - \frac{1}{2}\rho\mu_G$. We have $a \leq 1$ since $\rho \leq \frac{2}{\mu_G}$, so $(1+a)\rho^2 \leq 2\rho^2$. Finally, we also have $1 + a^{-1} \leq 2a^{-1} = \frac{4}{\rho\mu_G}$.

Inequality for δ_v. As for δ_z, the difference with the proof of Lemma 3.9 is that we use we use Young's inequality for some $b > 0$ to get

$$\delta_v^{t+1} \leq (1+b)\mathbb{E}[\|v^{t+1} - v^*(x^t)\|^2] + (1+b^{-1})\mathbb{E}[\|v^*(x^{t+1}) - v^*(x^t)\|^2] \ . \tag{150}$$

The remaining part of the proof is similar to the proof of Lemma 3.9. □

The main difference with Lemma 3.9 is that we have $O(\frac{\gamma^2}{\rho})$ in factor of V_x^t instead of $O(\gamma^2)$. As a consequence, we need that the ratio $\frac{\gamma}{\rho}$ goes to zero to get convergence, as in [24]. This prevent us in getting rates that match rates of single level algorithms.

Hence, for SOBA, we have to choose $\gamma = O(T^{-\frac{3}{5}})$ and $\rho = O(T^{-\frac{2}{5}})$ and we end up with a convergence rate in $O(T^{-\frac{2}{5}})$. For SABA, we get a $O((n+m)\epsilon^{-1})$ sample complexity, which is actually the sample complexity of SOBA used with full batch estimated directions.