Communication-Efficient Multi-Level Vertical Federated Learning

Anonymous Authors[1]

Abstract

Distributed methods have proven to be effective solutions to contemporary machine learning problems, both in theory and in practice. However, with the increasing popularity of large neural networks, it has become challenging to fit them into the memory of a single node. To address this issue, the idea of dividing a model across multiple devices was introduced in vertical federated learning approaches. Nevertheless, these methods still suffer from a bottleneck common to all distributed algorithms. Specifically, the substantial amount of information transmitted during the search for a solution can significantly slow down the training process. To tackle this problem, this study introduces the first algorithm designed for a multi-level vertical setup, enhanced with a compression technique the communication cost per iteration reduction. To address issues arising from data sparsification, we propose a reformulation of the optimization problem by introducing new variables that simulate the workers' outputs. We ensure the accuracy of this simulation by penalizing the model with a mean squared error loss between these new parameters and the local functions' values. In our experiments, we demonstrate superior empirical performance compared to other approaches operating in this setup.

1. Introduction

Recent advancements in deep learning models have yielded significant progress in the field of machine learning, particularly in tasks such as computer vision, speech recognition, and natural language processing (Russakovsky et al., 2015; Prabhavalkar et al., 2023; Achiam et al., 2023). However, modern neural networks require large datasets to achieve state-of-the-art performance (Shi et al., 2016). Training such models is a significantly more arduous process and can take weeks or even months in terms of computation time (Simonyan, 2014). Consequently, there is a need for faster methods to maintain the training pace and ensure the quality of predictions.

One approach to addressing this problem is to distribute subsamples across multiple devices and perform computations in parallel. This is the case in classical distributed learning or, when data cannot be directly exposed to other workers, horizontal federated learning (McMahan et al., 2017).

However, the process of scaling up deep learning models has affected not only the size of datasets but also the number of trainable parameters. For example, the next-generation model `GPT-3` (Brown et al., 2020) has approximately 100 times more parameters than its predecessor `GPT-2` (Radford et al., 2019). This has led to a situation where the entire model cannot fit into the memory of the GPU on which it is trained. To address this challenge, vertical federated learning (Yang et al., 2023; Liu et al., 2024) and split learning (Vepakomma et al., 2018) have emerged as potential solutions. In these approaches, the model is divided across nodes, with each worker possessing its own local model and unique features.

Regardless of how the problem is parallelized across workers, certain bottlenecks remain. These methods require a considerable amount of information to be transmitted, and limited network capacity can severely slow down the training process (Lian et al., 2017). Various approaches have been proposed to mitigate this issue, such as reducing the dimensionality of data (Khan et al., 2022) or implementing asynchronous training (Chen et al., 2020). However, this paper focuses exclusively on compression techniques, which involve mapping transmitted information to a representation that occupies less network space, thereby reducing the average cost of a single communication.

Compression in the horizontal regime has been extensively researched, as evidenced by works such as (Alistarh et al., 2017; Richtárik et al., 2021), and is typically applied to some form of the gradient of the local loss function. However, in vertical federated learning, the compressor operator maps the output of the local model to another vector, which is then passed as an argument to another function. To the best of our knowledge, all existing theoretical studies focus only

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

on the case of one level of compression, i.e., in this setting, the device that receives information from other nodes does not subsequently transmit its computed output to another worker as an argument. In this paper, we explore a multi-level vertical setup enhanced with compression techniques.

1.1. Related Works

• **Compression.** One of the first steps in the theoretical research of compression was the introduction of the class of *unbiased* operators. Such operators are assumed to return the same vector on average, and the formal definition was first given in (Alistarh et al., 2017). It is verified in (Beznosikov et al., 2023) that, e.g., a quantization operator or a compressor that uniformly selects random coordinates possesses this property.

However, the requirement of unbiasedness in this definition significantly narrows down the set of potential compressors. Operators without this property are called *biased* and often perform empirically better compared to their *unbiased* counterparts (Seide et al., 2014; Lin et al., 2017). The simplest example of such an operator is the Top-k compressor (Alistarh et al., 2018), which takes a vector and returns the k coordinates with the largest absolute values. However, the greedy approach offered by biased compressors comes with disadvantages. It is proven that the distributed version of gradient descent with the Top-k operator does not converge in a simple example (Beznosikov et al., 2023).

• **Error Feedback.** To overcome the aforementioned issue, the *error feedback* mechanism was introduced. The underlying concept is based on accumulating the difference between the actual and compressed vectors. By tracking this error, it is possible to adjust the calculations to more closely approximate the uncompressed value. This mechanism was initially used in practical papers (Seide et al., 2014). The first theoretical result for biased compression was obtained in (Karimireddy et al., 2019), but only for a single-node setup and under the assumption of bounded gradients. These restrictions were overcome in (Stich & Karimireddy, 2019), and later the error feedback mechanism was adapted for single-node non-convex optimization (Ajalloeian & Stich, 2020).

In all the aforementioned methods with error feedback, compression is applied to the gradient summed with a variable that accumulates the error. However, since the gradient is considered a "large" vector, such an operation can lead to a significant increase in total variance, thereby worsening the convergence rate. To mitigate this issue, the technique of applying compression to the difference between the true gradient and the step vector was introduced. It was first applied to unbiased compression in `DIANA` (Mishchenko et al., 2024) and later adapted for biased compression in the `EF21` algorithm (Richtárik et al., 2021). The paper (Fatkhullin et al., 2021) introduced the analysis of several practical extensions of this improved error feedback mechanism. However, it is important to note that all of these works study the setup where each worker has its own local function, which is suitable only for the horizontal federated learning setting and not the vertical case.

• **Compression in the vertical federated learning.** Unlike the horizontal regime, there are few works related to compression in vertical federated learning. Most of them are practical and lack theoretical guarantees (Chen et al., 2021; Cai et al., 2022; Sun et al., 2023). To the best of our knowledge, the first theoretical approach was given in (Castiglia et al., 2022), but no concrete convergence rate was proven. A theory for unbiased compression in the vertical setup was later developed for strongly convex settings in (Stanko et al., 2024) and for convex problems in (Beznosikov et al., 2024). In the work by (Valdeira et al., 2024), biased compression with the error feedback mechanism was adapted for the non-convex vertical setup. However, this study only considered one level of vertical data division. Thus, it remains an open question to develop an algorithm that works for settings where more than two nodes use the output of the previous device's model as an argument. In this paper, we address this problem.

1.2. Our Contributions

• **New Formulation for the Problem in the Multi-Level Vertical Federated Learning Setup.** The main obstacle in worker-to-worker consecutive communication arises from the accumulation of errors due to sparsification, as each node passes the compressed output of its local model to the next device. To mitigate this issue, we introduce new variables that imitate the values of the workers' functions. To ensure the accuracy of this approximation, we penalize the problem with the norm of the difference between these new parameters and the output of the local models. Under common assumptions, we prove the smoothness of this new formulation.

• **New Algorithm.** Utilizing this enhanced setup, we propose Algorithm 1 – the first method with biased compression designed for the multi-level vertical federated learning. We base this method on the error feedback mechanism, which, under assumptions that are realistic in practice, allows us to achieve state-of-the-art $\mathcal{O}\left(\frac{1}{T}\right)$ convergence rate, where T is the total number of iterations.

• **Numerical Experiments.** In our numerical experiments, we solve a classification problem on the F-MNIST dataset and demonstrate superior performance compared to baseline and uncompressed algorithms in terms of the amount of data transmitted. Additionally, we present results for a scenario where the model is trained only on input and target embeddings from the previous and subsequent models.

1.3. Notations

We denote the standard dot product between two vectors $x, y \in \mathbb{R}^d$ as $\langle x, y\rangle \in \mathbb{R}$. The index i in our paper is associated with the i-th worker. We refer to the Jacobian matrix of the function $F_i(x_i, z_i)$, where $x_i \in \mathbb{R}^{d_{x_i}}$ and $z_i \in \mathbb{R}^{d_{z_i}}$, as J_{F_i}. Additionally, J_{F_i,x_i} and J_{F_i,z_i} denote the parts of J_{F_i} corresponding to the variables x_i and z_i, respectively. The index m denotes the m-th coordinate of a vector. We use proj_Q to label the projection operator onto the set Q, and $\text{prox}_{\gamma r}$ to denote the proximal operator (see Section 3.3 for details). The indicator function of the set Q is denoted by $\delta_Q(x)$, which equals 0 if $x \in Q$ and ∞ otherwise.

2. Introduction of `SVFL-EF21` Algorithm

In this section, we provide assumptions and describe the vertical split learning problem. The standard formulation assumes that every worker has its own model F_i and holds its trainable parameters $x_i \in \mathbb{R}^{d_{x_i}}$. The output of the local model is then passed as an argument to the subsequent node. Therefore, the target function f is calculated as the composition of the workers' models. This can be written in the following form:

$$\min_{x_1,\ldots,x_n} \left[f(x_1,\ldots,x_n) := F_1(x_1, F_2(x_2, F_3(x_3,\ldots)))\right], \tag{1}$$

where n is the number of workers. However, if this scenario is enhanced with compression, the gradient would be calculated at the point of the **composition** of compressed arguments. For example, the step direction for coordinates associated with x_1 would be

$$\nabla_{x_1} F_1(x_1, \mathcal{C}(F_2(x_2, \mathcal{C}(F_3(x_3, \mathcal{C}(\ldots)))))),$$

where $\mathcal{C}$ is some compression operator. It is evident that the error would accumulate from worker to worker, resulting in a poor convergence rate, both in theory and in practice. To avoid this issue, we reformulate the problem as follows:

$$\min_{\substack{x_1\in\mathbb{R}^{d_{x_1}},\ldots,x_n\in\mathbb{R}^{d_{x_n}},\\ z_1\in\mathbb{R}^{d_{z_1}},\ldots,z_{n-1}\in\mathbb{R}^{d_{z_{n-1}}}}} \Bigg[f(x_1,\ldots,x_n,z_2,\ldots,z_{n-1}) := F_1(x_1,z_1) + \lambda_2\|F_2(x_2,z_2) - z_1\|^2 + \ldots + \lambda_n\|F_n(x_n,z_n) - z_{n-1}\|^2\Bigg], \tag{2}$$

where $\lambda_i > 0$ denotes the regularizer for the i-th worker, and $z_n \in \mathbb{R}^{s\times d_{z_n}}$ represents the constant dataset values.

This approach allows us to accurately approximate the problem (1) by introducing additional variables $\{z_i\}_{i=1}^{n-1}$. Each of these variables mimics the output of the model from the previous worker, and the mean squared error (MSE) ensures that this simulation is precise. Moreover, due to the independent minimization of both the "old" parameters x_i and the "new" parameters z_i, the issue of accumulated compression is no longer present.

However, for theoretical reasons listed in Section 3.1, we must also add constraints on the variables $\{z_i\}_{i=1}^{n-1}$. Specifically, we restrict the search for these parameters to the image of the corresponding function. Therefore, the final formulation can be written as:

$$\min_{\substack{x_1\in\mathbb{R}^{d_{x_1}},\ldots,x_n\in\mathbb{R}^{d_{x_n}},\\ z_1\in\text{Im}(F_2),\ldots,z_{n-1}\in\text{Im}(F_n)}} \Bigg[f(x_1,\ldots,x_n,z_2,\ldots,z_{n-1}) := F_1(x_1,z_1) + \lambda_2\|F_2(x_2,z_2) - z_1\|^2 + \ldots + \lambda_n\|F_n(x_n,z_n) - z_{n-1}\|^2\Bigg]. \tag{3}$$

This formulation is intuitive, as the introduction of $\{z_i\}_{i=1}^{n-1}$ is meant to approximate the output of the local models $\{F_i\}_{i=1}^{n}$, and it is natural to bound these parameters to the image of the corresponding functions.

2.1. Algorithm Description

In this section, we present an algorithm, which we call `SVFL-EF21`, for solving the problem (3). In this scenario, to update the local trainable parameters x_i^t and z_{i-1}^t, every worker with index $i > 1$ must both send a term connected to the output of its model F_i and exchange the vector z_{i-1}^t with the worker of index $i-1$. To enhance the algorithm with compression and the error feedback mechanism, additional sequences are introduced: $\{\mathcal{H}_i^t\}_{t=0}^T$ for transmitting the term connected to the local model output, and $\{\mathcal{Z}_{i-1}^t\}_{t=0}^T$ for communicating z_{i-1}^t. These variables accumulate the error from compression and allow workers to exchange the compressed difference between the true and compressed information, rather than the vector itself. After the necessary data is collected, updates to the parameters x_i^t and z_{i-1}^t are performed. Since we solve the constrained optimization problem (3), the update procedure for z_{i-1}^t is enhanced with the proximal operator, represented as the indicator function of the image of the worker's model.

In-depth description. In `SVFL-EF21`, the i-th worker stores the parameters x_i^t and z_i^t. Additionally, if $1 < i < n$, the worker initializes the variables $\mathcal{H}_i^t$, $\mathcal{H}_{i+1}^t$, $\mathcal{Z}_i^t$, and $\mathcal{Z}_{i-1}^t$. However, the last device ($i = n$) only holds $\mathcal{H}_n^t$ and $\mathcal{Z}_{n-1}^t$, while the first node ($i = 1$) only holds $\mathcal{Z}_1^t$ and $\mathcal{H}_2^t$.

Every iteration begins with a cycle where the index i changes from n to 3, with i representing the corresponding machine number (line 4). The i-th worker, using the local parameters x_i^t, z_i^t, and the sequence value $\mathcal{Z}_{i-1}^t$, calculates

Algorithm 1 SVFL-EF21

1: **Input:** Initial points $\left\{x_i^0 \in \mathbb{R}^{d_{x_i}}\right\}_{i=1}^n$ and $\left\{z_i^0 \in \mathbb{R}^{d_{z_i}}\right\}_{i=1}^{n-1}$, amount of iterations T, initial sequences of vectors $\left\{\mathcal{Z}_i^0 = \mathcal{C}(z_i^0)\right\}_{i=1}^{n-1}$, $\left\{\mathcal{H}_i^0 = \mathcal{C}\left(2\lambda_i(\mathcal{Z}_i^0 - F_i(x_i^0, z_i^0))\right)\right\}_{i=2}^{n}$
2: **Parameter:** Stepsize $\gamma > 0$
3: **for** $t = 0, \ldots, T-1$ **do**
4: **for** $i = n, \ldots, 3$ **do**
5: The i-th worker does the following actions.
$g_{x_i}^t \leftarrow 2\lambda_i J_{F_i,x_i}^T(x_i^t, z_i^t)\left(F_i(x_i^t, z_i^t) - \mathcal{Z}_{i-1}^t\right)$
$x_i^{t+1} \leftarrow x_i^t - \gamma g_{x_i}^t$
$h_i^t \leftarrow \mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^t - F_i(x_i^t, z_i^t)) - \mathcal{H}_i^t\right)$
Send h_i^t to the $(i-1)$-th device
6: The $(i-1)$-th worker does the following actions.
$g_{z_{i-1}}^t \leftarrow \mathcal{H}_i^t + 2\lambda_{i-1} J_{F_{i-1},z_{i-1}}^T(x_{i-1}^t, z_{i-1}^t)(F_{i-1}(x_{i-1}^t, z_{i-1}^t) - \mathcal{Z}_{i-2}^t)$
$z_{i-1}^{t+1} \leftarrow \text{prox}_{\gamma\delta_{\text{Im}F_i}}(z_{i-1}^t - \gamma g_{z_{i-1}}^t)$
$\mathcal{Z}_{i-1}^{t+1} \leftarrow \text{proj}_{\text{Im}F_i}\left(\mathcal{Z}_{i-1}^t + \mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)\right)$
$\mathcal{H}_i^{t+1} \leftarrow \mathcal{H}_i^t + h_i^t$
Send $\mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)$ to the i-th worker
7: The i-th worker does the following actions.
$\mathcal{Z}_{i-1}^{t+1} \leftarrow \text{proj}_{\text{Im}F_i}\left(\mathcal{Z}_{i-1}^t + \mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)\right)$
$\mathcal{H}_i^{t+1} \leftarrow \mathcal{H}_i^t + h_i^t$
8: **end for**
9: The second worker does the following actions.
$g_{x_2}^t \leftarrow 2\lambda_2 J_{F_2,x_2}^T(x_2^t, z_2^t)\left(F_2(x_2^t, z_2^t) - \mathcal{Z}_1^t\right)$
$x_2^{t+1} \leftarrow x_2^t - \gamma g_{x_2}^t$
$h_2^t \leftarrow \mathcal{C}\left(2\lambda_2(\mathcal{Z}_1^t - F_2(x_2^t, z_2^t)) - \mathcal{H}_2^t\right)$
Send h_2^t to the first device
10: The first worker does the following actions.
$g_{x_1}^t \leftarrow \nabla_{x_1} F_1(x_1^t, z_1^t)$
$x_1^{t+1} \leftarrow x_1^t - \gamma g_{x_1}^t$
$g_{z_1}^t = \mathcal{H}_2^t + \nabla_{z_1} F_1(x_1^t, z_1^t)$
$z_1^{t+1} = \text{prox}_{\gamma\delta_{\text{Im}F_2}}(z_1^t - \gamma g_{z_1}^t)$
$\mathcal{Z}_1^{t+1} \leftarrow \text{proj}_{\text{Im}F_2}\left(\mathcal{Z}_1^t + \mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)\right)$
$\mathcal{H}_2^{t+1} \leftarrow \mathcal{H}_2^t + h_2^t$
Send $\mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)$ to the second worker
11: The second worker does the following actions.
$\mathcal{Z}_1^{t+1} \leftarrow \text{proj}_{\text{Im}F_2}\left(\mathcal{Z}_1^t + \mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)\right)$
$\mathcal{H}_2^{t+1} \leftarrow \mathcal{H}_2^t + h_2^t$
12: **end for**
13: **Output:** x^T

the step vector $g_{x_i}^t$ needed to update x_i^{t+1} (line 5). Afterward, the worker computes the compressed difference h_i^t, which is necessary for updating $\mathcal{H}_i^{t+1}$, and sends it to the $(i-1)$-th node.

During the same iteration of the cycle for index i, the $(i-1)$-th worker is responsible for computing the step direction for the parameter z_{i-1}^{t+1} and updating it via the proximal operator (line 6). This worker then proceeds to update the sequences $\mathcal{H}_i^{t+1}$ and $\mathcal{Z}_{i-1}^{t+1}$ using h_i^t, received from the i-th device, and the newly calculated compressed difference between z_{i-1}^{t+1} and $\mathcal{Z}_{i-1}^t$. The updated $\mathcal{Z}_{i-1}^{t+1}$ is then transmitted back to the i-th worker, which updates both $\mathcal{H}_i^{t+1}$ and $\mathcal{Z}_{i-1}^{t+1}$ in the same manner as the $(i-1)$-th node (line 7).

After all devices from the n-th to the 3-rd have completed the aforementioned actions, the second and first nodes perform similar updates. The only difference lies in the slightly modified calculation of the step vector $g_{x_1}^t$ (lines 9, 10, 11).

3. Convergence Results

3.1. Assumptions

To estimate the convergence rate of Algorithm 1, we need to impose some restrictions on the workers' functions $\{F_i\}_{i=1}^n$ and define the properties of the compressor $\mathcal{C}$. For the latter, we assume a common assumption in the literature (Richtárik et al., 2021; Fatkhullin et al., 2021; Beznosikov et al., 2023) – a *biased contractive* operator with the following definition.

Definition 3.1. We say that the compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is *contractive* if there exists $\alpha \in (0, 1)$ such that

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leqslant (1-\alpha)\|x\|^2 \quad \text{for all } x \in \mathbb{R}^d.$$

The next step in evaluating the asymptotics of Algorithm 1 is formalizing the assumptions on local and global models. Specifically, adding additional trainable parameters $\{z_i\}_{i=1}^{n-1}$ to the problem (1) comes at a cost. Utilizing these auxiliary variables makes it less physically meaningful to assume the smoothness of the target function f – a standard prerequisite for other approaches utilizing the error feedback mechanism (Richtárik et al., 2021; Fatkhullin et al., 2021; Valdeira et al., 2024). Formally, the function f is called L-smooth if there exists $L \geqslant 0$ such that for any points $x, y \in \mathbb{R}^d$, the following inequality holds:

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

Nevertheless, the same property for the reformulated problem can be proved using assumptions that are often fulfilled in practice. Let us list them below.

The first assumption arises from the fact that in vertical deep learning, the local model F_i is a combination of linear layers and activation functions. It can be easily verified that such a

composition has bounded partial derivatives. Formally, the following statement applies to $\{F_i\}_{i=1}^n$.

Assumption 3.2. There exists $c \in \mathbb{R}$ such that for all $i = \overline{1,\dots,n}$, $k = \overline{1,\dots,d_{x_i}}$, $m = \overline{1,\dots,d_{z_{i-1}}}$, $x_i \in \mathbb{R}^{d_{x_i}}$, and $z_i \in \mathbb{R}^{d_{z_i}}$, the following inequality holds:

$$\left|\frac{\partial F_{im}(x_i, z_i)}{\partial x_{ik}}\right| \leqslant c.$$

Note that as a simple implication of this property, every function F_i is $c\sqrt{\frac{d_{z_{i-1}}}{d_{x_i}}}$-Lipschitz (we consider $d_{z_0} = 1$).

The second assumption requires the workers' functions to be coordinate-wise L-smooth, i.e., for every coordinate function F_{im}, the condition of smoothness holds.

Assumption 3.3. For $i \in \overline{1,n}$, every individual worker's function $F_i(y) : \mathbb{R}^{d_{x_i}+d_{z_i}} \to \mathbb{R}^{d_{z_{i-1}}}$ is coordinate-wise L-smooth, i.e., for any coordinate $m \in \overline{1, d_{z_{i-1}}}$ and for all $i \in \overline{1,n}$, there exists $L_{im} > 0$ such that

$$\|\nabla F_{im}(y_1) - \nabla F_{im}(y_2)\| \leqslant L_{im}\|y_1 - y_2\|.$$

The aforementioned assumptions are sufficient for the following estimation of the Lipschitz constant of the target function.

Lemma 3.4. *Consider the target function f defined as*

$$\begin{aligned} f(x_1,\dots,x_n,z_1,\dots,z_{n-1}) = F_1(x_1,z_1) \\ +\lambda_2\|F_2(x_2,z_2) - z_1\|^2 \\ +\dots \\ +\lambda_n\|F_n(x_n,z_n) - z_{n-1}\|^2. \end{aligned}$$

Suppose Assumptions 3.2 and 3.3 hold. If $i \in \overline{2,n}$, define L_{F_i} as

$$L_{F_i}^2 := 2\left(\sum_{m=1}^{d_{F_i}} L_{im}^2\right)\sup_{x_i,z_i,z_{i-1}}\|F_i(x_i,z_i) - z_{i-1}\|^2 + 2d_i^2c^4.$$

For $i = 1$, define L_{F_1} as

$$L_{F_1}^2 := \sum_{m=1}^{d_{F_1}} L_{1m}^2.$$

If $L_{F_i} < \infty$, then the target function f is L-smooth with

$$L = \sqrt{n}\max_{i\in\overline{1,n}} L_{F_i}.$$

Nevertheless, in the context of such a theorem, a problem arises where some L_{F_i} may not be finite. This happens because of the term under the supremum operator: $\|F_i(x_i,z_i) - z_{i-1}\|^2$, which can, in general, be unbounded from above. This issue is not present when F_i is a piecewise linear function, i.e., $L_{im} \equiv 0$. This implies that L_{F_i} does not depend on the term with the supremum and is identical to $2d_i^2c^4$. Another property that solves this problem is the boundedness of the $\|F_i(x_i,z_i) - z_{i-1}\|^2$ term. The sufficient conditions for this are: **(i)** the image of F_i is a bounded set, and **(ii)** z_{i-1} is located in it. This would imply that $L_{F_i} \leqslant 2\left(\sum_{m=1}^{d_{F_i}} L_{im}^2\right)(4D_i^2) + 2d_i^2c^4$, where D_i is the diameter of $\mathrm{Im}(F_i)$.

The restriction **(ii)** imposed on z_i is addressed by reformulating the problem as a constrained optimization problem with respect to $\{z_i\}_{i=1}^{n-1}$ variables, as formally written in the problem (3).

To surpass the restriction **(i)**, we need to overlook the nature of workers' function F_i. In deep learning, they are represented as a combination of linear and non-linear mappings. The last is called activation function and the question of boundedness or linearity of F_i depends on the boundedness or linearity of it. However, not all existing activation functions meet these restrictions. Nevertheless, the majority of them satisfy one property on one set of arguments and the second property on another, which is sufficient for the worker's function to be L-smooth. For example, the `ELU` activation is linear on the positive semi-axis and has a bounded range on the negative. We provide more examples of popular functions in Section 3.2. Considering this fact, we make the following assumption on the workers' functions F_i.

Assumption 3.5. We assume that the workers' functions F_i are either piecewise linear functions or have a bounded image over their entire domain.

Note that the requirement of linearity can be substituted with asymptotic convergence to a linear function. However, the convergence speed must be sufficiently fast.

3.2. Activation Functions Review

In this section, we provide a detailed analysis of popular activation functions. Namely, we prove that they follow Assumption 3.5.

ReLU. The most common activation function in DL is `ReLU` – Rectified Linear Unit, introduced in the work (Nair & Hinton, 2010).

$$\texttt{ReLU}(x) = \max(0,x) = \begin{cases} 0, & x \leqslant 0, \\ x, & x > 0. \end{cases}$$

It can be obviously seen, that despite being unbounded on $\mathbb{R}\setminus\{0\}$, `ReLU`'s second derivative on this set is equal to 0. In the point 0, the sub-differential of `ReLU` is equal to $[0,1]$, which after taking second sub-differential will also become

0. One can also notice, that LeakyReLU (Maas et al., 2013) will also share the same property.

ELU. Despite being popular, ReLU suffers from a major problem with differentiability at point 0. This issue was negated after introducing ELU (Exponential Linear Unit) activation function (Clevert et al., 2015).

$$\texttt{ELU}(x) = \begin{cases} \alpha(e^x - 1), & x \leqslant 0, \\ x, & x > 0, \end{cases}$$

where $\alpha > 0$. Note that

$$\forall x \in \mathbb{R}_-: \ -\alpha \leqslant \texttt{ELU}(x) \leqslant 0,$$

which means boundedness on the $\mathbb{R}_-$ set and for all $x \in \mathbb{R}_{++}$ the same reasoning as in ReLU can be applied.

Sigmoid/Tanh/Softmax. Classical activation functions, like the Sigmoid and Tanh, were widely employed in the early development of neural networks. However, these functions encountered challenges when it came to training deeper networks due to their tendency to produce saturated outputs. After the breakthrough of deep learning, the Softmax function is used as the activation function (Glorot et al., 2011). It produces the categorical probability distribution equivalent output.

$$\begin{aligned} \sigma(x) &= \frac{1}{1+\exp(-x)} = 1 - \sigma(-x), \\ \tanh(x) &= \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}, \\ [\texttt{Softmax}(x)]_i &= \frac{\exp(x_i)}{\sum\limits_i \exp(x_i)} \end{aligned}$$

All these activation functions are bounded, thus it is suitable for our purpose.

GELU. GELU, introduced in (Hendrycks & Gimpel, 2016), has gained prominence for its ability to enhance the learning capabilities of neural networks. Unlike its predecessors, GELU is derived from a smooth approximation of the cumulative distribution function (CDF) of the standard normal distribution

$$\texttt{GELU}(x) = x \cdot \text{cdf}(x) = \frac{x}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right).$$

Looking at the form of GELU activation function, one can note, that the following inequality holds for all $x \in \mathbb{R}_{++}$.

$$\texttt{GELU}(x) < \frac{x}{2}(1+1) = x = \texttt{ReLU}(x).$$

Its second derivative can be calculated in the following way.

$$\frac{\partial \texttt{GELU}(x)}{\partial x} = \frac{1}{2} + \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} + \frac{1}{2}\int_0^{\frac{x}{\sqrt{2}}} e^{-t^2} dt,$$

$$\frac{\partial^2 \texttt{GELU}(x)}{\partial x^2} = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\left(2 - x^2\right).$$

As exponent growth asymptotically faster, than power function, when $x \to \infty$ the second derivative will tend to zero. Moreover, GELU is bounded from below by -0.5.

Swish. This activation function was proposed by Google researchers in (Ramachandran et al., 2017). The idea behind Swish is to combine aspects of both ReLU and Sigmoid functions to provide better performance, particularly in deeper networks where vanishing or exploding gradients can impede learning:

$$\texttt{Swish}(x) = x \cdot \sigma(x).$$

The same analysis as for $\texttt{GELU}(x)$ function can be applied here, since it's obvious that

$$\forall x \in \mathbb{R}_{++}: \texttt{Swish}(x) < \texttt{ReLU}(x),$$

therefore the same approach works here.

3.3. Convergence Results

In order to estimate the asymptotics of Algorithm 1, let us overlook some general properties of constrained optimization. It can be noted that the problem of searching $\min\limits_{y \in Q} f(y)$ on some convex closed set Q can be reduced to the form of

$$\min_{y \in \mathbb{R}^d} \left[\Phi(y) := f(y) + \delta_Q(y)\right], \tag{4}$$

where δ_Q is the indicator function of the set Q. It can be noted, that the target function $\Phi(y)$ is represented as the sum of two functions, one of which $(f(y))$ is Lipschitz and the other $(\delta_Q(y))$ is convex. In general, such configuration, is called *composite* and formally looks like

$$\Phi(y) = f(y) + r(y).$$

To find the minimum of the function in this setup, the *proximal* operator with the following definition is used.

Definition 3.6. For any $\gamma > 0, y \in \mathbb{R}^d$, for the function $r: \mathbb{R}^d \to \mathbb{R}$ we define the proximal operator as

$$\text{prox}_{\gamma r}(y) := \underset{w \in \mathbb{R}^d}{\text{argmin}} \left\{ r(w) + \frac{1}{2\gamma}\|w - y\|^2 \right\}.$$

In our case, i.e, in the case of constrained optimization problem (3), the proximal operator can be viewed as generalized projection on an image of F_i, and for i -th worker looks like

$$\text{prox}_{\gamma \delta_{\text{Im}F_i}}(z_i) := \underset{w \in \text{Im}F_i}{\text{argmin}} \left\{ \frac{1}{2\gamma}\|w - z_i\|^2 \right\}.$$

The introduction of proximal operator makes possible for us to use lemmas from (Fatkhullin et al., 2021), where error

feedback mechanism, enhanced with proximal operators was researched (see Section B in Appendix).

However, with the transition to the composite case, the standard metric for non-convex optimization – the average of gradient norm of the target function $f(y)$ through every iteration is no longer a representative metric. Thus, we define the *generalized gradient mapping* at a point y with a parameter γ as

$$\mathcal{G}_\gamma(y) := \frac{1}{\gamma}\left(y - \text{prox}_{\gamma r}(y - \gamma \nabla f(y))\right).$$

The aforementioned metric is well-defined, as can be seen from the work (Beck, 2017). Specifically, the condition $\mathcal{G}_\gamma(y^*) = 0$ for some point $y^* \in \mathbb{R}^d$ is sufficient and necessary for y^* to be the stationary point of the problem (4).

Utilizing these facts, we are able to formulate the convergence theorem.

Theorem 3.7. *Denote y^t as a result of concatenation of vectors $\{x_i^t\}_{i=1}^n$ and $\{z_i^t\}_{i=1}^{n-1}$. Let Assumptions 3.2, 3.3 be fulfilled for the problem (3) and let compressor $\mathcal{C}$ be from Definition 3.1. Then after T iterations of Algorithm 1 the following statement is true:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\mathcal{G}_\gamma(y^t)\right\|^2 = \mathcal{O}\left(\frac{1}{T}\right).$$

Discussion. From Theorem 3.7, one can say that Algorithm 1, enjoys state-of-the-art convergence rate $\mathcal{O}\left(\frac{1}{T}\right)$ (Richtárik et al., 2021). The concrete guarantees can be found in Appendix. From them, it can be noted, that in the regime with no compression, i.e., with constant $\alpha = 1$ (see Definition 3.1), we recall the asymptotics, proportional to $\frac{L}{T}$, which is the same as in the GD algorithm with projection.

4. Stochastic Gradients

One of the popular techniques for reducing the amount of computations at each iteration is the batching method, in which the calculations are performed not on the whole dataset, but rather on some subset of it (Ruder, 2016). However, such a procedure introduces additional stochasticity, related to a random sample selection.

The usual assumption for such setting is the ability to represent the target function as the sum of other functions, where each one is associated with some subset of samples. It can be noted that the problem (3) can be equivalently reformulated in this way, if the local model at the first worker F_1 can be divided into this sum. Namely, it can be written in the following expression:

$$\min_{\substack{x_1\in\mathbb{R}^{d_{x_1}},\ldots,x_n\in\mathbb{R}^{d_{x_n}},\\ z_1\in\text{Im}(F_2),\ldots,z_{n-1}\in\text{Im}(F_n)}} \Bigg[f(x_1,\ldots,x_n,z_2,\ldots,z_{n-1}) := \sum_{j=1}^{s}\Big\{F_{1j}(x_1,z_1) + \lambda_2\|F_{2j}(x_2,z_2) - z_{1j}\|^2 + \ldots + \lambda_n\|F_{nj}(x_n,z_n) - z_{(n-1)j}\|^2\Big\}\Bigg]. \quad (5)$$

Here, $s \in \mathbb{R}$ denotes the total number of samples, and a vector with index j denotes a vector whose components corresponding to the j-th element of the dataset are equal to the original vector, and other values are nullified.

However, directly changing the function computation over the whole dataset to the computation on some fixed, small enough batch size, without careful stepsize tuning, can lead to theoretical non-convergence. To address this issue, we enhance Algorithm 1 with the PAGE update (Li et al., 2021), in which every iteration, with some probability, the whole gradient is computed, and otherwise, the step vector is adjusted with the difference between batch functions at the $(t+1)$-th and t-th points. We call this new approach SVFL-EF21-PAGE (see Algorithm 2 in the Appendix), and it remains identical to Algorithm 1, with the exception being the aforementioned update.

In order to estimate the asymptotics of this new approach, we introduce additional assumptions. Specifically, we consider the function representing the j-th sample, F_{1j}, to be L_{1j}-smooth. Formally, this can be written as the following statement.

Assumption 4.1. For $j \in \overline{1,s}$, every function F_{1j} is L_{1j}-smooth, i.e., there exists $L_{1j} > 0$ such that

$$\|\nabla F_{1j}(y_1) - \nabla F_{1j}(y_2)\| \leqslant L_{1j}\|y_1 - y_2\|.$$

Thus, we are able to write the following theorem.

Theorem 4.2. *Denote y^t as the result of concatenation of vectors $\{x_i^t\}_{i=1}^n$ and $\{z_i^t\}_{i=1}^{n-1}$. Let Assumptions 3.2, 3.3, and 4.1 be fulfilled for the problem (5), and let the compressor $\mathcal{C}$ be as in Definition 3.1. Then after T iterations of Algorithm 2 the following statement is true:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\mathcal{G}_\gamma(y^t)\right\|^2 = \mathcal{O}\left(\frac{1}{T}\right).$$

Discussion. As in non-stochastic case, our method achieves $\mathcal{O}\left(\frac{1}{T}\right)$ convergence rate. Specific dependencies of the asymptotics on parameters such as the update probabilities or the compression constant α can be found in Appendix. If compression is removed, i.e., if α is set to 1, we recall PAGE (Li et al., 2021) asymptotics.

5. Experiments

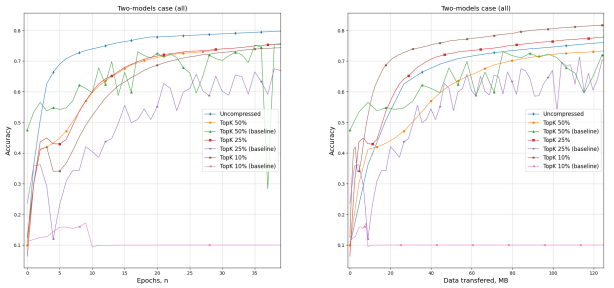
To reinforce our theoretical estimations, we conduct several experiments on Fashion-MNIST (F-MNIST) dataset (Xiao et al., 2017), consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 gray-scale image, associated with a label from 10 classes. We train a fully-connected neural network with four hidden layers, `ReLU`, `Sigmoid` and `Softmax` activation functions, divided between either two, or three sub-models. We compare the results of

- baseline, where gradient is calculated in the point of composition of compressed gradients;
- `SVFL-EF21` without any compression (`SGD`);
- `SVFL-EF21` with Top-$k\%$ compressor (leaves only $k\%$ of the biggest absolute values of the input).

5.1. Two-models case

In this scenario, we split the model into two sub-models, the first consists of 3 fully-connected layers and has access to the training samples, when the second one – two layers and access to the training labels. The results of accuracy metric on the F-MNIST dataset prediction are presented in the Fig. 1.

As one can notice, despite being detached from the whole computational graph, models achieve the expected values of the metric, compared to baseline, and also show superiority when comparing in terms of the number of information transmitted.



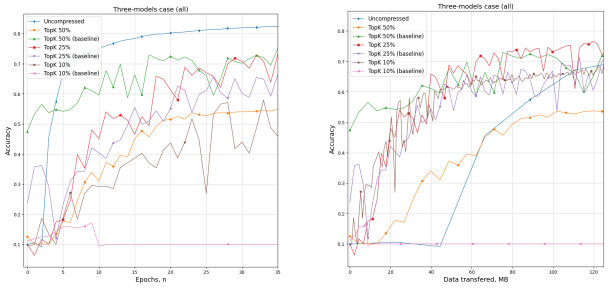
(a) Accuracy dependence on the number of epochs

(b) Accuracy dependence on the amount of data transmitted

Figure 1: The results of F-MNIST prediction for different methods on *two-models* case.

5.2. Three-models case

Here we split the model into three sub-models (the first and the second have two layers each and the third one has one layer), and emphasize, that the second model (that can be referred as F_2 in our notation) has no access to the data while training process. Its learning process is based completely on the embeddings z_2 and z_1, which are adjusted by the first and the third model. The results can be seen in the Fig. 2.



(a) Accuracy dependence on the number of epochs

(b) Accuracy dependence on the amount of data transmitted

Figure 2: The results of F-MNIST prediction for different methods on *three-models* case.

5.3. Discussion

The `SVFL-EF21` algorithm demonstrates comparable performance in accuracy/epochs graphs (Fig. 1a and Fig. 2a). Furthermore, as can be seen from the Fig. 1b and the Fig. 2b it surpasses the baseline results and achieves better score for less amount of data being transmitted. This confirms the theoretical computations and shows that split learning can be applied

6. Conclusion

This work introduces communication-efficient algorithm in multi-level vertical federated learning setup. We reformulate an optimization problem in order to surpass the obstacles, that appear in this setting. We achieve state-of-the-art theoretical results both in stochastic and deterministic cases, with convergence rate being equal to $\mathcal{O}\left(\frac{1}{T}\right)$. Moreover, we show a great practical performance, surpassing the non-compressed and baseline methods.

With this study, we hope to enhance distributed learning frameworks, as well as encourage other scientists, working in this field.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S.,

Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ajalloeian, A. and Stich, S. U. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Beznosikov, A., Kormakov, G., Grigorievskiy, A., Rudakov, M., Nazykov, R., Rogozin, A., Vakhrushev, A., Savchenko, A., Takáč, M., and Gasnikov, A. Exploring new frontiers in vertical federated learning: the role of saddle point reformulation. 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cai, D., Fan, T., Kang, Y., Fan, L., Xu, M., Wang, S., and Yang, Q. Accelerating vertical federated learning. *IEEE Transactions on Big Data*, pp. 1–10, 2022. ISSN 2372-2096. doi: 10.1109/tbdata.2022.3192898. URL http://dx.doi.org/10.1109/TBDATA.2022.3192898.

Castiglia, T. J., Das, A., Wang, S., and Patterson, S. Compressed-vfl: Communication-efficient learning with vertically partitioned data. In *International Conference on Machine Learning*, pp. 2738–2766. PMLR, 2022.

Chen, T., Jin, X., Sun, Y., and Yin, W. Vafl: a method of vertical asynchronous federated learning, 2020. URL https://arxiv.org/abs/2007.06081.

Chen, W., Ma, G., Fan, T., Kang, Y., Xu, Q., and Yang, Q. Secureboost+: A high performance gradient boosting tree framework for large scale vertical federated learning. *arXiv preprint arXiv:2110.10927*, 2021.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *Under Review of ICLR2016 (1997)*, 11 2015.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Khan, A., ten Thij, M., and Wilbik, A. Communication-efficient vertical federated learning. *Algorithms*, 15 (8), 2022. ISSN 1999-4893. doi: 10.3390/a15080273. URL https://www.mdpi.com/1999-4893/15/8/273.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., and Yang, Q. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, pp. 3. Atlanta, GA, 2013.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, pp. 1–16, 2024.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., and Watanabe, S. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pp. 1058–1062. Singapore, 2014.

Shi, S., Wang, Q., Xu, P., and Chu, X. Benchmarking state-of-the-art deep learning software tools. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pp. 99–104. IEEE, 2016.

Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Stanko, S., Karimullin, T., Beznosikov, A., and Gasnikov, A. Accelerated methods with compression for horizontal and vertical federated learning. *arXiv preprint arXiv:2404.13328*, 2024.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Sun, J., Xu, Z., Yang, D., Nath, V., Li, W., Zhao, C., Xu, D., Chen, Y., and Roth, H. R. Communication-efficient vertical federated learning with limited overlapping samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5203–5212, 2023.

Valdeira, P., Xavier, J., Soares, C., and Chi, Y. Communication-efficient vertical federated learning via compressed error feedback. *arXiv preprint arXiv:2406.14420*, 2024.

Vepakomma, P., Gupta, O., Swedish, T., and Raskar, R. Split learning for health: Distributed deep learning without sharing raw patient data. *CoRR*, abs/1812.00564, 2018. URL http://arxiv.org/abs/1812.00564.

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

Yang, L., Chai, D., Zhang, J., Jin, Y., Wang, L., Liu, H., Tian, H., Xu, Q., and Chen, K. A survey on vertical federated learning: From a layered perspective, 2023. URL https://arxiv.org/abs/2304.01829.

Appendix

A. Auxiliary Lemmas

Lemma A.1. *Suppose $A \in \mathbb{R}^{n\times m}$. Denote $\|A\|_p$ as the matrix norm, induced by vector p - norm. Then,*

$$\|A\|_2^2 \leqslant \|A\|_1 \|A\|_\infty. \tag{Spect}$$

The next lemma is a consequence of Jensen's inequality.

Lemma A.2. *For any vectors $a_1, a_2 \in \mathbb{R}^d$, for any $s > 0$ the following inequality holds*

$$\|a_1 + a_2\|^2 \leqslant (1+s)\|a_1\|^2 + (1+s^{-1})\|a_2\|^2 . \tag{Jensen}$$

The following statement is a generalization of the previous lemma for a sum of more than two vectors.

Lemma A.3. *For any vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ the following inequality holds*

$$\left\|\sum_{i=1}^n a_i\right\|^2 \leqslant n\sum_{i=1}^n \|a_i\|^2 . \tag{gen. Jensen}$$

Lemma A.4. *For any vectors $a, b \in \mathbb{R}^d$ the following inequality holds*

$$\left|\|a\|^2 - \|b\|^2\right| \leqslant \|a-b\|^2 . \tag{inv. triangle}$$

Lemma A.5. *For any vectors $a, b \in \mathbb{R}^d$ the following inequality holds*

$$2\langle a, b\rangle^2 \leqslant \|a\|\|b\| \tag{CBS}$$

The following Lemma is used for searching the solutions of quadratic inequalities.

Lemma A.6. *(Richtárik et al., 2021) If $a, b > 0$, then from*

$$0 \leqslant \gamma \leqslant \frac{1}{\sqrt{a}+b}$$

follows

$$a\gamma^2 + b\gamma \leqslant 1.$$

B. Main lemmas

Lemma B.1. *Consider the setup for optimization problem researched in this paper:*

$$f(x_1, \ldots, x_n, z_2, \ldots, z_n) = F_1(x_1, z_1) + \lambda_2\|F_2(x_2, z_2) - z_1\|^2 + \ldots + \lambda_n\|F_n(x_n, z_n) - z_{n-1}\|^2. \tag{6}$$

Denote L_{F_i} as the Lipschitz constant with respect to F_i and L as the Lipschitz constant of f. Then, for all $k > 1$ the following estimation holds:

$$L_{F_k} \geq 2\|F_k(x_k, z_k) - z_{k-1}\|^2 \left(\sum_{m=1}^{d_{F_k}} L_{F_k m}^2\right) + 2d_k^2 c_k^4,$$

where d_{F_k} stands for a dimension of function F_k, $L_{F_k m}$ denotes coordinate-wise Lipschitz constants of the function F_k and c_k is the maximum element in the Jacobian matrix J_{F_k}. For case of $k = 1$ the estimation has a slightly different kind:

$$L_{F_1}^2 \geq \sum_{m=1}^{d_{F_1}} L_{F_1 m}^2.$$

Moreover, function f is L-Lipschitz with

$$L = \sqrt{n}\max_i L_{F_i}.$$

Proof. We begin our proof with straightforward differentiation on x_k of the target function f with $k > 1$.

$$\nabla_{x_k} f(\dots, x_k, \dots, z_{k-1}, z_k, \dots) = 2\lambda_k \sum_{i=1}^{d_k} \langle \nabla F_{ki}(x_k, z_k), F_{ki}(x_k, z_k) - z_{k-1} \rangle = 2\lambda_k J_{F_k}^T (F_k(x_k, z_k) - z_{k-1})$$

where as J_{F_k} we denote the Jacobian of $F_k(x_k, z_k)$. Therefore, we can rewrite it in coordinate-wise view as

$$\frac{\partial f(\dots, x_k, \dots, z_{k-1}, z_k, \dots)}{\partial x_{ki}} = \left\langle \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}}, F_k(x_k, z_k) - z_{k-1} \right\rangle.$$

It is obvious, that if we take a mixed second derivative (with respect to x_l, for example), the result will be equal to 0, because

$$\frac{\partial^2 f(\dots, x_k, \dots, x_l, \dots, z_{k-1}, z_k, \dots, z_{l-1}, z_l, \dots)}{\partial x_{ki} \partial x_{lj}} = \frac{1}{\partial x_{lj}} \left\langle \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}}, F_k(x_k, z_k) - z_{k-1} \right\rangle = 0.$$

We estimate the Lipschitz-constant via calculating the Frobenius norm of the matrix $\frac{\partial^2 f(\dots, x_k, \dots, z_{k-1}, z_k, \dots)}{\partial x_{ki} \partial x_{kj}}$:

$$\begin{aligned}
\sum_{i=0}^{d_k} \sum_{j=0}^{d_k} \left| \frac{\partial^2 f(\dots, x_k, \dots, z_{k-1}, z_k, \dots)}{\partial x_{ki} \partial x_{kj}} \right|^2 &= \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} \left| \frac{\partial}{\partial x_{kj}} \left\langle \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}}, F_k(x_k, z_k) - z_{k-1} \right\rangle \right|^2 \\
&= \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} \left| \left\langle \frac{\partial^2 F_k(x_k, z_k)}{\partial x_{ki} \partial x_{kj}}, F_k(x_k, z_k) - z_{k-1} \right\rangle \right. \\
&+ \left. \left\langle \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}}, \frac{\partial F_k(x_k, z_k)}{\partial x_{kj}} \right\rangle \right|^2 \\
&\overset{(Jensen)}{\leqslant} \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} 2 \left\langle \frac{\partial^2 F_k(x_k, z_k)}{\partial x_{ki} \partial x_{kj}}, F_k(x_k, z_k) - z_{k-1} \right\rangle^2 \\
&+ \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} 2 \left\langle \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}}, \frac{\partial F_k(x_k, z_k)}{\partial x_{kj}} \right\rangle^2 \\
&\overset{(CBS)}{\leqslant} 2 \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} \left\| \frac{\partial^2 F_k(x_k, z_k)}{\partial x_{ki} \partial x_{kj}} \right\|^2 \|F_k(x_k, z_k) - z_{k-1}\|^2 \\
&+ 2 \sum_{i=0}^{d_k} \sum_{j=0}^{d_k} \left\| \frac{\partial F_k(x_k, z_k)}{\partial x_{ki}} \right\|^2 \left\| \frac{\partial F_k(x_k, z_k)}{\partial x_{kj}} \right\|^2 \\
&\overset{(\text{Ass.3.2})}{\leqslant} 2 \|F_k(x_k, z_k) - z_{k-1}\|^2 \left(\sum_{m=1}^{d_{F_k}} L_{F_k m}^2 \right) + 2 d_k^2 c_k^4
\end{aligned}$$

For case when $k = 1$ the easier estimation can be provided:

$$\begin{aligned}
\sum_{i=0}^{d_1} \sum_{j=0}^{d_1} \left| \frac{\partial^2 f(x_1, \dots, z_1, \dots)}{\partial x_{1i} \partial x_{1j}} \right|^2 &= \sum_{i=0}^{d_1} \sum_{j=0}^{d_1} \left| \frac{\partial^2 F(x_1, z_1)}{\partial x_{1i} \partial x_{1j}} \right|^2 \\
&\leqslant \sum_{m=1}^{d_{F_1}} L_m^2
\end{aligned}$$

In the similar way, we derive the Lipschitzness with respect to z_k. Note, that there are two terms, containing z_k (for $k > 1$): $\lambda_k \|F_k(x_k, z_k) - z_{k-1}\|^2$ and $\lambda_{k+1} \|F(x_{k+1}, z_{k+1}) - z_k\|^2$. It is obvious, that

$$\sum_{i=0}^{d_k} \sum_{j=1}^{d_k} \left| \frac{\partial^2 \left[\lambda_{k+1} \|F(x_{k+1}, z_{k+1}) - z_k\|^2 \right]}{\partial z_{ki} \partial z_{kj}} \right|^2 = 2\lambda_{k+1} \sum_{i=0}^{d_k} \sum_{j=1}^{d_k} 2 = 4\lambda_{k+1} d_k^2.$$

Due to symmetry, the first term has the same bound as previously made for x:

$$\sum_{i=0}^{d_k}\sum_{j=0}^{d_k}\left|\frac{\partial^2\left[\|F_k(x_k,z_k)-z_{k-1}\|^2\right]}{\partial x_{ki}\partial x_{kj}}\right|^2 \leq 2\|F_k(x_k,z_k)-z_{k-1}\|^2\left(\sum_{m=1}^{d_{F_k}}L_{F_km}^2\right)+2d_k^2c_k^4.$$

Finally, summing these results up, we obtain:

$$\sum_{i=0}^{d_k}\sum_{j=0}^{d_k}\left|\frac{\partial^2 f(\ldots,x_k,\ldots,z_{k-1},z_k,\ldots)}{\partial x_{ki}\partial x_{kj}}\right|^2 \leq 2\|F_k(x_k,z_k)-z_{k-1}\|^2\left(\sum_{m=1}^{d_{F_k}}L_{F_km}^2\right)+2d_k^2(2\lambda_{k+1}^2+\lambda_k^2c_k^4)$$

Let's denote $\{x_1,\ldots,x_n,\ldots,z_n\}$ as the first set of arguments y and similarly $\{\hat{x}_1,\ldots,\hat{x}_n,\ldots,\hat{z}_n\}$ as the second set of arguments $\hat{y}$. Then, the following holds:

$$\begin{aligned}
\|f(y)-f(\hat{y})\|^2 &= \|[F_1(x_1,z_1)-F_1(\hat{x}_1,\hat{z}_1)]+\ldots \\
&+ \lambda_n\left[\|F_n(x_n,z_n)-z_{n-1})\|^2-\|F_n(\hat{x}_n,\hat{z}_n)-\hat{z}_{n-1})\|^2\right]\Big\|^2 \\
&\overset{(gen.Jensen)}{\leqslant} n\Big(\|[F_1(x_1,z_1)-F_1(\hat{x}_1,\hat{z}_1)]\|^2 \\
&+ \sum_{i=2}^{n}\Big\|\|F_i(x_i,z_i)-z_{i-1})\|^2-\|F_i(\hat{x}_i,\hat{z}_i)-\hat{z}_{i-1})\|^2\Big\|^2\Big) \\
&\overset{(gen.Jensen)}{\leqslant} n\Big(\|[F_1(x_1,z_1)-F_1(\hat{x}_1,\hat{z}_1)]\|^2 \\
&+ \sum_{i=2}^{n}\|F_i(x_i,z_i)-z_{i-1})-F_i(\hat{x}_i,\hat{z}_i)-\hat{z}_{i-1})\|^2\Big) \\
&\leqslant n\Big(L_{F_1}^2\left[\|x_1-\hat{x}_1\|^2+\|z_1-\hat{z}_1\|^2\right] \\
&+ \ldots+L_{F_n}^2\left[\|x_n-\hat{x}_n\|^2+\|z_n-\hat{z}_{n-1}\|^2+\|z_n-\hat{z}_{n-1}\|^2\right]\Big) \\
&\leqslant n\max_i L_{F_i}^2\|y-\hat{y}\|^2.
\end{aligned}$$

□

We use lemmas from (Fatkhullin et al., 2021) for proximal EF21. We utilize such setup, as we need to use projection on z_i^{t+1} for our problem to be L -smooth. In the proximal setup, the target function looks like

$$\Phi(x) := f(x)+r(x),$$

where f is L-smooth and r is convex. For the projection on, Q we have to set proximal function as an indicator on needed set, i.e.,

$$r_{\text{proj}_Q}(x) = \delta_Q(x).$$

For such approach, we have to define a different convergence metric. For it, we use the squared norm of *generalized gradient mapping*, which is defined as

$$\mathcal{G}_\gamma(x) := \frac{1}{\gamma}\left(x-\text{prox}_{\gamma r}(x-\gamma\nabla f(x))\right)$$

Lemma B.2. *(Lemma 13 from (Fatkhullin et al., 2021)) Suppose the function $\Phi(x)=f(x)+r(x)$, where f is L-smooth and r is convex and let $y^{t+1}=prox_{\gamma r}(y^t-\gamma g^t)$, where $g^t\in\mathbb{R}^d$ and $\gamma>0$. Then we have*

$$\mathbb{E}\|\mathcal{G}_\gamma(x^t)\|^2 \leqslant \frac{2}{\gamma^2}\mathbb{E}\|x^{t+1}-x^t\|^2+2\mathbb{E}\|g^t-\nabla f(x^t)\|^2.$$

Lemma B.3. *(Lemma 14 from (Fatkhullin et al., 2021)) Suppose the function $\Phi(x) = f(x) + r(x)$, where f is L-smooth and r is convex and let $y^{t+1} = prox_{\gamma r}(y^t - \gamma g^t)$, where $g^t \in \mathbb{R}^d$ and $\gamma > 0$. Then we have for any $\lambda > 0$*

$$\Phi(y^{t+1}) \leqslant \Phi(y^t) - \left(\frac{1}{\gamma} - \frac{L}{2} - \frac{\lambda}{2}\right)\left\|y^{t+1} - y^t\right\|^2 + \frac{1}{2\lambda}\left\|g^t - \nabla f(y^t)\right\|^2.$$

Lemma B.4. *In Algorithm 1 the following inequality is fulfilled.*

$$\sum_{i=2}^{n}\left\|g_{x_i}^t - \nabla_{x_i} f(y^t)\right\|^2 \leqslant \sum_{i=2}^{n} 4\lambda_i^2 d_{x_i} d_{z_i} c^2 \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2$$

Proof.

$$\begin{aligned}
\sum_{i=2}^{n}\left\|g_{x_i}^t - \nabla_{x_i} f(y^t)\right\|^2 &= \sum_{i=2}^{n}\left\|2\lambda_i J_{F_i,x_i}^T\left(x_i^t, z_i^t\right)\mathcal{Z}_{i-1}^t - 2\lambda_i J_{F_i,x_i}^T\left(x_i^t, z_i^t\right) z_{i-1}^t\right\|^2 \\
&\leqslant \sum_{i=2}^{n} 4\lambda_i^2\left\|J_{F_i,x_i}^T\left(x_i^t, z_i^t\right)\right\|_2^2\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\overset{(Spect)}{\leqslant} \sum_{i=2}^{n} 4\lambda_i^2\left\|J_{F_i,x_i}^T\left(x_i^t, z_i^t\right)\right\|_1\left\|J_{F_i,x_i}^T\left(x_i^t, z_i^t\right)\right\|_\infty\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\overset{\text{(Ass. 3.2)}}{\leqslant} \sum_{i=2}^{n} 4\lambda_i^2 d_{x_i} d_{z_{i-1}} c^2\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2.
\end{aligned}$$

The first inequality comes from the definition of operator norm. □

Lemma B.5. *In Algorithm 1, the following inequality holds:*

$$\begin{aligned}
\sum_{i=1}^{n-1}\left\|g_{z_i}^t - \nabla_{z_i} f(y^t)\right\|^2 \leqslant 4\sum_{i=1}^{n-1}\left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(\mathcal{Z}_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
+ 8\sum_{i=2}^{n}\lambda_i^2\left(2 + c^2 d_{z_i} d_{z_{i-1}}\right)\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2.
\end{aligned}$$

Proof. We proceed as follows:

$$\begin{aligned}
\sum_{i=1}^{n-1}\left\|g_{z_i}^t - \nabla_{z_i} f(y^t)\right\|^2 &= \left\|\mathcal{H}_2^t - 2\lambda_2\left(z_1^t - F_2(x_2^t, z_2^t)\right)\right\|^2 \\
&\quad + \sum_{i=2}^{n-1}\Big\|\mathcal{H}_{i+1}^t + 2\lambda_i J_{F_i,z_i}^T(x_i^t, z_i^t)\left(F_i(x_i^t, z_i^t) - \mathcal{Z}_{i-1}^t\right) \\
&\quad - \left(2\lambda_{i+1}\left(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right) + 2\lambda_i J_{F_i,z_i}^T(x_i^t, z_i^t)\left(F_i(x_i^t, z_i^t) - z_{i-1}^t\right)\right)\Big\|^2 \\
&\overset{(Jensen)}{\leqslant} \left\|\mathcal{H}_2^t - 2\lambda_2\left(z_1^t - F_2(x_2^t, z_2^t)\right)\right\|^2 \\
&\quad + 2\sum_{i=2}^{n-1}\left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&\quad + 2\sum_{i=2}^{n-1}\left\|2\lambda_i J_{F_i,z_i}^T(x_i^t, z_i^t)\left(\mathcal{Z}_{i-1}^t - z_{i-1}^t\right)\right\|^2 \\
&\overset{(Spect)}{\leqslant} \left\|\mathcal{H}_2^t - 2\lambda_2\left(z_1^t - F_2(x_2^t, z_2^t)\right)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{i=2}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&+ 8\sum_{i=2}^{n-1} \lambda_i^2 \left\|J_{F_i,z_i}^T(x_i^t, z_i^t)\right\|_1 \left\|J_{F_i,z_i}^T(x_i^t, z_i^t)\right\|_\infty \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\overset{\text{(Ass. 3.2)}}{\leqslant} \left\|\mathcal{H}_2^t - 2\lambda_2\left(z_1^t - F_2(x_2^t, z_2^t)\right)\right\|^2 \\
&+ 2\sum_{i=2}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&+ 8\sum_{i=2}^{n-1} \lambda_i^2 c^2 d_{z_i} d_{z_{i-1}} \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\leqslant 2\sum_{i=1}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&+ 8\sum_{i=2}^{n-1} \lambda_i^2 c^2 d_{z_i} d_{z_{i-1}} \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\overset{(Jensen)}{\leqslant} 4\sum_{i=1}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(\mathcal{Z}_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&+ 16\sum_{i=1}^{n-1} \lambda_{i+1}^2 \left\|\mathcal{Z}_i^t - z_i^t\right\|^2 \\
&+ 8\sum_{i=2}^{n-1} \lambda_i^2 c^2 d_{z_i} d_{z_{i-1}} \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&= 4\sum_{i=1}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(\mathcal{Z}_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&+ 8\sum_{i=2}^{n} \lambda_i^2 \left(2 + c^2 d_{z_i} d_{z_{i-1}}\right) \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2.
\end{aligned}
$$

□

Lemma B.6. *For $i = 1, \ldots, n-1$, $\forall s_1 > 0$ the following inequality is fulfilled.*

$$
\mathbb{E}\left\|\mathcal{Z}_i^{t+1} - z_i^{t+1}\right\|^2 \leqslant (1-\alpha)(1+s_1)\mathbb{E}\left\|\mathcal{Z}_i^t - z_i^t\right\|^2 + (1-\alpha)(1+s_1^{-1})\mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2.
$$

Proof.

$$
\begin{aligned}
\mathbb{E}\left\|\mathcal{Z}_i^{t+1} - z_i^{t+1}\right\|^2 &= \mathbb{E}\left\|\text{proj}_{\text{Im}F_{i+1}}\left(\mathcal{Z}_i^t + \mathcal{C}\left(z_i^{t+1} - \mathcal{Z}_i^t\right)\right) - z_i^{t+1}\right\|^2 \\
&\overset{(*)}{=} \mathbb{E}\left\|\text{proj}_{\text{Im}F_{i+1}}\left(\mathcal{Z}_i^t + \mathcal{C}\left(z_i^{t+1} - \mathcal{Z}_i^t\right)\right) - \text{proj}_{\text{Im}F_{i+1}}\left(z_i^{t+1}\right)\right\|^2 \\
&\overset{(**)}{\leqslant} \mathbb{E}\left\|\mathcal{Z}_i^t + \mathcal{C}\left(z_i^{t+1} - \mathcal{Z}_i^t\right) - z_i^{t+1}\right\|^2 \\
&\overset{\text{(Ass. 3.1)}}{\leqslant} (1-\alpha)\mathbb{E}\left\|\mathcal{Z}_i^t - z_i^{t+1}\right\|^2 \\
&\overset{(Jensen)}{\leqslant} (1-\alpha)(1+s_1)\mathbb{E}\left\|\mathcal{Z}_i^t - z_i^t\right\|^2 + (1-\alpha)(1+s_1^{-1})\mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2.
\end{aligned}
$$

□

The equality * comes from the fact that the projection of a vector that already lies in the set equals to the vector itself. The inequality ** is followed from the 1-Lipschitzness of the projection operator.

Lemma B.7. *For $i = 2, \dots, n$, and for all $s_2 > 0$, the following inequality holds:*

$$
\begin{aligned}
&\mathbb{E}\left\|\mathcal{H}_i^{t+1} - 2\lambda_i(\mathcal{Z}_{i-1}^{t+1} - F_i(x_i^{t+1}, z_i^{t+1}))\right\|^2 \\
&\leqslant (1-\alpha)(1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\left((1-\alpha)(1+s_1)+1\right)\left\|\mathcal{Z}_{i-1}^{t} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right)\left\|z_{i-1}^{t+1} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\left\|x_i^{t+1} - x_i^{t}\right\|^2 + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\left\|z_i^{t+1} - z_i^{t}\right\|^2.
\end{aligned}
$$

Proof. We proceed as follows:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathcal{H}_i^{t+1} - 2\lambda_i(\mathcal{Z}_{i-1}^{t+1} - F_i(x_i^{t+1}, z_i^{t+1}))\right\|^2 \\
&= \mathbb{E}\left\|\mathcal{H}_i^{t} + \mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t})) - \mathcal{H}_i^{t}\right) - 2\lambda_i(\mathcal{Z}_{i-1}^{t+1} - F_i(x_i^{t+1}, z_i^{t+1}))\right\|^2 \\
&\overset{(Jensen)}{\leqslant} (1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} + \mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t})) - \mathcal{H}_i^{t}\right) - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + \left(1+s_2^{-1}\right)\mathbb{E}\left\|2\lambda_i(\mathcal{Z}_{i-1}^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})) - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\overset{\text{(Ass. 3.1)}}{\leqslant} (1-\alpha)(1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + 4\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|(\mathcal{Z}_{i-1}^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})) - (\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\overset{(gen.Jensen)}{\leqslant} (1-\alpha)(1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t+1} - z_{i-1}^{t+1}\right\|^2 + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^{t}\right\|^2 + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|F_i(x_i^{t+1}, z_i^{t+1}) - F_i(x_i^{t}, z_i^{t})\right\|^2 \\
&\overset{\text{(Ass. 3.2)}}{\leqslant} (1-\alpha)(1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t+1} - z_{i-1}^{t+1}\right\|^2 + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\mathbb{E}\left\|x_i^{t+1} - x_i^{t}\right\|^2 + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_i^{t+1} - z_i^{t}\right\|^2 \\
&\overset{(B.6)}{\leqslant} (1-\alpha)(1+s_2)\,\mathbb{E}\left\|\mathcal{H}_i^{t} - 2\lambda_i(\mathcal{Z}_{i-1}^{t} - F_i(x_i^{t}, z_i^{t}))\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\left((1-\alpha)(1+s_1)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t} - z_{i-1}^{t}\right\|^2 + (1-\alpha)(1+s_1^{-1})\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^{t}\right\|^2\right) \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^{t}\right\|^2 \\
&\quad + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\mathbb{E}\left\|x_i^{t+1} - x_i^{t}\right\|^2 + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_i^{t+1} - z_i^{t}\right\|^2.
\end{aligned}
$$

□

In the inequality with reference to Ass. 3.2 we utilize the definition of $c\sqrt{\frac{d_{z_{i-1}}}{d_{x_i}}}$-Lipschitz function.

C. Main theorems

Let us define some notations for the sequences.

$$\text{(sequence of } g^t_{z_i}) \qquad \Omega^t_{z_i} := \mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2 \tag{7}$$

$$\text{(sequence of } g^t_{x_i}) \qquad \Omega^t_{x_i} := \mathbb{E}\left\|x_i^{t+1} - x_i^t\right\|^2 \tag{8}$$

$$\text{(compression error for } \mathcal{H}^t_i) \qquad D^t_{\mathcal{H}_i} := \mathbb{E}\left\|\mathcal{H}_i^t - 2\lambda_i(\mathcal{Z}_{i-1}^t - F_i(x_i^t, z_i^t))\right\|^2 \tag{9}$$

$$\text{(compression error for } \mathcal{Z}^t_i) \qquad D^t_{\mathcal{Z}_i} := \mathbb{E}\left\|\mathcal{Z}_i^t - z_i^t\right\|^2 \tag{10}$$

C.1. Proof of Theorem 3.7

Proof. Starting from Lemma B.3, we have:

$$\begin{aligned}
\Phi(y^{t+1}) &\overset{(B.3)}{\leqslant} \Phi(y^t) - \left(\frac{1}{\gamma} - \frac{L}{2} - \frac{\lambda}{2}\right)\left\|y^{t+1} - y^t\right\|^2 + \frac{1}{2\lambda}\left\|g^t - \nabla f(y^t)\right\|^2 \\
&= \Phi(y^t) - \left(\frac{1}{\gamma} - \frac{L}{2} - \frac{\lambda}{2} - \frac{1}{4\gamma}\right)\left\|y^{t+1} - y^t\right\|^2 \\
&\quad + \left(\frac{1}{2\lambda} + \frac{\gamma}{4}\right)\left\|g^t - \nabla f(y^t)\right\|^2 - \frac{1}{4\gamma}\left\|y^{t+1} - y^t\right\|^2 - \frac{\gamma}{4}\left\|g^t - \nabla f(y^t)\right\|^2.
\end{aligned}$$

As L, we denote the Lipschitz constant from Lemma 3.4. Utilizing Lemma B.2 and setting $\lambda = \frac{1}{2\gamma}$, we obtain the following inequality:

$$\begin{aligned}
\Phi(y^{t+1}) &\overset{(B.2)}{\leqslant} \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{\gamma} - \frac{L}{2} - \frac{\lambda}{2} - \frac{1}{4\gamma}\right)\left\|y^{t+1} - y^t\right\|^2 \\
&\quad + \left(\frac{1}{2\lambda} + \frac{\gamma}{4}\right)\left\|g^t - \nabla f(y^t)\right\|^2 \\
&= \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 + \frac{5\gamma}{4}\left\|g^t - \nabla f(y^t)\right\|^2.
\end{aligned} \tag{11}$$

From the fact that the squared Euclidean norm of a vector can be written as the sum of the squared norms of its sub-vectors, the following expression holds:

$$\begin{aligned}
\Phi(y^{t+1}) = \Phi(y^t) &- \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 \\
&+ \frac{5\gamma}{4}\left(\sum_{i=2}^{n}\left\|g^t_{x_i} - \nabla_{x_i} f(y^t)\right\|^2 + \sum_{i=1}^{n-1}\left\|g^t_{z_i} - \nabla_{z_i} f(y^t)\right\|^2\right).
\end{aligned}$$

Note that according to Algorithm 1, the term $\left\|g^t_{x_1} - \nabla_{x_1} f(y^t)\right\|^2$ is equal to zero. Using Lemma B.4 and Lemma B.5, we proceed with another inequality:

$$\begin{aligned}
\Phi(y^{t+1}) &\overset{(B.4,B.5)}{\leqslant} \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 \\
&\quad + 5\gamma\sum_{i=2}^{n}\lambda_i^2 d_{x_i} d_{z_{i-1}} c^2\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&\quad + 5\gamma\sum_{i=1}^{n-1}\left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(\mathcal{Z}_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2 \\
&\quad + 10\gamma\sum_{i=2}^{n}\lambda_i^2\left(2 + c^2 d_{z_i} d_{z_{i-1}}\right)\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&= \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 \\
&\quad + 5\gamma \sum_{i=1}^{n-1} \lambda_{i+1}^2 \left(d_{z_i} c^2 \left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right) \left\|\mathcal{Z}_i^t - z_i^t\right\|^2 \\
&\quad + 5\gamma \sum_{i=1}^{n-1} \left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}\left(\mathcal{Z}_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t)\right)\right\|^2.
\end{aligned}
$$

Next, we take the mathematical expectation and redefine this inequality according to notations (7)–(10):

$$
\begin{aligned}
\mathbb{E}\Phi(y^{t+1}) &\leqslant \mathbb{E}\Phi(y^t) - \frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left(\sum_{i=1}^{n}\Omega_{x_i}^t + \sum_{i=1}^{n-1}\Omega_{z_i}^t\right) \\
&\quad + 5\gamma \sum_{i=1}^{n-1} \lambda_{i+1}^2 \left(d_{z_i} c^2 \left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right) D_{\mathcal{Z}_i}^t + 5\gamma \sum_{i=2}^{n} D_{\mathcal{H}_i}^t. \qquad (12)
\end{aligned}
$$

From Lemma B.6, we have:

$$
D_{\mathcal{Z}_i}^{t+1} \leqslant (1-\alpha)(1+s_1) D_{\mathcal{Z}_i}^t + (1-\alpha)(1+s_1^{-1})\Omega_{z_i}^t. \qquad (13)
$$

And from Lemma B.7, the following inequality is fulfilled:

$$
\begin{aligned}
D_{\mathcal{H}_i}^{t+1} &\leqslant (1-\alpha)\left(1+s_2\right) D_{\mathcal{H}_i}^t \\
&\quad + 16\lambda_i^2 \left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1\right)+1\right) D_{\mathcal{Z}_{i-1}}^t \\
&\quad + 16\lambda_i^2 \left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right) \Omega_{z_{i-1}}^t \\
&\quad + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}} \left(1+s_2^{-1}\right) \Omega_{x_i}^t + 16\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}} \left(1+s_2^{-1}\right) \Omega_{z_i}^t. \qquad (14)
\end{aligned}
$$

For $i = 1, \ldots, (n-1)$ we multiply (13) by ω_{z_i} and add to (12). Similarly, for inequality (14) we multiply it by ω_{h_i} for $i = 2, \ldots, n$ and add to (12). Thus, we obtain:

$$
\begin{aligned}
&\mathbb{E}\Phi(y^{t+1}) + \sum_{i=1}^{n-1} \omega_{z_i} D_{\mathcal{Z}_i}^{t+1} + \sum_{i=2}^{n} \omega_{h_i} D_{\mathcal{H}_i}^{t+1} \\
&\leqslant \mathbb{E}\Phi(y^t) - \frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2 \\
&\quad - \sum_{i=1}^{n} \left(\frac{1}{2\gamma} - \frac{L}{2} - 16\omega_{h_i}\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}} \left(1+s_2^{-1}\right)\right) \Omega_{x_i}^t \\
&\quad - \sum_{i=1}^{n-1} \left(\frac{1}{2\gamma} - \frac{L}{2}\right) \Omega_{z_i}^t \\
&\quad + 5\gamma \sum_{i=1}^{n-1} \lambda_{i+1}^2 \left(d_{z_i} c^2 \left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right) D_{\mathcal{Z}_i}^t \\
&\quad + 5\gamma \sum_{i=2}^{n} D_{\mathcal{H}_i}^t \\
&\quad + 16 \sum_{i=2}^{n} \omega_{h_i} \lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}} \left(1+s_2^{-1}\right) \Omega_{z_i}^t \\
&\quad + 16 \sum_{i=2}^{n} \omega_{h_i} \lambda_i^2 \left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right) \Omega_{z_{i-1}}^t
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=2}^{n}\omega_{h_i}\left(1-\alpha\right)\left(1+s_2\right)D_{\mathcal{H}_i}^t\\
&+16\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2\left(1+s_2^{-1}\right)\left(\left(1-\alpha\right)\left(1+s_1\right)+1\right)D_{\mathcal{Z}_{i-1}}^t\\
&+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)(1+s_1)D_{\mathcal{Z}_i}^t\\
&+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)(1+s_1^{-1})\Omega_{z_i}^t.
\end{aligned}
$$

Now we rearrange indices in sums of $\Omega_{z_{i-1}}^t$ and $D_{\mathcal{Z}_{i-1}}^t$. As $\Omega_{z_n}=0$ we reduce sum in Ω_{z_i} from n to $n-1$.

$$
\begin{aligned}
&\mathbb{E}\Phi(y^{t+1})+\sum_{i=1}^{n-1}\omega_{z_i}D_{\mathcal{Z}_i}^{t+1}+\sum_{i=2}^{n}\omega_{h_i}D_{\mathcal{H}_i}^{t+1}\\
&\leqslant\mathbb{E}\Phi(y^t)-\frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2\\
&-\sum_{i=1}^{n}\left(\frac{1}{2\gamma}-\frac{L}{2}-16\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\right)\Omega_{x_i}^t\\
&-\sum_{i=1}^{n-1}\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\Omega_{z_i}^t\\
&+5\gamma\sum_{i=1}^{n-1}\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)+4\right)D_{\mathcal{Z}_i}^t\\
&+5\gamma\sum_{i=2}^{n}D_{\mathcal{H}_i}^t\\
&+16\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\Omega_{z_i}^t\\
&+16\sum_{i=1}^{n-1}\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left(\left(1-\alpha\right)\left(1+s_1^{-1}\right)+1\right)\Omega_{z_i}^t\\
&+\sum_{i=2}^{n}\omega_{h_i}\left(1-\alpha\right)\left(1+s_2\right)D_{\mathcal{H}_i}^t\\
&+16\sum_{i=1}^{n-1}\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left(\left(1-\alpha\right)\left(1+s_1\right)+1\right)D_{\mathcal{Z}_i}^t\\
&+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)(1+s_1)D_{\mathcal{Z}_i}^t\\
&+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)\left(1+s_1^{-1}\right)\Omega_{z_i}^t.
\end{aligned}
$$

After adding the term $16\omega_{h_1}\lambda_1^2c^2\frac{d_{z_0}}{d_{x_1}}\left(1+s_2^{-1}\right)\Omega_{z_1}^t$ (we consider $d_{z_0}=1$) and regrouping, we obtain the following inequality:

$$
\mathbb{E}\Phi(y^{t+1})+\sum_{i=1}^{n-1}\left(\omega_{z_i}D_{\mathcal{Z}_i}^{t+1}-\psi_{z_i}D_{\mathcal{Z}_i}^t\right)+\sum_{i=2}^{n}\left(\omega_{h_i}D_{\mathcal{H}_i}^{t+1}-\psi_{h_i}D_{\mathcal{H}_i}^t\right)
$$

$$\leqslant \Phi(y^t) - \frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2 - \sum_{i=1}^{n}\phi_{x_i}\Omega^t_{x_i} - \sum_{i=1}^{n-1}\phi_{z_i}\Omega^t_{z_i}. \tag{15}$$

Where

$$\begin{aligned}
\psi_{z_i} &= 5\gamma\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right)\\
&\quad + 16\omega_{h_{i+1}}\lambda_{i+1}^2\left(1 + s_2^{-1}\right)\left((1-\alpha)\left(1+s_1\right) + 1\right)\\
&\quad + \omega_{z_i}(1-\alpha)(1+s_1),\\
\psi_{h_i} &= 5\gamma + \omega_{h_i}\left(1+s_2\right)(1-\alpha),\\
\phi_{z_i} &= \frac{1}{2\gamma} - \frac{L}{2} - 16\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)\\
&\quad - 16\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right) - \omega_{z_i}(1-\alpha)\left(1+s_1^{-1}\right),\\
\phi_{x_i} &= \begin{cases}\frac{1}{2\gamma} - \frac{L}{2} - 16\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right), i = \overline{2,n},\\ \frac{1}{2\gamma} - \frac{L}{2}, i = 1.\end{cases}
\end{aligned}$$

It can be noted that if for any possible i, $\phi_{x_i} \geqslant 0$ and $\phi_{z_i} \geqslant 0$, the sums of $\Omega^t_{x_i}$ and $\Omega^t_{z_i}$ can be dropped. Also, if for any i, $\psi_{z_i} \leqslant \omega_{z_i}$ and $\psi_{h_i} \leqslant \omega_{h_i}$, the terms $\omega_{z_i}D^{t+1}_{\mathcal{Z}_i} - \psi_{z_i}D^t_{\mathcal{Z}_i}$ and $\omega_{h_i}D^{t+1}_{\mathcal{H}_i} - \psi_{h_i}D^t_{\mathcal{H}_i}$ can be recursively reduced, using a telescopic sum on t.

Let us overlook these conditions. They bound $\omega_{h_i}, i = \overline{2,n}$ with the following inequalities:

$$\frac{5\gamma}{1-(1+s_2)(1-\alpha)} \leqslant \omega_{h_i} \leqslant \frac{1-\gamma L}{32\gamma\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)}. \tag{16}$$

In order for such ω_{h_i} to exist, the right bound of this inequality must be greater than the left bound. This formally can be written as:

$$\frac{5\gamma}{1-(1+s_2)(1-\alpha)} \leqslant \frac{1-\gamma L}{32\gamma\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)}. \tag{17}$$

Equivalently,

$$\gamma^2\left(\frac{32\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)}{1-(1+s_2)(1-\alpha)}\right) + \gamma L \leqslant 1.$$

Using Lemma A.6, we have the solution:

$$0 \leqslant \gamma \leqslant \frac{1}{\lambda_i c\sqrt{\frac{32d_{z_{i-1}}\left(1+s_2^{-1}\right)}{d_{x_i}(1-(1+s_2)(1-\alpha))}} + L}.$$

We also want ω_{h_i} to be greater than 0; thus, we must ensure that the left bound in (17) is also greater than 0. Therefore,

$$1-(1+s_2)(1-\alpha) > 0.$$

Or,

$$(1+s_2) < \frac{1}{1-\alpha}. \tag{18}$$

In order to maximize the right bound of possible steps, we need to minimize $\frac{\left(1+s_2^{-1}\right)}{1-(1+s_2)(1-\alpha)}$ under the condition of positive ω_{h_i} (see (18)). Formally, it can be written as the following constrained optimization problem:

$$\min_{s_2}\left\{\frac{\left(1+s_2^{-1}\right)}{1-(1+s_2)(1-\alpha)}\,\middle|\,(1+s_2) < \frac{1}{1-\alpha},\ s_2 > 0\right\}. \tag{19}$$

It was solved in (Richtárik et al., 2021), and the solution is $s_2^* = \frac{1}{(1-\alpha)^{\frac{1}{2}}} - 1$. The target expression itself equals to:

$$\frac{\left(1+(s_2^*)^{-1}\right)}{1-(1+s_2^*)(1-\alpha)} = \frac{1}{\left(1-\sqrt{1-\alpha}\right)^2}.$$

And for the stepsize γ^* with the optimal right bound:

$$0 \leqslant \gamma^* \leqslant \frac{1}{\lambda_i c \sqrt{\frac{32 d_{z_{i-1}}}{d_{x_i}\left(1-\sqrt{1-\alpha}\right)^2}} + L}. \tag{20}$$

For ϕ_{x_1}, we have the condition:

$$0 \leqslant \gamma^* \leqslant \frac{1}{L}.$$

Similarly, we have for $\omega_{z_i}, i = \overline{1, n-1}$:

$$\begin{aligned}\omega_{z_i} \leqslant & \frac{1-\gamma L - 32\gamma\omega_{h_i}\lambda_i^2 c^2 \frac{d_{z_{i-1}}}{d_{x_i}}\left(1+s_2^{-1}\right)}{2\gamma(1-\alpha)(1+s_1^{-1})} \\ & - \frac{32\gamma\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right)}{2\gamma(1-\alpha)(1+s_1^{-1})},\end{aligned} \tag{21}$$

and

$$\frac{5\gamma\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)+4\right)+16\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1\right)+1\right)}{1-(1-\alpha)(1+s_1)} \leqslant \omega_{z_i}. \tag{22}$$

We set for all i, $\omega_{h_i} = \frac{5\gamma}{1-(1+s_2)(1-\alpha)}$, i.e., as the left bound of (16). Substituting it into the condition of the lower bound (21) being less than the upper bound (22), we have:

$$a_{z_i}\gamma^2 + L\gamma \leqslant 1,$$

where

$$\begin{aligned}a_{z_i} = & \left(\frac{10\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)+4\right)}{1-(1-\alpha)(1+s_1)}\right)(1-\alpha)(1+s_1^{-1}) \\ & + \left(\frac{160\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1\right)+1\right)}{1-(1+s_2)(1-\alpha)}\right)(1-\alpha)(1+s_1^{-1}) \\ & + 160\frac{\lambda_{i+1}^2\left(1+s_2^{-1}\right)\left((1-\alpha)\left(1+s_1^{-1}\right)+1\right)}{1-(1+s_2)(1-\alpha)}.\end{aligned}$$

Using Lemma A.6 again, we have the solution for it, which is written as:

$$0 \leqslant \gamma \leqslant \frac{1}{\sqrt{a_{z_i}} + L}.$$

We take $s_1^* = s_2^* = \frac{1}{(1-\alpha)^{\frac{1}{2}}} - 1$. Therefore, we have the concrete $a_{z_i} = a_{z_i}^*$:

$$\begin{aligned}a_{z_i}^* = & \left(10\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)+4\right)+160\lambda_{i+1}^2\frac{1+\sqrt{1-\alpha}}{\left(1-\sqrt{1-\alpha}\right)^2}\right)\frac{1-\alpha}{\left(1-\sqrt{1-\alpha}\right)^2} \\ & + 160\lambda_{i+1}^2\frac{1+\sqrt{1-\alpha}}{\left(1-\sqrt{1-\alpha}\right)^2}.\end{aligned}$$

And

$$0 \leqslant \gamma \leqslant \frac{1}{\sqrt{a_{z_i}^*} + L}. \tag{23}$$

We take γ as the minimum of the higher bounds of (20) and (23), so both of these conditions are always fulfilled.

Returning to (15) and using recursion on t, we have

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\mathcal{G}_\gamma(y^t)\right\|^2 \leq \frac{2\Delta}{\gamma T} + \frac{2\sum_{i=2}^{n}\omega_{h_i}D_{\mathcal{H}_i}^0}{\gamma T} + \frac{2\sum_{i=1}^{n-1}\omega_{z_i}D_{\mathcal{Z}_i}^0}{\gamma T},$$

where $\Delta := \Phi(x^0) - \Phi(x^*)$. Inserting ω_{h_i} and ω_{z_i} as low as possible, and γ as high as possible, we obtain the asymptotics. The maximum possible γ, which we call γ^*, is the minimum of the right bound of (20) and the right bound of (23):

$$\gamma^* = \min\left\{\min_{i=\overline{2,n}}\frac{1}{\lambda_i c\sqrt{\frac{32d_{z_{i-1}}}{d_{x_i}\left(1-\sqrt{1-\alpha}\right)^2}} + L}, \min_{i=\overline{1,n-1}}\frac{1}{\sqrt{a_{z_i}^*} + L}\right\},$$

where

$$a_{z_i}^* = \left(10\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right) + 160\lambda_{i+1}^2\frac{1+\sqrt{1-\alpha}}{\left(1-\sqrt{1-\alpha}\right)^2}\right)\frac{1-\alpha}{\left(1-\sqrt{1-\alpha}\right)^2}$$
$$+ 160\lambda_{i+1}^2\frac{1+\sqrt{1-\alpha}}{\left(1-\sqrt{1-\alpha}\right)^2}.$$

The minimum possible ω_{h_i} is $\omega_{h_i}^*$, which equals to

$$\omega_{h_i}^* = \frac{5\gamma^*}{1-\sqrt{1-\alpha}}.$$

Similarly,

$$\omega_{z_i}^* = \frac{\gamma^*}{1-\sqrt{1-\alpha}}\left(5\lambda_{i+1}^2\left(d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right) + 4\right) + 16\lambda_{i+1}^2\frac{1+\sqrt{1-\alpha}}{\left(1-\sqrt{1-\alpha}\right)^2}\right).$$

□

D. Proofs for the stochastic case

Updates for Algorithm 2.
Step vectors for $\{x_i\}_{i=2}^n$.

$$\widetilde{g}_{x_i}^{t+1} \leftarrow \begin{cases} 2\lambda_i J_{F_i,x_i}^T(x_i^{t+1}, z_i^{t+1}) \left(F_i(x_i^{t+1}, z_i^{t+1}) - \mathcal{Z}_{i-1}^{t+1}\right), \text{ with probability } p_{x_i}. \\ \widetilde{g}_{x_i}^t + 2\lambda_i \frac{s}{\tau_t} \sum\limits_{j \in I^t} \left[J_{F_{ij},x_i}^T(x_i^{t+1}, z_i^{t+1}) \left(F_{ij}(x_i^{t+1}, z_i^{t+1}) - \mathcal{Z}_{(i-1)j}^{t+1}\right) - J_{F_{ij},x_i}^T(x_i^t, z_i^t) \left(F_{ij}(x_i^t, z_i^t) - \mathcal{Z}_{(i-1)j}^t\right)\right], \text{ with probability } 1 - p_{x_i}. \end{cases} \tag{24}$$

Approximation sequences for $\left\{F_i(x_i^{t+1}, z_i^{t+1})\right\}_{i=2}^n$.

$$u_i^{t+1} \leftarrow \begin{cases} F_i(x_i^{t+1}, z_i^{t+1}), \text{ with probability } p_{u_i} \\ u_i^t + \frac{s}{\tau_t} \sum\limits_{j \in I^t} \left[F_{ij}(x_i^{t+1}, z_i^{t+1}) - F_{ij}(x_i^t, z_i^t)\right], \text{ with probability } 1 - p_{u_i} \end{cases} \tag{25}$$

Step vectors for $\{z_{i-1}\}_{i=2}^n$.

$$\widetilde{g}_{z_{i-1}}^{t+1} \leftarrow \begin{cases} \mathcal{H}_i^{t+1} + 2\lambda_{i-1} J_{F_{i-1},z_{i-1}}^T(x_{i-1}^{t+1}, z_{i-1}^{t+1}) \left(F_{i-1}(x_{i-1}^{t+1}, z_{i-1}^{t+1}) - \mathcal{Z}_{i-2}^{t+1}\right), \text{ with probability } p_{z_{i-1}} \\ \widetilde{g}_{z_{i-1}}^t + \mathcal{H}_i^{t+1} - \mathcal{H}_i^t + 2\lambda_{i-1} \frac{s}{\tau_t} \sum\limits_{j \in I^t} \left[J_{F_{(i-1)j},z_{i-1}}^T(x_{i-1}^{t+1}, z_{i-1}^{t+1}) \left(F_{(i-1)j}(x_{i-1}^{t+1}, z_{i-1}^{t+1}) - \mathcal{Z}_{(i-2)j}^{t+1}\right) - J_{F_{(i-1)j},z_{i-1}}^T(x_{i-1}^t, z_{i-1}^t) \left(F_{(i-1)j}(x_{i-1}^t, z_{i-1}^t) - \mathcal{Z}_{(i-2)j}^t\right)\right], \text{ with probability } 1 - p_{z_{i-1}} \end{cases} \tag{26}$$

Step vector for x_1.

$$\widetilde{g}_{x_1}^{t+1} \leftarrow \begin{cases} \nabla_{x_1} F_1(x_1^{t+1}, z_1^{t+1}), \text{ with probability } p_{x_1} \\ \widetilde{g}_{x_1}^t + \frac{s}{\tau_t} \sum\limits_{j \in I^t} \left[\nabla_{x_1} F_{1j}(x_1^{t+1}, z_1^{t+1}) - \nabla_{x_1} F_{1j}(x_1^t, z_1^t)\right], \text{ with probability } 1 - p_{u_i} \end{cases} \tag{27}$$

Step vector for z_1.

$$\widetilde{g}_{z_1}^{t+1} \leftarrow \begin{cases} \mathcal{H}_2^{t+1} + \nabla_{z_1} F_1(x_1^{t+1}, z_1^{t+1}), \text{ with probability } p_{x_1} \\ \widetilde{g}_{x_1}^t + \mathcal{H}_2^{t+1} - \mathcal{H}_2^t + \frac{s}{\tau_t} \sum\limits_{j \in I^t} \left[\nabla_{z_1} F_{1j}(x_1^{t+1}, z_1^{t+1}) - \nabla_{z_1} F_{1j}(x_1^t, z_1^t)\right], \text{ with probability } 1 - p_{u_i} \end{cases} \tag{28}$$

Lemma D.1. *For $i = 2, \ldots, n$ the following inequality is fulfilled:*

$$\begin{aligned} \mathbb{E}\|\widetilde{g}_{x_i}^{t+1} - g_{x_i}^{t+1}\|^2 &\leqslant (1 - p_{x_i}) \mathbb{E} \left\|\widetilde{g}_{x_i}^t - g_{x_i}^t\right\|^2 \\ &+ 24\lambda_i^2 (1 - p_{x_i}) \left(\frac{s}{\tau_t} + 1\right) (2D_{ImF_i})^2 \sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 \mathbb{E} \left\|x_i^{t+1} - x_i^t\right\|^2 \\ &+ 24\lambda_i^2 (1 - p_{x_i}) \left(\frac{s}{\tau_t} + 1\right) (2D_{ImF_i})^2 \sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 \left\|z_i^{t+1} - z_i^t\right\|^2 \\ &+ 24\lambda_i^2 (1 - p_{x_i}) c^4 (s + d_{z_{i-1}}) d_{z_{i-1}} \mathbb{E} \left\|x_i^{t+1} - x_i^t\right\|^2 \\ &+ 24\lambda_i^2 (1 - p_{x_i}) c^4 (s + d_{z_{i-1}}) d_{z_{i-1}} \mathbb{E} \left\|z_i^{t+1} - z_i^t\right\|^2 \\ &+ 72\lambda_i^2 (1 - p_{x_i}) (d_{x_i} c^2) (s + d_{z_{i-1}}) \Big(((1 - \alpha)(1 + s_1) + 1) \mathbb{E} \left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\ &+ ((1 - \alpha)(1 + s_1^{-1}) + 1) \mathbb{E} \left\|z_{i-1}^{t+1} - z_{i-1}^t\right\|^2 \Big). \end{aligned}$$

Proof.

$$\mathbb{E}_{I^t} \mathbb{E}_{p_{x_i}} \|\widetilde{g}_{x_i}^{t+1} - g_{x_i}^{t+1}\|^2$$

$$
\begin{aligned}
&= (1-p_{x_i})\mathbb{E}_{I^t}\left\|g_{x_i}^{t+1}-\widetilde{g}_{x_i}^{t}-2\lambda_i\frac{s}{\tau_t}\sum_{j\in I^t}\left[J_{F_{ij}x_i}^T(x_i^{t+1},z_i^{t+1})\left(F_{ij}(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{(i-1)j}^{t+1}\right)-J_{F_{ij}x_i}^T(x_i^t,z_i^t)\left(F_{ij}(x_i^t,z_i^t)-\mathcal{Z}_{(i-1)j}^{t}\right)\right]\right\|^2\\
&\overset{(*)}{=} (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2\\
&+ (1-p_{x_i})\mathbb{E}_{I^t}\left\|g_{x_i}^{t+1}-g_{x_i}^{t}-2\lambda_i\frac{s}{\tau_t}\sum_{j\in I^t}\left[J_{F_{ij}x_i}^T(x_i^{t+1},z_i^{t+1})\left(F_{ij}(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{(i-1)j}^{t+1}\right)-J_{F_{ij}x_i}^T(x_i^t,z_i^t)\left(F_{ij}(x_i^t,z_i^t)-\mathcal{Z}_{(i-1)j}^{t}\right)\right]\right\|^2\\
&\overset{(Jensen)}{\leqslant} (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2\\
&+ 8\lambda_i^2(1-p_{x_i})\mathbb{E}_{I^t}\left\|\frac{s}{\tau_t}\sum_{j\in I^t}\left[J_{F_{ij}x_i}^T(x_i^{t+1},z_i^{t+1})\left(F_{ij}(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{(i-1)j}^{t+1}\right)-J_{F_{ij}x_i}^T(x_i^t,z_i^t)\left(F_{ij}(x_i^t,z_i^t)-\mathcal{Z}_{(i-1)j}^{t}\right)\right]\right\|^2\\
&+ 8\lambda_i^2(1-p_{x_i})\left\|J_{F_i,x_i}^T(x_i^{t+1},z_i^{t+1})\left(F_i(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{i-1}^{t+1}\right)-J_{F_i,x_i}^T(x_i^t,z_i^t)\left(F_i(x_i^t,z_i^t)-\mathcal{Z}_{i-1}^{t}\right)\right\|^2\\
&\overset{(**)}{=} (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t+1},z_i^{t+1})-J_{F_{ij}x_i}^T(x_i^t,z_i^t)\right\|_2^2\left\|F_{ij}(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{(i-1)j}^{t+1}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t},z_i^{t})\right\|_2^2\left\|F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t},z_i^{t})\right\|_2^2\left\|\mathcal{Z}_{(i-1)j}^{t+1}-\mathcal{Z}_{(i-1)j}^{t}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\left\|J_{F_i,x_i}^T(x_i^{t+1},z_i^{t+1})-J_{F_i,x_i}^T(x_i^t,z_i^t)\right\|_2^2\left\|F_i(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{i-1}^{t+1}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\left\|J_{F_i,x_i}^T(x_i^t,z_i^t)\right\|_2^2\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\left\|J_{F_ix_i}^T(x_i^t,z_i^t)\right\|_2^2\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^{t}\right\|^2,
\end{aligned}
$$

where the equality * follows from

$$
\mathbb{E}_{I^t}\left[2\lambda_i\frac{s}{\tau_t}\sum_{j\in I^t}\left[J_{F_{ij},x_i}^T(x_i^{t+1},z_i^{t+1})\left(F_{ij}(x_i^{t+1},z_i^{t+1})-\mathcal{Z}_{(i-1)j}^{t+1}\right)-J_{F_{ij},x_i}^T(x_i^t,z_i^t)\left(F_{ij}(x_i^t,z_i^t)-\mathcal{Z}_{(i-1)j}^{t}\right)\right]\right]=(g_{x_i}^{t+1}-g_{x_i}^t),
$$

and the equality ** is true from the non-intersection of non-zero coordinates of batch elements.
It can be noted from Assumption 3.5 that $F_{ij}(x_i^t,z_i^t)-\mathcal{Z}_{(i-1)j}^t$ for any batch I^t and for any t is either bounded or is multiplied by the term that is equal to zero (as in linear function $J_{F_{ij},x_i}^T(x_i^{t+1},z_i^{t+1})-J_{F_{ij},x_i}^T(x_i^t,z_i^t)=0$). Thus, we introduce the variable D_Q as a diameter of the set Q, if Q is bounded and $D_Q=0$ otherwise, which allows us to make the following inequality.

$$
\begin{aligned}
&\mathbb{E}_{I^t}\mathbb{E}_{p_{x_i}}\|\widetilde{g}_{x_i}^{t+1}-g_{x_i}^{t+1}\|^2\\
&\overset{\text{(Ass. 3.5)}}{\leqslant} (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t+1},z_i^{t+1})-J_{F_{ij}x_i}^T(x_i^t,z_i^t)\right\|_2^2(2D_{\mathrm{Im}F_i})^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t},z_i^{t})\right\|_2^2\left\|F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J_{F_{ij}x_i}^T(x_i^{t},z_i^{t})\right\|_2^2\left\|\mathcal{Z}_{(i-1)j}^{t+1}-\mathcal{Z}_{(i-1)j}^{t}\right\|^2\\
&+ 24\lambda_i^2(1-p_{x_i})\left\|J_{F_i,x_i}^T(x_i^{t+1},z_i^{t+1})-J_{F_i,x_i}^T(x_i^t,z_i^t)\right\|_2^2(2D_{\mathrm{Im}F_i})^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad 24\lambda_i^2(1-p_{x_i})\left\|J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_2^2\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left\|J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_2^2\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^t\right\|^2\\
&\overset{\text{(Ass. 3.2)}}{\leqslant} \quad (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^t-g_{x_i}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|J^T_{F_{ij},x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_{ij},x_i}(x_i^t,z_i^t)\right\|_F^2(2D_{\mathrm{Im}F_i})^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}(\tau_t d_{x_i}c^2)\left\|F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}(\tau_t d_{x_i}c^2)\left\|\mathcal{Z}_{(i-1)j}^{t+1}-\mathcal{Z}_{(i-1)j}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2(2D_{\mathrm{Im}F_i})^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{z_{i-1}}d_{x_i}c^2)\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{z_{i-1}}d_{x_i}c^2)\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^t\right\|^2\\
&= \quad (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^t-g_{x_i}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\frac{s}{\tau_t}\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2(2D_{\mathrm{Im}F_i})^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})s(d_{x_i}c^2)\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})s(d_{x_i}c^2)\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2(2D_{\mathrm{Im}F_i})^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{z_{i-1}}d_{x_i}c^2)\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{z_{i-1}}d_{x_i}c^2)\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^t\right\|^2\\
&= \quad (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^t-g_{x_i}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left\|\mathcal{Z}_{i-1}^{t+1}-\mathcal{Z}_{i-1}^t\right\|^2\\
&\overset{(gen.Jensen)}{\leqslant} \quad (1-p_{x_i})\mathbb{E}_{I^t}\left\|\widetilde{g}_{x_i}^t-g_{x_i}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 72\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left(\left\|\mathcal{Z}_{i-1}^{t+1}-z_{i-1}^{t+1}\right\|^2+\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2+\left\|z_{i-1}^{t+1}-z_{i-1}^t\right\|^2\right).
\end{aligned}
$$

After taking full mathematical expectancy and using Lemma B.6 we prove that the following inequality holds.

$$
\begin{aligned}
\mathbb{E}\|\widetilde{g}_{x_i}^{t+1}-g_{x_i}^{t+1}\|^2 \quad &\overset{(B.6)}{\leqslant} \quad (1-p_{x_i})\mathbb{E}\left\|\widetilde{g}_{x_i}^t-g_{x_i}^t\right\|^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\mathbb{E}\left\|J^T_{F_i,x_i}(x_i^{t+1},z_i^{t+1})-J^T_{F_i,x_i}(x_i^t,z_i^t)\right\|_F^2\\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\mathbb{E}\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 72\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\right.
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad ((1-\alpha)(1+s_1^{-1})+1)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^{t}\right\|^2\Big) \\
&\overset{\text{(Ass. 3.3)}}{\leqslant} \quad (1-p_{x_i})\mathbb{E}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\left\|x_i^{t+1}-x_i^t\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\left\|z_i^{t+1}-z_i^t\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\mathbb{E}\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2 \\
&+ \quad 72\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\Big(((1-\alpha)(1+s_1)+1)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t}-z_{i-1}^{t}\right\|^2 \\
&+ \quad ((1-\alpha)(1+s_1^{-1})+1)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^{t}\right\|^2\Big) \\
&\overset{\text{(Ass. 3.2)}}{\leqslant} \quad (1-p_{x_i})\mathbb{E}\left\|\widetilde{g}_{x_i}^{t}-g_{x_i}^{t}\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\left\|z_i^{t+1}-z_i^t\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2 \\
&+ \quad 24\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1}-z_i^t\right\|^2 \\
&+ \quad 72\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\Big(((1-\alpha)(1+s_1)+1)\mathbb{E}\left\|\mathcal{Z}_{i-1}^{t}-z_{i-1}^{t}\right\|^2 \\
&+ \quad ((1-\alpha)(1+s_1^{-1})+1)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^{t}\right\|^2\Big).
\end{aligned}
$$

□

Lemma D.2.

$$
\begin{aligned}
\mathbb{E}\|\widetilde{g}_{x_1}^{t+1}-g_{x_1}^{t+1}\|^2 &\leqslant (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^{t}-g_{x_1}^{t}\right\|^2 \\
&+ 2(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2+2(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\mathbb{E}\|\widetilde{g}_{x_1}^{t+1}-g_{x_1}^{t+1}\|^2 &= (1-p_{x_1})\left\|g_{x_1}^{t+1}-\widetilde{g}_{x_1}^{t}-\frac{s}{\tau_t}\sum_{j\in I^t}\left[\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right]\right\|^2 \\
&\overset{(*)}{=} (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^{t}-g_{x_1}^{t}\right\|^2 \\
&+ (1-p_{x_1})\mathbb{E}\left\|g_{x_1}^{t+1}-g_{x_1}^{t}-\frac{s}{\tau_t}\sum_{j\in I^t}\left[\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right]\right\|^2 \\
&\overset{(Jensen)}{\leqslant} (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^{t}-g_{x_1}^{t}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad 2(1-p_{x_1})\mathbb{E}\left\|\nabla_{x_1}F_1(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_1(x_1^t,z_1^t)\right\|^2\\
&+ \quad 2(1-p_{x_1})\frac{s^2}{\tau_t^2}\mathbb{E}\left\|\sum_{j\in I^t}\left[\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right]\right\|^2\\
&\overset{(gen.Jensen)}{\leqslant} \quad (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^t-g_{x_1}^t\right\|^2\\
&+ \quad 2(1-p_{x_1})\mathbb{E}\left\|\nabla_{x_1}F_1(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_1(x_1^t,z_1^t)\right\|^2\\
&+ \quad 2(1-p_{x_1})\frac{s^2}{\tau_t}\mathbb{E}\sum_{j\in I^t}\left\|\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right\|^2\\
&= \quad (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^t-g_{x_1}^t\right\|^2\\
&+ \quad 2(1-p_{x_1})\mathbb{E}\left\|\nabla_{x_1}F_1(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_1(x_1^t,z_1^t)\right\|^2\\
&+ \quad 2(1-p_{x_1})s\sum_{j=1}^s\mathbb{E}\left\|\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right\|^2\\
&\overset{\text{(Ass. 3.3)}}{\leqslant} \quad (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^t-g_{x_1}^t\right\|^2\\
&+ \quad 2(1-p_{x_1})L_1\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2+2(1-p_{x_1})L_1\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2\\
&+ \quad 2(1-p_{x_1})s\sum_{j=1}^s\mathbb{E}\left\|\nabla_{x_1}F_{1j}(x_1^{t+1},z_1^{t+1})-\nabla_{x_1}F_{1j}(x_1^t,z_1^t)\right\|^2\\
&\overset{\text{(Ass. 3.3)}}{\leqslant} \quad (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^t-g_{x_1}^t\right\|^2\\
&+ \quad 2(1-p_{x_1})L_1\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2+2(1-p_{x_1})L_1\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2\\
&+ \quad 2(1-p_{x_1})s\sum_{j=1}^s L_{1j}\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2+2(1-p_{x_1})s\sum_{j=1}^s L_{1j}\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2\\
&= \quad (1-p_{x_1})\mathbb{E}\left\|\widetilde{g}_{x_1}^t-g_{x_1}^t\right\|^2\\
&+ \quad 2(1-p_{x_1})\left(L_1+s\sum_{j=1}^s L_{1j}\right)\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2\\
&+ \quad 2(1-p_{x_1})\left(L_1+s\sum_{j=1}^s L_{1j}\right)\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2.
\end{aligned}
$$

□

Lemma D.3. *For $i=2,\ldots,n-1$ the following inequality is fulfilled:*

$$
\begin{aligned}
&\mathbb{E}\|\widetilde{g}_{z_i}^{t+1}-g_{z_i}^{t+1}\|^2\\
\leqslant\ & (1-p_{z_i})\mathbb{E}\left\|\widetilde{g}_{z_i}^t-g_{x_i}^t\right\|^2\\
+\ & 24\lambda_i^2(1-p_{z_i})\left(\frac{s}{\tau_t}+1\right)(2D_{ImF_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2\\
+\ & 24\lambda_i^2(1-p_{z_i})\left(\frac{s}{\tau_t}+1\right)(2D_{ImF_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2\left\|z_i^{t+1}-z_i^t\right\|^2\\
+\ & 24\lambda_i^2(1-p_{z_i})c^4(s+d_{z_{i-1}})d_{z_{i-1}}\frac{d_{z_i}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad 24\lambda_i^2(1-p_{z_i})c^4(s+d_{z_{i-1}})d_{z_{i-1}}\frac{d_{z_i}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1}-z_i^t\right\|^2\\
&+ \quad 72\lambda_i^2(1-p_{z_i})(d_{z_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\right.\\
&+ \quad \left.((1-\alpha)(1+s_1^{-1})+1)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^t\right\|^2\right).
\end{aligned}
$$

Proof. This proof repeats the proof for Lemma D.1 with the exception being the change of Jacobean J_{F_i} of x_i coordinates to z_i coordinates. □

Lemma D.4.

$$
\begin{aligned}
\mathbb{E}\|\widetilde{g}_{z_1}^{t+1}-g_{z_1}^{t+1}\|^2 &\leqslant (1-p_{z_1})\mathbb{E}\left\|\widetilde{g}_{z_1}^t-g_{z_1}^t\right\|^2\\
&+ 2(1-p_{z_1})\left(L_1+s\sum_{j=1}^s L_{1j}\right)\mathbb{E}\left\|x_1^{t+1}-x_1^t\right\|^2+2(1-p_{x_1})\left(L_1+s\sum_{j=1}^s L_{1j}\right)\mathbb{E}\left\|z_1^{t+1}-z_1^t\right\|^2.
\end{aligned}
$$

Proof. This proof repeats the proof for Lemma D.2 with the exception being the change of operator ∇_{x_1} to ∇_{z_1}. □

Lemma D.5.

$$
\begin{aligned}
\mathbb{E}\left\|u_i^{t+1}-F_i^{t+1}\right\|^2 &\leqslant (1-p_{u_i})\mathbb{E}\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
&+ 2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2\\
&+ 2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1}-z_i^t\right\|^2.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{p_{u_i}}\mathbb{E}_{I^t}\left\|u_i^{t+1}-F_i(x_i^{t+1},z_i^{t+1}))\right\|^2\\
=\quad &(1-p_{u_i})\mathbb{E}_{I^t}\left\|u_i^t+\frac{s}{\tau_t}\sum_{j\in I^t}\left[F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right]-F_i(x_i^{t+1},z_i^{t+1}))\right\|^2\\
=\quad &(1-p_{u_i})\mathbb{E}_{I^t}\left\|u_i^t-F_i(x_i^t,z_i^t)+\frac{s}{\tau_t}\sum_{j\in I^t}\left[F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right]-F_i(x_i^{t+1},z_i^{t+1})+F_i(x_i^t,z_i^t)\right\|^2\\
\overset{(*)}{=}\quad &(1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
+\quad &(1-p_{u_i})\mathbb{E}_{I^t}\left\|\frac{s}{\tau_t}\sum_{j\in I^t}\left[F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right]-F_i(x_i^{t+1},z_i^{t+1})+F_i(x_i^t,z_i^t)\right\|^2\\
\overset{(Jensen)}{\leqslant}\quad &(1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
+\quad &2(1-p_{u_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\left\|\sum_{j\in I^t}F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right\|^2+2(1-p_{u_i})\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
\overset{(**)}{=}\quad &(1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
+\quad &2(1-p_{u_i})\frac{s^2}{\tau_t^2}\mathbb{E}_{I^t}\sum_{j\in I^t}\left\|F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right\|^2+2(1-p_{u_i})\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
=\quad &(1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2+2(1-p_{u_i})\frac{s}{\tau_t}\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad 2(1-p_{u_i})\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&= \quad (1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2+2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\left\|F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t)\right\|^2\\
&\overset{\text{(Ass. 3.2)}}{\leqslant} \quad (1-p_{u_i})\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
&+ \quad 2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left\|x_i^{t+1}-x_i^t\right\|^2+2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left\|z_i^{t+1}-z_i^t\right\|^2.
\end{aligned}
$$

where equality * holds, because

$$
\mathbb{E}_{I^t}\frac{s}{\tau_t}\sum_{j\in I^t}\left[F_{ij}(x_i^{t+1},z_i^{t+1})-F_{ij}(x_i^t,z_i^t)\right]=F_i(x_i^{t+1},z_i^{t+1})-F_i(x_i^t,z_i^t),
$$

and the equality ** is true from the non-intersection of non-zero coordinates of F_{ij}. □

Lemma D.6.

$$
\begin{aligned}
&\mathbb{E}\left\|\mathcal{H}_i^{t+1}-2\lambda_i(z_{i-1}^{t+1}-F_i(x_i^{t+1},z_i^{t+1}))\right\|^2\\
\leqslant\ &(1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t-2\lambda_i(z_{i-1}^t-F_i(x_i^t,z_i^t))\right\|^2\\
+\ &8\lambda_i^2\left(2\left(1+s_2^{-1}\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\\
+\ &8\lambda_i^2\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\mathbb{E}\left\|u_i^t-F_i(x_i^t,z_i^t)\right\|^2\\
+\ &16\lambda_i^2\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2\\
+\ &16\lambda_i^2\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1}-z_i^t\right\|^2\\
+\ &16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^t\right\|^2\\
+\ &16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1}-x_i^t\right\|^2+16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1}-z_i^t\right\|^2.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\left\|\mathcal{H}_i^{t+1}-2\lambda_i(z_{i-1}^{t+1}-F_i(x_i^{t+1},z_i^{t+1}))\right\|^2\\
= \quad &\mathbb{E}\left\|\mathcal{H}_i^t+\mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^t-u_i^{t+1})-\mathcal{H}_i^t\right)-2\lambda_i(z_{i-1}^{t+1}-F_i(x_i^{t+1},z_i^{t+1}))\right\|^2\\
\overset{(Jensen)}{\leqslant} \quad &(1+s_2)\mathbb{E}\left\|\mathcal{H}_i^t+\mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^t-u_i^{t+1})-\mathcal{H}_i^t\right)-2\lambda_i(\mathcal{Z}_{i-1}^t-u_i^{t+1})\right\|^2\\
+\ &2\left(1+s_2^{-1}\right)\mathbb{E}\left\|2\lambda_i\mathcal{Z}_{i-1}^t-2\lambda_i z_{i-1}^{t+1}\right\|^2+8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^{t+1},z_i^{t+1})\right\|^2\\
\overset{\text{(Ass. 3.1)}}{\leqslant} \quad &(1+s_2)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t-2\lambda_i(\mathcal{Z}_{i-1}^t-u_i^{t+1})\right\|^2+8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^{t+1}\right\|^2\\
+\ &8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^{t+1},z_i^{t+1})\right\|^2\\
\overset{(Jensen)}{\leqslant} \quad &(1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t-2\lambda_i(z_{i-1}^t-F_i(x_i^t,z_i^t))\right\|^2+8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^{t+1}\right\|^2\\
+\ &8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^{t+1},z_i^{t+1})\right\|^2+8\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\\
+\ &8\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^t,z_i^t)\right\|^2\\
\overset{(Jensen)}{\leqslant} \quad &(1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t-2\lambda_i(z_{i-1}^t-F_i(x_i^t,z_i^t))\right\|^2+16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\\
+\ &8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^{t+1},z_i^{t+1})\right\|^2+8\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t-z_{i-1}^t\right\|^2\\
+\ &8\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|u_i^{t+1}-F_i(x_i^t,z_i^t)\right\|^2+16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1}-z_{i-1}^t\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(Jensen)}{\leqslant} && (1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t - 2\lambda_i(z_{i-1}^t - F_i(x_i^t, z_i^t))\right\|^2 + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&+ && 8\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|u_i^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})\right\|^2 + 8\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|u_i^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})\right\|^2 + 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|F_i(x_i^{t+1}, z_i^{t+1}) - F_i(x_i^t, z_i^t)\right\|^2 \\
&= && (1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t - 2\lambda_i(z_{i-1}^t - F_i(x_i^t, z_i^t))\right\|^2 \\
&+ && 8\lambda_i^2\left(2\left(1+s_2^{-1}\right) + (1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&+ && 8\lambda_i^2\left(\left(1+s_2^{-1}\right) + 2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|u_i^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})\right\|^2 \\
&+ && 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\mathbb{E}\left\|F_i(x_i^{t+1}, z_i^{t+1}) - F_i(x_i^t, z_i^t)\right\|^2 \\
&\overset{(\text{Ass. } 3.2)}{\leqslant} && (1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t - 2\lambda_i(z_{i-1}^t - F_i(x_i^t, z_i^t))\right\|^2 \\
&+ && 8\lambda_i^2\left(2\left(1+s_2^{-1}\right) + (1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&+ && 8\lambda_i^2\left(\left(1+s_2^{-1}\right) + 2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|u_i^{t+1} - F_i(x_i^{t+1}, z_i^{t+1})\right\|^2 \\
&+ && 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1} - x_i^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2 \\
&\overset{(D.5)}{\leqslant} && (1+s_2)(1+s_3)(1-\alpha)\mathbb{E}\left\|\mathcal{H}_i^t - 2\lambda_i(z_{i-1}^t - F_i(x_i^t, z_i^t))\right\|^2 \\
&+ && 8\lambda_i^2\left(2\left(1+s_2^{-1}\right) + (1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\mathbb{E}\left\|\mathcal{Z}_{i-1}^t - z_{i-1}^t\right\|^2 \\
&+ && 8\lambda_i^2\left(\left(1+s_2^{-1}\right) + 2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\mathbb{E}\left\|u_i^t - F_i(x_i^t, z_i^t)\right\|^2 \\
&+ && 16\lambda_i^2\left(\left(1+s_2^{-1}\right) + 2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1} - x_i^t\right\|^2 \\
&+ && 16\lambda_i^2\left(\left(1+s_2^{-1}\right) + 2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2 \\
&+ && 16\lambda_i^2\left(1+s_2^{-1}\right)\mathbb{E}\left\|z_{i-1}^{t+1} - z_{i-1}^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|x_i^{t+1} - x_i^t\right\|^2 \\
&+ && 16\lambda_i^2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\mathbb{E}\left\|z_i^{t+1} - z_i^t\right\|^2.
\end{aligned}
$$

□

Let us define additional notations, used in this section.

$$
\text{(error of batch stochastics of } g_{x_i}) \qquad G_{x_i}^t := \mathbb{E}\left\|\widetilde{g}_{x_i}^t - g_{x_i}^t\right\|^2, \tag{29}
$$

$$
\text{(error of batch stochastics of } g_{z_i}) \qquad G_{z_i}^t := \mathbb{E}\left\|\widetilde{g}_{z_i}^t - g_{z_i}^t\right\|^2. \tag{30}
$$

$$
\text{(error of batch stochastics of } F_i) \qquad U_i^t := \mathbb{E}\left\|u_i^t - F_i(x_i^t, z_i^t)\right\|^2. \tag{31}
$$

It can be noted, that Lemmas B.2, B.3, B.4, B.5, B.6 remain the same after the introduction of stochastic.

D.1. Proof for Theorem 4.2

Proof. As Lemmas B.3 and B.2 still work for $\widetilde{g}^t$ we have similar to (11)

$$\Phi(y^{t+1}) \overset{(B.2,B.3)}{\leqslant} \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 + \frac{5\gamma}{4}\left\|\widetilde{g}^t - \nabla f(y^t)\right\|^2.$$

Applying Jensen inequality to the last term, we obtain

$$\Phi(y^{t+1}) \overset{(Jensen)}{\leqslant} \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 + \frac{5\gamma}{2}\left\|g^t - \nabla f(y^t)\right\|^2 + \frac{5\gamma}{2}\left\|\widetilde{g}^t - g^t\right\|^2.$$

Using Lemma B.4 and Lemma B.5 we proceed with another inequality.

$$\begin{aligned}\Phi(y^{t+1}) \overset{(B.4,B.5)}{\leqslant}\ & \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 \\ &+ 10\gamma\sum_{i=1}^{n-1}\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right)\left\|\mathcal{Z}_i^t - z_i^t\right\|^2 \\ &+ 5\gamma\sum_{i=1}^{n-1}\left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t))\right\|^2 + \frac{5\gamma}{2}\left\|\widetilde{g}^t - g^t\right\|^2.\end{aligned}$$

Using the definition of the Euclidean norm, we have:

$$\begin{aligned}\Phi(y^{t+1}) \leqslant\ & \Phi(y^t) - \frac{\gamma}{8}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|y^{t+1} - y^t\right\|^2 \\ &+ 10\gamma\sum_{i=1}^{n-1}\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right)\left\|\mathcal{Z}_i^t - z_i^t\right\|^2 \\ &+ 5\gamma\sum_{i=1}^{n-1}\left\|\mathcal{H}_{i+1}^t - 2\lambda_{i+1}(z_i^t - F_{i+1}(x_{i+1}^t, z_{i+1}^t))\right\|^2 + \frac{5\gamma}{2}\sum_{i=1}^{n}\left\|\widetilde{g}_{x_i}^t - g_{x_i}^t\right\|^2 + \frac{5\gamma}{2}\sum_{i=1}^{n-1}\left\|\widetilde{g}_{z_i}^t - g_{z_i}^t\right\|^2.\end{aligned}$$

Taking expectancy and utilizing notations (7)-(10) and (29)-(30).

$$\begin{aligned}\mathbb{E}\Phi(y^{t+1}) \leqslant\ & \mathbb{E}\Phi(y^t) - \frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left(\sum_{i=1}^{n}\Omega_{x_i}^t + \sum_{i=1}^{n-1}\Omega_{z_i}^t\right) \\ &+ 10\gamma\sum_{i=1}^{n-1}\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}} + 2d_{z_{i+1}}\right)D_{\mathcal{Z}_i}^t + 5\gamma\sum_{i=1}^{n-1}D_{\mathcal{H}_i}^t \\ &+ \frac{5\gamma}{2}\sum_{i=1}^{n}G_{x_i}^t + \frac{5\gamma}{2}\sum_{i=1}^{n-1}G_{z_i}^t. \qquad (32)\end{aligned}$$

□

From Lemma D.1 we have for $i = \overline{2,n}$:

$$\begin{aligned}G_{x_i}^{t+1} \leqslant\ & (1-p_{x_i})G_{x_i}^t \\ &+ 24\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{x_i}^t \\ &+ 24\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{z_i}^t\end{aligned}$$

$$+ \quad 72\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)D^t_{\mathcal{Z}_{i-1}}+((1-\alpha)(1+s_1^{-1})+1)\Omega^t_{z_{i-1}}\right).$$

From Lemma D.2:

$$G^{t+1}_{x_1} \leqslant (1-p_{x_1})G^t_{x_1}+2(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega^t_{x_1}+2(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega^t_{z_1}.$$

Similarly, for variables $\{z_i\}_{i=2}^{n-1}$ from Lemma D.3:

$$\begin{aligned}
G^{t+1}_{z_i} \leqslant\ & (1-p_{z_i})G^t_{z_i}\\
&+ 24\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega^t_{x_i}\\
&+ 24\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega^t_{z_i}\\
&+ 72\lambda_i^2(1-p_{z_i})(d_{z_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)D^t_{\mathcal{Z}_{i-1}}+((1-\alpha)(1+s_1^{-1})+1)\Omega^t_{z_{i-1}}\right).
\end{aligned}$$

And we have Lemma D.4 for z_1:

$$G^{t+1}_{z_1} \leqslant (1-p_{z_1})G^t_{z_1}+2(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega^t_{x_1}+2(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega^t_{z_1}.$$

From Lemma D.5:

$$U_i^{t+1} \leqslant (1-p_{u_i})U_i^t+2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\Omega^t_{x_i}+2(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\Omega^t_{z_i}.$$

From Lemma D.6:

$$\begin{aligned}
D^{t+1}_{\mathcal{H}_i} \leqslant\ & (1+s_2)(1+s_3)(1-\alpha)D^t_{\mathcal{H}_i}\\
&+ 8\lambda_i^2\left(2\left(1+s_2^{-1}\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)D^t_{\mathcal{Z}_{i-1}}\\
&+ 8\lambda_i^2\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})U_i^t\\
&+ 16\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right.\\
&\quad\left.+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]\Omega^t_{x_i}\\
&+ 16\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right.\\
&\quad\left.+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]\Omega^t_{z_i}\\
&+ 16\lambda_i^2\left(1+s_2^{-1}\right)\Omega^t_{z_{i-1}}.
\end{aligned}$$

Lemma B.6 remains the same from non-stochastic case:

$$D^{t+1}_{\mathcal{Z}_i} \leqslant (1-\alpha)(1+s_1)D^t_{\mathcal{Z}_i}+(1-\alpha)(1+s_1^{-1})\Omega^t_{z_i}.$$

Multiplying each of the aforementioned inequalities by some positive values $\omega_{\widetilde{x}_i}, \omega_{\widetilde{x}_1}, \omega_{\widetilde{z}_i}, \omega_{z_1}, \omega_{u_i}, \omega_{h_i}, \omega_{z_i}$ respectfully, and adding them to (32) we have:

$$\mathbb{E}\Phi(y^{t+1})+\sum_{i=1}^{n}\omega_{\widetilde{x}_i}G^{t+1}_{x_i}+\sum_{i=1}^{n-1}\omega_{\widetilde{z}_i}G^{t+1}_{z_i}+\sum_{i=2}^{n}\omega_{u_i}U_i^{t+1}+\sum_{i=1}^{n-1}\omega_{z_i}D^{t+1}_{\mathcal{Z}_i}+\sum_{i=2}^{n}\omega_{h_i}D^{t+1}_{\mathcal{H}_i}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}\Phi(y^t) - \frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2 - \sum_{i=1}^{n}\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\Omega_{x_i}^t - \sum_{i=1}^{n-1}\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\Omega_{z_i}^t \\
&\quad + \sum_{i=2}^{n}\omega_{\widetilde{x}_i}(1-p_{x_i})G_{x_i}^t \\
&\quad + 24\sum_{i=2}^{n}\omega_{\widetilde{x}_i}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{x_i}^t \\
&\quad + 24\sum_{i=2}^{n}\omega_{\widetilde{x}_i}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{z_i}^t \\
&\quad + 72\sum_{i=2}^{n}\omega_{\widetilde{x}_i}\lambda_i^2(1-p_{x_i})(d_{x_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)D_{\mathcal{Z}_{i-1}}^t + ((1-\alpha)(1+s_1^{-1})+1)\Omega_{z_{i-1}}^t\right) \\
&\quad + \omega_{\widetilde{x}_1}(1-p_{x_1})G_{x_1}^t + 2\omega_{\widetilde{x}_1}(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega_{x_1}^t + 2\omega_{\widetilde{x}_1}(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega_{z_1}^t \\
&\quad + \sum_{i=2}^{n-1}\omega_{\widetilde{z}_i}(1-p_{z_i})G_{z_i}^t \\
&\quad + 24\sum_{i=2}^{n-1}\omega_{\widetilde{z}_i}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{x_i}^t \\
&\quad + 24\sum_{i=2}^{n-1}\omega_{\widetilde{z}_i}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\Omega_{z_i}^t \\
&\quad + 72\sum_{i=2}^{n-1}\omega_{\widetilde{z}_i}\lambda_i^2(1-p_{z_i})(d_{z_i}c^2)(s+d_{z_{i-1}})\left(((1-\alpha)(1+s_1)+1)D_{\mathcal{Z}_{i-1}}^t + ((1-\alpha)(1+s_1^{-1})+1)\Omega_{z_{i-1}}^t\right) \\
&\quad + \omega_{\widetilde{z}_1}(1-p_{z_1})G_{z_1}^t + 2\omega_{\widetilde{z}_1}(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega_{x_1}^t + 2\omega_{\widetilde{z}_1}(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\Omega_{z_1}^t \\
&\quad + \sum_{i=2}^{n}\omega_{u_i}(1-p_{u_i})U_i^t + 2\sum_{i=2}^{n}\omega_{u_i}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\Omega_{x_i}^t + 2\sum_{i=2}^{n}\omega_{u_i}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\Omega_{z_i}^t \\
&\quad + \sum_{i=2}^{n}\omega_{h_i}\left(1+s_2\right)\left(1+s_3\right)(1-\alpha)D_{\mathcal{H}_i}^t \\
&\quad + 8\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2\left(2\left(1+s_2^{-1}\right)+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)D_{\mathcal{Z}_{i-1}}^t \\
&\quad + 8\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})U_i^t \\
&\quad + 16\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right]\Omega_{x_i}^t \\
&\quad + 16\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right]\Omega_{z_i}^t \\
&\quad + 16\sum_{i=2}^{n}\omega_{h_i}\lambda_i^2\left(1+s_2^{-1}\right)\Omega_{z_{i-1}}^t
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)(1+s_1)D^t_{\mathcal{Z}_i}+\sum_{i=1}^{n-1}\omega_{z_i}(1-\alpha)(1+s_1^{-1})\Omega^t_{z_i}\\
&+10\gamma\sum_{i=1}^{n-1}\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)D^t_{\mathcal{Z}_i}+5\gamma\sum_{i=1}^{n-1}D^t_{\mathcal{H}_i}\\
&+\frac{5\gamma}{2}\sum_{i=1}^{n}G^t_{x_i}+\frac{5\gamma}{2}\sum_{i=1}^{n-1}G^t_{z_i}.
\end{aligned}
$$

The next step is gathering these terms together.

$$
\begin{aligned}
&\mathbb{E}\Phi(y^{t+1})+\sum_{i=1}^{n}\left(\omega_{\widetilde{x}_i}G^{t+1}_{x_i}-\psi_{\widetilde{x}_i}G^t_{x_i}\right)+\sum_{i=1}^{n-1}\left(\omega_{\widetilde{z}_i}G^{t+1}_{z_i}-\psi_{\widetilde{z}_i}G^t_{z_i}\right)+\sum_{i=2}^{n}\left(\omega_{u_i}U^{t+1}_i-\psi_{u_i}U^t_i\right)\\
&+\sum_{i=1}^{n-1}\left(\omega_{z_i}D^{t+1}_{\mathcal{Z}_i}-\psi_{z_i}D^t_{\mathcal{Z}_i}\right)+\sum_{i=2}^{n}\left(\omega_{h_i}D^{t+1}_{\mathcal{H}_i}-\psi_{h_i}D^t_{\mathcal{H}_i}\right)\\
&\leqslant\Phi(y^t)-\frac{\gamma}{8}\mathbb{E}\|\mathcal{G}_\gamma(y^t)\|^2-\sum_{i=1}^{n}\phi_{x_i}\Omega^t_{x_i}-\sum_{i=1}^{n-1}\phi_{z_i}\Omega^t_{z_i},
\end{aligned}
$$

where for $i=\overline{1,n}$:

$$
\psi_{\widetilde{x}_i}=\omega_{\widetilde{x}_i}(1-p_{x_i})+\frac{5\gamma}{2}.
$$

For $i=\overline{2,n}$:

$$
\begin{aligned}
\psi_{u_i}&=\omega_{u_i}(1-p_{u_i})+8\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i}),\\
\psi_{h_i}&=5\gamma+\omega_{h_i}\left(1+s_2\right)\left(1+s_3\right)(1-\alpha).
\end{aligned}
$$

For $i=\overline{1,n-1}$:

$$
\psi_{\widetilde{z}_i}=\omega_{\widetilde{z}_i}(1-p_{z_i})+\frac{5\gamma}{2},
$$

$$
\begin{aligned}
\psi_{z_i}&=10\gamma\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)\\
&\quad+8\omega_{h_{i+1}}\lambda_{i+1}^2\left(2\left(1+s_2^{-1}\right)+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)\\
&\quad+\omega_{z_i}(1-\alpha)(1+s_1)\\
&\quad+72\omega_{\widetilde{x}_{i+1}}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1)+1)\\
&\quad+72\omega_{\widetilde{z}_{i+1}}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1)+1).
\end{aligned}
$$

For $i=\overline{2,n}$:

$$
\begin{aligned}
\phi_{x_i}&=\frac{1}{2\gamma}-\frac{L}{2}\\
&\quad-24\omega_{\widetilde{x}_i}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&\quad-24\omega_{\widetilde{z}_i}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&- 2\omega_{u_i}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}} \\
&- 16\omega_{h_i}\lambda_i^2 c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right. \\
&\left.\quad+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]. \qquad (33)
\end{aligned}
$$

For ϕ_{x_1}:

$$
\begin{aligned}
\phi_{x_1} = \frac{1}{2\gamma} - \frac{L}{2} \\
&- 2\omega_{\widetilde{x}_1}(1-p_{x_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) \\
&- 2\omega_{\widetilde{z}_1}(1-p_{z_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right).
\end{aligned}
$$

For $i = \overline{2, n-1}$:

$$
\begin{aligned}
\phi_{z_i} &= \frac{1}{2\gamma} - \frac{L}{2} && (34)\\
&- 24\omega_{\widetilde{x}_i}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) && (35)\\
&- 72\omega_{\widetilde{x}_{i+1}}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1) && (36)\\
&- 24\omega_{\widetilde{z}_i}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) && (37)\\
&- 72\omega_{\widetilde{z}_{i+1}}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1) && (38)\\
&- 2\omega_{u_i}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}} && (39)\\
&- 16\omega_{h_i}\lambda_i^2 c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right. && (40)\\
&\left.\quad+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right] && (41)\\
&- 16\omega_{h_{i+1}}\lambda_{i+1}^2\left(1+s_2^{-1}\right) - \omega_{z_i}(1-\alpha)(1+s_1^{-1}). && (42)
\end{aligned}
$$

For ϕ_{z_1}:

$$
\begin{aligned}
\phi_{z_1} = \frac{1}{2\gamma} - \frac{L}{2} \\
&- 72\omega_{\widetilde{x}_2}\lambda_2^2(1-p_{x_2})(d_{x_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1) \\
&- 2\omega_{\widetilde{x}_1}(1-p_{x_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) \\
&- 72\omega_{\widetilde{z}_2}\lambda_2^2(1-p_{z_2})(d_{z_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1) \\
&- 2\omega_{\widetilde{z}_1}(1-p_{z_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) \\
&- 16\omega_{h_2}\lambda_2^2\left(1+s_2^{-1}\right) - \omega_{z_1}(1-\alpha)(1+s_1^{-1}).
\end{aligned}
$$

Let us overlook conditions on aforementioned coefficient. It can be noticed, that conditions are similar to Theorem 3.7. Denote $()$ as some sequence index. If $\psi_{()} \leqslant \omega_{()}$, then the corresponding sequence can be reduced, using telescoping sum. We firstly apply this inequality to $\omega_{h_i}, i = \overline{2,n}$:

$$\frac{5\gamma}{1-(1+s_2)\,(1+s_3)(1-\alpha)} \leqslant \omega_{h_i}. \tag{43}$$

Similarly, for $\omega_{\widetilde{x}_i}, i = \overline{1,n}$, $\omega_{\widetilde{z}_i}, i = \overline{1,n-1}$:

$$\frac{5\gamma}{2p_{x_i}} \leqslant \omega_{\widetilde{x}_i}, \tag{44}$$

$$\frac{5\gamma}{2p_{z_i}} \leqslant \omega_{\widetilde{z}_i}. \tag{45}$$

For $\omega_{u_i}, i = \overline{2,n}$:

$$\frac{8\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\,(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})}{p_{u_i}} \leqslant \omega_{u_i}. \tag{46}$$

And for $\omega_{z_i}, i = \overline{1,n-1}$:

$$\frac{\gamma b_{z_i}}{1-(1+s_1)(1-\alpha)} \leqslant \omega_{z_i},$$

where

$$\begin{aligned}
b_{z_i} \;:=\;& 10\lambda_{i+1}^2 d_{z_i} c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)+8\frac{\omega_{h_{i+1}}}{\gamma}\lambda_{i+1}^2\left(2\left(1+s_2^{-1}\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)\\
+\;& 72\frac{\omega_{\widetilde{x}_{i+1}}}{\gamma}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1)+1)\\
+\;& 72\frac{\omega_{\widetilde{z}_{i+1}}}{\gamma}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1)+1)
\end{aligned}$$

Inserting lower bound of (44) and (46) into (33), we obtain the concrete value for $\phi_{x_i}, i = \overline{2,n}$:

$$\begin{aligned}
\phi_{x_i} =\;& \frac{1}{2\gamma}-\frac{L}{2}\\
&-60\frac{\gamma}{p_{x_i}}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&-60\frac{\gamma}{p_{z_i}}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&-\frac{16\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\,(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})}{p_{u_i}}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\\
&-\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\,(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}.
\end{aligned}$$

It can be noticed that if for $i = \overline{2,n}$, $\phi_{x_i} \geqslant 0$, then Ω_{x_i} can be dropped by this inequality. Thus, we have:

$$\begin{aligned}
0 \leqslant\;& \frac{1}{2\gamma}-\frac{L}{2}\\
&-60\frac{\gamma}{p_{x_i}}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)
\end{aligned}$$

$$
\begin{aligned}
&- 60\frac{\gamma}{p_{z_i}}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&- \frac{16\omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})}{p_{u_i}}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\\
&- \omega_{h_i}\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}.
\end{aligned}
$$

Expressing ω_{h_i}:

$$
\omega_{h_i}\leqslant\frac{1-\gamma L-120\gamma^2\lambda_i^2\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum\limits_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\left(\frac{1-p_{x_i}}{p_{x_i}}+\frac{1-p_{z_i}}{p_{z_i}}\right)}{32\gamma\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]} \tag{47}
$$

Now, we have ω_{h_i} bounded from both above and below. In order for at least a one possible value to exist, the right bound should be greater than the left bound. There, the following inequality needs to hold:

$$
\frac{5\gamma}{1-(1+s_2)(1+s_3)(1-\alpha)}\leq\frac{1-\gamma L-120\gamma^2\lambda_i^2\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum\limits_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\left(\frac{1-p_{x_i}}{p_{x_i}}+\frac{1-p_{z_i}}{p_{z_i}}\right)}{32\gamma\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]}
$$

Equivalently:

$$
a_{h_i}\gamma^2+\gamma L\leqslant 1,
$$

where

$$
\begin{aligned}
a_{h_i} \quad := \quad & \frac{10\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\left(1+s_2^{-1}\right)+2(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+(1+s_2)\left(1+s_3^{-1}\right)(1-\alpha)\right]}{1-(1+s_2)(1+s_3)(1-\alpha)}\\
+ \quad & 120\lambda_i^2\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\left(\frac{1-p_{x_i}}{p_{x_i}}+\frac{1-p_{z_i}}{p_{z_i}}\right).
\end{aligned}
$$

Lemma A.6 gives a solution to this inequality:

$$
0\leqslant\gamma\leqslant\frac{1}{\sqrt{a_{h_i}}+L}.
$$

We choose $s_2=s_3=\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1$. Thus, we have the concrete value:

$$
\begin{aligned}
a_{h_i}^* \quad = \quad & \frac{10\lambda_i^2c^2\frac{d_{z_{i-1}}}{d_{x_i}}\left[\left(\frac{(1-\alpha)^{\frac{1}{3}}}{1-(1-\alpha)^{\frac{1}{3}}}+\frac{2(1-\alpha)^{\frac{2}{3}}}{1-(1-\alpha)^{\frac{1}{3}}}\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+\frac{(1-\alpha)^{\frac{2}{3}}}{(1-(1-\alpha)^{\frac{1}{3}})}\right]}{1-(1-\alpha)^{\frac{1}{3}}}\\
+ \quad & 120\lambda_i^2\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\left(\frac{1-p_{x_i}}{p_{x_i}}+\frac{1-p_{z_i}}{p_{z_i}}\right).
\end{aligned}
$$

Considering condition on ϕ_{x_1}, we have:

$$
0\leqslant\frac{1}{2\gamma}-\frac{L}{2}-2\omega_{\widetilde{x}_1}(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)-2\omega_{\widetilde{z}_1}(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)
$$

$$= \frac{1}{2\gamma} - \frac{L}{2} - 5\frac{\gamma}{p_{x_1}}(1-p_{x_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) - 5\frac{\gamma}{p_{z_1}}(1-p_{z_1})\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right).$$

Or

$$10\gamma^2\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right)\left(\frac{(1-p_{x_1})}{p_{x_1}} + \frac{(1-p_{z_1})}{p_{z_1}}\right) + \gamma L \leqslant 1$$

From Lemma A.6:

$$0 \leqslant \gamma \leqslant \frac{1}{\sqrt{\left(L_1 + s\sum\limits_{j=1}^{s} L_{1j}\right)\left(\frac{(1-p_{x_1})}{p_{x_1}} + \frac{(1-p_{z_1})}{p_{z_1}}\right)} + L}.$$

Similarly, inserting $\omega_{\widetilde{x}_i}, \omega_{\widetilde{z}_i}, \omega_{u_i}, \omega_{h_i}$ into (34) for $i = \overline{2, n-1}$:

$$\begin{aligned}
\phi_{z_i} = \frac{1}{2\gamma} - \frac{L}{2} & \\
& - 60\frac{\gamma}{p_{x_i}}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{lm}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) \\
& - 180\frac{\gamma}{p_{x_{i+1}}}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1) \\
& - 60\frac{\gamma}{p_{z_i}}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{lm}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) \\
& - 180\frac{\gamma}{p_{z_{i+1}}}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1) \\
& - 80\frac{\gamma\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})}{p_{u_i}\left(1-(1+s_2)(1+s_3)(1-\alpha)\right)}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}} \\
& - 80\frac{\gamma\lambda_i^2c^2d_{z_{i-1}}}{d_{x_i}(1-(1+s_2)(1+s_3)(1-\alpha))}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right] \\
& - 80\frac{\gamma}{1-(1+s_2)(1+s_3)(1-\alpha)}\lambda_{i+1}^2\left(1+s_2^{-1}\right) - \omega_{z_i}(1-\alpha)(1+s_1^{-1}).
\end{aligned}$$

As we have $\phi_{z_i} \geqslant 0$, we have for $i = \overline{2, n-1}$ the following inequality:

$$\omega_{z_i} \leqslant \frac{1-\gamma L-\gamma^2 a_{z_i}}{2\gamma(1-\alpha)\left(1+s_1^{-1}\right)},$$

where

$$\begin{aligned}
a_{z_i} := 2\Bigg[& \frac{60}{p_{x_i}}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{lm}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) \\
& + \frac{180}{p_{x_{i+1}}}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1) \\
& + \frac{60}{p_{z_i}}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{lm}F_i})^2\sum_{m=1}^{d_{z_{i-1}}} L_{im}^2 + c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right) \\
& + \frac{180}{p_{z_{i+1}}}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})((1-\alpha)(1+s_1^{-1})+1)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{80\lambda_i^2\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})}{p_{u_i}\left(1-(1+s_2)\left(1+s_3\right)(1-\alpha)\right)}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\\
&+\frac{80\lambda_i^2c^2d_{z_{i-1}}}{d_{x_i}\left(1-(1+s_2)\left(1+s_3\right)(1-\alpha)\right)}\left[\left(\left(1+s_2^{-1}\right)+2\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right.\\
&\left.+\left(1+s_2\right)\left(1+s_3^{-1}\right)(1-\alpha)\right]\\
&+\frac{80}{1-(1+s_2)\left(1+s_3\right)(1-\alpha)}\lambda_{i+1}^2\left(1+s_2^{-1}\right)\Bigg].
\end{aligned}
$$

Writing an inequality between lower bound and higher bound of $\omega_{z_i}, i=\overline{2,n-1}$, we have:

$$
\frac{\gamma b_{z_i}}{1-(1+s_1)(1-\alpha)}\leqslant\frac{1-\gamma L-\gamma^2a_{z_i}}{2\gamma(1-\alpha)\left(1+s_1^{-1}\right)}.
$$

Thus, the following inequality holds:

$$
\gamma^2\left[\frac{2(1-\alpha)\left(1+s_1^{-1}\right)b_{z_i}}{1-(1+s_1)(1-\alpha)}+a_{z_i}\right]+\gamma L\leqslant 1.
$$

The solution comes from Lemma A.6:

$$
0\leqslant\gamma\leqslant\frac{1}{\sqrt{\frac{2(1-\alpha)\left(1+s_1^{-1}\right)b_{z_i}}{1-(1+s_1)(1-\alpha)}+a_{z_i}}+L}.
$$

Substituting $s_1=\frac{1}{\sqrt{1-\alpha}}-1$, $s_2=s_3=\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1$:

$$
0\leqslant\gamma\leqslant\frac{1}{\sqrt{\frac{2(1-\alpha)b_{z_i}^*}{(1-\sqrt{1-\alpha})^2}+a_{z_i}^*}+L},
$$

where, if we denote $\beta=(1-\alpha)^{\frac{1}{3}}$:

$$
\begin{aligned}
a_{z_i}^*:=2\Bigg[&\frac{60}{p_{x_i}}\lambda_i^2(1-p_{x_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&+\frac{180}{p_{x_{i+1}}}\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})\left(\beta^3\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)+1\right)\\
&+\frac{60}{p_{z_i}}\lambda_i^2(1-p_{z_i})\left(\left(\frac{s}{\tau_t}+1\right)(2D_{\mathrm{Im}F_i})^2\sum_{m=1}^{d_{z_{i-1}}}L_{im}^2+c^4(s+d_{z_{i-1}})d_{z_{i-1}}\right)\\
&+\frac{180}{p_{z_{i+1}}}\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})\left(\beta^3\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)+1\right)\\
&+\frac{80\lambda_i^2\left(\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)+2\left(1+\frac{1}{\beta}-1\right)\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)\beta^3\right)(1-p_{u_i})}{p_{u_i}\left(1-\left(1+\frac{1}{\beta}-1\right)\left(1+\frac{1}{\beta}-1\right)\beta^3\right)}(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)c^2\frac{d_{z_{i-1}}}{d_{x_i}}\\
&\frac{80\lambda_i^2c^2d_{z_{i-1}}}{d_{x_i}\left(1-\left(1+\frac{1}{\beta}-1\right)\left(1+\frac{1}{\beta}-1\right)\beta^3\right)}\left[\left(\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)+2\left(1+\frac{1}{\beta}-1\right)\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)\beta^3\right)(1-p_{u_i})\left(\frac{s}{\tau_t}+1\right)\right.\\
&\left.+\left(1+\frac{1}{\beta}-1\right)\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)\beta^3\right]
\end{aligned}
$$

$$+\frac{80}{1-\left(1+\frac{1}{\beta}-1\right)\left(1+\frac{1}{\beta}-1\right)\beta^3}\lambda_{i+1}^2\left(1+\left(\frac{1}{\beta}-1\right)^{-1}\right)\Bigg],$$

$$\begin{aligned}
b_{z_i}^* &:= 10\lambda_{i+1}^2 d_{z_i}c^2\left(d_{x_{i+1}}+2d_{z_{i+1}}\right)\\
&+ 40\frac{\lambda_{i+1}^2}{1-(1-\alpha)^{\frac{1}{3}}}\left(2\left(1+\left(\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)^{-1}\right)+\left(1+\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)\left(1+\left(\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)^{-1}\right)(1-\alpha)\right)\\
&+ 72\frac{\lambda_{i+1}^2(1-p_{x_{i+1}})(d_{x_{i+1}}c^2)(s+d_{z_i})\left((1-\alpha)^{\frac{1}{2}}+1\right)}{p_{x_{i+1}}}\\
&+ 72\frac{\lambda_{i+1}^2(1-p_{z_{i+1}})(d_{z_{i+1}}c^2)(s+d_{z_i})\left((1-\alpha)^{\frac{1}{2}}+1\right)}{p_{z_{i+1}}}.
\end{aligned}$$

One of the last steps is to write similar conditions for ϕ_{z_1}:

$$\begin{aligned}
0\leqslant\phi_{z_1} = \frac{1}{2\gamma}-\frac{L}{2}&\\
&-180\frac{\gamma}{p_{x_2}}\lambda_2^2(1-p_{x_2})(d_{x_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1)\\
&-5\frac{\gamma}{p_{x_1}}(1-p_{x_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\\
&-180\frac{\gamma}{p_{z_2}}\lambda_2^2(1-p_{z_2})(d_{z_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1)\\
&-5\frac{\gamma}{p_{z_1}}(1-p_{z_1})\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\\
&-80\frac{\gamma\lambda_2^2\left(1+s_2^{-1}\right)}{1-(1+s_2)(1+s_3)(1-\alpha)}-\omega_{z_1}(1-\alpha)(1+s_1^{-1}).
\end{aligned}$$

Therefore,

$$\omega_{z_1}\leqslant\frac{1-\gamma L-\gamma^2 a_{z_1}}{2\gamma(1-\alpha)\left(1+s_1^{-1}\right)},$$

where

$$\begin{aligned}
a_{z_1} := 180\frac{\lambda_2^2(1-p_{x_2})}{p_{x_2}}&(d_{x_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1)\\
&+5\frac{1-p_{x_1}}{p_{x_1}}\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\\
&+180\frac{\lambda_2^2(1-p_{z_2})}{p_{z_2}}(d_{z_2}c^2)(s+d_{z_1})((1-\alpha)(1+s_1^{-1})+1)\\
&+5\frac{1-p_{z_1}}{p_{z_1}}\left(L_1+s\sum_{j=1}^{s}L_{1j}\right)\\
&+80\frac{\lambda_2^2\left(1+s_2^{-1}\right)}{1-(1+s_2)(1+s_3)(1-\alpha)}.
\end{aligned}$$

Comparing with lower bound, we obtain:

$$\frac{\gamma b_{z_1}}{1-(1+s_1)(1-\alpha)} \leqslant \frac{1-\gamma L-\gamma^2 a_{z_1}}{2\gamma(1-\alpha)\left(1+s_1^{-1}\right)}$$

Substituting $s_1 = \frac{1}{\sqrt{1-\alpha}} - 1$, $s_2 = s_3 = \frac{1}{(1-\alpha)^{\frac{1}{3}}} - 1$ and using Lemma A.6:

$$0 \leqslant \gamma \leqslant \frac{1}{\sqrt{\frac{2(1-\alpha)b^*_{z_1}}{(1-\sqrt{1-\alpha})^2} + a^*_{z_1}} + L},$$

where

$$\begin{aligned}
a_{z_1} := {} & 180\frac{\lambda_2^2(1-p_{x_2})}{p_{x_2}}(d_{x_2}c^2)(s+d_{z_1})\left(\frac{1-\alpha}{1-\sqrt{1-\alpha}}+1\right) \\
& + 5\frac{1-p_{x_1}}{p_{x_1}}\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) \\
& + 180\frac{\lambda_2^2(1-p_{z_2})}{p_{z_2}}(d_{z_2}c^2)(s+d_{z_1})\left(\frac{1-\alpha}{1-\sqrt{1-\alpha}}+1\right) \\
& + 5\frac{1-p_{z_1}}{p_{z_1}}\left(L_1 + s\sum_{j=1}^{s} L_{1j}\right) \\
& + 80\frac{\lambda_2^2\left(1+\left(\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)^{-1}\right)}{1-(1-\alpha)^{\frac{1}{3}}},
\end{aligned}$$

$$\begin{aligned}
b^*_{z_1} := {} & 10\lambda_2^2 d_{z_1}c^2\left(d_{x_2}+2d_{z_2}\right) \\
& + 40\frac{\lambda_2^2}{1-(1-\alpha)^{\frac{1}{3}}}\left(2\left(1+\left(\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)^{-1}\right)+\left(1+\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)\left(1+\left(\frac{1}{(1-\alpha)^{\frac{1}{3}}}-1\right)^{-1}\right)(1-\alpha)\right) \\
& + 72\frac{\lambda_2^2(1-p_{x_2})(d_{x_2}c^2)(s+d_{z_1})\left((1-\alpha)^{\frac{1}{2}}+1\right)}{p_{x_2}} \\
& + 72\frac{\lambda_2^2(1-p_{z_2})(d_{z_2}c^2)(s+d_{z_1})\left((1-\alpha)^{\frac{1}{2}}+1\right)}{p_{z_2}}.
\end{aligned}$$

Returning to (15) and using recursion on t we have

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\mathcal{G}_\gamma(y^t)\right\|^2 \leqslant \frac{2\Delta}{\gamma T} + \frac{2\sum\limits_{i=2}^{n}\omega_{h_i}D^0_{\mathcal{H}_i}}{\gamma T} + \frac{2\sum\limits_{i=1}^{n-1}\omega_{z_i}D^0_{\mathcal{Z}_i}}{\gamma T} + \frac{2\sum\limits_{i=1}^{n}\omega_{\widetilde{x}_i}G^0_{x_i}}{\gamma T} + \frac{2\sum\limits_{i=1}^{n-1}\omega_{\widetilde{z}_i}G^0_{z_i}}{\gamma T} + \frac{2\sum\limits_{i=2}^{n}\omega_{u_i}U^0_i}{\gamma T},$$

Algorithm 2 SVFL-EF21-PAGE

Input: Initial points $\left\{x_i^0 \in \mathbb{R}^{d_{x_i}}\right\}_{i=1}^n, \left\{z_i^0 \in \mathbb{R}^{d_{z_i}}\right\}_{i=1}^{n-1}$, amount of iterations T, batch sizes $\{\tau_t\}_{t=0}^T$, initial sequences of vectors $\left\{\mathcal{Z}_i^0 = \mathcal{C}(z_i^0)\right\}_{i=1}^{n-1}$, $\left\{\mathcal{H}_i^0 = \mathcal{C}\left(2\lambda_i(\mathcal{Z}_i^0 - F_i(x_i^0, z_i^0))\right)\right\}_{i=2}^n$, $\widetilde{g}_{x_1}^0 = \nabla_{x_1} F_1(x_1^0, z_1^0)$, $\left\{\widetilde{g}_{x_i}^0 = 2\lambda_i J_{F_i,x_i}^T(x_i^0, z_i^0)\left(F_i(x_i^0, z_i^0) - \mathcal{Z}_{i-1}^0\right)\right\}_{i=2}^n$, $\widetilde{g}_{z_1}^0 = \mathcal{H}_2^0 + \nabla_{z_1} F_1(x_1^0, z_1^0)$, $\left\{\widetilde{g}_{z_i}^0 = \mathcal{H}_{i+1}^0 + 2\lambda_i J_{F_i,z_i}^T(x_i^0, z_i^0)\left(F_i(x_i^0, z_i^0) - \mathcal{Z}_{i-1}^0\right)\right\}_{i=2}^{n-1}$, $\left\{u_i^0 = F_i(x_i^0, z_i^0)\right\}_{i=2}^n$.

Parameter: Stepsize $\gamma > 0$

for $t = 0, \ldots, T-1$ **do**

  for $i = n, \ldots, 3$ **do**

    The i-th worker does the following actions.

    $x_i^{t+1} \leftarrow x_i^t - \gamma \widetilde{g}_{x_i}^t$

    Update $\widetilde{g}_{x_i}^{t+1}$ as in (24)

    Update u_i^{t+1} as in (25)

    $h_i^t \leftarrow \mathcal{C}\left(2\lambda_i(\mathcal{Z}_{i-1}^t - u_i^{t+1}) - \mathcal{H}_i^t\right)$

    Send h_i^t to the $(i-1)$-th device

    The $(i-1)$-th worker does the following actions.

    $z_{i-1}^{t+1} \leftarrow \text{prox}_{\gamma\delta_{\text{Im}F_i}}(z_{i-1}^t - \gamma \widetilde{g}_{z_{i-1}}^t)$

    Update $\widetilde{g}_{z_{i-1}}^{t+1}$ as in (26)

    $\mathcal{Z}_{i-1}^{t+1} \leftarrow \text{proj}_{\text{Im}F_i}\left(\mathcal{Z}_{i-1}^t + \mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)\right)$

    $\mathcal{H}_i^{t+1} \leftarrow \mathcal{H}_i^t + h_i^t$

    Send $\mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)$ to the i-th worker

    The i-th worker does the following actions.

    $\mathcal{Z}_{i-1}^{t+1} \leftarrow \text{proj}_{\text{Im}F_i}\left(\mathcal{Z}_{i-1}^t + \mathcal{C}(z_{i-1}^{t+1} - \mathcal{Z}_{i-1}^t)\right)$

    $\mathcal{H}_i^{t+1} \leftarrow \mathcal{H}_i^t + h_i^t$

  end for

  The second worker does the following actions.

  $x_2^{t+1} \leftarrow x_2^t - \gamma \widetilde{g}_{x_2}^t$

  Update $\widetilde{g}_{x_2}^{t+1}$ as in (24)

  Update u_2^{t+1} as in (25)

  $h_2^t \leftarrow \mathcal{C}\left(2\lambda_2(\mathcal{Z}_1^t - u_2^{t+1}) - \mathcal{H}_2^t\right)$

  Send h_2^t to the first device

  The first worker does the following actions.

  $x_1^{t+1} \leftarrow x_1^t - \gamma \widetilde{g}_{x_1}^t$

  Update $\widetilde{g}_{x_1}^{t+1}$ as in (27)

  $z_1^{t+1} = \text{prox}_{\gamma\delta_{\text{Im}F_2}}(z_1^t - \gamma \widetilde{g}_{z_1}^t)$

  Update $\widetilde{g}_{z_1}^{t+1}$ as in (28)

  $\mathcal{Z}_1^{t+1} \leftarrow \text{proj}_{\text{Im}F_2}\left(\mathcal{Z}_1^t + \mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)\right)$

  $\mathcal{H}_2^{t+1} \leftarrow \mathcal{H}_2^t + h_2^t$

  Send $\mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)$ to the second worker

  The second worker does the following actions.

  $\mathcal{Z}_1^{t+1} \leftarrow \text{proj}_{\text{Im}F_2}\left(\mathcal{Z}_1^t + \mathcal{C}(z_1^{t+1} - \mathcal{Z}_1^t)\right)$

  $\mathcal{H}_2^{t+1} \leftarrow \mathcal{H}_2^t + h_2^t$

end for

Output: x^T