Contents

Supplementary Material

A Optima algorithms (additional)

This is an addendum to Section 4 on optimal algorithms. Here we consider:

• Sliding in case with $\lambda\lambda_{\max}(W) \leq L\sqrt{\chi(W)}$ but with summand gradient oracle – Appendix A.1.1;

• Sliding in case with $\lambda\lambda_{\max}(W) \geq L\sqrt{\chi(W)}$ – Algorithm 2 in Appendix A.1.2;

• other approach based not on Sliding but on Randomized Local method – Algorithm 3 in Appendix A.2.

A.1 PF Min-Max via Sliding

A.1.1 Case $\lambda\lambda_{\max}(W) \leq L\sqrt{\chi(W)}$

• **S1DMM + Rand. Extra Step Method.** We apply this algorithm when $f_m(x_m, y_m) = \frac{1}{r}\sum_{j=1}^{r} f_{m,r}(x_m, y_m)$, where each $f_{m,r}$ is convex-concave and each f_i is L-average smooth. The following theorem states the convergence rate of S1DMM with Extragradient with Variance Reduction [1] as a local algorithm.

Theorem 3 *Let Algorithm 1 be applied for solving (2) with convex-concave and L-average smooth local functions f_m. Then one can choose a constant step γ, a precision δ, so that we need $\mathcal{O}(\frac{\lambda\lambda_{\max}(W)\Omega^2}{\varepsilon})$ communication rounds and $\tilde{\mathcal{O}}(\frac{(r\lambda\lambda_{\max}(W)+\sqrt{r}L)\Omega^2}{\varepsilon})$ local computation on each node, to obtain $\hat{z} = (\hat{x}, \hat{y})$ such that $[\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \min_{x\in\mathcal{X}} f(x, \hat{y})] \leq \varepsilon$.*
If we additionally assume that global objective function f is μ-strongly-convex-strongly-concave, then after $\mathcal{O}(\frac{\lambda\lambda_{\max}(W)}{\mu}\log\frac{1}{\varepsilon})$ communication rounds and $\tilde{\mathcal{O}}((r\frac{\lambda\lambda_{\max}(W)}{\mu} + \frac{\sqrt{r}L}{\mu})\log\frac{1}{\varepsilon})$ local computations on each node we will obtain $\hat{z}$, such that $\|\hat{z} - z^*\|^2 \leq \varepsilon$.
One can find the γ and δ settings for Algorithm 1 in Appendix C.

A.1.2 Case $\lambda\lambda_{\max}(W) \geq L\sqrt{\chi(W)}$

Algorithm 2 Sliding 2 for Decentralized Min-Max(S2DMM)

Parameters: stepsize γ
Initialization: choose $x^0, y^0 \in \mathcal{X}\times\mathcal{Y}$, $x_m^0 = x^0$, $y_m^0 = y^0$ for all m

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $V_x^k = X^k - \gamma\cdot\nabla_X f(X^k, Y^k)$
$V_y^k = Y^k + \gamma\cdot\nabla_Y f(X^k, Y^k)$
3: Find U^k, such that $\|U^k - \hat{U}^k\|_F^2 \leq \delta$, where $\hat{U}^k$ is a solution of:

$$\min_{U_x} \frac{\lambda\|\sqrt{W}U_x\|_F^2}{2} + \frac{\|U_x - V_x^k\|_F^2}{2\gamma}$$
$$\max_{U_y} -\frac{\lambda\|\sqrt{W}U_y\|_F^2}{2} - \frac{\|U_y - V_y^k\|_F^2}{2\gamma} \tag{4}$$

4: $X^{k+1} = \text{proj}_{\mathcal{X}}\left(U_x^k + \gamma\cdot\left(\nabla_X f(X^k, Y^k) - \nabla_X f(U_x^k, U_y^k)\right)\right)$
5: $Y^{k+1} = \text{proj}_{\mathcal{Y}}\left(U_y^k - \gamma\cdot\left(\nabla_Y f(X^k, Y^k) - \nabla_Y f(U_x^k, U_y^k)\right)\right)$
6: **end for**

Note that we only need to communicate with other devices on lines 3 when computing the prox operator for $\frac{\lambda\gamma}{2}\left\|\sqrt{W}X\right\|_F^2$. This problem is divided into two minimization sub-problems, by X, and by Y. Hence, the problem (4) is solved by Fast Gradient Descent. Further, we note that the algorithm's steps in lines 2, 4, and 5 are local and separable on each machine. The following theorem states the convergence rate of S2DMM with FGD.

Theorem 4 *Let Algorithm 2 be applied for solving (2) with convex-concave and L - smooth local functions f_m. Then one can choose a constant step γ, so that we need $\tilde{\mathcal{O}}\left(\min\left\{\frac{L\Omega^2}{\varepsilon}\sqrt{\chi(W)}, \frac{\sqrt{\lambda\lambda_{\max}(W)L}\Omega^2}{\varepsilon}\right\}\right)$ comm. rounds and $\mathcal{O}\left(\frac{L\Omega^2}{\varepsilon}\right)$ local comp. on each node to obtain $\hat{z} = (\hat{x}, \hat{y})$ s.t. $[\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \min_{x\in\mathcal{X}} f(x, \hat{y})] \leq \varepsilon$.*
If we additionally assume that global objective function f is μ_f-strongly-convex-strongly-concave, then it holds that we have $\|\hat{z} - z^*\|^2 \leq \varepsilon$ after $\tilde{\mathcal{O}}(\min\{\frac{L}{\mu}\sqrt{\chi(W)}, \frac{\sqrt{\lambda\lambda_{\max}(W)L}}{\mu}\}\log\frac{1}{\varepsilon})$ comm. rounds and $\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon})$ local comp. on each node.
As expected, the communication complexity of S2DMM + FGD is $O\left(\frac{L}{\mu}\sqrt{\chi(W)}\log\frac{1}{\varepsilon}\right)$ in the strongly-convex-strongly-concave case. It is optimal when $\lambda\lambda_{\max}(W) = O(L\sqrt{\chi(W)})$. The local gradient complexity is $\tilde{O}(\frac{L}{\mu})$, which is, up to log and constant factors identical to the lower bound on the local gradient calls.
One can find the γ and δ settings for Algorithm 2 in Appendix C. Below we will discuss and compare both methods. Next, we move on to the second method.

A.2 PF Min-Max via Randomized Local Extra Step Method

Our first two methods (Algorithms 1 and 2) make several iterations between communications when λ is small (or vice versa, for big λ make some communications between one local iteration). The following method (Algorithm 3) is also sharpened on the alternation of local iterations and communications, but it makes them more evenly. Our method is similar to the randomized local methods (for example, as the method from [10]), but it uses not only importance sampling, but also implicit variance reduction technique [1].

Algorithm 3 Randomized for Decentralized Min-Max (RDMM)

parameters: stepsize γ, probability p, probability ρ, probability distribution Q
initialization: choose $x^0, y^0 \in \mathcal{X}\times\mathcal{Y}$, $x_m^0 = x^0$, $y_m^0 = y^0$ for all m

for $k = 0, 1, 2, \ldots$ **do**
  $\bar{X}^k = (1-\rho)X^k + \rho U_x^k, \quad \bar{Y}^k = (1-\rho)Y^k + \rho U_y^k$
  $X^{k+1/2} = \text{proj}_{\mathcal{X}}\left(\bar{X}^k - \gamma\cdot(\nabla_X f(U_x^k, U_y^k) + \lambda W U_x^k)\right)$,
  $Y^{k+1/2} = \text{proj}_{\mathcal{Y}}\left(\bar{Y}^k + \gamma\cdot(\nabla_Y f(U_x^k, U_y^k) - \lambda W U_y^k)\right)$
  Generate $\xi^k = \begin{cases}1, & \text{with probability } 1-p\\ 0, & \text{with probability } p\end{cases}$,
  If $\xi^k = 0$:
    $X^{k+1} = \text{proj}_{\mathcal{X}}\left(\bar{X}^k - \gamma\cdot\left(\frac{\lambda}{p}WX^{k+1/2} - \frac{\lambda(1-p)}{p}WU_x^k + \nabla_X f(U_x^k, U_y^k)\right)\right)$,
    $Y^{k+1} = \text{proj}_{\mathcal{Y}}\left(\bar{Y}^k + \gamma\cdot\left(-\frac{\lambda}{p}WY^{k+1/2} + \frac{\lambda(1-p)}{p}WU_y^k + \nabla_Y f(U_x^k, U_y^k)\right)\right)$
  If $\xi^k = 1$:
    Generate an vector of indexes $\hat{\xi}_k$ according to Q
    $X^{k+1} = \text{proj}_{\mathcal{X}}\Bigg(\bar{X}^k - \gamma\cdot\Big(\frac{1}{(1-p)N}\left(\nabla_X f_{\hat{\xi}_k}(X^{k+1/2}, Y^{k+\frac{1}{2}}) - \nabla_X f_{\hat{\xi}_k}(U_x^k, U_y^k)\right)$
    $+\nabla_X f(U_x^k, U_y^k) + \lambda W U_x^k\Big)\Bigg)$,
    $Y^{k+1} = \text{proj}_{\mathcal{Y}}\Bigg(\bar{Y}^k + \gamma\cdot\Big(\frac{1}{(1-p)N}\left(\nabla_Y f_{\hat{\xi}_k}(X^{k+1/2}, Y^{k+\frac{1}{2}}) - \nabla_Y f_{\hat{\xi}_k}(U_x^k, U_y^k)\right)$
    $+\nabla_Y f(U_x^k, U_y^k) - \lambda W U_y^k\Big)\Bigg)$
  Generate $\xi^{k+1/2} = \begin{cases}1, & \text{with prob. } 1-\rho\\ 0, & \text{with prob. } \rho\end{cases}$,
  $U_x^{k+1} = \xi^{k+1/2}\cdot U_x^k + (1-\xi^{k+1/2})\cdot X^{k+1}$
  $U_y^{k+1} = \xi^{k+1/2}\cdot U_y^k + (1-\xi^{k+1/2})\cdot Y^{k+1}$
end for

The following theorem states the convergence rate of RDMM.

Theorem 5 *Let Algorithm 3 be applied for solving (2) with convex-concave and L -smooth local functions f_{mj} with sum structure. Then one can choose a constant step γ, probabilities ρ and p so that we need (in average) $\mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)\Omega^2}{\varepsilon}\right)$ comm. rounds while the local average complexity for some ρ and p is $\mathcal{O}\left(\frac{\sqrt{r}\bar{L}\Omega^2}{\varepsilon}\right)$ to obtain $\hat{z} = (\hat{x}, \hat{y})$ s.t. $[\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \min_{x\in\mathcal{X}} f(x, \hat{y})] \leq \varepsilon$. Here $\bar{L} = \sqrt{L^2 + \lambda^2\lambda^2_{\max}(W)}$.*
If we additionally assume that global objective function f is μ-strongly-convex-strongly-concave, then for ρ and p it holds that we have $\|\hat{z} - z^*\|^2 \leq \varepsilon$ after (in average) $\mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)}{\mu}\log\frac{1}{\varepsilon}\right)$ comm. rounds while the local average complexity for some ρ and p is $\mathcal{O}\left(\left(r + \frac{\sqrt{r}\bar{L}}{\mu}\right)\log\frac{1}{\varepsilon}\right)$.
One can find the ρ and p settings for Algorithm 3 in Appendix C.

A.3 Discussions of the methods

Let us note some remarks about the obtained methods. Here we compare two different approaches: deterministic – Algorithm 1 and 2, and stochastic – Algorithm 3.
• Both approaches allow combining and interleaving local iterations and decentralized gossip communications. As far as we know, there are no such algorithms in the literature for saddle point problems.
• Algorithm 1 and 2 have better convergence estimates than Algorithm 3. Moreover, we assume that Algorithm 1 is optimal for (1) in case, when λ is small, and Algorithm 2 – when λ is big.
• It seems that Algorithm 3 is more robust than Algorithms 2 and 2. This is primarily due to the fact that Algorithm 1 solves an auxiliary problem with a given precision. In the case of convex-concave functions, there are no problems here. The stopping criterion is clear. But if we try to extend Algorithm 1 to more complex problems, for example, to a non-convex one, the question of the precision and stopping criteria becomes open. Algorithm 2 do not face such problems.
• On the other hand, in Algorithm 1 it is straightforward to change the method for the auxiliary problem, taking into account the peculiarities of f_m. For example, stochastic methods [13], or variance reduction technique [1] can be easily used.
• In Algorithm 3 on lines 5, 6 and 9, the choice is made synchronously on all nodes, i.e., we need to transfer information to all nodes somehow, that is, at the current step you need to choose the 1st or 2nd option. There are no problems in a centralized case. Still, this slightly slows down the algorithm for a decentralized architecture until the package reaches all addressees, although we only need to send 1 bit. In practice, this problem can be solved using a schedule, making communication or a local step not random, but according to some predefined schedule that all nodes know. For example, once in $1/p$ iterations, make a communication.

B Additional experiments

We implement all methods in Python 3.8 using PyTorch 1.4 [26] and run on a machine with 56 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz cores, and on NVIDIA TITAN X GPU with 11,264MBs RAM memory (Cuda 11.1).

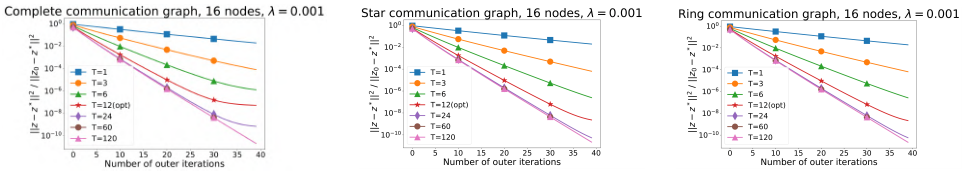
B.1 Toy experiments

We start from toy experiments, the purpose of this experiment is to verify the theoretical results. We conduct our toy experiments on bilinear problem:

$$f_m(x, y) = x^\top A_m y + a_m^\top x + b_m^\top y + \frac{\beta}{2}\|x\|^2 - \frac{\beta}{2}\|y\|^2, \tag{5}$$

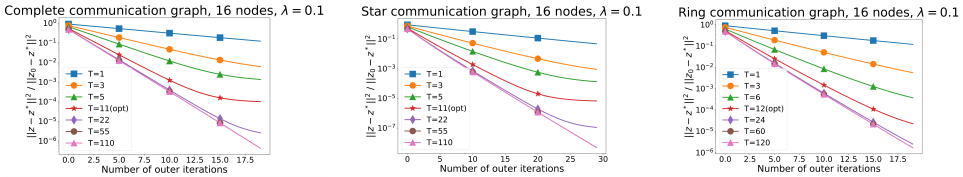
where $A_m \in \mathbb{R}^{n\times n}$, $a_m, b_m \in \mathbb{R}^d$. We take $n = 100$ and generate positive definite matrices A_m and vectors a_m, b_m randomly, such that $L = 5$. We take $M = 16$ and $\beta = 0, 1$. We use three topologies of network: complete graph, star and ring. In all experiments, we compare the Algorithms in the rate of convergence the solution in terms of the number of communications (for Algorithm 1 – outer iterations, for Algorithm 2 – inner iterations).

• In the first experiment we compare Algorithm 1 with $\gamma = \frac{1}{2\lambda\lambda_{\max}(W)}$ (as in theory) and the different numbers of inner iterations T (for subproblem (3)). See results on Figure 3. We see that from the



(a) complete (b) star (c) ring

Figure 2: Comparison of Algorithm 1 with different T on different networks for (2)+(5) with $\lambda = 0, 001$.

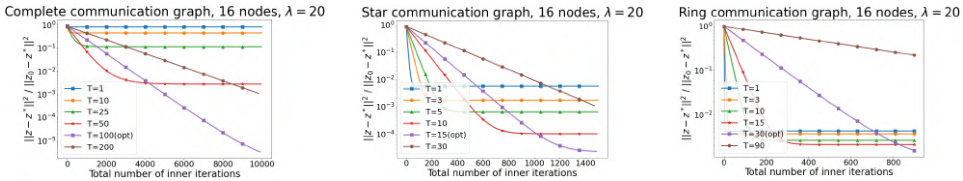


(a) complete (b) star (c) ring

Figure 3: Comparison of Algorithm 1 with different T on different networks for (2)+(5) with $\lambda = 0, 1$.

point of view of the number of communications, the theoretically optimal number of innner steps T_{opt} is almost optimal in practice. It is also seen that there is a certain limiting T after which an increase in the number of inner iterations does not give a particular acceleration of convergence in terms of communications (outer iterations).

• In the second experiment we compare Algorithm 2 with $\gamma = \frac{1}{2L}$ (as in theory) and the different numbers of inner iterations T (for subproblem (4)). See results on Figure 4.



(a) complete (b) star (c) ring

Figure 4: Comparison of Algorithm 2 with different T on different networks for (2)+(5) with $\lambda = 20$.

We see that from the point of view of the number of communications, the large number innner steps $T > T_{\text{opt}}$ only slows down the convergence. On the contrary, a small number of inner iterations $T < T_{\text{opt}}$ accelerates, but degrades the accuracy of the solution. The optimal T_{opt} gives a good balance of accuracy and rate.

• In the third experiment we compare Algorithm 3 for problem with $r = 1$ with $\gamma = \frac{\sqrt{p}}{2(L+\lambda\lambda_{\max}(W))}$ (as in theory) and the different probabilities $p = \rho$. See results on Figures 6 and 7.

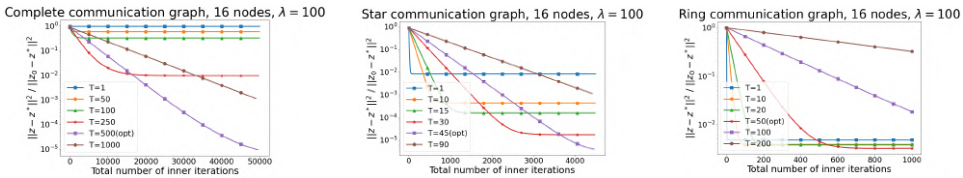
In terms of the number of communications, the optimal value is $\rho = p = \frac{\lambda^2\lambda^2_{\max}(W)}{\lambda^2\lambda^2_{\max}(W)+L^2}$ (see Section C.2.4). But it can be seen on Figures 6 (bottom) and 7 (bottom), the probability (frequency of communications) can be reduced. But the optimal $\rho = p = \frac{\lambda^2\lambda^2_{\max}(W)}{\lambda^2\lambda^2_{\max}(W)+L^2}$ outperforms the smaller probabilities in terms of the total number of iterations.

B.2 Neural networks

Our additional experiments are aimed at comparing the operation of Algorithms 1 and 3 with other parameters r and q (see Table 3).

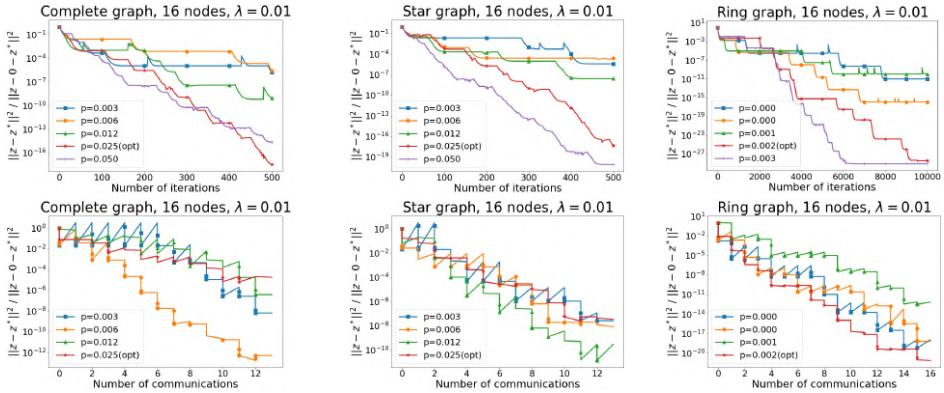
On all plots, then vertical axis – accuracy, the horizontal axis – the number of epochs. 1 epoch = calculation of the gradient for all batches on the local dataset.

The results show that for other r and θ Algorithm 1 is superior to Algorithm 3 in training the global model. Only in the case of frequent communications with the server Algorithm 3 shows better results.



(a) complete (b) star (c) ring

Figure 5: Comparison of Algorithm 2 with different T on different networks for (2)+(5) with $\lambda = 100$.



(a) complete (b) star (c) ring

Figure 6: Comparison of Algorithm 3 with different $p = \rho$ on different networks for (2)+(5) with $\lambda = 0,01$.
Top: in terms of all iterations, **bottom:** in terms of communications.

λ	T	θ	q	Results
$1/4$	40 (1 epoch)	1	1	Figure 8 (a)
$1/40$	400 (10 epochs)	1	1	Figure 8 (b)
$1/80$	800 (20 epochs)	1	1	Figure 8 (c)
$1/4$	40 (1 epoch)	4	4	Figure 9 (a)
$1/40$	400 (10 epochs)	4	4	Figure 9 (b)
$1/80$	800 (20 epochs)	4	4	Figure 9 (c)

Table 3: Additional parameters for comparison of Algorithm 1 and Algorithm 3.

But in the case of Federated Learning, communications are a bottleneck and we want to reduce their number. Algorithm 1 copes with this problem without losing quality.

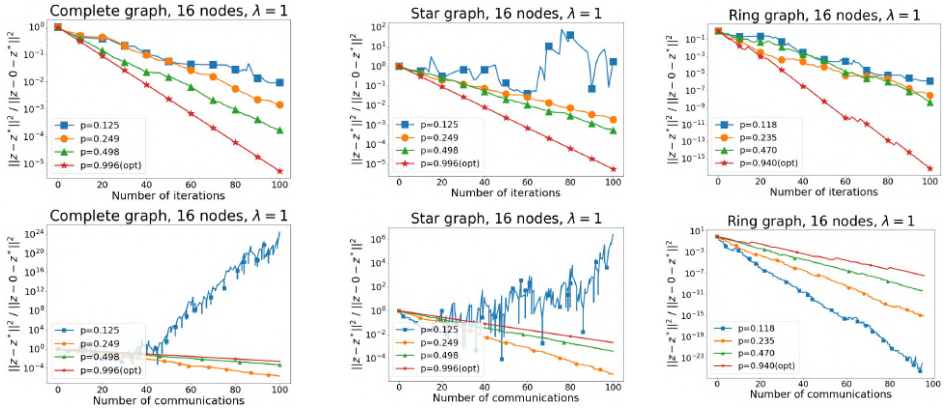
B.3 Intuition for Neural Networks

Despite the fact that our theoretical analysis captures only convex-concave case, we would like to adapt the proposed Algorithms for federated training for training neural networks. For simplicity, we will consider only the centralized case. In this situations, the regularizer $\frac{\lambda}{2}\|\sqrt{W}X\|^2 = \frac{\lambda}{2}\sum_{m=1}^{M}\|x_m - \bar{x}\|^2$ and $\lambda_{\max}(W) = 1$. Additionally, let us define the smoothness constant of the neural network for particular dataset as follows: $L = \frac{1}{\gamma_{\text{opt}}}$, where γ_{opt} is the learning rate that is recommended to be used by concrete algorithm for given dataset and neural network architecture. (for example, such information can be found in tutorials or open-source codes). This definition comes from the standard results, that for smooth functions the step-size $\sim \frac{1}{L}$. We do not say that this is a good definition of L, but we only need it for the intuition. We also mention that we are interested in the case when $\lambda \ll L$.

Now, let us start parsing and adapting the Algorithms 1 and 3.

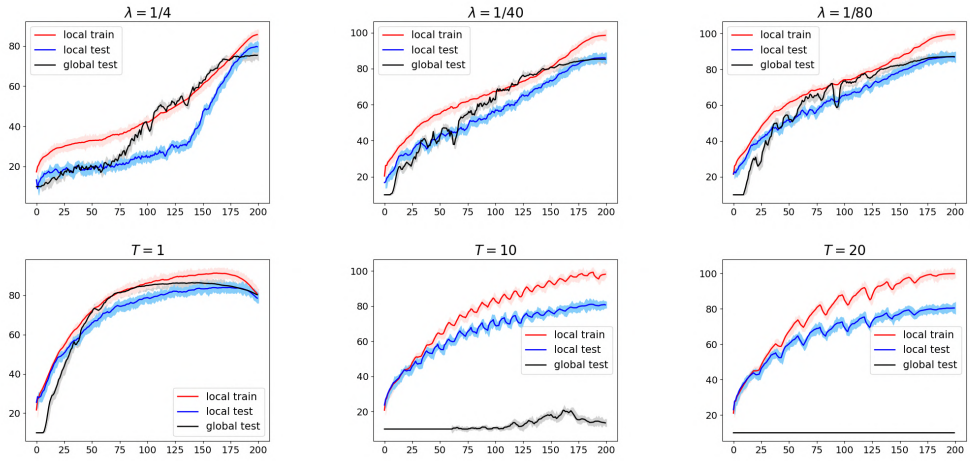
Algorithm 1. Let us take $\gamma = \frac{1}{\theta\lambda}$ in Algorithm 1, with constant $\theta \geq 1$. Then line 2 can be rewritten as follows:

$$v_{x,m}^k = x_m^k - \tfrac{1}{\theta}(x_m^k - \bar{x}^k) = \left(1 - \tfrac{1}{\theta}\right) x_m^k + \tfrac{1}{\theta}\bar{x}^k. \tag{6}$$



(a) complete (b) star (c) ring

Figure 7: Comparison of Algorithm 3 with different $p = \rho$ on different networks for (2)+(5) with $\lambda = 1$.
Top: in terms of all iterations, **bottom:** in terms of communications.



(a) (b) (c)

Figure 8: Evolution of average accuracy during training process for different parameters of λ and T. We run this experiment with $q = \theta = 1$.
Top: presents the results of Algorithm 1, **bottom:** presents the results obtained by Algorithm 3. Red line – accuracy of the local model on local train data, blue line - accuracy of the local model on local test data, black line – accuracy of the global model on global test data. The experiment was repeated 5 times, the confidence intervals are shown in lighter color.

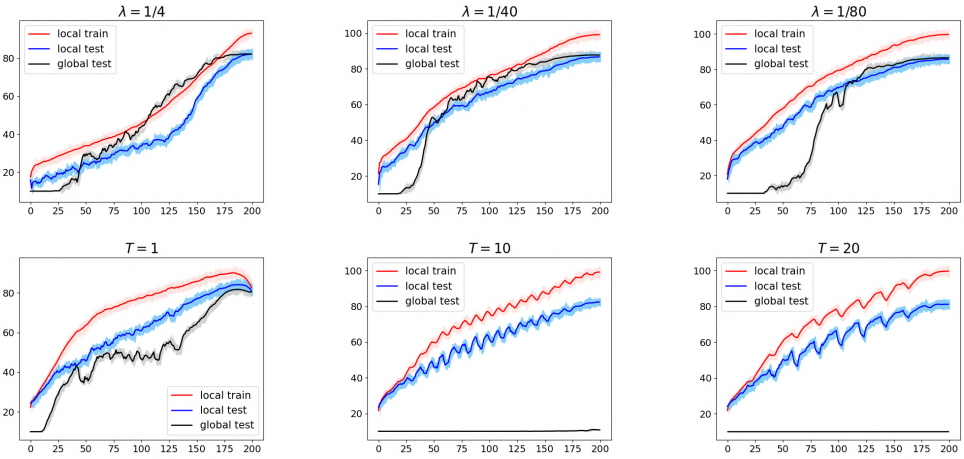
It is easy to see that the degree of how much we trust the average model depends on θ. Next, on line 3, local models are trained taking into account the regularizer:

$$\min_{u_{x,m}} \max_{u_{y,m}} \left[f_m(u_{x,m}, u_{y,m}) + \frac{\theta\lambda}{2}\|u_{x,m} - v_{x,m}^k\|^2 - \frac{\theta\lambda}{2}\|u_{y,m} - v_{y,m}^k\|^2 \right].$$

In this case, we do not know how long (number of iterations T) it takes to train the model, but one can adapt various empirical stopping criteria. The only thing that we note is that we need to solve this problem with the learning rate γ_{opt}. This is due to the fact that $L \gg \lambda$; therefore, we can assume that the "smoothness" constant of problem (3) is equal to L and we can take learning rate $\sim \frac{1}{L} = \gamma_{\text{opt}}$. On lines 4 and 5 we again make averaging:

$$\begin{aligned} x_m^{k+1} &= u_{x,m}^k + \tfrac{1}{\theta}\left(\tfrac{1}{\theta}(x_m^k - \bar{x}^k) - \tfrac{1}{\theta}(u_{x,m}^k - \bar{u}^k)\right) \\ &= \left(1 - \tfrac{1}{\theta^2}\right) u_{x,m}^k + \tfrac{1}{\theta^2} x_m^k + \tfrac{1}{\theta^2}\bar{u}^k - \tfrac{1}{\theta^2}\bar{x}^k. \end{aligned}$$

This is the whole essence of Algorithm 1. The only hyper-parameters that we can tune are θ and the number of iterations T.



(a) (b) (c)

Figure 9: Average accuracy in during process of learning with different average parameters λ, T and $r = \theta = 4$.
Top: presents the results of Algorithm 1, **bottom:** presents the results obtained by Algorithm 3. Red line – accuracy of the local model on local train data, blue line - accuracy of the local model on local test data, black line – accuracy of the global model on global test data. The experiment was repeated 5 times, the deviations are reflected.

Algorithm 3. The Algorithm 3 with $r = 1$ looks rather long, let us for simplicity discuss the main idea in a centralized setting. We have

$$x_m^{k+1} = \begin{cases} x_m^k - \frac{\gamma}{1-p} \cdot \nabla_x f(x_m^k, y_m^k), & \text{with prob. } 1-p \\ x_m^k - \frac{\gamma\lambda}{p} \cdot (x_m^k - \bar{x}^k), & \text{with prob. } p, \end{cases}$$

where p can be chosen according to importance sampling, in which case we have $p = \frac{\lambda}{\lambda+L} \approx \frac{\lambda}{L}$ (see Section C.2.4). Further, let $\gamma = \frac{\gamma_{\text{opt}}}{q}$ with $q \geq 1$. Then we get an update:

$$x_m^{k+1} = \begin{cases} x_m^k - \frac{\gamma_{\text{opt}}}{q} \cdot \nabla_x f(x_m^k, y_m^k), & \text{with prob. } 1-p \\ x_m^k - \frac{1}{q} \cdot (x_m^k - \bar{x}^k), & \text{with prob. } p \end{cases}$$

Afterward, the algorithm makes a communication: $x_m^{k+1} = x_m^k - \frac{1}{q} \cdot (x_m^k - \bar{x}^k)$, once in $\frac{1}{p} = \frac{1}{\lambda\gamma_{\text{opt}}}$ iterations, otherwise make a local step: $x_m^{k+1} = x_m^k - \frac{\gamma_{\text{opt}}}{q} \cdot \nabla_x f(x_m^k, y_m^k)$.

Remark. We wrote only about update of x_m^k, the update for y is easy to get in a similar way.
Remark. In order for the comparison of Algorithm 1 and Algorithm 3 to be fair, it is nessessary to balance the number of communications and local iterations for both algorithms, that is why we take $T = \frac{1}{p} = \frac{1}{\gamma_{\text{opt}}\lambda} = \frac{10}{\lambda}$, where T – parameter of Algorithm 1 and p – of Algorithm 3. Next, we want to take $\theta = r = 2$. This is due to the fact that we want to level how much each of the local models trusts the average model and relies on it (see (6)).

C Missing proofs

C.1 Lower bounds

C.1.1 Notation

In this section we present a proof of lower bounds for a class of algorithms satisfying Assumption 3. However, the problem solved by the algorithm $\mathcal{A}$ has a decentralized structure for only one group of variables, without limiting generality, let these be variables $\mathbf{y}$:

$$\min_{\mathbf{x}\in\mathbb{R}^{Mn_x}} \max_{\mathbf{y}\in\mathbb{R}^{Mn_y}} \left\{ F(\mathbf{x},\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} f_m(x_m, y_m) - \frac{\lambda}{2}\mathbf{y}^\top \mathbf{W}\mathbf{y} \right\}, \tag{7}$$

where $\mathbf{x}^\top = (x_1^\top, \ldots, x_M^\top)$ and $\mathbf{y}^\top = (y_1^\top, \ldots, y_M^\top)$, $x_m \in \mathbb{R}^{n_x}$ and $y_m \in \mathbb{R}^{n_y}$ for any $m \in \{1, \ldots, M\}$, matrix $\mathbf{W} = W \otimes I_M$, where W is the gossip matrix (see Assumption 2). Moreover, $\mathbf{W}$ satisfies the same properties of gossip matrix:

1. $\mathbf{W}$ is symmetric and positive semi-defined matrix;
2. $\mathbf{W}_{i,j} \neq 0$ if and only if $i = j$ or $(i, j) \in \mathcal{E}$;
3. ker$\mathbf{W}$ is consensus space;
4. $\lambda_{\max}(\mathbf{W}) = \lambda_{\max}(W)$ and $\lambda^+_{\min}(\mathbf{W}) = \lambda^+_{\min}(W)$, where $\lambda^+_{\min}$ is the smallest positive eigenvalue and $\lambda_{\max}$ is the largest eigenvalue;

It is worth noting that for the simplicity of the proof, we work in vector space, unlike other sections where the proofs are carried out in matrix notation. But we would like to assure the reader that the structure of variables does not affect the results in any way. We would also like to draw the attention to the structure of the problem (7). We consider a simplified formulation of the problem when compared to the original one (2), but this only means that the lower bounds for the original problem (2) will look either similar or more complicated than for a simple formulation (7). Fortunately, the upper bounds for the problem (2) coincide with the lower bounds for the problem (7), which leads us to the following conclusion: **the lower bounds are the same as for the problem (2).**

C.1.2 Proof of Theorem 1

As in many papers we give an example of a "bad" function on which algorithms satisfying Assumption 3 converge at least at a rate that coincides with the lower bounds. We consider a linear graph with the number of nodes equal to $M = 3\left\lfloor \frac{\chi}{3} \right\rfloor$, where $\chi = \lambda_{\max}(\mathbf{W})/\lambda_{\min}(\mathbf{W})$ is condition number of communication network $\mathcal{G}$. Let us divide the nodes into three types: the first type includes $\mathcal{V}_1 = \left\{1, 2, \ldots, \frac{M}{3}\right\}$, the second type includes $\mathcal{V}_2 = \left\{\frac{M}{3} + 1, \frac{M}{3} + 2, \ldots, \frac{2M}{3}\right\}$, the third type includes $\mathcal{V}_3 = \left\{\frac{2M}{3} + 1, \frac{2M}{3} + 2, \ldots, M\right\}$. Let $n_x = n_y = n = 2T$ dimension and T more than M. Next, we define

$$f_m(x, y) = \begin{cases} \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2 - ay(1) + \frac{\sqrt{\lambda}}{2}x^\top A_1 y, & \text{if } m \in \mathcal{V}_1 \\ \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2, & \text{if } m \in \mathcal{V}_2 \\ \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{\sqrt{\lambda}}{2}x^\top A_2 y, & \text{if } m \in \mathcal{V}_3, \end{cases} \tag{8}$$

where matrices A_1, A_2 are defined as follows

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & \begin{pmatrix} c & -c \\ -c & c \end{pmatrix} & 0 & \ddots & \vdots \\ 0 & 0 & \begin{pmatrix} c & -c \\ -c & c \end{pmatrix} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & \sqrt{2}b \end{pmatrix} \text{ and}$$

$$A_2 = \begin{pmatrix} \begin{pmatrix} c & -c \\ -c & c \end{pmatrix} & 0 & \ldots \\ 0 & \begin{pmatrix} c & -c \\ -c & c \end{pmatrix} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

For the gossip matrix, we take the Laplacian of a linear graph. Then, for our problem, we get that the matrix $\mathbf{W}$ will have the following form:

$$\mathbf{W} = \hat{W} \otimes I_M,$$

where matrix $\hat{W} \in \mathbb{R}^{n \times n}$ has following form

$$\hat{W} = \frac{1}{2M} \cdot \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & -1 & 2 & -1 & & & \\ & & & -1 & 2 & -1 & & \\ & & & & & \dots & & \\ & & & & & -1 & 2 & -1 \\ & & & & & & -1 & 1 \end{pmatrix}.$$

It is easy to make sure that this matrix $\mathbf{W}$ satisfies the assumption of the gossip matrix. According to the definition of functions, now we can write the form of the objective function of problem:

$$F(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \left(\frac{\mu}{2} \|\mathbf{x}\|^2 - \frac{\mu}{2} \|\mathbf{y}\|^2 + \frac{\sqrt{\lambda}}{2} \mathbf{x}^\top \mathbf{M} \mathbf{y} - a \sum_{m \in \mathcal{V}_1} y_m(1) \right) - \frac{\lambda}{2} \mathbf{y}^\top \mathbf{W} \mathbf{y}, \tag{9}$$

where $\mathbf{M}$ looks like

$$\mathbf{M} = \begin{pmatrix} I \otimes A_1 & 0 & 0 \\ 0 & \mathbf{0}_{n|\mathcal{V}_2|} & 0 \\ 0 & 0 & I \otimes A_2 \end{pmatrix}.$$

To obtain lower bounds, we will find the dual function of our objective function$F(\mathbf{x}, \mathbf{y})$. But before that, using properties of Kronecker product, we note that

$$\begin{aligned} \mathbf{M}^\top \mathbf{M} &= \begin{pmatrix} I \otimes A_1^\top A_1 & 0 & 0 \\ 0 & \mathbf{0}_{n|\mathcal{V}_2|} & 0 \\ 0 & 0 & I \otimes A_2^\top A_2 \end{pmatrix} = \begin{pmatrix} I \otimes A_1^2 & 0 & 0 \\ 0 & \mathbf{0}_{n|\mathcal{V}_2|} & 0 \\ 0 & 0 & I \otimes A_2^2 \end{pmatrix} \\ &= 2 \begin{pmatrix} I \otimes B_1 & 0 & 0 \\ 0 & \mathbf{0}_{n|\mathcal{V}_2|} & 0 \\ 0 & 0 & I \otimes B_2 \end{pmatrix}, \end{aligned}$$

where matrices B_1, B_2 are defined in such way

$$B_1 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & \begin{pmatrix} c^2 & -c^2 \\ -c^2 & c^2 \end{pmatrix} & 0 & \ddots & \vdots \\ 0 & 0 & \begin{pmatrix} c^2 & -c^2 \\ -c^2 & c^2 \end{pmatrix} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & b^2 \end{pmatrix} \text{ and}$$

$$B_2 = \begin{pmatrix} \begin{pmatrix} c^2 & -c^2 \\ -c^2 & c^2 \end{pmatrix} & 0 & \dots \\ 0 & \begin{pmatrix} c^2 & -c^2 \\ -c^2 & c^2 \end{pmatrix} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

The dual function $\Psi(\mathbf{y})$ of our objective function $F(\mathbf{x}, \mathbf{y})$ is

$$\Psi(\mathbf{y}) = \min_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}) = -\frac{1}{M} \left(\frac{\lambda}{8\mu} \mathbf{y}^\top M^\top M \mathbf{y} - a \sum_{i \in \mathcal{V}_1} y_i(1) - \frac{\mu}{2} \mathbf{y}^\top \mathbf{y} \right) - \frac{\lambda}{2} \mathbf{y}^\top \mathbf{W} \mathbf{y}.$$

Therefore, we obtain

$$\frac{M}{\lambda} \Psi(\mathbf{y}) = -\frac{1}{2} \mathbf{y}^\top \mathbf{C} \mathbf{y} - a \sum_{i \in \mathcal{V}_1} y_i(1),$$

where $\mathbf{C} = M\mathbf{W} + \frac{\mu}{\lambda}\mathbf{I} + \frac{1}{2\mu}\mathbf{B}$ and

$$\mathbf{B} = \begin{pmatrix} I \otimes B_1 & 0 & 0 \\ 0 & \mathbf{0}_{n|\mathcal{V}_2|} & 0 \\ 0 & 0 & I \otimes B_2 \end{pmatrix}.$$

Now, we are ready to present the problem

$$\max_{\mathbf{y}} \left\{ \frac{M}{\lambda} \Psi(\mathbf{y}) = -\frac{1}{2} \mathbf{y}^\top \mathbf{C} \mathbf{y} - a \sum_{i \in \mathcal{V}_1} y_i(1) \right\}. \tag{10}$$

Due to the fact that the functions of each type are similar, their maximums coincide. Therefore, we denote them as follows

$$\underset{v}{\operatorname{argmax}} \left\{ \min_{u} f_m(u, v) \right\} = \begin{cases} x^\star, & \text{if } m \in \mathcal{V}_1 \\ y^\star, & \text{if } m \in \mathcal{V}_2 \\ z^\star, & \text{if } m \in \mathcal{V}_3 \end{cases}. \tag{11}$$

Now we give a proof of the lemma that indicates a recursive connection

Lemma 1 (see [9]) *Let*

$$w_i = \begin{cases} \begin{pmatrix} z_i^\star \\ x_i^\star \end{pmatrix} & \textit{if } i \textit{ is even} \\ \begin{pmatrix} x_i^\star \\ z_i^\star \end{pmatrix} & \textit{if } i \textit{ is odd} \end{cases}. \tag{12}$$

Then, we have

$$w_{i+1} = \mathbf{Q} w_i, \tag{13}$$

where

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{2\left(\frac{c^2}{2\mu}\right)\left(1+\frac{2\mu}{\lambda}\right)} & \frac{\left(\frac{c^2}{2\mu}\right)+\frac{\mu}{\lambda}+\frac{1}{2}}{\left(\frac{c^2}{2\mu}\right)} \\ -\frac{\left(\frac{c^2}{2\mu}\right)+\frac{\mu}{\lambda}+\frac{1}{2}}{\left(\frac{c^2}{2\mu}\right)} & \left(1+\frac{2\mu}{\lambda}\right)\left(\frac{2\left(\left(\frac{c^2}{2\mu}\right)+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\left(\frac{c^2}{2\mu}\right)} - 2\left(\frac{c^2}{2\mu}\right)\right) \end{pmatrix}. \tag{14}$$

Proof Let's write down the first-order optimality conditions for the problem (10)

$$\left(\frac{c^2}{2\mu} + \frac{\mu}{\lambda} + \frac{1}{2}\right) x_{2i+1}^* - \frac{c^2}{2\mu} x_{2i}^* - \frac{1}{2} y_{2i+1}^* = 0, \quad \text{for } 0 \le i \le T-1 \tag{15}$$

$$\left(\frac{c^2}{2\mu} + \frac{\mu}{\lambda} + \frac{1}{2}\right) x_{2i}^* - \frac{c^2}{2\mu} x_{2i+1}^* - \frac{1}{2} y_{2i}^* = 0, \quad \text{for } 0 \le i \le T-1 \tag{16}$$

$$\left(\frac{c^2}{2\mu} + \frac{\mu}{\lambda} + \frac{1}{2}\right) z_{2i-1}^* - \frac{c^2}{2\mu} z_{2i}^* - \frac{1}{2} y_{2i-1}^* = 0, \quad \text{for } 1 \le i \le T-1 \tag{17}$$

$$\left(\frac{c^2}{2\mu} + \frac{\mu}{\lambda} + \frac{1}{2}\right) z_{2i}^* - \frac{c^2}{2\mu} z_{2i-1}^* - \frac{1}{2} y_{2i}^* = 0, \quad \text{for } 1 \le i \le T-1 \tag{18}$$

$$\left(1 + \frac{2\mu}{\lambda}\right) y_i^* - x_i^* = 0, \quad \text{for } 1 \le i \le 2T-1 \tag{19}$$

$$\left(1 + \frac{2\mu}{\lambda}\right) y_i^* - z_i^* = 0, \quad \text{for } 1 \le i \le 2T-1 \tag{20}$$

Combining (17) and (18), we get for all $1 \le i \le T$

$$\begin{pmatrix} \frac{c^2}{2\mu} & 0 \\ -\frac{c^2}{2\mu} - \frac{1}{2} - \frac{\mu}{\lambda} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} z_{2i}^\star \\ y_{2i}^\star \end{pmatrix} = \begin{pmatrix} \frac{c^2}{2\mu} + \frac{1}{2} + \frac{\mu}{\lambda} & -\frac{1}{2} \\ -\frac{c^2}{2\mu} & 0 \end{pmatrix} \begin{pmatrix} z_{2i-1}^\star \\ y_{2i-1}^\star \end{pmatrix} \tag{21}$$

Rewriting this equation, we get

$$\begin{pmatrix} z_{2i}^\star \\ y_{2i}^\star \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\frac{c^2}{2\mu}} & \frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} \\ -\frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} & \frac{2\left(\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\frac{c^2}{2\mu}} - 2\frac{c^2}{2\mu} \end{pmatrix} \begin{pmatrix} y_{2i-1}^\star \\ z_{2i-1}^\star \end{pmatrix}.$$

Similarly, using (15), (16) we get for all $1 \leq i \leq T$

$$\begin{pmatrix} x^\star_{2i+1} \\ y^\star_{2i+1} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\frac{c^2}{2\mu}} & \frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} \\ -\frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} & \frac{2\left(\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\frac{c^2}{2\mu}} - 2\frac{c^2}{2\mu} \end{pmatrix} \begin{pmatrix} y^\star_{2i} \\ x^\star_{2i} \end{pmatrix}.$$

Using (19) and (20), we obtain

$$\begin{pmatrix} x^\star_{2i+1} \\ z^\star_{2i+1} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\frac{c^2}{2\mu}\left(1+\frac{2\mu}{\lambda}\right)} & \frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} \\ -\frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} & \left(1+\frac{2\mu}{\lambda}\right)\left(\frac{2\left(\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\frac{c^2}{2\mu}} - 2\frac{c^2}{2\mu}\right) \end{pmatrix} \begin{pmatrix} z^\star_{2i} \\ x^\star_{2i} \end{pmatrix}.$$

$$\begin{pmatrix} z^\star_{2i} \\ x^\star_{2i} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\frac{c^2}{2\mu}\left(1+\frac{2\mu}{\lambda}\right)} & \frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} \\ -\frac{\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}}{\frac{c^2}{2\mu}} & \left(1+\frac{2\mu}{\lambda}\right)\left(\frac{2\left(\frac{c^2}{2\mu}+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\frac{c^2}{2\mu}} - 2\frac{c^2}{2\mu}\right) \end{pmatrix} \begin{pmatrix} x^\star_{2i-1} \\ z^\star_{2i-1} \end{pmatrix}.$$

For simplicity, we introduce the following replacement $\bar{c} = \frac{c^2}{2\mu}$. Then we have

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{2\bar{c}\left(1+\frac{2\mu}{\lambda}\right)} & \frac{\bar{c}+\frac{\mu}{\lambda}+\frac{1}{2}}{\bar{c}} \\ -\frac{\bar{c}+\frac{\mu}{\lambda}+\frac{1}{2}}{\bar{c}} & \left(1+\frac{2\mu}{\lambda}\right)\left(\frac{2\left(\bar{c}+\frac{\mu}{\lambda}+\frac{1}{2}\right)^2}{\bar{c}} - 2\bar{c}\right) \end{pmatrix}. \tag{22}$$

The following lemma would be very difficult to prove without using Mathematica due to very cumbersome expressions. The following statement shows a recursive relationship between coordinates in the following sense $w_i = \gamma^{i-1} w_1$, where γ is the eigenvalue of the matrix $\mathbf{Q}$.

Lemma 2 (see [9]) *Choose* $c = \begin{cases} \delta \frac{\mu^{5/2}}{(\lambda\lambda_{\max}(\mathbf{W}))^2} & \text{if } L \geq \lambda\lambda_{\max}(\mathbf{W}) + \mu \\ \delta \frac{\mu^{5/2}}{(\lambda\lambda_{\max}(\mathbf{W}))^2} & \text{if } L < \lambda\lambda_{\max}(\mathbf{W}) + \mu \end{cases}, \delta \geq 1, (\bar{c} = \frac{c^2}{2\mu})$ *and*

$$b = \frac{1+\frac{2\mu}{\lambda}}{2}\left(\frac{-(1+2\frac{\mu}{\lambda})\nu}{2(1+2\bar{c}+2\frac{\mu}{\lambda})} + \frac{-\alpha+2\sqrt{\alpha^2-4\beta}}{4(1+2\frac{\mu}{\lambda})(1+2\bar{c}+2\frac{\mu}{\lambda})}\right) - \frac{1}{2} - \frac{\mu}{\lambda} \tag{23}$$

Then, we have $b \geq 0$ *and*

$$w_i = \gamma^{i-1} w_1 \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{for } i = 1, 2, \ldots, d,$$

where

$$\gamma = \frac{\alpha}{8\bar{c}(1+2\frac{\mu}{\lambda})} - \frac{\sqrt{(\alpha^2-4\beta)}}{8\bar{c}(1+2\frac{\mu}{\lambda})} \geq 1 - 10\sqrt{\frac{1}{\delta}}, \tag{24}$$

$$\begin{aligned} \alpha = & -1 + 4\bar{c} - 4\bar{c}^2 + 8\frac{\mu}{\lambda} + 24\bar{c}x - 16\bar{c}^2\frac{\mu}{\lambda} + 24\left(\frac{\mu}{\lambda}\right)^2 + 48\bar{c}\left(\frac{\mu}{\lambda}\right)^2 \\ & -16\bar{c}^2\left(\frac{\mu}{\lambda}\right)^2 + 32\left(\frac{\mu}{\lambda}\right)^3 + 32\bar{c}\left(\frac{\mu}{\lambda}\right)^3 + 16\left(\frac{\mu}{\lambda}\right)^4 \end{aligned}$$

$$\begin{aligned} \beta = & \ 2 + 8 + 24\bar{c}^2 + 16\left(\frac{\mu}{\lambda}\right) + 48c\left(\frac{\mu}{\lambda}\right) + 96\bar{c}^2\left(\frac{\mu}{\lambda}\right) + 48\left(\frac{\mu}{\lambda}\right)^2 \\ & +96\bar{c}\left(\frac{\mu}{\lambda}\right)^2 + 96\bar{c}^2\left(\frac{\mu}{\lambda}\right)^2 + 64\left(\frac{\mu}{\lambda}\right)^3 + 64\bar{c}\left(\frac{\mu}{\lambda}\right)^3 + 32\left(\frac{\mu}{\lambda}\right)^4 \end{aligned}$$

and

$$\nu = -1 - 4\bar{c} + 4\bar{c}^2 - 4\left(\frac{\mu}{\lambda}\right) - 8\bar{c}\left(\frac{\mu}{\lambda}\right) - 4\left(\frac{\mu}{\lambda}\right)^2$$

Proof. First, using the Mathematica software package, we calculate the minimum eigenvalue γ of the matrix $\mathbf{W}$ and prove that it satisfies the inequality. For a detailed study of the proof, see the file or screenshots

```
Eigenvalues[{{-1 / (2 * (c * (1 + 2 * x))), (c + x + 1 / 2) / c},
  {-(c + x + 1 / 2) / c, (1 + 2 * x) * (((c + x + 1 / 2) ^ 2) / c - 2 * c)}}]
```

$$\left\{\frac{1}{8\,c\,(1+2\,x)}\left(-1+4\,c-4\,c^2+8\,x+24\,c\,x-16\,c^2\,x+24\,x^2+48\,c\,x^2-16\,c^2\,x^2+32\,x^3+32\,c\,x^3+16\,x^4-\sqrt{\left(\left(1-4\,c+4\,c^2-8\,x-24\,c\,x+16\,c^2\,x-24\,x^2-48\,c\,x^2+16\,c^2\,x^2-32\,x^3-32\,c\,x^3-16\,x^4\right)^2-4\left(2+8\,c+24\,c^2+16\,x+48\,c\,x+96\,c^2\,x+48\,x^2+96\,c\,x^2+96\,c^2\,x^2+64\,x^3+64\,c\,x^3+32\,x^4\right)\right)}\right),\ \frac{1}{8\,c\,(1+2\,x)}\left(-1+4\,c-4\,c^2+8\,x+24\,c\,x-16\,c^2\,x+24\,x^2+48\,c\,x^2-16\,c^2\,x^2+32\,x^3+32\,c\,x^3+16\,x^4+\sqrt{\left(\left(1-4\,c+4\,c^2-8\,x-24\,c\,x+16\,c^2\,x-24\,x^2-48\,c\,x^2+16\,c^2\,x^2-32\,x^3-32\,c\,x^3-16\,x^4\right)^2-4\left(2+8\,c+24\,c^2+16\,x+48\,c\,x+96\,c^2\,x+48\,x^2+96\,c\,x^2+96\,c^2\,x^2+64\,x^3+64\,c\,x^3+32\,x^4\right)\right)}\right)\right\}$$

```
In[25]:= FindInstance[(1 / (8 c (1 + 2 x))) (-1 + 4 c - 4 c^2 + 8 x + 24 c x - 16 c^2 x + 24 x^2 + 48 c x^2 - 16 c^2 x^2 + 32 x^3 + 32 c x^3 + 16 x^4 - ((1 -
          4 c + 4 c^2 - 8 x - 24 c x + 16 c^2 x - 24 x^2 - 48 c x^2 + 16 c^2 x^2 - 32 x^3 - 32 c x^3 - 16 x^4) ^ 2 - 4 (2 + 8 c + 24 c^2 + 16 x + 48 c x +
          96 c^2 x + 48 x^2 + 96 c x^2 + 96 c^2 x^2 + 64 x^3 + 64 c x^3 + 32 x^4)) ^ (0.5)) < 1 - 10 * Sqrt[1 / d] && d ≥ 1 && c ≤ d * x && d * x ≤
      1 && x > 0 && c > 0, {c, x, d}]

Out[25]= {}
```

Using Mathematica, we find the eigenvector v of the matrix Q corresponding to the eigenvalue γ

$$v = \begin{pmatrix} \frac{-(1+2\frac{\mu}{\lambda})\nu}{2(1+2c+2\frac{\mu}{\lambda})} + \frac{-\alpha+2\sqrt{\alpha^2-4\beta}}{4(1+2\frac{\mu}{\lambda})(1+2c+2\frac{\mu}{\lambda})} \\ 1 \end{pmatrix} \tag{25}$$

Now, using Mathematica, we prove that $b \geq 0$

```
In[31]:= FindInstance[(1 / (4 (1 + 2 c + 2 x))) (-(1 + 2 x) ^ 2 (-1 - 4 c + 4 c^2 - 4 x - 8 c * x - 4 x^2) + 0.5 (1 - 4 c + 4 c^2 - 8 x - 24 c x + 16 c^2 x - 24 x^2 - 48
        c x^2 + 16 c^2 x^2 - 32 x^3 - 32 c * x^3 - 16 x^4 + ((1 - 4 c + 4 c^2 - 8 x - 24 c x + 16 c^2 x - 24 x^2 - 48 c x^2 + 16 c^2 x^2 - 32 x^3 - 32 c
          x^3 - 16 x^4) ^ 2 - 4 (2 + 8 c + 24 c^2 + 16 x + 48 c x + 96 c^2 x + 48 x^2 + 96 c x^2 + 96 c^2 x^2 + 64 x^3 + 64 c x^3 + 32 x^4)) ^ (0.5))) - 1 /
      2 - x < 0 && d ≥ 1 && c ≤ d * x && d * x ≤ 1 && x > 0 && c > 0, {c, x, d}]

Out[31]= {}
```

It is easy to see that when choosing the parameter b according to (23), the vector w_i is proportional to the eigenvector v of the matrix Q.

□

Now we are ready to complete the proof of the theorem. Let $\mathbf{y}^0 = 0 \in \mathbb{R}^{Mn}$. It is worth noting that after N iterations of the algorithm $\mathcal{A}$ satisfying the Assumption 3, $\mathbf{y}^N$ has no more than S/Δ nonzero coordinates, where S is the number of rounds of communication (consensus), Δ is the diameter of the network $\mathcal{G}$. In our case, $\Delta \leq \sqrt{\chi}$. Then, using the equations (19), (20), we get

$$\begin{aligned} \frac{\|\mathbf{y}^N - \mathbf{y}^*\|^2}{\|\mathbf{y}^0 - \mathbf{y}^*\|} &\geq \frac{1}{2}\frac{\sum_{i=S/\Delta+1}^n \|w_j\|^2 + (y_j^*)^2}{\sum_{i=1}^n \|w_j\|^2 + (y_j^*)^2} = \frac{1}{2}\frac{\sum_{i=S/\Delta+1}^n \|w_j\|^2 + \left(\frac{\lambda}{\mu+\lambda}\right)^2 (y_j^*)^2}{\sum_{i=1}^n \|w_j\|^2 + \left(\frac{\lambda}{\mu+\lambda}\right)^2 (z_j^*)^2} \\ &= \frac{1}{2}\frac{\sum_{i=S/\Delta+1}^n \|R\tilde{w}_j\|^2}{\sum_{i=1}^n \|R\tilde{w}_j\|^2}, \end{aligned}$$

where $R = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1+\left(\frac{\lambda}{\mu+\lambda}\right)^2} \end{pmatrix}$ and $\tilde{w}_j = (x_j^*, z_j^*)^\top$ ($\tilde{w}_j = \gamma^{j-1}\tilde{w}_1$). It is not difficult to see that the matrix is positive definite and symmetric. Therefore, you can use it to define a new norm $\|w\|_R = \sqrt{\langle x, Rx\rangle}$ and use the properties of the norm. Then for large enough n we have

$$\begin{aligned}
\frac{\|\mathbf{y}^N - \mathbf{x}^*\|^2}{\|\mathbf{y}^0 - \mathbf{x}^*\|} &\geq \frac{1}{2}\frac{\sum_{i=s/\Delta+1}^n \|R\tilde{w}_j\|^2}{\sum_{i=1}^n \|R\tilde{w}_j\|^2} = \frac{1}{2}\frac{\sum_{i=s/\Delta+1}^n \gamma^{j-1}\|R\tilde{w}_1\|^2}{\sum_{i=1}^n \gamma^{j-1}\|R\tilde{w}_1\|^2} \\
&= \frac{1}{2}\frac{\sum_{i=s/\Delta+1}^n \gamma^{j-1}}{\sum_{i=1}^n \gamma^{j-1}} = \frac{1}{2}\frac{\gamma^{s/\Delta}\sum_{i=0}^{n-s/\Delta-1}\gamma^j}{\sum_{i=0}^{n-1}\gamma^j} \\
&= \frac{1}{2}\gamma^{s/\Delta}\frac{1-\gamma^{n-s/\Delta}}{1-\gamma^n} \geq \frac{1}{4}\left(1 - 10\sqrt{\frac{1}{\delta}}\right)^{s/\Delta}.
\end{aligned}$$

It is worth noting that according to our choice of the structure of communication network $\mathcal{G}$, we have that $\Delta \leq \sqrt{\chi}$, then we obtain

$$\begin{aligned}
\|\mathbf{y}^N - \mathbf{y}^*\|^2 &\geq \frac{1}{4}\left(1 - 10\sqrt{\frac{1}{\delta}}\right)^{s/\Delta}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 \\
&\geq \frac{1}{4}\left(1 - 10\sqrt{\frac{1}{\delta}}\right)^{s/\sqrt{\chi}}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 \\
&\geq \frac{1}{4}\left(1 - 10\sqrt{\frac{1}{\delta\chi}}\right)^{s}\|\mathbf{y}^0 - \mathbf{y}^*\|^2.
\end{aligned}$$

Now we have to consider two cases: one when $L \geq \lambda\lambda_{\max}(\mathbf{W}) + \mu$, the other when $L < \lambda\lambda_{\max}(\mathbf{W}) + \mu$. Moreover, we must select the parameter δ in such a way that $\delta \geq 1$ but also $\bar{c} \leq 1$.

- $L \geq \lambda\lambda_{\max}(\mathbf{W}) + \mu$: if we take $\delta = \frac{\lambda\lambda_{\max}(\mathbf{W})\lambda\lambda^+_{\min}(\mathbf{W})}{\mu^2}$, then we have $\bar{c} \leq 1$ and

$$\|\mathbf{y}^N - \mathbf{y}^*\|^2 \geq \frac{1}{4}\left(1 - 10\frac{\mu}{\lambda\lambda_{\max}(\mathbf{W})}\right)^{s}\|\mathbf{y}^0 - \mathbf{y}^*\|^2$$

- $L < \lambda\lambda_{\max}(\mathbf{W}) + \mu$: if we take $\delta = \left(\frac{L-\mu}{\mu}\right)^2$, then we have $\bar{c} \leq 1$ and

$$\|\mathbf{y}^N - \mathbf{y}^*\|^2 \geq \frac{1}{4}\left(1 - 10\frac{\mu}{(L-\mu)\sqrt{\chi}}\right)^{s}\|\mathbf{y}^0 - \mathbf{y}^*\|^2$$

Summing up the results obtained above, we can conclude

$$\|\mathbf{y}^N - \mathbf{y}^*\|^2 \geq \left(1 - 10\max\left\{\frac{\mu}{\lambda\lambda_{\max}(\mathbf{W})}, \frac{\mu}{(L-\mu)\sqrt{\chi}}\right\}\right)^{s}\frac{\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{4}.$$

C.2 Optimal algorithms

C.2.1 Notation

We will use matrix notation in our proofs, as in the main body of the paper. It is easy to check that the matrix notation in this case will not differ greatly from the standard vector notation; it is enough just to move from the Euclidean norm to the Frobenius norm and from the scalar product to the trace:

$$\|X\|_F^2 = \|x_1\|^2 + \ldots + \|x_M\|^2, \quad \text{tr}\left[(X)^T(Y)\right] = \langle x_1, y_1\rangle + \ldots + \langle x_M, y_M\rangle,$$

where $x_1, \ldots, x_M$ and $y_1, \ldots, y_M$ – vectors from which matrices X and Y are composed according to Section 2.1. Additionally, we use matrix of the solution $X = [x^*, \ldots, x^*]^T$ and $Y = [y^*, \ldots, y^*]^T$.

It is also easy to make sure that the functions f and φ are

• smooth: f is L - smooth, φ is $\lambda\lambda_{\max}(W)$ - smooth;

• smooth: f is μ - strongly-convex-strongly-concave, φ is 0 - strongly-convex-strongly-concave.

C.2.2 Proof of Theorems 2 and 3

Strongly-convex-strongly-concave case

Lemma 3 (for Theorems 2 and 3) *For Algorithm 1 it holds:*

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\qquad\leq (1-\gamma\mu)\left\|X^k-X^*\right\|_F^2+(1-\gamma\mu)\left\|Y^k-Y^*\right\|_F^2 \\
&\qquad\quad+\left(2+\frac{4\gamma\lambda\lambda_{\max}(W)}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad\quad-(1-3\gamma\mu-4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad\quad+\left(2+\frac{4\gamma\lambda\lambda_{\max}(W)}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2 \\
&\qquad\quad-(1-3\gamma\mu-4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

Proof: Let us use additional notation $W_x^k=U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)$ and $W_y^k=U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)$ for short. Then by non-expansiveness of the Euclidean projection, we get

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\qquad=\left\|\text{proj}_{\mathcal{X}}\left[W_x^k\right]-\text{proj}_{\mathcal{X}}\left[X^*\right]\right\|_F^2+\left\|\text{proj}_{\mathcal{Y}}\left[W_y^k\right]-\text{proj}_{\mathcal{Y}}\left[Y^*\right]\right\|_F^2 \\
&\qquad\leq\left\|W_x^k-X^*\right\|_F^2+\left\|W_y^k-Y^*\right\|_F^2 \\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(W_x^k-X^k\right)^T\left(X^k-X^*\right)\right]+\left\|W_x^k-X^k\right\|_F^2 \\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(W_y^k-Y^k\right)^T\left(Y^k-Y^*\right)\right]+\left\|W_y^k-Y^k\right\|_F^2 \\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(W_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad\quad+2\,\text{tr}\left[\left(W_x^k-X^k\right)^T\left(X^k-\hat{U}_x^k\right)\right]+\left\|W_x^k-X^k\right\|_F^2 \\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(W_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad\quad+2\,\text{tr}\left[\left(W_y^k-Y^k\right)^T\left(Y^k-\hat{U}_y^k\right)\right]+\left\|W_y^k-Y^k\right\|_F^2 \\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(W_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|W_x^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad\quad-\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(W_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad\quad+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2 \\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad\quad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad\quad+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_x^k + \gamma\cdot\lambda W X^k - X^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_y^k + \gamma\cdot\lambda W Y^k - Y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With expressions for V_x^k and V_y^k in Algorithm 1, we have

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \le \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_y^k - V_y^k\right)^T\left(\hat{U}_y^k - X^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad = \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\hat{U}_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\hat{U}_y^k - V_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

According to the optimal condition for $\hat{U}_x^k$ and $\hat{U}_y^k$: for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $X = [x, \ldots, x]^T$, $Y = [y, \ldots, y]^T$

$$
\begin{aligned}
&\mathrm{tr}\left[\left(\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k) + \hat{U}_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X\right)\right] \\
&+\ \mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k) + \hat{U}_y^k - V_y^k\right)^T\left(\hat{U}_y^k - Y\right)\right] \le 0,
\end{aligned}
$$

we get

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \le \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W(U_x^k - \hat{U}_x^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W(U_y^k - \hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With property of the solution X^*, Y^*: for all $X \in \mathcal{X}, Y \in \mathcal{Y}$

$$
\begin{aligned}
&\mathrm{tr}\left[\left(\nabla_X f(X^*;Y^*) + \lambda W X^*\right)^T(X^* - X)\right] \\
&+\ \mathrm{tr}\left[\left(-\left(\nabla_Y f(X^*;Y^*) - \lambda W Y^*\right)\right)^T(Y^* - Y)\right] \le 0.
\end{aligned}
$$

And then μ-strong convexity - strong concavity of f, we obtain

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \le \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_X f(\hat{U}_x^k;\hat{U}_y^k) + \lambda W\hat{U}_x^k - \nabla_X f(X^*;Y^*) - \lambda W X^*\right)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\lambda W(U_x^k - \hat{U}_x^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad - \left\|X^k - \hat{U}_x^k\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(-\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k) + \lambda W\hat{U}_y^k + \nabla_Y f(X^*;Y^*) - \lambda W Y^*\right)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W(U_y^k - \hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad - \left\|Y^k - \hat{U}_y^k\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad \le \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - 2\gamma\mu\left\|\hat{U}_x^k - X^*\right\|_F^2 - 2\gamma\mu\left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\lambda W(U_x^k - \hat{U}_x^k) - (U_x^k - \hat{U}_x^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\lambda W(U_y^k - \hat{U}_y^k) - (U_y^k - \hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

By Young's inequality, we have

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-2\gamma\mu\left\|\hat{U}_x^k-X^*\right\|_F^2-2\gamma\mu\left\|\hat{U}_y^k-Y^*\right\|_F^2\\
&\qquad+\frac{2}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_x^k-\hat{U}_x^k)-(U_x^k-\hat{U}_x^k)\right\|_F^2+\frac{\gamma\mu}{2}\left\|U_x^k-X^*\right\|_F^2\\
&\qquad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+\frac{2}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_y^k-\hat{U}_y^k)-(U_y^k-\hat{U}_y^k)\right\|_F^2+\frac{\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2\\
&\qquad+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_x^k-X^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2\\
&\qquad+\frac{4}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_x^k-\hat{U}_x^k)\right\|_F^2+\frac{4}{\gamma\mu}\left\|U_x^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+\left\|U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+\frac{4}{\gamma\mu}\left\|U_y^k-\hat{U}_y^k\right\|_F^2+\frac{4}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_y^k-\hat{U}_y^k)\right\|_F^2\\
&\qquad+\left\|U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_x^k-X^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2\\
&\qquad+\frac{4}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_x^k-\hat{U}_x^k)\right\|_F^2+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+2\left\|\gamma\cdot\lambda\left(WX^k-WU_x^k\right)\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2+\frac{4}{\gamma\mu}\left\|\gamma\cdot\lambda W(U_y^k-\hat{U}_y^k)\right\|_F^2\\
&\qquad+2\left\|\gamma\cdot\lambda\left(WY^k-WU_y^k\right)\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

Then we use $\lambda\lambda_{\max}(W)$-smoothness of φ

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_x^k-X^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2\\
&\qquad+\frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu}\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+2\gamma^2\lambda^2\lambda_{\max}^2(W)\left\|X^k-U_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2+\frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu}\left\|U_y^k-\hat{U}_y^k\right\|_F^2\\
&\qquad+2\gamma^2\lambda^2\lambda_{\max}^2(W)\left\|Y^k-U_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_x^k - X^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_x^k - \hat{U}_x^k\right\|_F^2 \\
&\quad - (1 - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_y^k - \hat{U}_y^k\right\|_F^2 \\
&\quad - (1 - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

By inequality $\|A+B\|_F^2 \geq \frac{2}{3}\|A\|_F^2 - 2\|B\|_F^2$, we have

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad \leq (1-\gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1-\gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&\qquad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_x^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|U_y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

□

Lemma 4 (for Theorem 2) *Assume that for problem (3) we use extragradient method with starting points X^k, Y^k and number of iterations:*

$$
T = \mathcal{O}\left((1+\gamma L)\log\frac{1}{\tilde{\delta}}\right). \tag{26}
$$

Then for an output u^k it holds that

$$
\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2 \leq \tilde{\delta}\left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).
$$

Proof: The proof follows from the convergence estimates for the extragradient method and the fact that Problem (3) is 1-strongly-convex-strongly-concave, as well as $(1+\gamma L)$-smooth.

□

Remark. As mentioned earlier, problem (3) is divided into M local problems, which are solved on each of the machines separately.

Note that Algorithm 1 involves the absolute precision δ, $\tilde{\delta}$ is the relative precision - a more convenient value for practice.

Theorem 6 (Theorem 2) *Assume that problem (3) be solved by extragradient with precision $\tilde{\delta}$:*

$$
\tilde{\delta} = \frac{1}{2\left(2 + \frac{4\gamma\lambda\lambda_{\max}(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda\lambda_{\max}(W)\right)} \tag{27}
$$

and the number of iterations T (see (26)). Additionally, stepsize γ is

$$
\gamma = \min\left\{\frac{1}{12\mu}; \frac{1}{16\lambda\lambda_{\max}(W)}\right\}. \tag{28}
$$

Then Algorithm 1 converges linearly to the solution z^* and it holds that $\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \leq \varepsilon$ after

$$K = \mathcal{O}\left(\frac{1}{\gamma\mu}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \quad \textit{iterations.} \tag{29}$$

Proof: Combining results from Lemma 3 and 4 gives

$$\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \leq (1-\gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1-\gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&\qquad\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\tilde{\delta}\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\tilde{\delta}\left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}$$

With the choice e from (27), we obtain

$$\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \leq (1-\gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1-\gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&\qquad\quad - \left(\frac{1}{2} - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad - \left(\frac{1}{2} - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}$$

The proof is completed by choosing γ from (28).

□

Remark. (29) also corresponds to calls of φ gradients which in turn corresponds to the number of communication rounds. Substitution of (28) in (29) gives that

$$K = \mathcal{O}\left(\left(1 + \frac{\lambda\lambda_{\max}(W)}{\mu}\right)\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right).$$

It is also easy to estimate the total number of local iterations on each node:

$$\begin{aligned}
K \times T &= \mathcal{O}\left(\frac{1}{\gamma\mu}(1+\gamma L)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \\
&= \mathcal{O}\left(\left(\frac{1}{\gamma\mu} + \frac{L}{\mu}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \\
&= \mathcal{O}\left(\left(1 + \frac{\lambda\lambda_{\max}(W)}{\mu} + \frac{L}{\mu}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \\
&= \mathcal{O}\left(\left(1 + \frac{L}{\mu}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right).
\end{aligned}$$

The last follows from the fact that we consider the case of small λ.

Lemma 5 (for Theorem 3) *Assume that for solving the problem (3) we use Randomized Extra Step Method from [1] with starting points X^k, Y^k and number of iterations:*

$$T = \mathcal{O}\left(\left(r + \frac{\sqrt{r}\left(L + \frac{1}{\gamma}\right)}{\mu + \frac{1}{\gamma}}\right)\log\frac{1}{\tilde{\delta}}\right). \tag{30}$$

Then for an output u^k it holds that

$$\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2 \leq \tilde{\delta}\left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).$$

Proof: The proof follows from the convergence estimates for the Algorithm 1 from [1] and the fact that Problem (3) is $\left(\mu + \frac{1}{\gamma}\right)$-strongly-convex-strongly-concave, as well as $\left(\frac{1}{\gamma} + L\right)$-smooth.

□

Remark. As mentioned earlier, the problem (3) is divided into M local problems, which are solved on each of the machines separately.

Note that Algorithm 1 involves the absolute precision δ, $\tilde{\delta}$ is the relative precision - a more convenient value for practice.

Theorem 7 (Theorem 3) *Assume that the problem (3) is solved by Randomized Extra Step Method with variance reduction method from [1] with precision $\tilde{\delta}$:*

$$\tilde{\delta} = \frac{1}{2\left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)} \tag{31}$$

and the number of iterations T (see (30)). Additionally, stepsize γ is

$$\gamma = \min\left\{\frac{1}{12\mu}; \frac{1}{16\lambda\lambda_{\max}(W)}\right\}. \tag{32}$$

Then Algorithm 1 converges linearly to the solution z^* and it holds that $\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \leq \varepsilon$ after

$$K = \mathcal{O}\left(\frac{1}{\gamma\mu}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \quad \textit{iterations}. \tag{33}$$

Proof: Combining results from Lemma 3 and 5 gives

$$\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \leq (1 - \gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1 - \gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&\qquad\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\tilde{\delta}\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left(2 + \frac{4\gamma\lambda^2\lambda_{\max}^2(W)}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\tilde{\delta}\left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad - (1 - 3\gamma\mu - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}$$

With the choice $\tilde{\delta}$ from (31), we obtain

$$\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\qquad\leq (1-\gamma\mu)\left\|X^k-X^*\right\|_F^2+(1-\gamma\mu)\left\|Y^k-Y^*\right\|_F^2 \\
&\qquad\quad-\left(\frac{1}{2}-3\gamma\mu-4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad\quad-\left(\frac{1}{2}-3\gamma\mu-4\gamma^2\lambda^2\lambda_{\max}^2(W)\right)\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}$$

The proof is completed by choosing γ from (32).

□

Remark. (33) also corresponds to communication rounds. By plugging (32) into (33) gives that

$$K=\mathcal{O}\left(\left(1+\frac{\lambda\lambda_{\max}(W)}{\mu}\right)\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right).$$

It is also easy to estimate the total number of local computation on each node:

$$\begin{aligned}
K\times T&=\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\frac{\sqrt{r}\left(L+\frac{1}{\gamma}\right)}{\mu+\frac{1}{\gamma}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\sqrt{r}+\frac{\sqrt{r}L}{\mu+\frac{1}{\gamma}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&=\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\frac{\sqrt{r}L}{\mu+\frac{1}{\gamma}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\frac{\sqrt{r}L}{\max\left\{\mu,\frac{1}{\gamma}\right\}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&=\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\frac{\sqrt{r}L}{\max\left\{\mu,\lambda\lambda_{\max}(W)\right\}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq\mathcal{O}\left(\frac{1}{\gamma\mu}\left(r+\frac{\sqrt{r}L}{\lambda\lambda_{\max}(W)}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&=\mathcal{O}\left(\left(1+\frac{\lambda\lambda_{\max}(W)}{\mu}\right)\left(r+\frac{\sqrt{r}L}{\lambda\lambda_{\max}(W)}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&=\mathcal{O}\left(\left(r\frac{\lambda\lambda_{\max}(W)}{\mu}+\frac{\sqrt{r}L}{\mu}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right).
\end{aligned}$$

If $\sqrt{r}\lambda\lambda_{\max}(W)=\mathcal{O}(L)$, the total expected number of local gradients becomes optimal.

Convex-Concave case

Lemma 6 (for Theorems 2 and 3) *For Algorithm 1 it holds:*

$$
\begin{aligned}
&2\gamma\ tr\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X^*\right)\right]\\
&\quad+2\gamma\ tr\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y^*\right)\right]\\
&\qquad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X^*\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad\quad+4(1+\gamma L+\gamma\lambda\lambda_{\max}(W))\sqrt{M}(\Omega+G)\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&\qquad\quad+(2+4\gamma^2\lambda^2\lambda_{\max}^2(W))\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right)\\
&\qquad\quad-(1-4\gamma^2\lambda^2\lambda_{\max}^2(W))\left(\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-\hat{U}_y^k\right\|_F^2\right). \qquad (34)
\end{aligned}
$$

Proof: Let us use additional notation $W_x^k=U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)$ and $W_y^k=U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)$ for short. Then by non-expansiveness of the Euclidean projection, we get

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad=\left\|\text{proj}_{\mathcal{X}}\left[W_x^k\right]-\text{proj}_{\mathcal{X}}\left[X^*\right]\right\|_F^2+\left\|\text{proj}_{\mathcal{Y}}\left[W_y^k\right]-\text{proj}_{\mathcal{Y}}\left[Y^*\right]\right\|_F^2\\
&\qquad\le\left\|W_x^k-X^*\right\|_F^2+\left\|W_y^k-Y^*\right\|_F^2\\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\ \text{tr}\left[\left(W_x^k-X^k\right)^T\left(X^k-X^*\right)\right]+\left\|W_x^k-X^k\right\|_F^2\\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\ \text{tr}\left[\left(W_y^k-Y^k\right)^T\left(Y^k-Y^*\right)\right]+\left\|W_y^k-Y^k\right\|_F^2\\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\ \text{tr}\left[\left(W_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\
&\qquad\quad+2\ \text{tr}\left[\left(W_x^k-X^k\right)^T\left(X^k-\hat{U}_x^k\right)\right]+\left\|W_x^k-X^k\right\|_F^2\\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\ \text{tr}\left[\left(W_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\
&\qquad\quad+2\ \text{tr}\left[\left(W_y^k-Y^k\right)^T\left(Y^k-\hat{U}_y^k\right)\right]+\left\|W_y^k-Y^k\right\|_F^2\\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\ \text{tr}\left[\left(W_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\
&\qquad\quad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\ \text{tr}\left[\left(W_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\
&\qquad\quad+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\
&\qquad=\left\|X^k-X^*\right\|_F^2+2\ \text{tr}\left[\left(U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\
&\qquad\quad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad\quad+\left\|Y^k-Y^*\right\|_F^2+2\ \text{tr}\left[\left(U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\
&\qquad\quad+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_x^k + \gamma\cdot\lambda W X^k - X^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_y^k + \gamma\cdot\lambda W Y^k - Y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With expressions for V_x^k and V_y^k in Algorithm 1, we have

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad\le \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(U_y^k - V_y^k\right)^T\left(\hat{U}_y^k - X^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\quad= \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\hat{U}_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\hat{U}_y^k - V_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

According to the optimal condition for $\hat{U}_x^k$ and $\hat{U}_y^k$: for all $X \in \mathcal{X}, Y \in \mathcal{Y}$

$$
\begin{aligned}
&\mathrm{tr}\left[\left(\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k) + \hat{U}_x^k - V_x^k\right)^T\left(\hat{U}_x^k - X\right)\right] \\
&\quad + \mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k) + \hat{U}_y^k - V_y^k\right)^T\left(\hat{U}_y^k - Y\right)\right] \le 0,
\end{aligned}
$$

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad\le \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad + \left\|W_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad + \left\|W_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

Small rearrangement gives

$$
\begin{aligned}
&2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X^*\right)\right]\\
&\quad+2\gamma \operatorname{tr}\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y^*\right)\right]\\
&\qquad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X^*\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad\quad+2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k;U_y^k)-\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\
&\qquad\quad+2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k;U_y^k)\right)^T\left(U_x^k-\hat{U}_x^k\right)\right]\\
&\qquad\quad+2\gamma \operatorname{tr}\left[\left(\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)-\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\
&\qquad\quad+2\gamma \operatorname{tr}\left[\left(\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-U_y^k\right)\right]\\
&\qquad\quad+2 \operatorname{tr}\left[\left(U_x^k-\hat{U}_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]+2 \operatorname{tr}\left[\left(-\gamma\cdot\lambda W U_x^k\right)^T\left(\hat{U}_x^k-U_x^k\right)\right]\\
&\qquad\quad+2 \operatorname{tr}\left[\left(U_y^k-\hat{U}_y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+2 \operatorname{tr}\left[\left(-\gamma\cdot\lambda W U_y^k\right)^T\left(\hat{U}_y^k-U_y^k\right)\right]\\
&\qquad\quad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\
&\qquad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X^*\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad\quad+2\gamma\|\nabla_X f(U_x^k;U_y^k)-\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\|_F\cdot\|\hat{U}_x^k-X^*\|_F\\
&\qquad\quad+2\gamma\|\nabla_X f(U_x^k;U_y^k)\|\cdot\|U_x^k-\hat{U}_x^k\|_F\\
&\qquad\quad+2\gamma\|\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)-\nabla_Y f(U_x^k;U_y^k)\|_F\cdot\|\hat{U}_y^k-Y^*\|_F\\
&\qquad\quad+2\gamma\|\nabla_Y f(U_x^k;U_y^k)\|_F\cdot\|\hat{U}_y^k-U_y^k\|_F\\
&\qquad\quad+2\|U_x^k-\hat{U}_x^k\|_F\cdot\|\hat{U}_x^k-X^*\|_F+2\gamma\|\lambda W U_x^k\|_F\cdot\|\hat{U}_x^k-U_x^k\|_F\\
&\qquad\quad+2\|U_y^k-\hat{U}_y^k\|_F\cdot\|\hat{U}_y^k-Y^*\|_F+2\gamma\|\lambda W U_y^k\|_F\cdot\|\hat{U}_y^k-U_y^k\|_F\\
&\qquad\quad+\left\|W_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|W_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With definition of W_x^k and W_y^k we get

$$
\begin{aligned}
&2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X^*\right)\right]\\
&\quad+2\gamma \operatorname{tr}\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y^*\right)\right]\\
&\qquad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad\quad+2\gamma\|\nabla_X f(U_x^k;U_y^k)-\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\|_F\cdot\|\hat{U}_x^k-X^*\|_F\\
&\qquad\quad+2\gamma\|\nabla_X f(U_x^k;U_y^k)\|\cdot\|U_x^k-\hat{U}_x^k\|_F\\
&\qquad\quad+2\gamma\|\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)-\nabla_Y f(U_x^k;U_y^k)\|_F\cdot\|\hat{U}_y^k-Y^*\|_F\\
&\qquad\quad+2\gamma\|\nabla_Y f(U_x^k;U_y^k)\|_F\cdot\|\hat{U}_y^k-U_y^k\|_F\\
&\qquad\quad+2\|U_x^k-\hat{U}_x^k\|_F\cdot\|\hat{U}_x^k-X^*\|_F+2\gamma\|\lambda W U_x^k\|_F\cdot\|\hat{U}_x^k-U_x^k\|_F\\
&\qquad\quad+2\|U_y^k-\hat{U}_y^k\|_F\cdot\|\hat{U}_y^k-Y^*\|_F+2\gamma\|\lambda W U_y^k\|_F\cdot\|\hat{U}_y^k-U_y^k\|_F\\
&\qquad\quad+\left\|U_x^k+\gamma\cdot\lambda\left(WX^k-WU_x^k\right)-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\
&\qquad\quad+\left\|U_y^k+\gamma\cdot\lambda\left(WY^k-WU_y^k\right)-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad + 2\gamma\|\nabla_X f(U_x^k; U_y^k) - \nabla_X f(\hat{U}_x^k; \hat{U}_y^k)\|_F \cdot \|\hat{U}_x^k - X^*\|_F \\
&\quad + 2\gamma\|\nabla_X f(U_x^k; U_y^k)\| \cdot \|U_x^k - \hat{U}_x^k\|_F \\
&\quad + 2\gamma\|\nabla_Y f(\hat{U}_x^k; \hat{U}_y^k) - \nabla_Y f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_y^k - Y^*\|_F \\
&\quad + 2\gamma\|\nabla_Y f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_y^k - U_y^k\|_F \\
&\quad + 2\|U_x^k - \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\lambda W U_x^k\|_F \cdot \|\hat{U}_x^k - U_x^k\|_F \\
&\quad + 2\|U_y^k - \hat{U}_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\lambda W U_y^k\|_F \cdot \|\hat{U}_y^k - U_y^k\|_F \\
&\quad + 2\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + 2\gamma^2\left\|\lambda\left(WX^k - WU_x^k\right)\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + 2\left\|U_y^k - \hat{U}_y^k\right\|_F^2 + 2\gamma^2\left\|\lambda\left(WY^k - WU_y^k\right)\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad + 2\gamma\sqrt{M}\Omega\|\nabla_X f(U_x^k; U_y^k) - \nabla_X f(\hat{U}_x^k; \hat{U}_y^k)\|_F \\
&\quad + 2\gamma\sqrt{M}\Omega\|\nabla_Y f(\hat{U}_x^k; \hat{U}_y^k) - \nabla_Y f(U_x^k; U_y^k)\|_F \\
&\quad + 2(1 + \gamma L + \gamma\lambda\lambda_{\max}(W))\sqrt{M}(\Omega + G)\left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\quad + 2\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + 2\gamma^2\left\|\lambda\left(WX^k - WU_x^k\right)\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + 2\left\|U_y^k - \hat{U}_y^k\right\|_F^2 + 2\gamma^2\left\|\lambda\left(WY^k - WU_y^k\right)\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2
\end{aligned}
$$

Here we additionally used the diameter Ω of $\mathcal{Z}$ and assume point $\hat{x}, \hat{y}$, s.t. $\|\hat{x}\| \leq G, \|\hat{y}\| \leq G$ for some constant G.

Then we use smoothness of f, φ and obtain

$$
\begin{aligned}
&2\gamma \text{ tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T\left(U_x^k - X^*\right)\right] \\
&\quad + 2\gamma \text{ tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T\left(U_y^k - Y^*\right)\right] \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad\quad + 4(1 + \gamma L + \gamma\lambda\lambda_{\max}(W))\sqrt{M}(\Omega + G)\left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad\quad + (2 + 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right) \\
&\qquad\quad - (1 - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}
$$

□

Theorem 8 (Theorem 3) *Assume that problem (3) be solved by extragradient with precision δ:*

$$
\delta = \min\left\{\frac{\sqrt{\varepsilon}}{9\lambda\lambda_{\max}(W)}; \frac{\varepsilon}{24\left(\lambda\lambda_{\max}(W) + L\right)\sqrt{M}(\Omega + G)}\right\} \tag{35}
$$

and number of iterations T:

$$
T = \mathcal{O}\left((1 + \gamma L)\log\frac{\Omega^2}{\delta}\right). \tag{36}
$$

Additionally, let us choose stepsize γ as follows

$$
\gamma = \frac{1}{2\lambda\lambda_{\max}(W)}. \tag{37}
$$

It holds that $gap(X_{avg}^K, Y_{avg}^K) \leq \varepsilon$ *after*

$$K = \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{\varepsilon}\right) \quad \textit{iterations}, \tag{38}$$

where

$$gap(X, Y) := \max_{Y' \in \mathcal{Y}} f(X, Y') - \min_{X' \in \mathcal{X}} f(X', Y).$$

and $X_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_x^k$, $Y_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_y^k$.

Proof: Summing (34) over all k from 0 to K

$$\begin{aligned}
&2\gamma \sum_{k=0}^K \Big(\text{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&\quad + \text{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right] \Big) \\
&\qquad \leq \left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2 \\
&\qquad\quad + 4(1 + \gamma L + \gamma\lambda\lambda_{\max}(W))\sqrt{M}(\Omega + G)\sum_{k=0}^K \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad\quad + (2 + 4\gamma^2\lambda^2\lambda_{\max}^2(W))\sum_{k=0}^K \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right) \\
&\qquad\quad - (1 - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\sum_{k=0}^K \left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

With our choice of γ

$$\begin{aligned}
&\frac{1}{\lambda\lambda_{\max}(W)}\sum_{k=0}^K \Big(\text{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&\quad + \text{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right] \Big) \\
&\qquad \leq \left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2 \\
&\qquad\quad + 4\left(2 + \frac{L}{\lambda\lambda_{\max}(W)}\right)\sqrt{M}(\Omega + G)\sum_{k=0}^K \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad\quad + 3\sum_{k=0}^K \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

Then, by $X_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_x^k$ and $Y_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_y^k$, Jensen's inequality and convexity-concavity of F:

$$\begin{aligned}
\text{gap}(X_{avg}^K, Y_{avg}^K) &\leq \max_{Y' \in \mathcal{Y}} F\left(\frac{1}{K}\left(\sum_{k=0}^K U_x^k\right), Y'\right) - \min_{X' \in \mathcal{X}} F\left(X', \frac{1}{K}\left(\sum_{k=0}^K U_y^k\right)\right) \\
&\leq \max_{Y' \in \mathcal{Y}} \frac{1}{K}\sum_{k=0}^K F(U_x^k, Y') - \min_{X' \in \mathcal{X}} \frac{1}{K}\sum_{k=0}^K F(X', U_y^k).
\end{aligned}$$

Given the fact of linear independence of X' and Y':

$$\text{gap}(X_{avg}^K, Y_{avg}^K) \leq \max_{(X', Y')} \frac{1}{K}\sum_{k=0}^K \left(F(X^K, Y') - F(X', Y^k)\right).$$

Using convexity and concavity of the function F:

$$\begin{aligned}
\text{gap}(X_{avg}^K, Y_{avg}^K) &\leq \max_{(X',Y')} \frac{1}{K} \sum_{k=0}^{K} \left(F(U_x^k, Y') - F(X', U_y^k)\right) \\
&= \max_{(X',Y')} \frac{1}{K} \sum_{k=0}^{K} \left(F(U_x^k, Y') - F(U_x^k, U_y^k) + F(U_x^k, U_y^k) - F(X', U_y^k)\right) \\
&\leq \max_{(X',Y')} \frac{1}{K} \sum_{k=0}^{K} \Big(\text{tr}\left[\left(\nabla_Y F(U_x^k; U_y^k)\right)^T \left(Y' - U_y^k\right)\right] \\
&\quad + \text{tr}\left[\left(\nabla_X F(U_x^k; U_y^k)\right)^T \left(U_x^k - X'\right)\right] \Big).
\end{aligned}$$

Then

$$\begin{aligned}
\text{gap}(X_{avg}^K, Y_{avg}^K) \leq & \frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{K} \\
& + 8\left(\lambda\lambda_{\max}(W) + L\right)\sqrt{M}(\Omega + G) \sum_{k=0}^{K} \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
& + 3\lambda\lambda_{\max}(W) \sum_{k=0}^{K} \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

δ from (35) and K from (38) are completed the proof.

□

Remark. (38) also corresponds to the number of communication rounds. It is also easy to estimate the total number of local iterations on server:

$$\begin{aligned}
K \times T &= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{\varepsilon}(1 + \gamma L)\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{\varepsilon}\left(1 + \frac{L}{\delta}\right)\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{(L + \lambda\lambda_{\max}(W))(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{\varepsilon}\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{L(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{\varepsilon}\log\frac{\Omega^2}{\delta}\right).
\end{aligned}$$

The last follows from the fact that we consider the case of small λ.

Lemma 7 (for Theorem 3) *Assume that for solving the problem (3) we use Randomized Extra Step Method from [1] with starting points X^k, Y^k and number of iterations:*

$$T = \mathcal{O}\left(r + \frac{\sqrt{r}\left(L + \frac{1}{\gamma}\right)}{\frac{1}{\gamma}}\log\frac{1}{\tilde{\delta}}\right) = \mathcal{O}\left(\left(r + \sqrt{r}\left(\gamma L + 1\right)\right)\log\frac{1}{\tilde{\delta}}\right). \tag{39}$$

Then for an output u^k it holds that

$$\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2 \leq \tilde{\delta}\left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).$$

Proof: The proof follows from the convergence estimates for the Algorithm 1 from [1] and the fact that Problem (3) is $\frac{1}{\gamma}$-strongly-convex-strongly-concave, as well as $\left(\frac{1}{\gamma} + L\right)$-smooth.

□

Remark. As mentioned earlier, the problem (3) is divided into M local problems, which are solved on each of the machines separately.

Note that Algorithm 1 involves the absolute precision δ, $\tilde{\delta}$ is the relative precision - a more convenient value for practice.

Theorem 9 (for Theorem 3) *Assume that the problem (3) is solved by Randomized Extra Step Method with variance reduction method from [1] with precision $\tilde{\delta}$:*

$$\delta = \min\left\{\frac{\sqrt{\varepsilon}}{9\lambda\lambda_{\max}(W)}; \frac{\varepsilon}{(24\lambda\lambda_{\max}(W)+L)\sqrt{M}(\Omega+G)}\right\} \tag{40}$$

and the number of iterations T (see (30)). Additionally, stepsize γ is

$$\gamma = \frac{1}{2\lambda\lambda_{\max}(W)}. \tag{41}$$

It holds that $gap(X_{avg}^K, Y_{avg}^K) \leq \varepsilon$ after

$$K = \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\right) \quad \textit{iterations}, \tag{42}$$

where

$$gap(X,Y) := \max_{Y' \in \mathcal{Y}} f(X,Y') - \min_{X' \in \mathcal{X}} f(X',Y).$$

and $X_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_x^k$, $Y_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_y^k$.

Proof: Summing (34) over all k from 0 to K

$$\begin{aligned}
&2\gamma \sum_{k=0}^K \Big(\text{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&+ \text{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right]\Big) \\
&\quad \leq \left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2 \\
&\qquad + 4(1 + \gamma L + \gamma\lambda\lambda_{\max}(W))\sqrt{M}(\Omega+G)\sum_{k=0}^K \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad + (2 + 4\gamma^2\lambda^2\lambda_{\max}^2(W))\sum_{k=0}^K \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right) \\
&\qquad - (1 - 4\gamma^2\lambda^2\lambda_{\max}^2(W))\sum_{k=0}^K \left(\left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|Y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

With our choice of γ

$$\begin{aligned}
&\frac{1}{\lambda\lambda_{\max}(W)}\sum_{k=0}^K \Big(\text{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&+ \text{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right]\Big) \\
&\quad \leq \left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2 \\
&\qquad + 4\left(2 + \frac{L}{\lambda\lambda_{\max}(W)}\right)\sqrt{M}(\Omega+G)\sum_{k=0}^K \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad + 3\sum_{k=0}^K \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

Then, by $X_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_x^k$ and $Y_{avg}^K = \frac{1}{K}\sum_{k=0}^K U_y^k$, Jensen's inequality and convexity-concavity of F:

$$\begin{aligned}\text{gap}(X_{avg}^K, Y_{avg}^K) &\leq \max_{Y' \in \mathcal{Y}} F\left(\frac{1}{K}\left(\sum_{k=0}^K U_x^k\right), Y'\right) - \min_{X' \in \mathcal{X}} F\left(X', \frac{1}{K}\left(\sum_{k=0}^K U_y^k\right)\right) \\ &\leq \max_{Y' \in \mathcal{Y}} \frac{1}{K}\sum_{k=0}^K F(U_x^k, Y') - \min_{X' \in \mathcal{X}} \frac{1}{K}\sum_{k=0}^K F(X', U_y^k).\end{aligned}$$

Given the fact of linear independence of X' and Y':

$$\text{gap}(X_{avg}^K, Y_{avg}^K) \leq \max_{(X',Y')} \frac{1}{K}\sum_{k=0}^K \left(F(X^K, Y') - F(X', Y^k)\right).$$

Using convexity and concavity of the function F:

$$\begin{aligned}\text{gap}(X_{avg}^K, Y_{avg}^K) &\leq \max_{(X',Y')} \frac{1}{K}\sum_{k=0}^K \left(F(U_x^k, Y') - F(X', U_y^k)\right) \\ &= \max_{(X',Y')} \frac{1}{K}\sum_{k=0}^K \left(F(U_x^k, Y') - F(U_x^k, U_y^k) + F(U_x^k, U_y^k) - F(X', U_y^k)\right) \\ &\leq \max_{(X',Y')} \frac{1}{K}\sum_{k=0}^K \Big(\text{tr}\left[\left(\nabla_Y F(U_x^k; U_y^k)\right)^T \left(Y' - U_y^k\right)\right] \\ &\quad + \text{tr}\left[\left(\nabla_X F(U_x^k; U_y^k)\right)^T \left(U_x^k - X'\right)\right]\Big).\end{aligned}$$

Then

$$\begin{aligned}\text{gap}(X_{avg}^K, Y_{avg}^K) \leq & \frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2)}{K} \\ & + 8\left(\lambda\lambda_{\max}(W) + L\right)\sqrt{M}(\Omega + G)\sum_{k=0}^K \left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\ & + 3\lambda\lambda_{\max}(W)\sum_{k=0}^K \left(\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + \left\|U_y^k - \hat{U}_y^k\right\|_F^2\right).\end{aligned}$$

δ from (40) and K from (42) are completed the proof.

□

Remark. (42) also corresponds to communication rounds. It is also easy to estimate the total number of local computation on each node:

$$\begin{aligned}K \times T &= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\left(r + \sqrt{r}\left(\gamma L + 1\right)\right)\log\frac{\Omega^2}{\delta}\right) \\ &= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)\Omega^2}{\varepsilon}\left(r + \frac{\sqrt{r}L}{\lambda\lambda_{\max}(W)}\right)\log\frac{\Omega^2}{\delta}\right) \\ &= \mathcal{O}\left(\frac{(r\lambda\lambda_{\max}(W) + \sqrt{r}L)\,\Omega^2}{\varepsilon}\log\frac{\Omega^2}{\delta}\right).\end{aligned}$$

If $\sqrt{r}\lambda\lambda_{\max}(W) = \mathcal{O}(L)$, the total expected number of local gradients becomes optimal.

C.2.3 Proof of Theorem 4

Strongly-convex-strongly-concave case

Lemma 8 *For Algorithm 2 it holds:*

$$\begin{aligned}\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \leq{}& (1-\gamma\mu)\left\|X^k-X^*\right\|_F^2+(1-\gamma\mu)\left\|Y^k-Y^*\right\|_F^2\\ &+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2\\ &-(1-3\gamma\mu-4\gamma^2L^2)\left\|X^k-\hat{U}_x^k\right\|_F^2\\ &+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2\\ &-(1-3\gamma\mu-4\gamma^2L^2)\left\|Y^k-\hat{U}_y^k\right\|_F^2.\end{aligned}$$

Proof: Let us use additional notation $F_x^k=U_x^k+\gamma\cdot\left(\nabla_X f(X^k,Y^k)-\nabla_X f(U_x^k,U_y^k)\right)$ and $F_y^k=U_y^k-\gamma\cdot\left(\nabla_Y f(X^k,Y^k)-\nabla_Y f(U_x^k,U_y^k)\right)$ for short. Then by non-expansiveness of the Euclidean projection, we get

$$\begin{aligned}&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2\\ &\quad=\left\|\text{proj}_{\mathcal{X}}\left[F_x^k\right]-\text{proj}_{\mathcal{X}}\left[X^*\right]\right\|_F^2+\left\|\text{proj}_{\mathcal{Y}}\left[F_y^k\right]-\text{proj}_{\mathcal{Y}}\left[Y^*\right]\right\|_F^2\\ &\quad\leq\left\|F_x^k-X^*\right\|_F^2+\left\|F_y^k-Y^*\right\|_F^2\\ &\quad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(F_x^k-X^k\right)^T\left(X^k-X^*\right)\right]+\left\|F_x^k-X^k\right\|_F^2\\ &\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(F_y^k-Y^k\right)^T\left(Y^k-Y^*\right)\right]+\left\|F_y^k-Y^k\right\|_F^2\\ &\quad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(F_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\ &\qquad+2\,\text{tr}\left[\left(F_x^k-X^k\right)^T\left(X^k-\hat{U}_x^k\right)\right]+\left\|F_x^k-X^k\right\|_F^2\\ &\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(F_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\ &\qquad+2\,\text{tr}\left[\left(F_y^k-Y^k\right)^T\left(Y^k-\hat{U}_y^k\right)\right]+\left\|F_y^k-Y^k\right\|_F^2\\ &\quad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(F_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\ &\qquad+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\ &\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(F_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\ &\qquad+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\ &\quad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(U_x^k+\gamma\cdot\left(\nabla_X f(X^k)^T\left(Y^k\right)-\nabla_X f(U_x^k,U_y^k)\right)-X^k,\hat{U}_x^k-X^*\right)\right]\\ &\qquad+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\ &\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(U_y^k-\gamma\cdot\left(\nabla_Y f(X^k;Y^k)-\nabla_Y f(U_x^k;U_y^k)\right)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\ &\qquad+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2\\ &\quad=\left\|X^k-X^*\right\|_F^2+2\,\text{tr}\left[\left(U_x^k+\gamma\cdot\nabla_X f(X^k;Y^k)-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]\\ &\qquad+2\,\text{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2\\ &\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\text{tr}\left[\left(U_y^k+\gamma\cdot\nabla_Y f(X^k;Y^k)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]\\ &\qquad+2\,\text{tr}\left[\left(-\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2.\end{aligned}$$

With expressions for V_x^k and V_y^k in Algorithm 2, we have

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad\le \left\|X^k-X^*\right\|_F^2+2\,\mathrm{tr}\left[\left(U_x^k-V_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\mathrm{tr}\left[\left(U_y^k-V_y^k\right)^T\left(\hat{U}_y^k-X^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2 \\
&\quad= \left\|X^k-X^*\right\|_F^2+2\,\mathrm{tr}\left[\left(\hat{U}_x^k-V_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(U_x^k-\hat{U}_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\mathrm{tr}\left[\left(\hat{U}_y^k-V_y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(U_y^k-\hat{U}_y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

According to the optimal condition for $\hat{U}_x^k$ and $\hat{U}_y^k$: for all $x\in\mathcal{X}$, $y\in\mathcal{Y}$ and $X=[x,\ldots,x]^T$, $Y=[y,\ldots,y]^T$

$$
\mathrm{tr}\left[\left(\gamma\cdot\lambda W\hat{U}_x^k+\hat{U}_x^k-V_x^k\right)^T\left(\hat{U}_x^k-X\right)\right]+\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_y^k+\hat{U}_y^k-V_y^k\right)^T\left(\hat{U}_y^k-Y\right)\right]\le 0,
$$

we get

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad\le \left\|X^k-X^*\right\|_F^2+2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(U_x^k-\hat{U}_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+2\,\mathrm{tr}\left[\left(U_y^k-\hat{U}_y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|X^k - X^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\left(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left\|Y^k - Y^*\right\|_F^2 + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\lambda W\hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right] + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2\,\mathrm{tr}\left[\left(-\gamma\cdot\left(\nabla_Y f(U_x^k;U_y^k) - \nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With property of the solution X^*, Y^*: for all $X \in \mathcal{X}, Y \in \mathcal{Y}$

$$
\begin{aligned}
&\mathrm{tr}\left[\left(\nabla_X f(X^*;Y^*) + \lambda W X^*\right)^T\left(X^* - X\right)\right] \\
&+\ \mathrm{tr}\left[\left(-\left(\nabla_Y f(X^*;Y^*) - \lambda W Y^*\right)\right)^T\left(Y^* - Y\right)\right] \le 0.
\end{aligned}
$$

And then μ-strong convexity - strong concavity of f and 0-strong convexity - strong concavity of φ, we obtain

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \le \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_X f(\hat{U}_x^k;\hat{U}_y^k) + \lambda W\hat{U}_x^k - \nabla_X f(X^*;Y^*) - \lambda W X^*\right)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(-\nabla_Y f(\hat{U}_x^k;\hat{U}_y^k) + \lambda W\hat{U}_y^k + \nabla_Y f(X^*;Y^*) - \lambda W Y^*\right)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_Y f(U_x^k;U_y^k) - \nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\qquad\quad + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad + 2\,\mathrm{tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T\left(\hat{U}_y^k - Y^*\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - 2\gamma\mu \left\|\hat{U}_x^k - X^*\right\|_F^2 - 2\gamma\mu \left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k)\right) - (U_x^k - \hat{U}_x^k)\right)^T\left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad - 2\,\mathrm{tr}\left[\left(\gamma\cdot\left(\nabla_Y f(U_x^k;U_y^k) - \nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)\right) - (U_y^k - \hat{U}_y^k)\right)^T\left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

By Young's inequality, we have

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - 2\gamma\mu \left\|\hat{U}_x^k - X^*\right\|_F^2 - 2\gamma\mu \left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\qquad\quad + \frac{2}{\gamma\mu}\left\|\gamma\cdot(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k)) - (U_x^k - \hat{U}_x^k)\right\|_F^2 \\
&\qquad\quad + \frac{\gamma\mu}{2}\left\|U_x^k - X^*\right\|_F^2 + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \frac{2}{\gamma\mu}\left\|\gamma\cdot(\nabla_Y f(U_x^k;U_y^k) - \nabla_Y f(\hat{U}_x^k;\hat{U}_y^k)) - (U_y^k - \hat{U}_y^k)\right\|_F^2 \\
&\qquad\quad + \frac{\gamma\mu}{2}\left\|\hat{U}_y^k - Y^*\right\|_F^2 + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_x^k - X^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\qquad\quad + \frac{4}{\gamma\mu}\left\|\gamma(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k))\right\|_F^2 + \frac{4}{\gamma\mu}\left\|U_x^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left\|U_x^k + \gamma(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k)) - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \frac{4}{\gamma\mu}\left\|\gamma(\nabla_Y f(U_x^k;U_y^k) - \nabla_Y f(\hat{U}_x^k;\hat{U}_y^k))\right\|_F^2 + \frac{4}{\gamma\mu}\left\|U_y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad + \left\|U_y^k + \gamma(\nabla_Y f(X^k;Y^k) - \nabla_Y f(U_x^k;U_y^k)) - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_x^k - X^*\right\|_F^2 - \frac{3\gamma\mu}{2}\left\|\hat{U}_y^k - Y^*\right\|_F^2 \\
&\qquad\quad + \frac{4}{\gamma\mu}\left\|\gamma\cdot(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k))\right\|_F^2 + \left(2 + \frac{4}{\gamma\mu}\right)\left\|U_x^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + 2\left\|\gamma\cdot\left(\nabla_X f(X^k;Y^k) - \nabla_X f(U_x^k;U_y^k)\right)\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \frac{4}{\gamma\mu}\left\|\gamma\cdot(\nabla_X f(U_x^k;U_y^k) - \nabla_X f(\hat{U}_x^k;\hat{U}_y^k))\right\|_F^2 + \left(2 + \frac{4}{\gamma\mu}\right)\left\|U_y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad\quad + 2\left\|\gamma\cdot\left(\nabla_Y f(X^k;Y^k) - \nabla_Y f(U_x^k;U_y^k)\right)\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

Then we use L-smoothness of f

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_x^k-X^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2 \\
&\qquad+\frac{4\gamma L^2}{\mu}\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+2\gamma^2L^2\left\|X^k-U_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\frac{4\gamma L^2}{\mu}\left\|U_y^k-\hat{U}_y^k\right\|_F^2+\left(2+\frac{4}{\gamma\mu}\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2 \\
&\qquad+2\gamma^2L^2\left\|Y^k-U_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2 \\
&\quad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_x^k-X^*\right\|_F^2-\frac{3\gamma\mu}{2}\left\|\hat{U}_y^k-Y^*\right\|_F^2 \\
&\qquad+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2-(1-4\gamma^2L^2)\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2-(1-4\gamma^2L^2)\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

By inequality $\|A+B\|_F^2\ge\frac{2}{3}\|A\|_F^2-2\|B\|_F^2$, we have

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad\le (1-\gamma\mu)\left\|X^k-X^*\right\|_F^2+(1-\gamma\mu)\left\|Y^k-Y^*\right\|_F^2 \\
&\qquad+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_x^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad-(1-3\gamma\mu-4\gamma^2L^2)\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left(2+\frac{4\gamma L^2}{\mu}+\frac{4}{\gamma\mu}+4\gamma^2L^2\right)\left\|U_y^k-\hat{U}_y^k\right\|_F^2 \\
&\qquad-(1-3\gamma\mu-4\gamma^2L^2)\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

□

Lemma 9 *Assume that for problems (4) we use fast gradient method with starting points X^k, Y^k and number of iterations:*

$$
T=\mathcal{O}\left(\left(1+\sqrt{\frac{\lambda\lambda_{\max}(W)+\frac{1}{\gamma}}{\lambda\lambda_{\min}^+(W)+\frac{1}{\gamma}}}\right)\log\frac{1}{\tilde{\delta}}\right). \tag{43}
$$

Then for an output u^k it holds that

$$
\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\le\tilde{\delta}\left(\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-\hat{U}_y^k\right\|_F^2\right).
$$

Proof: We compute prox for $\frac{\gamma\lambda}{2}\|\sqrt{W}X\|_F^2-\frac{\gamma\lambda}{2}\|\sqrt{W}Y\|_F^2$ on the third step of Algorithm 2. This saddle point problem is equivalent to the minimization problems (4). It is solved by fast gradient method. The complexity of solving any of these two problems is $\tilde{\mathcal{O}}(1)$ on the **Ker** W and is $\tilde{\mathcal{O}}\left(1+\sqrt{\frac{\lambda\lambda_{\max}(W)+\frac{1}{\gamma}}{\lambda\lambda_{\min}^+(W)+\frac{1}{\gamma}}}\right)$ on the $(\textbf{Ker}\ W)^\perp$. It follows from the convergence estimates for the fast gradient method and the fact that Problems (4) are $\frac{1}{\gamma}$-strongly-convex-strongly-concave, as well as $\frac{1}{\gamma}$-smooth on the **Ker** W and $\lambda\lambda_{\min}^+(W)+\frac{1}{\gamma}$-strongly-convex-strongly-concave, as well as $\lambda\lambda_{\max}(W)+\frac{1}{\gamma}$-smooth on the $(\textbf{Ker}\ W)^\perp$.

□

Note that Algorithm 2 involves the absolute precision δ, $\tilde{\delta}$ is the relative precision - a more convenient value for practice.

Theorem 10 (Theorem 4) *Assume that problem (4) be solved by fast gradient method with precision $\tilde{\delta}$:*

$$\tilde{\delta} = \frac{1}{2\left(2 + \frac{4\gamma L^2}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2 L^2\right)} \tag{44}$$

and the number of iterations T (see (43)). Additionally, stepsize γ is

$$\gamma = \min\left\{\frac{1}{12\mu}; \frac{1}{16L}\right\}. \tag{45}$$

Then Algorithm 2 converges linearly to the solution z^* and it holds that $\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \leq \varepsilon$ after

$$K = \mathcal{O}\left(\frac{1}{\gamma\mu}\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right) \quad \textit{iterations.} \tag{46}$$

Proof: Combining results from Lemma 8 and 9 gives

$$\begin{aligned}
\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \leq{}& (1-\gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1-\gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&+ \left(2 + \frac{4\gamma L^2}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2 L^2\right)\tilde{\delta}\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&- (1 - 3\gamma\mu - 4\gamma^2 L^2)\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&+ \left(2 + \frac{4\gamma L^2}{\mu} + \frac{4}{\gamma\mu} + 4\gamma^2 L^2\right)\tilde{\delta}\left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&- (1 - 3\gamma\mu - 4\gamma^2 L^2)\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}$$

With the choice $\tilde{\delta}$ from (44), we obtain

$$\begin{aligned}
\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \leq{}& (1-\gamma\mu)\left\|X^k - X^*\right\|_F^2 + (1-\gamma\mu)\left\|Y^k - Y^*\right\|_F^2 \\
&- \left(\frac{1}{2} - 3\gamma\mu - 4\gamma^2 L^2\right)\left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&- \left(\frac{1}{2} - 3\gamma\mu - 4\gamma^2 L^2\right)\left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}$$

The proof is completed by choosing γ from (45).

□

Remark. (46) also corresponds to calls of f gradients which in turn corresponds to the number of local computation on each node. Substitution of (45) in (46) gives that

$$K = \mathcal{O}\left(\left(1 + \frac{L}{\mu}\right)\log\frac{\left\|X^0 - X^*\right\|_F^2 + \left\|Y^0 - Y^*\right\|_F^2}{\varepsilon}\right).$$

It is also easy to estimate the total number of communication rounds:

$$
\begin{aligned}
K \times T &= \mathcal{O}\left(\frac{1}{\gamma\mu}\left(1+\sqrt{\frac{\lambda\lambda_{\max}(W)+\frac{1}{\gamma}}{\lambda\lambda^{+}_{\min}(W)+\frac{1}{\gamma}}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq \mathcal{O}\left(\frac{1}{\gamma\mu}\left(1+\sqrt{1+\frac{\lambda\lambda_{\max}(W)}{\lambda\lambda^{+}_{\min}(W)+\frac{1}{\gamma}}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq \mathcal{O}\left(\frac{1}{\gamma\mu}\left(1+\sqrt{1+\frac{\lambda\lambda_{\max}(W)}{\max\left\{\lambda\lambda^{+}_{\min}(W),\frac{1}{\gamma}\right\}}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&= \mathcal{O}\left(\frac{1}{\gamma\mu}\left(1+\sqrt{\min\left\{\chi(W),\gamma\lambda\lambda_{\max}(W)\right\}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&\leq \mathcal{O}\left(\frac{1}{\gamma\mu}\left(1+\sqrt{\min\left\{\chi(W),\frac{\lambda\lambda_{\max}(W)}{L}\right\}}\right)\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&= \mathcal{O}\left(\frac{1}{\gamma\mu}\sqrt{\min\left\{\chi(W),\frac{\lambda\lambda_{\max}(W)}{L}\right\}}\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&= \mathcal{O}\left(\left(1+\frac{L}{\mu}\right)\sqrt{\min\left\{\chi(W),\frac{\lambda\lambda_{\max}(W)}{L}\right\}}\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&= \mathcal{O}\left(\frac{L}{\mu}\sqrt{\min\left\{\chi(W),\frac{\lambda\lambda_{\max}(W)}{L}\right\}}\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right)\\
&= \mathcal{O}\left(\min\left\{\frac{L}{\mu}\sqrt{\chi(W)},\frac{\sqrt{\lambda\lambda_{\max}(W)L}}{\mu}\right\}\log\frac{1}{\tilde{\delta}}\log\frac{\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2}{\varepsilon}\right).
\end{aligned}
$$

The third inequality follows from the inequality: $\gamma \leq \frac{1}{L}$.

Convex-Concave case

Lemma 10 *For Algorithm 2 it holds:*

$$
\begin{aligned}
&2\gamma\, tr\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X^*\right)\right]\\
&\quad+2\gamma\, tr\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y^*\right)\right]\\
&\qquad\leq \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X^*\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad+4(1+\gamma\lambda\lambda_{\max}(W)+\gamma L)\sqrt{M}(\Omega+G)\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&\qquad+(2+4\gamma^2L^2)\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right)\\
&\qquad-(1-4\gamma^2L^2)\left(\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-\hat{U}_y^k\right\|_F^2\right). \qquad (47)
\end{aligned}
$$

Proof: Let us use additional notation $F_x^k = U_x^k+\gamma\cdot\left(\nabla_X f(X^k;Y^k)-\nabla_X f(U_x^k;U_y^k)\right)$ and $F_y^k = U_y^k-\gamma\cdot\left(\nabla_Y f(X^k;Y^k)-\nabla_Y f(U_x^k;U_y^k)\right)$ for short. Then by non-expansiveness of the

Euclidean projection, we get

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad=\left\|\operatorname{proj}_{\mathcal{X}}\left[F_x^k\right]-\operatorname{proj}_{\mathcal{X}}\left[X^*\right]\right\|_F^2+\left\|\operatorname{proj}_{\mathcal{Y}}\left[F_y^k\right]-\operatorname{proj}_{\mathcal{Y}}\left[Y^*\right]\right\|_F^2 \\
&\quad\leq\left\|F_x^k-X^*\right\|_F^2+\left\|F_y^k-Y^*\right\|_F^2 \\
&\quad=\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(F_x^k-X^k\right)^T\left(X^k-X^*\right)\right]+\left\|F_x^k-X^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(F_y^k-Y^k\right)^T\left(Y^k-Y^*\right)\right]+\left\|F_y^k-Y^k\right\|_F^2 \\
&\quad=\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(F_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]+2\operatorname{tr}\left[\left(F_x^k-X^k\right)^T\left(X^k-\hat{U}_x^k\right)\right] \\
&\qquad+\left\|F_x^k-X^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(F_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+2\operatorname{tr}\left[\left(F_y^k-Y^k\right)^T\left(Y^k-\hat{U}_y^k\right)\right] \\
&\qquad+\left\|F_y^k-Y^k\right\|_F^2 \\
&\quad=\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(F_x^k-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(F_y^k-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2 \\
&\quad=\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(U_x^k+\gamma\cdot\left(\nabla_X f(X^k;Y^k)-\nabla_X f(U_x^k;U_y^k)\right)-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(U_y^k-\gamma\cdot\left(\nabla_Y f(X^k;Y^k)-\nabla_Y f(U_x^k;U_y^k)\right)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2 \\
&\quad=\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(U_x^k+\gamma\cdot\nabla_X f(X^k;Y^k)-X^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\operatorname{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(U_y^k+\gamma\cdot\nabla_Y f(X^k;Y^k)-Y^k\right)^T\left(\hat{U}_y^k-Y^*\right)\right] \\
&\qquad+2\operatorname{tr}\left[\left(\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With expressions for V_x^k and V_y^k in Algorithm 2, we have

$$
\begin{aligned}
&\left\|X^{k+1}-X^*\right\|_F^2+\left\|Y^{k+1}-Y^*\right\|_F^2 \\
&\quad\leq\left\|X^k-X^*\right\|_F^2+2\operatorname{tr}\left[\left(U_x^k-V_x^k\right)^T\left(\hat{U}_x^k-X^*\right)\right] \\
&\qquad+2\operatorname{tr}\left[\left(-\gamma\cdot\nabla_X f(U_x^k;U_y^k)\right)^T\left(\hat{U}_x^k-X^*\right)\right]+\left\|F_x^k-\hat{U}_x^k\right\|_F^2-\left\|X^k-\hat{U}_x^k\right\|_F^2 \\
&\qquad+\left\|Y^k-Y^*\right\|_F^2+2\operatorname{tr}\left[\left(U_y^k-V_y^k\right)^T\left(\hat{U}_y^k-X^*\right)\right] \\
&\qquad+2\operatorname{tr}\left[\left(\gamma\cdot\nabla_Y f(U_x^k;U_y^k)\right)^T\left(\hat{U}_y^k-Y^*\right)\right]+\left\|F_y^k-\hat{U}_y^k\right\|_F^2-\left\|Y^k-\hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left\|X^k - X^*\right\|_F^2 + 2 \text{ tr}\left[\left(\hat{U}_x^k - V_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + 2 \text{ tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] + 2 \text{ tr}\left[\left(-\gamma \cdot \nabla_X f(U_x^k; U_y^k)\right)^T \left(\hat{U}_x^k - X^*\right)\right] \\
&\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad + \left\|Y^k - Y^*\right\|_F^2 + 2 \text{ tr}\left[\left(\hat{U}_y^k - V_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + 2 \text{ tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] + 2 \text{ tr}\left[\left(\gamma \cdot \nabla_Y f(U_x^k; U_y^k)\right)^T \left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

According to the optimal condition for $\hat{U}_x^k$ and $\hat{U}_y^k$: for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $X = [x, \ldots, x]^T$, $Y = [y, \ldots, y]^T$

$$
\text{tr}\left[\left(\gamma \cdot \lambda W \hat{U}_x^k + \hat{U}_x^k - V_x^k\right)^T \left(\hat{U}_x^k - X\right)\right] + \text{tr}\left[\left(-\gamma \cdot \lambda W \hat{U}_y^k + \hat{U}_y^k - V_y^k\right)^T \left(\hat{U}_y^k - Y\right)\right] \leq 0,
$$

we get

$$
\begin{aligned}
&\left\|X^{k+1} - X^*\right\|_F^2 + \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad \leq \left\|X^k - X^*\right\|_F^2 + 2 \text{ tr}\left[\left(-\gamma \cdot \lambda W \hat{U}_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] \\
&\quad\quad + 2 \text{ tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] + 2 \text{ tr}\left[\left(-\gamma \cdot \nabla_X f(U_x^k; U_y^k)\right)^T \left(\hat{U}_x^k - X^*\right)\right] \\
&\quad\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\quad\quad + \left\|Y^k - Y^*\right\|_F^2 + 2 \text{ tr}\left[\left(\gamma \cdot \lambda W \hat{U}_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad\quad + 2 \text{ tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] + 2 \text{ tr}\left[\left(\gamma \cdot \nabla_Y f(U_x^k; U_y^k)\right)^T \left(\hat{U}_y^k - Y^*\right)\right] \\
&\quad\quad + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

Small rearrangement gives

$$
\begin{aligned}
&2\gamma \text{ tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&\quad + 2\gamma \text{ tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right] \\
&\quad\quad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad\quad\quad + 2\gamma \text{ tr}\left[\left(\lambda W U_x^k - \lambda W \hat{U}_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] + 2\gamma \text{ tr}\left[\left(\lambda W U_x^k\right)^T \left(U_x^k - \hat{U}_x^k\right)\right] \\
&\quad\quad\quad + 2\gamma \text{ tr}\left[\left(\lambda W \hat{U}_y^k - \lambda W U_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] + 2\gamma \text{ tr}\left[\left(\lambda W U_y^k\right)^T \left(U_y^k - \hat{U}_y^k\right)\right] \\
&\quad\quad\quad + 2 \text{ tr}\left[\left(U_x^k - \hat{U}_x^k\right)^T \left(\hat{U}_x^k - X^*\right)\right] + 2 \text{ tr}\left[\left(-\gamma \cdot \nabla_X f(U_x^k; U_y^k)\right)^T \left(\hat{U}_x^k - U_x^k\right)\right] \\
&\quad\quad\quad + 2 \text{ tr}\left[\left(U_y^k - \hat{U}_y^k\right)^T \left(\hat{U}_y^k - Y^*\right)\right] + 2 \text{ tr}\left[\left(-\gamma \cdot \nabla_Y f(U_x^k; U_y^k)\right)^T \left(\hat{U}_y^k - U_y^k\right)\right] \\
&\quad\quad\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\quad + 2\gamma\|\lambda W U_x^k - \lambda W \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\lambda W U_x^k\|_F \cdot \|U_x^k - \hat{U}_x^k\|_F \\
&\quad + 2\gamma\|\lambda W \hat{U}_y^k - \lambda W U_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\lambda W U_y^k\|_F \cdot \|U_y^k - \hat{U}_y^k\|_F \\
&\quad + 2\|U_x^k - \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\nabla_X f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_x^k - U_x^k\|_F \\
&\quad + 2\|U_y^k - \hat{U}_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\nabla_Y f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_y^k - U_y^k\|_F \\
&\quad + \left\|F_x^k - \hat{U}_x^k\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 + \left\|F_y^k - \hat{U}_y^k\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

With definition of F_x^k and F_y^k we get

$$
\begin{aligned}
&2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&\quad + 2\gamma \operatorname{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right] \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad\quad + 2\gamma\|\lambda W U_x^k - \lambda W \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\lambda W U_x^k\|_F \cdot \|U_x^k - \hat{U}_x^k\|_F \\
&\qquad\quad + 2\gamma\|\lambda W \hat{U}_y^k - \lambda W U_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\lambda W U_y^k\|_F \cdot \|U_y^k - \hat{U}_y^k\|_F \\
&\qquad\quad + 2\|U_x^k - \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\nabla_X f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_x^k - U_x^k\|_F \\
&\qquad\quad + 2\|U_y^k - \hat{U}_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\nabla_Y f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_y^k - U_y^k\|_F \\
&\qquad\quad + \left\|U_x^k - \hat{U}_x^k - \gamma \cdot (\nabla_X f(X^k; Y^k) - \nabla_X f(U_x^k; U_y^k))\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + \left\|U_y^k - \hat{U}_y^k - \gamma \cdot (\nabla_Y f(X^k; Y^k) - \nabla_Y f(U_x^k; U_y^k))\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2 \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad\quad + 2\gamma\|\lambda W U_x^k - \lambda W \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\lambda W U_x^k\|_F \cdot \|U_x^k - \hat{U}_x^k\|_F \\
&\qquad\quad + 2\gamma\|\lambda W \hat{U}_y^k - \lambda W U_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\lambda W U_y^k\|_F \cdot \|U_y^k - \hat{U}_y^k\|_F \\
&\qquad\quad + 2\|U_x^k - \hat{U}_x^k\|_F \cdot \|\hat{U}_x^k - X^*\|_F + 2\gamma\|\nabla_X f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_x^k - U_x^k\|_F \\
&\qquad\quad + 2\|U_y^k - \hat{U}_y^k\|_F \cdot \|\hat{U}_y^k - Y^*\|_F + 2\gamma\|\nabla_Y f(U_x^k; U_y^k)\|_F \cdot \|\hat{U}_y^k - U_y^k\|_F \\
&\qquad\quad + 2\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + 2\gamma^2\left\|\nabla_X f(X^k; Y^k) - \nabla_X f(U_x^k; U_y^k)\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + 2\left\|U_y^k - \hat{U}_y^k\right\|_F^2 + 2\gamma^2\left\|\nabla_Y f(X^k; Y^k) - \nabla_Y f(U_x^k; U_y^k)\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2.
\end{aligned}
$$

$$
\begin{aligned}
&2\gamma \operatorname{tr}\left[\left(\nabla_X f(U_x^k; U_y^k) + \lambda W U_x^k\right)^T \left(U_x^k - X^*\right)\right] \\
&\quad + 2\gamma \operatorname{tr}\left[\left(-\nabla_Y f(U_x^k; U_y^k) + \lambda W U_y^k\right)^T \left(U_y^k - Y^*\right)\right] \\
&\qquad \leq \left\|X^k - X^*\right\|_F^2 + \left\|Y^k - Y^*\right\|_F^2 - \left\|X^{k+1} - X^*\right\|_F^2 - \left\|Y^{k+1} - Y^*\right\|_F^2 \\
&\qquad\quad + 2\gamma\sqrt{M}\Omega\left(\|\lambda W U_x^k - \lambda W \hat{U}_x^k\|_F + \|\lambda W \hat{U}_y^k - \lambda W U_y^k\|_F\right) \\
&\qquad\quad + 2(1 + \gamma\lambda\lambda_{\max}(W) + \gamma L)\sqrt{M}(\Omega + G)\left(\|U_x^k - \hat{U}_x^k\|_F + \|\hat{U}_y^k - U_y^k\|_F\right) \\
&\qquad\quad + 2\left\|U_x^k - \hat{U}_x^k\right\|_F^2 + 2\gamma^2\left\|\nabla_X f(X^k; Y^k) - \nabla_X f(U_x^k; U_y^k)\right\|_F^2 - \left\|X^k - \hat{U}_x^k\right\|_F^2 \\
&\qquad\quad + 2\left\|U_y^k - \hat{U}_y^k\right\|_F^2 + 2\gamma^2\left\|\nabla_Y f(X^k; Y^k) - \nabla_Y f(U_x^k; U_y^k)\right\|_F^2 - \left\|Y^k - \hat{U}_y^k\right\|_F^2
\end{aligned}
$$

Here we additionally used the diameter Ω of $\mathcal{Z}$ and assume point $\hat{x}, \hat{y}$, s.t. $\|\hat{x}\| \leq G, \|\hat{y}\| \leq G$ for some constant G.

Then we use smoothness of f, φ and obtain

$$
\begin{aligned}
&2\gamma \text{ tr}\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X^*\right)\right]\\
&\quad+2\gamma \text{ tr}\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y^*\right)\right]\\
&\qquad\le \left\|X^k-X^*\right\|_F^2+\left\|Y^k-Y^*\right\|_F^2-\left\|X^{k+1}-X^*\right\|_F^2-\left\|Y^{k+1}-Y^*\right\|_F^2\\
&\qquad\quad+4(1+\gamma\lambda\lambda_{\max}(W)+\gamma L)\sqrt{M}(\Omega+G)\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&\qquad\quad+(2+4\gamma^2L^2)\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right)\\
&\qquad\quad-(1-4\gamma^2L^2)\left(\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-\hat{U}_y^k\right\|_F^2\right).
\end{aligned}
$$

□

Theorem 11 (Theorem 4) *Assume that problem (4) be solved by fast gradient method with precision δ:*

$$
\delta=\min\left\{\frac{\sqrt{\varepsilon}}{9L};\frac{\varepsilon}{(24\lambda\lambda_{\max}(W)+L)\sqrt{M}(\Omega+G)}\right\} \tag{48}
$$

and number of iterations T:

$$
T=\mathcal{O}\left(\left(1+\sqrt{\frac{\lambda\lambda_{\max}(W)+\frac{1}{\gamma}}{\lambda\lambda^+_{\min}(W)+\frac{1}{\gamma}}}\right)\log\frac{\Omega^2}{\delta}\right). \tag{49}
$$

Additionally, let us choose stepsize γ as follows

$$
\gamma=\frac{1}{2L}. \tag{50}
$$

It holds that $gap(X_{avg}^K,Y_{avg}^K)\le\varepsilon$ after

$$
K=\mathcal{O}\left(\frac{L(\left\|X^0-X\right\|_F^2+\left\|Y^0-Y\right\|_F^2)}{\varepsilon}\right)\quad \textit{iterations}, \tag{51}
$$

where

$$
gap(X,Y):=\max_{Y'\in\mathcal{Y}}f(X,Y')-\min_{X'\in\mathcal{X}}f(X',Y).
$$

and $X_{avg}^K=\frac{1}{K}\sum_{k=0}^K U_x^k$, $Y_{avg}^K=\frac{1}{K}\sum_{k=0}^K U_y^k$.

Proof: Summing (47) over all k from 0 to K

$$
\begin{aligned}
&2\gamma\sum_{k=0}^K\left(\text{ tr}\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X\right)\right]+\text{ tr}\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y\right)\right]\right)\\
&\quad\le\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2\\
&\qquad+4(1+\gamma\lambda\lambda_{\max}(W)+\gamma L)\sqrt{M}(\Omega+G)\sum_{k=0}^K\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&\qquad+(2+4\gamma^2L^2)\sum_{k=0}^K\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right)\\
&\qquad-(1-4\gamma^2L^2)\sum_{k=0}^K\left(\left\|X^k-\hat{U}_x^k\right\|_F^2+\left\|Y^k-\hat{U}_y^k\right\|_F^2\right).
\end{aligned}
$$

With our choice of γ

$$\begin{aligned}
\frac{1}{L}\sum_{k=0}^{K}\Big(&\text{tr}\left[\left(\nabla_X f(U_x^k;U_y^k)+\lambda W U_x^k\right)^T\left(U_x^k-X\right)\right]\\
&+\text{tr}\left[\left(-\nabla_Y f(U_x^k;U_y^k)+\lambda W U_y^k\right)^T\left(U_y^k-Y\right)\right]\Big)\\
&\le \left\|X^0-X\right\|_F^2+\left\|Y^0-Y\right\|_F^2\\
&\quad+4\left(2+\frac{\lambda\lambda_{\max}(W)}{L}\right)\sqrt{M}(\Omega+G)\sum_{k=0}^{K}\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&\quad+3\sum_{k=0}^{K}\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

Then, by $X_{avg}^K=\frac{1}{K}\sum_{k=0}^{K}U_x^k$ and $Y_{avg}^K=\frac{1}{K}\sum_{k=0}^{K}U_y^k$, Jensen's inequality and convexity-concavity of F:

$$\begin{aligned}
\text{gap}(X_{avg}^K,Y_{avg}^K)&\le\max_{Y'\in\mathcal{Y}}F\left(\frac{1}{K}\left(\sum_{k=0}^{K}U_x^k\right),Y'\right)-\min_{X'\in\mathcal{X}}F\left(X',\frac{1}{K}\left(\sum_{k=0}^{K}U_y^k\right)\right)\\
&\le\max_{Y'\in\mathcal{Y}}\frac{1}{K}\sum_{k=0}^{K}F(U_x^k,Y')-\min_{X'\in\mathcal{X}}\frac{1}{K}\sum_{k=0}^{K}F(X',U_y^k).
\end{aligned}$$

Given the fact of linear independence of X' and Y':

$$\text{gap}(X_{avg}^K,Y_{avg}^K)\le\max_{(X',Y')}\frac{1}{K}\sum_{k=0}^{K}\left(F(X^K,Y')-F(X',Y^k)\right).$$

Using convexity and concavity of the function F:

$$\begin{aligned}
\text{gap}(X_{avg}^K,Y_{avg}^K)&\le\max_{(X',Y')}\frac{1}{K}\sum_{k=0}^{K}\left(F(U_x^k,Y')-F(X',U_y^k)\right)\\
&=\max_{(X',Y')}\frac{1}{K}\sum_{k=0}^{K}\left(F(U_x^k,Y')-F(U_x^k,U_y^k)+F(U_x^k,U_y^k)-F(X',U_y^k)\right)\\
&\le\max_{(X',Y')}\frac{1}{K}\sum_{k=0}^{K}\Big(\text{tr}\left[\left(\nabla_Y F(U_x^k;U_y^k)\right)^T\left(Y'-U_y^k\right)\right]\\
&\quad+\text{tr}\left[\left(\nabla_X F(U_x^k;U_y^k)\right)^T\left(U_x^k-X'\right)\right]\Big).
\end{aligned}$$

Then

$$\begin{aligned}
\text{gap}(X_{avg}^K,Y_{avg}^K)\le&\frac{L(\left\|X^0-X^*\right\|_F^2+\left\|Y^0-Y^*\right\|_F^2)}{K}\\
&+8\left(L+\lambda\lambda_{\max}(W)\right)\sqrt{M}(\Omega+G)\sum_{k=0}^{K}\left(\|U_x^k-\hat{U}_x^k\|_F+\|\hat{U}_y^k-U_y^k\|_F\right)\\
&+3L\sum_{k=0}^{K}\left(\left\|U_x^k-\hat{U}_x^k\right\|_F^2+\left\|U_y^k-\hat{U}_y^k\right\|_F^2\right).
\end{aligned}$$

the choice of δ from (48) and K from (51) completes the proof.

□

Remark. (51) also corresponds to the total number of local iterations. It is also easy to estimate the total number of communication rounds:

$$
\begin{aligned}
K \times T &= \mathcal{O}\left(\frac{L(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\left(1 + \sqrt{\frac{\lambda\lambda_{\max}(W) + \frac{1}{\gamma}}{\lambda\lambda_{\min}^+(W) + \frac{1}{\gamma}}}\right)\log\frac{\Omega^2}{\delta}\right) \\
&\leq \mathcal{O}\left(\frac{L(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\sqrt{\frac{\lambda\lambda_{\max}(W) + \frac{1}{\gamma}}{\lambda\lambda_{\min}^+(W) + \frac{1}{\gamma}}}\log\frac{\Omega^2}{\delta}\right) \\
&\leq \mathcal{O}\left(\frac{L(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\sqrt{1 + \frac{\lambda\lambda_{\max}(W)}{\lambda\lambda_{\min}^+(W) + \frac{1}{\gamma}}}\log\frac{\Omega^2}{\delta}\right) \\
&\leq \mathcal{O}\left(\frac{L(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\sqrt{1 + \frac{\lambda\lambda_{\max}(W)}{\max\left\{\lambda\lambda_{\min}^+(W), \frac{1}{\gamma}\right\}}}\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{L(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\sqrt{\min\left\{\chi(W), \gamma\lambda\lambda_{\max}(W)\right\}}\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{\min\left\{L\sqrt{\chi(W)}, \sqrt{L\lambda\lambda_{\max}(W)}\right\}(\left\|X^0 - X\right\|_F^2 + \left\|Y^0 - Y\right\|_F^2)}{\varepsilon}\log\frac{\Omega^2}{\delta}\right) \\
&= \mathcal{O}\left(\frac{\min\left\{L\sqrt{\chi(W)}, \sqrt{L\lambda\lambda_{\max}(W)}\right\}\Omega^2}{\varepsilon}\log\frac{\Omega^2}{\delta}\right).
\end{aligned}
$$

C.2.4 Proof of Theorem 5

Let us use additional notation $G_x(X,Y) = \frac{\xi^k}{(1-p)r}\cdot\nabla_X f_{\hat{\xi}_k}(X,Y) + \frac{1-\xi^k}{p}\cdot\lambda WX$ and $G_y(X,Y) = \frac{\xi^k}{(1-p)r}\cdot\nabla_Y f_{\hat{\xi}_k}(X,Y) - \frac{1-\xi^k}{p}\cdot\lambda WY$ for short, where $\mathbb{P}\left\{\hat{\xi}_k = j\right\} = \frac{1}{r}$.

Let us consider our problem as a finite sum problem with $r+1$ terms.

$$
F(X,Y) = \frac{1}{r}\sum_{j=1}^{r} g_j(X,Y) + g_{r+1}(X,Y),
$$

where $g_j(x,y) = \sum\limits_{m=1}^{M} f_{mj}(x_m, y_m)$ and $g_{r+1}(X,Y) = \frac{\lambda}{2}\|\sqrt{W}X\|_F^2 - \frac{\lambda}{2}\|\sqrt{W}Y\|_F^2$. For such a problem, one can use the results of the convergence of the variance reduction method [1] on which our method is based. **Convex-concave case** In the convex-concave case (Section 2.3.1 from [1]), the estimates on the number of iterations is

$$
\mathcal{O}\left(\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right),
$$

where L_{eff} can be computed as follows:

$$
\begin{aligned}
&\mathbb{E}\Big[\|G_x(X_1,Y_1)-G_x(X_2,Y_2)\|_F^2+\|G_y(X_1,Y_1)-G_y(X_2,Y_2)\|_F^2\Big]\\
&\quad=(1-p)\mathbb{E}\Bigg[\frac{1}{r^2(1-p)^2p_{\hat{\xi}_k}^2}\left\|\nabla_X f_{\hat{\xi}_k}(X_1,Y_1)-\nabla_X f_{\hat{\xi}_k}(X_2,Y_2)\right\|_F^2\\
&\qquad+\frac{1}{r^2(1-p)^2p_{\hat{\xi}_k}^2}\left\|\nabla_Y f_{\hat{\xi}_k}(X_1,Y_1)-\nabla_Y f_{\hat{\xi}_k}(X_2,Y_2)\right\|_F^2\Bigg]\\
&\qquad+p\left[\frac{1}{p^2}\|\lambda WX_1-\lambda WX_2\|_F^2+\frac{1}{p^2}\|\lambda WY_1-\lambda WY_2\|_F^2\right]\\
&\quad=\frac{1}{1-p}\Bigg[\sum_{i=1}^r p_i\cdot\frac{1}{r^2p_i^2}\|\nabla_X f_i(X_1,Y_1)-\nabla_X f_i(X_2,Y_2)\|_F^2\\
&\qquad+\sum_{i=1}^r p_i\cdot\frac{1}{r^2p_i^2}\|\nabla_Y f_i(X_1,Y_1)-\nabla_Y f_i(X_2,Y_2)\|_F^2\Bigg]\\
&\qquad+\frac{1}{p}\left[\|\lambda WX_1-\lambda WX_2\|_F^2+\|\lambda WY_1-\lambda WY_2\|_F^2\right]\\
&\quad\le\frac{1}{1-p}\cdot\sum_{i=1}^r\frac{L^2}{r^2p_i}\cdot\left[\|X_1-X_2\|_F^2+\|Y_1-Y_2\|_F^2\right]\\
&\qquad+\frac{1}{p}\cdot\lambda^2\lambda_{\max}^2(W)\cdot\left[\|X_1-X_2\|_F^2+\|Y_1-Y_2\|_F^2\right].
\end{aligned}
$$

Choose $p_i=\frac{1}{r}$:

$$
\begin{aligned}
&\mathbb{E}\left[\|G_x(X_1,Y_1)-G_x(X_2,Y_2)\|_F^2+\|G_y(X_1,Y_1)-G_y(X_2,Y_2)\|_F^2\right]\\
&\quad\le\left(\frac{L^2}{1-p}+\frac{\lambda^2\lambda_{\max}^2(W)}{p}\right)\left[\|X_1-X_2\|_F^2+\|Y_1-Y_2\|_F^2\right]=L_{\text{eff}}^2\left[\|X_1-X_2\|_F^2+\|Y_1-Y_2\|_F^2\right]
\end{aligned}
$$

Note that optimal complexities in Algorithm 3 for local computations and communications are achieved on **different sets of** p **and** ρ. Let us get them separately

• The local stochastic gradient complexity of a single iteration of Algorithm 3 is 0 if $\xi^k=0$, $\xi^{k+\frac{1}{2}}=1$, 1 if $\xi^k=1$, $\xi^{k+\frac{1}{2}}=1$, $r+1$ if $\xi^k=1$, $\xi^{k+\frac{1}{2}}=0$ and r if $\xi^k=0$, $\xi^{k+\frac{1}{2}}=0$. Thus, the total expected local stochastic gradient complexity is bounded by

$$
\mathcal{O}\left(((1-p)(1-\rho)+(r+1)(1-p)\rho+rp\rho)\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right)=\mathcal{O}\left((1-p+r\rho)\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right),
$$

For $\rho=\frac{1}{r}$, $p=\frac{\lambda\lambda_{\max}(W)}{\bar{L}+\lambda\lambda_{\max}(W)}$ the total expected local stochastic gradient complexity of Algorithm 3 becomes

$$
\mathcal{O}\left((1-p+r\rho)\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right)\le\mathcal{O}\left(2\frac{\sqrt{r}L_{\text{eff}}\Omega^2}{\varepsilon}\right)=\mathcal{O}\left(\frac{\sqrt{r}\bar{L}\Omega^2}{\varepsilon}\right),
$$

where $\bar{L}=L+\lambda\lambda_{\max}(W)$.

• The total communication complexity of Algorithm 3 is the sum of communication complexity coming from the full gradient computation (if statement that includes $\xi^{k+\frac{1}{2}}$) and the rest (if statement that includes ξ^k). The former requires a communication if $\xi^{k+\frac{1}{2}}=0$, the latter if ξ^k is equal to 0. The expected total communication $\mathcal{O}(\rho+p)$ per iteration. Thus, the total communication complexity is bounded by

$$
\mathcal{O}\left((p+\rho)\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right),
$$

For $\rho = p$, $p = \frac{\lambda^2\lambda^2_{\max}(W)}{\bar{L}^2+\lambda^2\lambda^2_{\max}(W)}$ the total communication complexity of Algorithm 3 becomes

$$\mathcal{O}\left((\rho+p)\frac{L_{\text{eff}}\Omega^2}{\sqrt{\rho}\varepsilon}\right) = \mathcal{O}\left(\sqrt{\rho}\frac{L_{\text{eff}}\Omega^2}{\varepsilon}\right) = \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)}{\sqrt{L^2+\lambda^2\lambda^2_{\max}(W)}}\frac{\sqrt{L^2+\lambda^2\lambda^2_{\max}(W)}\Omega^2}{\varepsilon}\right)$$
$$= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)\Omega^2}{\varepsilon}\right).$$

Strongly-convex-strongly-concave case

In the strongly-convex-strongly-concave case (Section 4.3 from [1]) the estimates on the number of iterations is

$$\mathcal{O}\left(\left(\frac{1}{\rho}+\frac{L_{\text{eff}}}{\sqrt{\rho}\mu}\right)\log\frac{1}{\varepsilon}\right),$$

• The total expected local stochastic gradient complexity is bounded by

$$\mathcal{O}\left((1-p+r\rho)\left(\frac{1}{\rho}+\frac{L_{\text{eff}}}{\sqrt{\rho}\mu}\right)\log\frac{1}{\varepsilon}\right),$$

For $\rho = \frac{1}{r}$, $p = \frac{\lambda\lambda_{\max}(W)}{\bar{L}+\lambda\lambda_{\max}(W)}$ the total expected local stochastic gradient complexity of Algorithm 3 becomes

$$\mathcal{O}\left((1-p+r\rho)\left(\frac{1}{\rho}+\frac{L_{\text{eff}}}{\sqrt{\rho}\mu}\right)\log\frac{1}{\varepsilon}\right) \leq \mathcal{O}\left(2\left(r+\frac{\sqrt{r}L_{\text{eff}}}{\mu}\right)\log\frac{1}{\varepsilon}\right) = \mathcal{O}\left(\left(r+\frac{\sqrt{r}\bar{L}}{\mu}\right)\log\frac{1}{\varepsilon}\right),$$

where $\bar{L} = \sqrt{L+\lambda\lambda_{\max}(W)}$.

• Thus, the total communication complexity is bounded by

$$\mathcal{O}\left((p+\rho)\left(\frac{1}{\rho}+\frac{L_{\text{eff}}}{\sqrt{\rho}\mu}\right)\log\frac{1}{\varepsilon}\right).$$

For $\rho = p$, $p = \frac{\lambda^2\lambda^2_{\max}(W)}{\bar{L}^2+\lambda^2\lambda^2_{\max}(W)}$ the total communication complexity of Algorithm 3 becomes

$$\mathcal{O}\left((\rho+p)\left(\frac{1}{\rho}+\frac{L_{\text{eff}}}{\sqrt{\rho}\mu}\right)\log\frac{1}{\varepsilon}\right) = \mathcal{O}\left(\left(1+\frac{\sqrt{\rho}L_{\text{eff}}}{\mu}\right)\log\frac{1}{\varepsilon}\right)$$
$$= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)}{\sqrt{L^2+\lambda^2\lambda^2_{\max}(W)}}\frac{\sqrt{L^2+\lambda^2\lambda^2_{\max}(W)}}{\mu}\log\frac{1}{\varepsilon}\right)$$
$$= \mathcal{O}\left(\frac{\lambda\lambda_{\max}(W)}{\mu}\log\frac{1}{\varepsilon}\right).$$

□