Fast decentralized gradient tracking for federated learning with local updates

author names withheld

Under Review for OPT 2024

Abstract

Federated learning (FL) for mini and minimax optimization has emerged as a powerful paradigm for training models across distributed nodes/clients while preserving data privacy and model robustness on data heterogeneity. In this work, we delve into the decentralized implementation of federated minimax optimization by proposing `K-GT-Minimax`, a novel decentralized minimax optimization algorithm that combines local updates and gradient tracking techniques. Our analysis showcases the algorithm's communication efficiency and convergence rate for nonconvex-strongly-concave (NC-SC) minimax optimization, demonstrating a superior convergence rate compared to existing methods. `K-GT-Minimax`'s ability to handle data heterogeneity and ensure robustness underscores its significance in advancing federated learning research and applications.

1. Introduction

In this paper, we delve into the realm of federated minimax optimization, focusing on a decentralized network comprising n agents tasked with optimizing the objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^{d_x}} \max_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\mathbf{x},\mathbf{y}) \equiv \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \tag{1}$$

Here, $f_i(\mathbf{x},\mathbf{y})$ represents the local function associated with client $i \in \mathcal{V} = [n] = 1,\ldots,n$, where $\mathcal{D}_i$ denotes the distribution of the data. Our focus lies on the challenging domain of *nonconvex-strongly-concave (NC-SC) minimax problems*, particularly in the context of federated learning setups, where $f(x,y)$ admits μ-strong concavity with respect to y, while each local function $f_i(x,y)$ is in expectation form indexed by random vector ξ_i. Problem (1) finds rich applications in adversarial training, distributionally robust optimization, reinforcement learning, AUC maximization, and learning with non-decomposable loss [3–5, 23].

Decentralized minimax optimization has emerged as a crucial area of research in machine learning, addressing complex optimization challenges within decentralized networks. In contrast to centralized methods, the decentralized approach facilitates efficient collaboration among agents while mitigating communication bottlenecks. Recent advancements in nonconvex minimization and minimax optimization have led to the exploration of achieving stationary points in the primal function $\Phi(\mathbf{x}) \equiv \max_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\mathbf{x},\mathbf{y})$, enhancing scalability and model robustness in decentralized optimization algorithms tailored for federated learning environments. Significant advancements have been made in addressing the unique challenges of federated minimax optimization, such as data heterogeneity, model robustness, and communication challenges.

The significance of decentralized minimax optimization spans diverse machine learning applications, showcasing its versatility and potential in improving learning efficiency and scalability. Noteworthy advancements such as variants of decentralized stochastic gradient methods have demonstrated promising results in both online and offline scenarios [2, 7, 8]. These developments underline ongoing efforts to address communication bottlenecks and enhance collaboration efficiency in decentralized optimization frameworks, paving the way for novel techniques that navigate challenging nonconvex optimization landscapes in distributed environments.

Related Work Decentralized minimax optimization, particularly in the context of federated learning setups, has gained traction for applications like adversarial training, distributionally robust optimization, and reinforcement learning [3, 5, 23]. Recent theoretical-front research focuses on ϵ-stationary points in nonconvex-strongly-concave (NC-SC) minimax problems and the importance of distributed optimization techniques for large-scale machine learning [8]. While centralized minimax optimization remains prominent, particularly in adversarial training and GANs, leveraging techniques for specific stochastic gradient complexities [9, 12, 23], decentralized minimax optimization in federated learning from convex-concave to nonconvex-(non)concave objectives still face challenges regarding decentralization and effective gradient tracking [15, 17]. Various decentralized optimization algorithms like SGDA, SREDA, GT-DA, and GT-GDA offer trade-offs in computational and communication complexities [11, 14, 20, 22]. Building upon variance-reduced minimax optimization, DREAM shows superior performance in decentralized minimax optimization, yet a comprehensive understanding of NC-SC minimax problems remains an active research area [2].

Addressing data heterogeneity is crucial in federated learning, prompting studies like FedPAGE, Federated Bose-Einstein Optimization, and Federated Learning with Decentralized Gradient Tracking [7, 18, 24]. However, these approaches often lack decentralization and may make restrictive gradient assumptions. Similarly, decentralized algorithms like K-GT [10] and LU-GT [13] primarily target minimization tasks, limiting their applicability in federated learning. Notably, algorithms like K-GT have shown promise in improving communication efficiency and robustness in federated minimax optimization tasks. It is a novel decentralized tracking mechanism that improves communication efficiency in Gradient Tracking algorithms, overcoming data heterogeneity between clients, and demonstrating model robustness in solving non-convex optimization problems, including neural network training tasks. Our work also heavily uses gradient tracking techniques, whereas our algorithm name `K-GT-Minimax` originates from K-GT proposed by Liu et al. [10] for decentralized single-agent optimization, although it should not be viewed as a straightforward generalization.

Our Contribution This paper introduces `K-GT-Minimax`, a novel decentralized minimax optimization algorithm designed specifically for federated learning environments. By combining gradient tracking and local updates, `K-GT-Minimax` addresses challenges related to data heterogeneity, model robustness, and communication efficiency. Furthermore, our algorithm demonstrates superior performance in terms of convergence rates, scalability, and stochastic gradient complexity compared to existing methods.

Notations. Throughout this paper, we use $\|\cdot\|$ for norms, $\mathbf{I}$ for identity matrices, and $\mathbf{1}_n$ for a vector of all ones. Aggregated variables are denoted by $\mathbf{x}$, $\mathbf{y}$, representing agents' local variables and gradients. We define $\mathcal{V} = [n] \equiv \{1, \dots n\}$ as the set of agents, with communication edges denoted by $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Additionally, we introduce d by n real matrix $\bar{\mathbf{X}} = [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}] = \mathbf{XJ}$ where

Algorithm 1 Gradient Tracking for Minimax Optimization (`K-GT-Minimax`)

Initialize: Communication round T; Number of local steps K; Local stepsize $\eta_c^{\mathbf{x}}, \eta_c^{\mathbf{y}}$; Communication stepsizes $\eta_s^{\mathbf{x}}, \eta_s^{\mathbf{y}}$; Mixing matrix $\mathbf{W} = (w_{ij})_{n\times n}$; $\forall i, j \in [n]$, $\mathbf{x}_i^{(0)} = \mathbf{x}_j^{(0)}$, $\mathbf{y}_i^{(0)} = \mathbf{y}_j^{(0)}$, $\mathbf{c}_i^{\mathbf{x},(0)} = -\nabla_{\mathbf{x}} F_i\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_i\right) + \frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{x}} F_j\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_j\right)$, $\mathbf{c}_i^{\mathbf{y},(0)} = -\nabla_{\mathbf{y}} F_i\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_i\right) + \frac{1}{n}\sum_{j=1}^n \nabla_{\mathbf{y}} F_j\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_j\right)$

for *client* $i \in [n]$ *parallel* **do**

  for ***communication:*** $t \leftarrow 0$ *to* $T-1$ **do**

    for ***local step:*** $k \leftarrow 0$ *to* $K-1$ **do**

      $\mathbf{x}_i^{(t)+k+1} = \mathbf{x}_i^{(t)+k} - \eta_c^{\mathbf{x}}\left(\nabla F_i(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi_i^{(t)+k}) + \mathbf{c}_i^{\mathbf{x},(t)}\right)$

      $\mathbf{y}_i^{(t)+k+1} = \mathbf{y}_i^{(t)+k} + \eta_c^{\mathbf{y}}\left(\nabla F_i(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi_i^{(t)+k}) + \mathbf{c}_i^{\mathbf{y},(t)}\right)$ ▷ variable update

    end

  $\mathbf{c}_i^{\mathbf{x},(t+1)} = \mathbf{c}_i^{\mathbf{x},(t)} + \frac{1}{K\eta_c^{\mathbf{x}}}\sum_{j=1}^n (\delta_{ij} - w_{ij})[\mathbf{x}_j^{(t)+K} - \mathbf{x}_j^{(t)}]$

  $\mathbf{c}_i^{\mathbf{y},(t+1)} = \mathbf{c}_i^{\mathbf{y},(t)} - \frac{1}{K\eta_c^{\mathbf{y}}}\sum_{j=1}^n (\delta_{ij} - w_{ij})[\mathbf{y}_j^{(t)+K} - \mathbf{y}_j^{(t)}]$ ▷ tracking variable update

  $\mathbf{x}_i^{(t+1)} = \sum_{j=1}^n w_{ij}\left(\mathbf{x}_j^{(t)} + \eta_s^{\mathbf{x}}[\mathbf{x}_i^{(t)+K} - \mathbf{x}_i^{(t)}]\right)$

  $\mathbf{y}_i^{(t+1)} = \sum_{j=1}^n w_{ij}\left(\mathbf{y}_j^{(t)} + \eta_s^{\mathbf{y}}[\mathbf{y}_i^{(t)+K} - \mathbf{y}_i^{(t)}]\right)$ ▷ model parameter update

  end

end

Output: $\mathbf{x}_{\text{out}} = \bar{\mathbf{x}}^{(\mathcal{T})} \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(\mathcal{T})}$ for randomized $\mathcal{T} \in \{1, 2, \cdots, T\}$

$\mathbf{J} \equiv \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and δ_{ij} represents the Kronecker delta with $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Other notations will be introduced at their first appearances.

2. Settings and Main Results

We propose Algorithm 1 for solving this problem in a distributed manner. To prepare for our main result, we introduce the following assumptions.

Assumption 1 (Lower Bound of $\Phi(\cdot)$) *The function $\Phi(\mathbf{x}) \equiv \max_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\mathbf{x}, \mathbf{y})$ is lower bounded, that is*

$$\Phi^* = \inf_{\mathbf{x}} \Phi(\mathbf{x}) > -\infty$$

Assumption 2 (Smoothness and Strong Concavity) *For each $i \in [n]$ let each local objective $f_i : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be twice differentiable and L-smooth for some constant $L > 0$, i.e. for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_x}$, $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_y}$*

$$\left\|\nabla f_i(\mathbf{x}, \mathbf{y}) - \nabla f_i(\mathbf{x}', \mathbf{y}')\right\|^2 \le L^2\left(\|\mathbf{x} - \mathbf{x}'\|^2 + \|\mathbf{y} - \mathbf{y}'\|^2\right)$$

Assume further $f_i(\mathbf{x}, \cdot)$ is μ-strongly concave for some shared $\mu > 0$ across all $x \in \mathbb{R}^p$, i.e. for all $\mathbf{x} \in \mathbb{R}^{d_x}$, $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_y}$

$$f_i(\mathbf{x}, \mathbf{y}') \le f_i(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y})^\top(\mathbf{y}' - \mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}' - \mathbf{y}\|^2$$

Call $\kappa \equiv L/\mu$ the condition number *of problem* (1).

Assumption 3 (Unbiaseness and Bounded Variance) *We assume that the stochastic gradients are unbiased and have bounded variance*

$$\mathbb{E}\left[\nabla F_i(\mathbf{x},\mathbf{y};\xi_i)\right] = \nabla f_i(\mathbf{x},\mathbf{y}) \qquad \mathbb{E}\|\nabla F_i(\mathbf{x},\mathbf{y};\xi_i) - \nabla f_i(\mathbf{x},\mathbf{y})\|^2 \le \sigma^2$$

Assumption 4 ($(1-p)$-Mixing) *The $n \times n$ mixing matrix $\mathbf{W}$ is symmetric, element-wise non-negative,*[1] *doubly stochastic in the sense that $\mathbf{W}\mathbf{1}_n = \mathbf{W}^\top \mathbf{1}_n = 1$, and there exists a constant $p \in [0,1]$ such that for any $\mathbf{X} \in \mathbb{R}^{d\times n}$*

$$\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\|_F^2 \le (1-p)\|\mathbf{X} - \bar{\mathbf{X}}\|_F^2$$

We are ready to present our main theorem:

Theorem 1 (Algorithm Complexity of `K-GT-Minimax`) *Let Assumptions 1, 2, 3 and 4 hold. There exists a global constant $v > 0$ such that, running `K-GT-Minimax` as in Algorithm 1 with stepsizes choice $\eta_c^{\mathbf{y}} = \frac{p}{300v\cdot\kappa KL}, \eta_c^{\mathbf{x}} = \frac{\eta_c^{\mathbf{y}}}{\kappa^2}$ and $\eta_s^{\mathbf{x}} = \eta_s^{\mathbf{y}} = v\cdot p$ gives $\mathbb{E}\|\nabla\Phi(\bar{\mathbf{x}}^{(\mathcal{T})})\|^2 \le \varepsilon^2$ for T communication rounds, each with K local updates, where*

$$T = O\left(\frac{\sigma^2}{nK}\frac{1}{\varepsilon^4} + \frac{\sigma}{p^2\sqrt{K}}\frac{1}{\varepsilon^3} + \frac{\kappa^3}{p^2}\frac{1}{\varepsilon^2}\right)\cdot L\mathscr{H}_0 \qquad K = \Omega\left(\left(1+\frac{\kappa}{\sqrt{np}}\right)\frac{\sigma}{\varepsilon}\right) \tag{2}$$

To interpret the bound in (2), note we have when $\Phi(\mathbf{x}_0) - \Phi^* = O(1)$ that $\mathscr{H}_0 = O\left(1 + \frac{1}{\mu^2 K\kappa p}\right)$. Further balancing T and K gives

$$T = O\left(\frac{\kappa^3}{p^2\varepsilon^2}\right)L\mathscr{H}_0 \qquad K = O\left(\frac{p^2\sigma^2}{\kappa^2 n\varepsilon^2} \vee \left(1+\frac{\kappa}{\sqrt{np}}\right)\frac{\sigma}{\varepsilon}\right)$$

The theorem above demonstrates how the convergence rate is affected by the accuracy parameter $\varepsilon > 0$ which ε vanishes to zero, `K-GT-Minimax` converges to an ε-stationary point within $T = O(1/\varepsilon^2)$ communication rounds, each round comprising $K = O(1/\varepsilon^2)$ local updates. Such convergence rates incorporate and balance among heterogeneity, local updates and model robustness. We list a table of comparison in Table 1.

3. Conclusion

In conclusion, this work introduces `K-GT-Minimax`, a pioneering decentralized minimax optimization algorithm for federated learning environments. By incorporating gradient tracking and local updates, `K-GT-Minimax` achieves state-of-the-art theoretical communication efficiency, showcasing its superiority over existing methods in terms of convergence rates and scalability. We hope our work addresses the critical challenges of data heterogeneity, communication efficiency and data heterogeneity robustness in federated learning setups.

1. $W_{ij} > 0$ if and only if i and j are connected.

Algorithm	Query	Communication	Decentralized	LU	DH
MLSGDA [15]	$\kappa^4/n\epsilon^4$	κ^3/ϵ^3	$\times$	✓	$\times$
SAGDA [21]	$\kappa^4/n\epsilon^4$	κ^2/ϵ^2	$\times$	✓	$\times$
Fed-Norm-SGDA [16]	$\kappa^4/n\epsilon^4$	κ^2/ϵ^2	$\times$	✓	$\times$
DM-HSGD [19]	κ^3/ϵ^3	κ^3/ϵ^3	✓	$\times$	✓
DREAM [2]	κ^3/ϵ^3	κ^2/ϵ^2	✓	$\times$	✓
`K-GT-Minimax` (This work)	$\kappa/n\epsilon^4$	κ^3/ϵ^2	✓	✓	✓

Table 1: Comparison of `K-GT-Minimax` with related algorithms for decentralized minimax optimization, highlighting stochastic gradient oracle complexity, communication rounds, decentralization, local updates (LU) and data heterogeneity (DH) robustness.

References

[1] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[2] Lesi Chen, Haishan Ye, and Luo Luo. An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2024.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[4] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[5] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[6] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[7] Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34:11422–11435, 2021.

[8] Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research*, 21(180):1–51, 2020.

[9] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[10] Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, pages 1–28, 2024.

[11] Zhuqing Liu, Xin Zhang, Songtao Lu, and Jia Liu. PRECISION: Decentralized constrained min-max learning with low communication and sample complexities. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 191–200, 2023.

[12] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

[13] Edward Duc Hien Nguyen, Sulaiman A Alghunaim, Kun Yuan, and César A Uribe. On the performance of gradient tracking with local updates. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4309–4313. IEEE, 2023.

[14] Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, Dmitry Kovalev, Pavel Dvurechensky, and Alexander Gasnikov. Decentralized saddle point problems via non-euclidean mirror prox. *Optimization Methods and Software*, pages 1–26, 2024.

[15] Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR, 2022.

[16] Pranay Sharma, Rohan Panda, and Gauri Joshi. Federated minimax optimization with client heterogeneity. *Transactions on machine learning research*, 2023.

[17] Zhenyu Sun and Ermin Wei. A communication-efficient algorithm with linear convergence for federated minimax learning. *Advances in Neural Information Processing Systems*, 35: 6060–6073, 2022.

[18] Lun Wang, Yang Xu, Hongli Xu, Zhida Jiang, Min Chen, Wuyang Zhang, and Chen Qian. BOSE: Block-wise federated learning in heterogeneous edge computing. *IEEE/ACM Transactions on Networking*, 2023.

[19] Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34: 25865–25877, 2021.

[20] Yangyang Xu. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *SIAM Journal on Optimization*, 34(1):1006–1044, 2024.

[21] Haibo Yang, Zhuqing Liu, Xin Zhang, and Jia Liu. SAGDA: Achieving $\mathcal{O}(\epsilon^{-2})$ communication complexity in federated min-max learning. *Advances in Neural Information Processing Systems*, 35:7142–7154, 2022.

[22] Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.

[23] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

[24] Guangmeng Zhou, Qi Li, Yang Liu, Yi Zhao, Qi Tan, Su Yao, and Ke Xu. FedPAGE: Pruning adaptively toward global efficiency of heterogeneous federated learning. *IEEE/ACM Transactions on Networking*, 2023.

Appendix A. Proof of Main Theorem

To prepare for the proof, we introduce some notions.

- The *client variance* quantifies how much variables $\mathbf{x}$ and $\mathbf{y}$ deviate from its average model across global steps:

$$\Xi_t^{\mathbf{x}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2$$

and

$$\Xi_t^{\mathbf{y}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{y}_i^{(t)} - \bar{\mathbf{y}}^{(t)}\|^2$$

- The *client drift* quantifies how variables $\mathbf{x}$ and $\mathbf{y}$ deviate from its averaged model across local steps:

$$e_{k,t}^{\mathbf{x}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}\|^2 \qquad \mathcal{E}_t^{\mathbf{x}} \equiv \sum_{k=0}^{K-1} e_{k,t}^{\mathbf{x}}$$

and

$$e_{k,t}^{\mathbf{y}} \equiv \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|\mathbf{y}_j^{(t)+k} - \bar{\mathbf{y}}^{(t)}\|^2 \qquad \mathcal{E}_t^{\mathbf{y}} \equiv \sum_{k=0}^{K-1} e_{k,t}^{\mathbf{y}}$$

where $\mathcal{E}_t^{\mathbf{x}}$ characterizes the accumulation of local steps for variable $\mathbf{x}$ and analogously $\mathcal{E}_t^{\mathbf{y}}$ for $\mathbf{y}$.

- The *quality of correction* that assesses the accuracy of the gradient correction across local steps, aiming to closely align local updates with global updates:

$$\gamma_t^{\mathbf{x}} = \frac{1}{nL^2}\mathbb{E}\left\|\mathbf{C}^{x,(t)} + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right) - \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\mathbf{J}\right\|_F^2$$

and

$$\gamma_t^{\mathbf{y}} = \frac{1}{nL^2}\mathbb{E}\left\|\mathbf{C}^{y,(t)} + \nabla_{\mathbf{y}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right) - \nabla_{\mathbf{y}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\mathbf{J}\right\|_F^2$$

- The *consensus distance for variable* $\mathbf{y}$ which quantifies the difference between the averaged value $\bar{\mathbf{y}}^{(t)}$ and the optimal value (when $\mathbf{x}$ equals its average $\bar{\mathbf{x}}^{(t)}$) of $\hat{\mathbf{y}}^{(t)} = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\bar{\mathbf{x}}^{(t)}, \mathbf{y})$:

$$\varepsilon_t = \|\bar{\mathbf{y}}^{(t)} - \hat{\mathbf{y}}^{(t)}\|^2$$

With these notions we present recursion bounds for the aforementioned client variance and client drift, along with the quality of correction, for both variables $\mathbf{x}$ and $\mathbf{y}$. We finally discuss the consensus distance specifically for variable $\mathbf{y}$.

We first bound the local drift for variables $\mathbf{x}$ and $\mathbf{y}$ as

Lemma 2 *Suppose* $\eta_c^{\mathbf{x}}, \eta_c^{\mathbf{y}} \le \frac{1}{8KL}$ *we have*

$$\mathcal{E}_t^{\mathbf{x}} \le 3K\Xi_t^{\mathbf{x}} + 12K^2(\eta_c^{\mathbf{x}})^2L^2\mathcal{E}_t^{\mathbf{y}} + 12K^3(\eta_c^{\mathbf{x}})^2L^2\gamma_t^{\mathbf{x}} + 12K^3(\eta_c^{\mathbf{x}})^2L^2\varepsilon_t + 12K^3(\eta_c^{\mathbf{x}})^2\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + 3K^2(\eta_c^{\mathbf{x}})^2\sigma^2$$

and

$$\mathcal{E}_t^{\mathbf{y}} \le 3K\Xi_t^{\mathbf{y}} + 12K^2(\eta_c^{\mathbf{y}})^2L^2\mathcal{E}_t^{\mathbf{x}} + 12K^3(\eta_c^{\mathbf{y}})^2L^2\gamma_t^{\mathbf{y}} + 6K^3(\eta_c^{\mathbf{y}})^2L^2\varepsilon_t + 3K^2(\eta_c^{\mathbf{y}})^2\sigma^2$$

Set $\eta^{\mathbf{x}} \equiv \eta_s^{\mathbf{x}}\eta_c^{\mathbf{x}}$ and $\eta^{\mathbf{y}} \equiv \eta_s^{\mathbf{y}}\eta_c^{\mathbf{y}}$. We now bound the client variance for variables $\mathbf{x}$ and $\mathbf{y}$, as follows

Lemma 3 *We have*

$$\Xi_{t+1}^{\mathbf{x}} \leq \left(1-\frac{p}{2}\right)\Xi_t^{\mathbf{x}} + \frac{6K(\eta^{\mathbf{x}})^2L^2}{p}\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + \frac{6K^2(\eta^{\mathbf{x}})^2L^2}{p}\gamma_t^{\mathbf{x}} + K(\eta^{\mathbf{x}})^2\sigma^2$$

$$\Xi_{t+1}^{\mathbf{y}} \leq \left(1-\frac{p}{2}\right)\Xi_t^{\mathbf{y}} + \frac{6K(\eta^{\mathbf{y}})^2L^2}{p}\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + \frac{6K^2(\eta^{\mathbf{y}})^2L^2}{p}\gamma_t^{\mathbf{y}} + K(\eta^{\mathbf{y}})^2\sigma^2$$

In the upcoming we bound the quality of correction for variables $\mathbf{x}$ and $\mathbf{y}$

Lemma 4 *Suppose $\eta^{\mathbf{x}}, \eta^{\mathbf{y}} \leq \frac{\sqrt{p}}{2\sqrt{6}KL}$ we have*

$$\gamma_{t+1}^{\mathbf{x}} \leq \left(1-\frac{p}{2}\right)\gamma_t^{\mathbf{x}} + \frac{30}{pK}\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + \frac{12K^2L^2}{p}\left(2(\eta^{\mathbf{x}})^2+(\eta^{\mathbf{y}})^2\right)\varepsilon_t + \frac{24K^2(\eta^{\mathbf{x}})^2}{p}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{2\sigma^2}{KL^2} \tag{3}$$

$$\gamma_{t+1}^{\mathbf{y}} \leq \left(1-\frac{p}{2}\right)\gamma_t^{\mathbf{y}} + \frac{30}{pK}\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + \frac{12K^2L^2}{p}\left(2(\eta^{\mathbf{x}})^2+(\eta^{\mathbf{y}})^2\right)\varepsilon_t + \frac{24K^2(\eta^{\mathbf{x}})^2}{p}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{2\sigma^2}{KL^2} \tag{4}$$

In the following we bound on the consensus distance for variable $\mathbf{y}$

Lemma 5 *Suppose $\eta^{\mathbf{x}} \leq \frac{\eta^{\mathbf{y}}}{4\sqrt{6}\kappa^2}$ and $\eta^{\mathbf{y}} \leq \frac{1}{KL}$ we have*

$$\varepsilon_{t+1} \leq \left(1-\frac{K\eta^{\mathbf{y}}L}{6\kappa}\right)\varepsilon_t + 12\eta^{\mathbf{y}}L\kappa\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + \frac{16\kappa^3K(\eta^{\mathbf{x}})^2}{\eta^{\mathbf{y}}L}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{8\eta^{\mathbf{y}}\kappa}{nL}\sigma^2$$

We further have the following bound on $\mathbb{E}\Phi(\bar{\mathbf{x}}^{(t+1)})$

Lemma 6 *Suppose $\eta^{\mathbf{x}} \leq \frac{1}{16KL\kappa}$ we have the following*

$$\mathbb{E}\left[\Phi(\bar{\mathbf{x}}^{(t+1)}) - \Phi(\bar{\mathbf{x}}^{(t)})\right] \leq -\frac{\eta^{\mathbf{x}}K}{4}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + 2\eta^{\mathbf{x}}L^2\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) + 2L^2\eta^{\mathbf{x}}K\varepsilon_t + \frac{K(\eta^{\mathbf{x}})^2L\kappa}{n}\sigma^2$$

Let $v > 1$ be a global constant to be determined in our upcoming Lyapunov analysis. In light of Lemma 6 we set the Lyapunov function as

$$\mathscr{H}_t = \mathbb{E}\left[\Phi(\bar{\mathbf{x}}^{(t)}) - \Phi(\mathbf{x}^*)\right] + B^{\mathbf{x}}\eta_c^{\mathbf{y}}L\Xi_t^{\mathbf{x}} + B^{\mathbf{y}}\eta_c^{\mathbf{y}}L\Xi_t^{\mathbf{y}} + A^{\mathbf{x}}K^2L^3(\eta_c^{\mathbf{y}})^3\gamma_t^{\mathbf{x}} + A^{\mathbf{y}}K^2L^3(\eta_c^{\mathbf{y}})^3\gamma_t^{\mathbf{y}} + C\frac{1}{K\kappa p}\varepsilon_t \tag{5}$$

Then we have the following recursion for (5):

Lemma 7 *Suppose $\eta_c^{\mathbf{x}} = \frac{\eta_c^{\mathbf{y}}}{\kappa^2}$ with $\eta_c^{\mathbf{y}} = \frac{p}{300v\cdot\kappa KL}$, $\eta_s^{\mathbf{x}} = \eta_s^{\mathbf{y}} = v\cdot p$, then one can choose global constant $v > 1$ such that $A^{\mathbf{x}} = A^{\mathbf{y}} = \frac{108v^3+54v}{p}$, $B^{\mathbf{x}} = B^{\mathbf{y}} = \frac{6v}{p}$ and $C = \frac{1}{24}$ so as to ensure $C_4^{\mathbf{x}} = \Theta(1)$ and $C_4^{\mathbf{y}} = O(1)$. We have*

$$\mathscr{H}_{t+1} - \mathscr{H}_t \leq -C_4^{\mathbf{x}}K\eta^{\mathbf{x}}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{C_4^{\mathbf{y}}}{p}(KL)^2(\eta_c^{\mathbf{y}})^3\sigma^2 + \frac{K(\eta^{\mathbf{x}})^2L\kappa}{n}\sigma^2 + C\frac{8\eta^{\mathbf{y}}}{nLKp}\sigma^2 \tag{6}$$

With all preliminary lemmas at hand we are ready for the final proof of our main theorem.

Proof of Theorem 1. Taking telescoping sum on both sides of (6) for $t = 0, 1, \ldots, T-1$ gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathscr{H}_{t+1}-\mathscr{H}_t\right) = \frac{1}{T}\left(\mathscr{H}_T - \mathscr{H}_0\right) \\
&\leq -C_4^{\mathbf{x}} K\eta^{\mathbf{x}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{C_4^{\mathbf{y}}}{p}(KL)^2(\eta_c^{\mathbf{x}})^3\sigma^2 + \frac{K(\eta^{\mathbf{x}})^2 L\kappa}{n}\sigma^2 + C\frac{8\eta^{\mathbf{y}}}{nLKp}\sigma^2
\end{aligned}$$

yielding

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 &\leq \frac{\mathscr{H}_0 - \mathscr{H}_T}{TC_4^{\mathbf{x}}}\frac{1}{K\eta^{\mathbf{x}}} + \frac{\eta^{\mathbf{x}}L\kappa}{nC_4^{\mathbf{x}}}\sigma^2 + \frac{\frac{C_4^{\mathbf{y}}}{p}KL^2(\eta_c^{\mathbf{x}})^3}{C_4^{\mathbf{x}}\eta^{\mathbf{x}}}\sigma^2 + C\frac{8\eta^{\mathbf{y}}}{nLC_4^{\mathbf{x}}K^2p\eta^{\mathbf{x}}}\sigma^2 \\
&\leq \frac{\mathscr{H}_0}{TC_4^{\mathbf{x}}}\frac{1}{K\eta^{\mathbf{x}}} + \frac{\eta^{\mathbf{x}}L\kappa}{nC_4^{\mathbf{x}}}\sigma^2 + \frac{C_4^{\mathbf{y}}KL^2(\eta^{\mathbf{x}})^2}{C_4^{\mathbf{x}}v^3p^4}\sigma^2 + \frac{8C\eta^{\mathbf{y}}}{nLC_4^{\mathbf{x}}K^2p\eta^{\mathbf{x}}}\sigma^2
\end{aligned} \tag{7}$$

Given desired accuracy $\varepsilon > 0$ since we want the expected squared gradient norm of the randomized output $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \leq 4\varepsilon^2$ we need to calibrate given the choice of $C_4^{\mathbf{x}}$, $C_4^{\mathbf{y}}$, $\eta^{\mathbf{x}}$ and $\eta^{\mathbf{y}}$. We make $K \gtrsim \frac{\kappa}{\sqrt{np}}\frac{\sigma}{\varepsilon}$ so the last term is bounded by ε^2. Using stepsize tuning lemma exemplified by Koloskova et al. [6, Lemma 17] guarantees the existence of constant stepsize such that the average of accumulation of gradient is upper bounded by

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2 L\mathscr{H}_0}{nKT}} + \left(\frac{\sigma L\mathscr{H}_0}{p^2\sqrt{K}T}\right)^{\frac{2}{3}} + \frac{\kappa^3 L\mathscr{H}_0}{p^2 T}\right)$$

so the complexity of communication T given local steps K is $\mathcal{O}\left(\frac{\sigma^2}{nK}\frac{1}{\varepsilon^4} + \frac{\sigma}{p^2\sqrt{K}}\frac{1}{\varepsilon^3} + \frac{\kappa^3}{p^2}\frac{1}{\varepsilon^2}\right)\cdot L\mathscr{H}_0$. Here since the initialization is shared across clients we have $\mathbf{x}^{(0)} = \mathbf{x}_i^{(0)}$ for all $i \in [n]$ and the way correction terms $\mathbf{c}_i^{\mathbf{x},(0)}, \mathbf{c}_i^{\mathbf{y},(0)}$ are defined ensures $\mathscr{H}_0 = O\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star) + \frac{\varepsilon_0}{K\kappa p}\right)$ and $\varepsilon_0 = O\left(\frac{q}{\mu^2}\right)$.

Appendix B. Deferred Auxiliary Proofs

B.1. Proof of Lemma 2

Lemma 8 *Using Assumption 2 and Young's Inequality we have*

$$\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 \leq 2L^2\varepsilon_t + 2\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \qquad \mathbb{E}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 \leq L^2\varepsilon_t$$

Proof [Proof of Lemma 8] We can write

$$\begin{aligned}
\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 &= \mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right) - \nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \hat{\mathbf{y}}^{(t)}\right) + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \hat{\mathbf{y}}^{(t)}\right)\right\|^2 \\
&\leq 2L^2\mathbb{E}\left\|\bar{\mathbf{y}}^{(t)} - \hat{\mathbf{y}}^{(t)}\right\|^2 + 2\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 = 2L^2\varepsilon_t + 2\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2
\end{aligned}$$

Moreover,

$$\mathbb{E}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 = \mathbb{E}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right) - \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \hat{\mathbf{y}}^{(t)}\right)\right\|^2 \leq L^2 \varepsilon_t \tag{8}$$

The equality in (8) holds due to the fact that $\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \hat{\mathbf{y}}^{(t)}\right) = 0$. ■

Proof [Proof of Lemma 2] For $K = 1$ the inequalities obviously hold since $\mathcal{E}_t^{\mathbf{x}} = \Xi_t^{\mathbf{x}} = \frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2$ and $\mathcal{E}_t^{\mathbf{y}} = \Xi_t^{\mathbf{y}} = \frac{1}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)} - \bar{\mathbf{Y}}^{(t)}\right\|_F^2$ and other terms on the RHSs are positive. For $K \geq 2$ we have

$$\begin{aligned}
&ne_{k,t}^{\mathbf{x}} \equiv \mathbb{E}\left\|\mathbf{X}^{(t)+k} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 \\
&= \mathbb{E}\left\|\mathbf{X}^{(t)+k-1} - \eta_c^{\mathbf{x}}\left(\nabla_{\mathbf{x}} F\left(\mathbf{X}^{(t)+k-1}, \mathbf{Y}^{(t)+k-1}; \xi^{(t)+k-1}\right) + \mathbf{C}^{x,(t)}\right) - \bar{\mathbf{X}}^{(t)}\right\|_F^2 \\
&\leq \left(1 + \frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{X}^{(t)+k-1} - \bar{\mathbf{X}}^{(t)}\right\|^2 + n(\eta_c^{\mathbf{x}})^2\sigma^2 \\
&\quad + K(\eta_c^{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\mathbf{X}^{(t)+k-1}, \mathbf{Y}^{(t)+k-1}\right) - \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right) + \mathbf{C}^{x,(t)}\right. \\
&\qquad\qquad \left. + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)(\mathbf{I} - \mathbf{J}) + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\mathbf{J}\right\|_F^2 \\
&\leq \underbrace{\left(1 + \frac{1}{K-1} + 4K(\eta_c^{\mathbf{x}})^2L^2\right)}_{\equiv q}\mathbb{E}\left\|\mathbf{X}^{(t)+k-1} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + 4K(\eta_c^{\mathbf{x}})^2L^2\mathbb{E}\left\|\mathbf{Y}^{(t)+k-1} - \bar{\mathbf{Y}}^{(t)}\right\|_F^2 \\
&\quad + 4K(\eta_c^{\mathbf{x}})^2L^2n\gamma_t^{\mathbf{x}} + 2K(\eta_c^{\mathbf{x}})^2n\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 + n(\eta_c^{\mathbf{x}})^2\sigma^2 \\
&\leq q^k\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 \\
&\quad + \sum_{r=0}^{k-1} q^r\left(4K(\eta_c^{\mathbf{x}})^2L^2\mathbb{E}\left\|\mathbf{Y}^{(t)+k-1} - \bar{\mathbf{Y}}^{(t)}\right\|_F^2 + 4K(\eta_c^{\mathbf{x}})^2L^2n\gamma_t^{\mathbf{x}} + 2K(\eta_c^{\mathbf{x}})^2n\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 + n(\eta_c^{\mathbf{x}})^2\sigma^2\right)
\end{aligned}$$

If the condition $\eta_c^{\mathbf{x}} \leq \frac{1}{8KL}$ holds, then it follows that $4K\left(\eta_c^{\mathbf{x}}L\right)^2 \leq \frac{1}{16K} < \frac{1}{16(K-1)}$. Given $q > 1$, it can be established that $q^k \leq q^K \leq \left(1 + \frac{1}{K-1} + \frac{1}{16(K-1)}\right)^K \leq e^{1+\frac{1}{16}} \leq 3$, and $\sum_r^{k-1} q^r \leq Kq^K \leq 3K$. Now, we can obtain a bound on client drift for variable $\mathbf{x}$

$$\mathcal{E}_t^{\mathbf{x}} = \sum_{k=0}^{K-1} e_{k,t}^{\mathbf{x}} \leq 3K\Xi_t^{\mathbf{x}} + 12K^2(\eta_c^{\mathbf{x}})^2L^2\mathcal{E}_t^{\mathbf{y}} + 12K^3(\eta_c^{\mathbf{x}})^2L^2\gamma_t^{\mathbf{x}} + 6K^3(\eta_c^{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 + 3K^2(\eta_c^{\mathbf{x}})^2\sigma^2 \tag{9}$$

Similarly, a bound on client drift for variable $\mathbf{y}$ can be formulated by

$$\mathcal{E}_t^{\mathbf{y}} = \sum_{k=0}^{K-1} e_{k,t}^{\mathbf{y}} \leq 3K\Xi_t^{\mathbf{y}} + 12K^2(\eta_c^{\mathbf{y}})^2L^2\mathcal{E}_t^{\mathbf{x}} + 12K^3(\eta_c^{\mathbf{y}})^2L^2\gamma_t^{\mathbf{y}} + 6K^3(\eta_c^{\mathbf{y}})^2\mathbb{E}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}\right)\right\|^2 + 3K^2(\eta_c^{\mathbf{y}})^2\sigma^2 \tag{10}$$

Using Lemma 8 in (9) and (10) will complete the proof. ■

B.2. Proof of Lemma 3

Proof [Proof of Lemma 3] Using the update rule from Algorithm 1, we can bound the client variance for variable $\mathbf{x}$

$$
\begin{aligned}
n\Xi_{t+1}^{\mathbf{x}} &= \mathbb{E}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\right\|^2 \\
&= \mathbb{E}\left\|\left(\mathbf{X}^{(t)} - \eta^{\mathbf{x}}\sum_{k=0}^{K-1}\left(\nabla_{\mathbf{x}} F\left(\mathbf{X}^{(t)+k}, \mathbf{Y}^{(t)+k}; \xi^{(t)+k}\right) + \mathbf{C}^{x,(t)}\right)\right)(\mathbf{W}-\mathbf{J})\right\|_F^2 \\
&\overset{(a)}{\leq} (1-p)\mathbb{E}\left\|\left(\mathbf{X}^{(t)} - \eta^{\mathbf{x}}\sum_{k=0}^{K-1}\left(\nabla_{\mathbf{x}} f\left(\mathbf{X}^{(t)+k}, \mathbf{Y}^{(t)+k}\right) + \mathbf{C}^{x,(t)}\right)\right)(\mathbf{I}-\mathbf{J})\right\|_F^2 + nK(\eta^{\mathbf{x}})^2\sigma^2 \\
&\leq nK(\eta^{\mathbf{x}})^2\sigma^2 + (1+\alpha)(1-p)\mathbb{E}\left\|\mathbf{X}^{(t)}(\mathbf{I}-\mathbf{J})\right\|_F^2 \\
&\quad + \left(1+\frac{1}{\alpha}\right)(\eta^{\mathbf{x}})^2\mathbb{E}\left\|\left[\sum_{k=0}^{K-1}\nabla_{\mathbf{x}} f\left(\mathbf{X}^{(t)+k}, \mathbf{Y}^{(t)+k}\right) - K\nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right) + K\nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\right](\mathbf{I}-\mathbf{J}) + K\mathbf{C}^{x,(t)}\right\|_F^2 \\
&\overset{(b)}{\leq} nK(\eta^{\mathbf{x}})^2\sigma^2 + \left(1-\frac{p}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 \\
&\quad + \frac{6}{p}\left(K(\eta^{\mathbf{x}})^2L^2\|\mathbf{I}-\mathbf{J}\|^2\left(\sum_{k=0}^{K-1}\left\|\mathbf{X}^{(t)+k} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{Y}^{(t)+k} - \bar{\mathbf{Y}}^{(t)}\right\|_F^2\right)\right. \\
&\qquad \left. + K^2(\eta^{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)(\mathbf{I}-\mathbf{J}) + \mathbf{C}^{x,(t)}\right\|_F^2\right) \\
&\leq \left(1-\frac{p}{2}\right)n\Xi_t^{\mathbf{x}} + \frac{6K(\eta^{\mathbf{x}})^2L^2}{p}n\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{6K^2(\eta^{\mathbf{x}})^2L^2}{p}n\gamma_t^{\mathbf{x}} + nK(\eta^{\mathbf{x}})^2\sigma^2
\end{aligned}
$$

where we used Assumption 4 in (a) and $\alpha = \frac{p}{2}, p \leq 1$ in (b). Similarly, we can derive an upper bound on client variance for variable $\mathbf{y}$, thereby concluding the proof. ■

B.3. Proof of Lemma 4

Lemma 9 *If we initialize $\mathbf{C}^{x,(0)}$ and $\mathbf{C}^{y,(0)}$ as below*

$$
\begin{aligned}
\mathbf{c}_i^{\mathbf{x},(0)} &= -\nabla_{\mathbf{x}} F_i\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_i\right) + \frac{1}{n}\sum_{j=1}^{n}\nabla_{\mathbf{x}} F_j\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_j\right) \\
\mathbf{c}_i^{\mathbf{y},(0)} &= -\nabla_{\mathbf{y}} F_i\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_i\right) + \frac{1}{n}\sum_{j=1}^{n}\nabla_{\mathbf{y}} F_j\left(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}; \xi_j\right)
\end{aligned}
\tag{11}
$$

then the averaged correction for variables $\mathbf{x}$ and $\mathbf{y}$ in any communication round equals to zero.

Proof [Proof of Lemma 9] According to Algorithm 1 we have

$$
\mathbf{C}^{x,(t+1)}\mathbf{J} = \mathbf{C}^{x,(t)}\mathbf{J} + \frac{1}{K\eta_c^{\mathbf{x}}}\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K}\right)(\mathbf{W}-\mathbf{I})\mathbf{J} = \mathbf{C}^{x,(t)}\mathbf{J}
$$

Using the initialization assumption in (11), we have $\mathbf{C}^{x,(t)}\mathbf{J} = \mathbf{C}^{x,(0)}\mathbf{J} = \mathbf{0}$. Similarly, we have $\mathbf{C}^{y,(t)}\mathbf{J} = \mathbf{C}^{y,(0)}\mathbf{J} = \mathbf{0}$. ■

Let $\Delta_{t+1}^{\mathbf{x}} \equiv \mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|^2$, $\Delta_{t+1}^{\mathbf{y}} \equiv \mathbb{E}\left\|\bar{\mathbf{y}}^{(t+1)} - \bar{\mathbf{y}}^{(t)}\right\|^2$. We have

Lemma 10 *The sum of averaged progress between communications for variables* $\mathbf{x}$ *and* $\mathbf{y}$ *can be bounded by*

$$\Delta_{t+1}^{\mathbf{x}} + \Delta_{t+1}^{\mathbf{y}} \leq 2KL^2\left((\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 2K^2L^2\left(2(\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)\varepsilon_t + 4K^2(\eta^{\mathbf{x}})^2\mathbb{E}\left\|\nabla\Phi\left(\bar{x}^{(t)}\right)\right\|^2 + \frac{K\sigma^2}{n}\left((\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)$$

Proof [Proof of Lemma 10] First, we derive an upper bound on the averaged progress for variable $\mathbf{x}$ as follows

$$\begin{aligned}
\Delta_{t+1}^{\mathbf{x}} &= \mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|^2 = (\eta^{\mathbf{x}})^2\mathbb{E}\left\|\frac{1}{n}\sum_{i,k}\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi^{(t)+k}\right) + \frac{K}{n}\sum_i \mathbf{c}_i^{\mathbf{x},(t)}\right\|^2 \\
&\overset{(a)}{\leq} \frac{2K(\eta^{\mathbf{x}})^2}{n}\sum_{i,k}\mathbb{E}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}\right) - \nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}_i^{(t)}, \bar{\mathbf{y}}_i^{(t)}\right)\right\|^2 + 2K^2(\eta^{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_i^{(t)}, \bar{\mathbf{y}}_i^{(t)}\right)\right\|^2 + \frac{K\eta_x^2\sigma^2}{n} \\
&\leq \frac{2K(\eta^{\mathbf{x}})^2L^2}{n}\sum_{i,k}\left(\mathbb{E}\left\|\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \mathbb{E}\left\|\mathbf{y}_i{}^{(t)+k} - \bar{\mathbf{y}}^{(t)}\right\|^2\right) + 2K^2(\eta^{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_i{}^{(t)}, \bar{\mathbf{y}}_i^{(t)}\right)\right\|^2 + \frac{K\eta_x^2\sigma^2}{n} \\
&\overset{(b)}{\leq} 2K(\eta^{\mathbf{x}})^2L^2\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 2K^2(\eta^{\mathbf{x}})^2\left(2L^2\varepsilon_t + 2\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2\right) + \frac{K\eta_x^2\sigma^2}{n}
\end{aligned} \tag{12}$$

Similar to the above derivations, we have

$$\begin{aligned}
\Delta_{t+1}^{\mathbf{y}} &= \mathbb{E}\left\|\bar{\mathbf{y}}^{(t+1)} - \bar{\mathbf{y}}^{(t)}\right\|^2 \leq 2K^2(\eta^{\mathbf{y}})^2L^2\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 2K^2(\eta^{\mathbf{y}})^2\mathbb{E}\left\|\nabla_{\mathbf{y}}f\left(\bar{\mathbf{x}}_i^{(t)}, \bar{\mathbf{y}}_i^{(t)}\right)\right\|^2 + \frac{K(\eta^{\mathbf{y}})^2\sigma^2}{n} \\
&\overset{(c)}{\leq} 2K(\eta^{\mathbf{y}})^2L^2\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 2K^2(\eta^{\mathbf{y}})^2L^2\varepsilon_t + \frac{K(\eta^{\mathbf{y}})^2\sigma^2}{n}
\end{aligned} \tag{13}$$

We used Lemma 9, 8, and 8 in (a), (b), and (c), respectively. Combining (12) and (13) completes the proof. ■

Proof [Proof of Lemma 4] We have

$$
\begin{aligned}
&nL^2\gamma_{t+1}^{\mathbf{x}} - \frac{n\sigma^2}{K} \\
&\equiv \mathbb{E}\left\|\mathbf{C}^{x,(t+1)} + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t+1)}, \bar{\mathbf{Y}}^{(t+1)}\right)(\mathbf{I}-\mathbf{J})\right\|_F^2 - \frac{n\sigma^2}{K} \\
&= \mathbb{E}\left\|\mathbf{C}^{x,(t)}\mathbf{W} + \frac{1}{K}\sum_{k=0}^{K-1}\nabla_{\mathbf{x}} F\left(\mathbf{X}^{(t)+k}, \mathbf{Y}^{(t)+k}; \xi^{(t)+k}\right)(\mathbf{W}-\mathbf{I}) + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t+1)}, \bar{\mathbf{Y}}^{(t+1)}\right)(\mathbf{I}-\mathbf{J})\right\|_F^2 - \frac{n\sigma^2}{K} \\
&\le \mathbb{E}\left\|\left(\mathbf{C}^{x,(t)} + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)(\mathbf{I}-\mathbf{J})\right)\mathbf{W} + \left(\frac{1}{K}\sum_{k=0}^{K-1}\nabla_{\mathbf{x}} f\left(\mathbf{X}^{(t)+k}, \mathbf{Y}^{(t)+k}\right) - \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\right)(\mathbf{W}-\mathbf{I})\right. \\
&\qquad \left. + \left(\nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t+1)}, \bar{\mathbf{Y}}^{(t+1)}\right) - \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)\right)(\mathbf{I}-\mathbf{J})\right\|_F^2 \\
&\overset{(a)}{\le} (1+\alpha)(1-p)nL^2\gamma_t^{\mathbf{x}} \\
&\quad + 2\left(1+\frac{1}{\alpha}\right)\left[\|\mathbf{W}-\mathbf{I}\|^2\frac{L^2}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}\left\|\mathbf{X}^{(t)+k} - \bar{\mathbf{X}}^{(t)}\right\|^2 + \mathbb{E}\left\|\mathbf{Y}^{(t)+k} - \bar{\mathbf{Y}}^{(t)}\right\|^2\right)\right. \\
&\qquad \left. + \|\mathbf{I}-\mathbf{J}\|^2 nL^2\left(\mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \mathbb{E}\left\|\bar{\mathbf{y}}^{(t+1)} - \bar{\mathbf{y}}^{(t)}\right\|^2\right)\right] \\
&\overset{(b)}{\le} \left(1-\frac{p}{2}\right)nL^2\gamma_t^{\mathbf{x}} + \frac{6}{p}\left(\frac{4L^2 n}{K}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + nL^2\left(\Delta_{t+1}^{\mathbf{x}} + \Delta_{t+1}^{\mathbf{y}}\right)\right)
\end{aligned}
$$

where in (b) we applied $\alpha = \frac{p}{2}, \frac{1}{p} \ge 1$, and in (a) we applied Assumption 4 and the fact that

$$
\left(\mathbf{C}^{x,(t)} + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)(\mathbf{I}-\mathbf{J})\right)\mathbf{J} = \mathbf{C}^{x,(t)}\mathbf{J} + \nabla_{\mathbf{x}} f\left(\bar{\mathbf{X}}^{(t)}, \bar{\mathbf{Y}}^{(t)}\right)(\mathbf{J}-\mathbf{J}) = \mathbf{0}
$$

where in the last equality we used Lemma 9. Using Lemma 10 to bound $\Delta_{t+1}^{\mathbf{x}} + \Delta_{t+1}^{\mathbf{y}}$ we have

$$
\begin{aligned}
\gamma_{t+1}^{\mathbf{x}} &\le \left(1-\frac{p}{2}\right)\gamma_t^{\mathbf{x}} + \frac{1}{p}\left(\frac{24}{K} + 12K(\eta^{\mathbf{x}})^2L^2 + 12K(\eta^{\mathbf{y}})^2L^2\right)\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) \\
&\quad + \frac{12K^2L^2}{p}\left(2(\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)\varepsilon_t + \frac{24K^2(\eta^{\mathbf{x}})^2}{p}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{6K\sigma^2\left((\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)}{np} + \frac{\sigma^2}{KL^2}
\end{aligned}
$$

Applying the conditions on the step sizes will result in (3). In a similar fashion, we can show (4). ■

B.4. Proof of Lemma 5

Lemma 11 *Using Proposition 13 and assuming that $\eta^{\mathbf{y}} \le \frac{1}{KL}$, we have the following bound on $\mathbb{E}\left\|\hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t+1)}\right\|^2$ for any $\alpha > 0$:*

$$
\mathbb{E}\left\|\hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t+1)}\right\|^2 \le (1+\alpha)\left(1 - K\eta^{\mathbf{y}}\mu\right)\varepsilon_t + \left(1+\frac{1}{\alpha}\right)(\eta^{\mathbf{y}})^2L^2K\left(\mathcal{E}^{\mathbf{x}} + \mathcal{E}^{\mathbf{y}}\right) + \frac{K\eta_y^2\sigma^2}{n}
$$

Proof [Proof of Lemma 11] If we replace $\mathbf{x} = \bar{\mathbf{x}}^{(t)}$, $\mathbf{y} = \bar{\mathbf{y}}^{(t)}$, and $\mathbf{y}' = \hat{\mathbf{y}}^{(t)}$ in Proposition 13, we have

$$\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)})^\top (\bar{\mathbf{y}}^{(t)} - \hat{\mathbf{y}}^{(t)}) + \frac{1}{2L} \left\| \nabla_{\mathbf{y}} f(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)}) \right\|^2 + \frac{\mu}{2} \|\bar{\mathbf{y}}^{(t)} - \hat{\mathbf{y}}^{(t)}\|^2 \leq 0 \tag{14}$$

We can also write that

$$\begin{aligned}
&\mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} - K\eta^{\mathbf{y}} \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \\
&= \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|^2 - 2K\eta y \mathbb{E} \left\langle \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)}, \nabla y f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\rangle + K^2 (\eta^{\mathbf{y}})^2 \mathbb{E} \left\| \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \\
&= \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|^2 + 2K\eta^{\mathbf{y}} \left(\mathbb{E} \left\langle \bar{\mathbf{y}}^{(t)} - \hat{\mathbf{y}}(t), \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\rangle + \frac{K\eta^{\mathbf{y}}}{2} \mathbb{E} \left\| \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \right) \\
&\overset{(a)}{\leq} \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|^2 + 2K\eta^{\mathbf{y}} \left(-\frac{\mu}{2} \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|^2 \right) = (1 - K\eta^{\mathbf{y}} \mu)\, \varepsilon_t
\end{aligned}$$

In (a), we used the assumption that $\eta^{\mathbf{y}} \leq \frac{1}{KL}$ and (14). Now, we can write

$$\begin{aligned}
&\mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t+1)} \right\|^2 - \frac{K\eta_y^2 \sigma^2}{n} \overset{(b)}{=} \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} - \frac{\eta^{\mathbf{y}}}{n} \sum_{i,k} \nabla_{\mathbf{y}} F_i \left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi^{(t)+k} \right) \right\|^2 - \frac{K\eta_y^2 \sigma^2}{n} \\
&\leq \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} - K\eta^{\mathbf{y}} \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) - \frac{\eta^{\mathbf{y}}}{n} \sum_{i,k} \nabla_{\mathbf{y}} f_i \left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k} \right) + \frac{\eta^{\mathbf{y}}}{n} \sum_{i,k} \nabla_{\mathbf{y}} f_i \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \\
&\leq (1+\alpha) \mathbb{E} \left\| \hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t)} - K\eta^{\mathbf{y}} \nabla_{\mathbf{y}} f \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \\
&\quad + \left(1 + \frac{1}{\alpha} \right) \frac{(\eta^{\mathbf{y}})^2 K}{n} \sum_{i,k} \mathbb{E} \left\| \nabla_{\mathbf{y}} f_i \left(\mathbf{x}_i^{(t)+k}, y_i^{(t)+k} \right) - \nabla_{\mathbf{y}} f_i \left(\bar{\mathbf{x}}^{(t)}, \bar{\mathbf{y}}^{(t)} \right) \right\|^2 \\
&\leq (1+\alpha)\, (1 - K\eta^{\mathbf{y}} \mu)\, \varepsilon_t + \left(1 + \frac{1}{\alpha} \right) (\eta^{\mathbf{y}})^2 L^2 K \left(\mathcal{E}^{\mathbf{x}} + \mathcal{E}^{\mathbf{y}} \right)
\end{aligned}$$

where in step (b) we used Lemma 9 i.e., $\frac{1}{n} \sum_i \mathbf{c}_i^{\mathbf{y},(t)} = \mathbf{0}$. ■

Proof [Proof of Lemma 5] We have

$$
\begin{aligned}
\varepsilon_{t+1} &\overset{(a)}{\leq} (1+\beta)\mathbb{E}\left\|\hat{\mathbf{y}}^{(t)} - \bar{\mathbf{y}}^{(t+1)}\right\|^2 + \left(1+\frac{1}{\beta}\right)\mathbb{E}\left\|\hat{\mathbf{y}}^{(t+1)} - \hat{\mathbf{y}}^{(t)}\right\|^2 \\
&\leq (1+\beta)(1+\alpha)\left(1 - K\eta^{\mathbf{y}}\mu\right)\varepsilon_t \\
&\quad + (1+\beta)\left(1+\frac{1}{\alpha}\right)(\eta^{\mathbf{y}})^2 L^2 K\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + (1+\frac{1}{\beta})\kappa^2\mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + (1+\beta)\frac{K(\eta^{\mathbf{y}})^2\sigma^2}{n} \\
&\overset{(b)}{\leq} \left(1 - \frac{K\eta^{\mathbf{y}}\mu}{3}\right)\varepsilon_t + \frac{6\eta^{\mathbf{y}}L^2}{\mu}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{4\eta^{\mathbf{y}}\sigma^2}{n\mu} \\
&\quad + \frac{4\kappa^2}{K\eta^{\mathbf{y}}\mu}\left(2K(\eta^{\mathbf{x}})^2L^2\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 4K^2L^2(\eta^{\mathbf{x}})^2\varepsilon_t + 4K^2(\eta^{\mathbf{x}})^2\mathbb{E}\|\nabla\Phi(\bar{\mathbf{x}}(t))\|^2 + \frac{K(\eta^{\mathbf{x}})^2\sigma^2}{n}\right) \\
&= \left(1 - \frac{K\eta^{\mathbf{y}}L}{3\kappa} + \frac{16L\kappa^3K(\eta^{\mathbf{x}})^2}{\eta^{\mathbf{y}}}\right)\varepsilon_t \\
&\quad + \left(\frac{8L\kappa^3(\eta^{\mathbf{x}})^2}{\eta^{\mathbf{y}}} + 6\eta^{\mathbf{y}}L\kappa\right)\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{16\kappa^3K(\eta^{\mathbf{x}})^2}{\eta^{\mathbf{y}}L}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{4\kappa^3(\eta^{\mathbf{x}})^2\sigma^2}{n\eta^{\mathbf{y}}L} + \frac{4\eta^{\mathbf{y}}\sigma^2\kappa}{nL}
\end{aligned}
$$

Using the assumption $\eta^{\mathbf{x}} \leq \frac{\eta^{\mathbf{y}}}{4\sqrt{6}\kappa^2}$ completes the proof. In (a), we used the bound in Lemma 11 for the first term and Proposition 12 for the second term. In (b), we replaced $\alpha = \beta = \frac{K\eta^{\mathbf{y}}\mu}{3}$ and used (12) in Lemma 10. ■

B.5. Proof of Lemma 6

Proof [Proof of Lemma 6] Proposition 12 indicates that $\Phi(\cdot)$ is $2\kappa L$-smooth, and hence yields

$$
\begin{aligned}
\Phi(\bar{\mathbf{x}}^{(t+1)}) &= \Phi\left(\bar{\mathbf{x}}^{(t)} - \frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi_i^{(t)+k}\right) + \mathbf{c}_i^{\mathbf{x},(t)}\right)\right) \\
&\leq \Phi(\bar{\mathbf{x}}^{(t)}) + \underbrace{\left\langle \nabla\Phi(\bar{\mathbf{x}}^{(t)}), -\frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_i^{(t)+k}, \mathbf{y}_i^{(t)+k}; \xi_i^{(t)+k}\right) + \mathbf{c}_i^{\mathbf{x},(t)}\right)\right\rangle}_{\equiv U} + \kappa L\mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|^2
\end{aligned}
$$

Further derivation of an upper bound for $\mathbb{E}[U]$ as follows gives

$$
\begin{aligned}
\mathbb{E}[U] &\equiv \mathbb{E}\left\langle \nabla\Phi(\bar{\mathbf{x}}^{(t)}), -\frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_i^{(t)+k},\mathbf{y}_i^{(t)+k};\xi_i^{(t)+k}\right)+\mathbf{c}_i^{\mathbf{x},(t)}\right)\right\rangle \\
&= \mathbb{E}\left\langle \nabla\Phi(\bar{\mathbf{x}}^{(t)}), -\frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\mathbb{E}_{\xi_i^{(t)+k}}\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_i^{(t)+k},\mathbf{y}_i^{(t)+k};\xi_i^{(t)+k}\right)\right\rangle \\
&= -\eta^{\mathbf{x}}\mathbb{E}\left\langle \nabla\Phi(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i,k}\left(\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_i^{(t)+k},\mathbf{y}_i^{(t)+k}\right)\right.\right. \\
&\qquad\qquad \left.\left.-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)+\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}^{(t)}\right)+\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}^{(t)}\right)\right)\right\rangle \\
&= -K\eta^{\mathbf{x}}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \\
&\quad -\frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\left\langle \nabla\Phi(\bar{\mathbf{x}}^{(t)}), \nabla_{\mathbf{x}}f_i\left(\mathbf{x}_i^{(t)+k},\mathbf{y}_i^{(t)+k}\right)-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)+\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}^{(t)}\right)\right\rangle \\
&\le -\frac{K\eta^{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \\
&\quad +\frac{\eta^{\mathbf{x}}}{n}\sum_{i,k}\left(\mathbb{E}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_i^{(t)+k},\mathbf{y}_i^{(t)+k}\right)-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)\right\|^2+\mathbb{E}\left\|\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\bar{\mathbf{y}}^{(t)}\right)-\nabla_{\mathbf{x}}f_i\left(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}^{(t)}\right)\right\|^2\right) \\
&\le -\frac{K\eta^{\mathbf{x}}}{2}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2+\eta^{\mathbf{x}}L^2\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right)+K\eta^{\mathbf{x}}L^2\varepsilon_t
\end{aligned}
$$

Now, we apply the above upper bound for $\mathbb{E}[U]$ and (12) in the proof of Lemma 10 as follows

$$
\begin{aligned}
&\mathbb{E}\left[\Phi(\bar{\mathbf{x}}^{(t+1)})-\Phi(\bar{\mathbf{x}}^{(t)})\right] \le \eta^{\mathbf{x}}L^2\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right)+L^2\eta^{\mathbf{x}}K\varepsilon_t-\frac{\eta^{\mathbf{x}}K}{2}\mathbb{E}\left\|\nabla\phi(\bar{\mathbf{x}}^{(t)})\right\|^2+\kappa L\mathbb{E}\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|^2 \\
&\le \mathbb{E}\Phi(\bar{\mathbf{x}}^{(t)})+\left(\eta^{\mathbf{x}}L^2+2K(\eta^{\mathbf{x}})^2L^3\kappa\right)\left(\mathcal{E}_t^{\mathbf{x}}+\mathcal{E}_t^{\mathbf{y}}\right) \\
&\quad +\frac{K(\eta^{\mathbf{x}})^2L\kappa\sigma^2}{n}+\left(L^2\eta^{\mathbf{x}}K+4K^2L^3(\eta^{\mathbf{x}})^2\kappa\right)\varepsilon_t+\left(4K^2(\eta^{\mathbf{x}})^2L\kappa-\frac{\eta^{\mathbf{x}}K}{2}\right)\mathbb{E}\left\|\nabla\phi(\bar{\mathbf{x}}^{(t)})\right\|^2
\end{aligned}
$$

Applying the assumption $\eta^{\mathbf{x}} \le \frac{1}{16KL\kappa}$ completes the proof. ■

B.6. Proof of Lemma 7

Proof [Proof of Lemma 7] According to the Lemma 2, we have

$$
\begin{aligned}
0 &\le -E^{\mathbf{x}}\frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{x}}+3E^{\mathbf{x}}L\eta_c^{\mathbf{y}}\Xi_t^{\mathbf{x}}+12E^{\mathbf{x}}K(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\mathcal{E}_t^{\mathbf{y}}+12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\gamma_t^{\mathbf{x}} \\
&\qquad +12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\varepsilon_t+12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2+3E^{\mathbf{x}}K(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L\sigma^2 \\
0 &\le -E^{\mathbf{y}}\frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{y}}+3E^{\mathbf{y}}L\eta_c^{\mathbf{y}}\Xi_t^{\mathbf{y}}+12E^{\mathbf{y}}K(\eta_c^{\mathbf{y}})^3L^3\mathcal{E}_t^{\mathbf{x}}+12E^{\mathbf{y}}K^2(\eta_c^{\mathbf{y}})^3L^3\gamma_t^{\mathbf{y}}+6E^{\mathbf{y}}K^2(\eta_c^{\mathbf{y}})^3L^3\varepsilon_t+3E^{\mathbf{y}}K(\eta_c^{\mathbf{y}})^3L\sigma^2
\end{aligned}
\tag{15}
$$

By applying the definition of $\mathscr{H}_t$ from (5) and using (15), Lemmas 2, 3, 4, 5 and 6, we have

$$
\begin{aligned}
&\mathscr{H}_{t+1} - \mathscr{H}_t \\
&= \mathbb{E}\left[\Phi(\bar{\mathbf{x}}^{(t+1)}) - \Phi(\bar{\mathbf{x}}^{(t)})\right] \\
&\quad + B^{\mathbf{x}}\eta_c^{\mathbf{y}} L(\Xi_{t+1}^{\mathbf{x}} - \Xi_t^{\mathbf{x}}) + B^{\mathbf{y}}\eta_c^{\mathbf{y}} L(\Xi_{t+1}^{\mathbf{y}} - \Xi_t^{\mathbf{y}}) \\
&\quad + A^{\mathbf{x}}K^2L^3(\eta_c^{\mathbf{y}})^3(\gamma_{t+1}^{\mathbf{x}} - \gamma_t^{\mathbf{x}}) + A^{\mathbf{y}}K^2L^3(\eta_c^{\mathbf{y}})^3(\gamma_{t+1}^{\mathbf{y}} - \gamma_t^{\mathbf{y}}) \\
&\quad + C\frac{1}{K\kappa p}(\varepsilon_{t+1} - \varepsilon_t) + 0 + 0 \\
&\le 2\eta^{\mathbf{x}}L^2\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + 2L^2\eta^{\mathbf{x}}K\varepsilon_t - \frac{\eta^{\mathbf{x}}K}{4}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{K(\eta^{\mathbf{x}})^2L\sigma^2\kappa}{n} \\
&\quad + B^{\mathbf{x}}\eta_c^{\mathbf{y}} L\left(-\frac{p}{2}\Xi_t^{\mathbf{x}} + \frac{6K(\eta^{\mathbf{x}})^2L^2}{p}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{6K^2(\eta^{\mathbf{x}})^2L^2}{p}\gamma_t^{\mathbf{x}} + K(\eta^{\mathbf{x}})^2\sigma^2\right) \\
&\quad + B^{\mathbf{y}}\eta_c^{\mathbf{y}} L\left(-\frac{p}{2}\Xi_t^{\mathbf{y}} + \frac{6K(\eta^{\mathbf{y}})^2L^2}{p}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{6K^2(\eta^{\mathbf{y}})^2L^2}{p}\gamma_t^{\mathbf{y}} + K(\eta^{\mathbf{y}})^2\sigma^2\right) \\
&\quad + A^{\mathbf{x}}K^2L^3(\eta_c^{\mathbf{y}})^3\left(-\frac{p}{2}\gamma_t^{\mathbf{x}} + \frac{30}{pK}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{12K^2L^2}{p}\left(2(\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)\varepsilon_t + \frac{24K^2(\eta^{\mathbf{x}})^2}{p}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{2\sigma^2}{KL^2}\right) \\
&\quad + A^{\mathbf{y}}K^2L^3(\eta_c^{\mathbf{y}})^3\left(-\frac{p}{2}\gamma_t^{\mathbf{y}} + \frac{30}{pK}\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{12K^2L^2}{p}\left(2(\eta^{\mathbf{x}})^2 + (\eta^{\mathbf{y}})^2\right)\varepsilon_t + \frac{24K^2(\eta^{\mathbf{x}})^2}{p}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{2\sigma^2}{KL^2}\right) \\
&\quad + C\frac{1}{K\kappa p}\left(-\frac{K\eta^{\mathbf{y}}L}{6\kappa}\varepsilon_t + 12\eta^{\mathbf{y}}L\kappa\left(\mathcal{E}_t^{\mathbf{x}} + \mathcal{E}_t^{\mathbf{y}}\right) + \frac{16\kappa^3K(\eta^{\mathbf{x}})^2}{\eta^{\mathbf{y}}L}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{8\eta^{\mathbf{y}}\sigma^2\kappa}{nL}\right) \\
&\quad - E^{\mathbf{x}}\frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{x}} + 3E^{\mathbf{x}}L\eta_c^{\mathbf{y}}\Xi_t^{\mathbf{x}} + 12E^{\mathbf{x}}K(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\mathcal{E}_t^{\mathbf{y}} + 12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\gamma_t^{\mathbf{x}} \\
&\qquad\qquad + 12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L^3\varepsilon_t + 12E^{\mathbf{x}}K^2(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 + 3E^{\mathbf{x}}K(\eta_c^{\mathbf{x}})^2\eta_c^{\mathbf{y}}L\sigma^2 \\
&\quad - E^{\mathbf{y}}\frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{y}} + 3E^{\mathbf{y}}L\eta_c^{\mathbf{y}}\Xi_t^{\mathbf{y}} + 12E^{\mathbf{y}}K(\eta_c^{\mathbf{y}})^3L^3\mathcal{E}_t^{\mathbf{x}} + 12E^{\mathbf{y}}K^2(\eta_c^{\mathbf{y}})^3L^3\gamma_t^{\mathbf{y}} + 6E^{\mathbf{y}}K^2(\eta_c^{\mathbf{y}})^3L^3\varepsilon_t + 3E^{\mathbf{y}}K(\eta_c^{\mathbf{y}})^3L\sigma^2
\end{aligned}
$$

Further rearranging gives

$$
\begin{aligned}
\mathscr{H}_{t+1} - \mathscr{H}_t \le & C_1^{\mathbf{x}}\cdot(\eta_c^{\mathbf{y}})^3K^2L^3\gamma_t^{\mathbf{x}} + C_1^{\mathbf{y}}\cdot(\eta_c^{\mathbf{y}})^3K^2L^3\gamma_t^{\mathbf{y}} + C_2^{\mathbf{x}}\cdot\Xi_t^{\mathbf{x}}\eta_c^{\mathbf{y}}L + C_2^{\mathbf{y}}\cdot\Xi_t^{\mathbf{y}}\eta_c^{\mathbf{y}}L + \frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{x}} \\
& + C_3^{\mathbf{y}}\cdot\frac{L}{K}\eta_c^{\mathbf{y}}\mathcal{E}_t^{\mathbf{y}} + C_4^{\varepsilon}\cdot\frac{L\eta_c^{\mathbf{y}}}{\kappa^2}\varepsilon_t + C_4^{\mathbf{x}}\cdot K\eta^{\mathbf{x}}\mathbb{E}\left\|\nabla\Phi(\bar{\mathbf{x}}^{(t)})\right\|^2 \\
& + \frac{C_4^{\mathbf{y}}}{p}\cdot KL(\eta_c^{\mathbf{y}})^3\sigma^2 + \frac{K(\eta^{\mathbf{x}})^2L\kappa}{n}\sigma^2 + C\frac{8\eta^{\mathbf{y}}}{nLKp}\sigma^2
\end{aligned}
$$

where constants $C_1^{\mathbf{x}}, C_1^{\mathbf{y}}, C_2^{\mathbf{x}}, C_2^{\mathbf{y}}, C_3^{\mathbf{x}}, C_4^{\varepsilon}, C_4^{\mathbf{x}}$ and $C_4^{\mathbf{y}}$ are chosen such that

$$
\begin{aligned}
&-A^{\mathbf{x}}\frac{p}{2}+B^{\mathbf{x}}\frac{6(\eta_s^{\mathbf{x}})^2}{p}+12E^{\mathbf{x}}\leq C_1^{\mathbf{x}}\\
&-A^{\mathbf{y}}\frac{p}{2}+B^{\mathbf{y}}\frac{6(\eta_s^{\mathbf{y}})^2}{p}+12E^{\mathbf{y}}\leq C_1^{\mathbf{y}}\\
&-B^{\mathbf{x}}\frac{p}{2}+3E^{\mathbf{x}}\leq C_2^{\mathbf{x}}\\
&-B^{\mathbf{y}}\frac{p}{2}+3E^{\mathbf{y}}\leq C_2^{\mathbf{y}}\\
&-E^{\mathbf{x}}+B^{\mathbf{x}}\frac{6K^2L^2(\eta^{\mathbf{x}})^2}{p}+B^{\mathbf{y}}\frac{6K^2L^2(\eta^{\mathbf{y}})^2}{p}+A^{\mathbf{x}}\frac{30(\eta_c^{\mathbf{y}})^2L^2K^2}{p}+A^{\mathbf{y}}\frac{30(\eta_c^{\mathbf{y}})^2L^2K^2}{p}\\
&\qquad\qquad+12E^{\mathbf{y}}K^2L^2(\eta_c^{\mathbf{y}})^2+2\eta^{\mathbf{x}}KL+C\frac{12\eta_s^{\mathbf{y}}}{p}\leq C_3^{\mathbf{x}}\\
&-C\frac{\eta_s^{\mathbf{y}}}{6p}+A^{\mathbf{x}}\frac{12K^4L^4}{p}(\eta_c^{\mathbf{y}})^2\cdot 3(\eta^{\mathbf{y}})^2\kappa^2+A^{\mathbf{y}}\frac{12K^4L^4}{p}(\eta_c^{\mathbf{y}})^2\cdot 3(\eta^{\mathbf{y}})^2\kappa^2\\
&\qquad\qquad+12E^{\mathbf{x}}K^2L^2(\eta_c^{\mathbf{x}})^2\kappa^2+6E^{\mathbf{y}}K^2L^2(\eta_c^{\mathbf{y}})^2\kappa^2+2LK\kappa^2\frac{\eta^{\mathbf{x}}}{\eta_c^{\mathbf{y}}}\leq C_4^{\varepsilon}\\
&-\frac{1}{4}+A^{\mathbf{x}}\frac{24K^3L^3}{p}(\eta_c^{\mathbf{y}})^3\eta^{\mathbf{x}}+A^{\mathbf{y}}\frac{24K^3L^3}{p}(\eta_c^{\mathbf{y}})^3\eta^{\mathbf{x}}+C\frac{16\kappa^2\eta^{\mathbf{x}}}{KL\eta^{\mathbf{y}}p}\\
&\qquad\qquad+12E^{\mathbf{x}}KL\eta_c^{\mathbf{y}}\frac{\eta_c^{\mathbf{x}}}{\eta_s^{\mathbf{x}}}\leq C_4^{\mathbf{x}}\\
&B^{\mathbf{x}}(\eta_s^{\mathbf{x}})^2+B^{\mathbf{y}}(\eta_s^{\mathbf{y}})^2+2A^{\mathbf{x}}+2A^{\mathbf{y}}+3E^{\mathbf{x}}+3E^{\mathbf{y}}\leq\frac{C_4^{\mathbf{y}}}{p}
\end{aligned}
$$

By letting $E^{\mathbf{x}}=E^{\mathbf{y}}=v$, $\eta_c^{\mathbf{y}}\leq\frac{p}{300v\kappa KL}$, $\eta_c^{\mathbf{x}}\leq\frac{\eta_c^{\mathbf{y}}}{\kappa^2}$, $\eta_s^{\mathbf{x}}=\eta_s^{\mathbf{y}}=pv$, $B^{\mathbf{x}}=B^{\mathbf{y}}=\frac{6v}{p}$, $A^{\mathbf{x}}=A^{\mathbf{y}}=\frac{1}{p}(72v^3+24v)$, and $C=\frac{1}{24}$, there exists a global constant $v>1$ that ensures $C_1^{\mathbf{x}}, C_1^{\mathbf{y}}, C_2^{\mathbf{x}}, C_2^{\mathbf{y}}, C_3^{\mathbf{x}}$, $C_3^{\mathbf{y}}, C_4^{\varepsilon}\leq 0$, $C_4^{\mathbf{x}}<0$ and $C_4^{\mathbf{y}}\geq 0$, hence proving Lemma 7. ■

Appendix C. Toolbox Lemmas

We introduce some technical lemmas we will use from time to time in this paper. Proofs are either standard or can be found in provided references.

Proposition 12 ([9]) *Under Assumption 2, $\Phi(\cdot)$ is $L(1+\kappa)$-smooth. Furthermore, $\mathbf{y}^*(\cdot)=\arg\max_{\mathbf{y}\in\mathbb{R}^{d_y}} f(\cdot,\mathbf{y})$ is κ-Lipschitz in the sense that for any $\mathbf{x}$ and $\mathbf{x}'$*

$$\left\|\mathbf{y}^*(\mathbf{x})-\mathbf{y}^*(\mathbf{x}')\right\|\leq\kappa\left\|\mathbf{x}-\mathbf{x}'\right\|$$

Proposition 13 ([1]) *Under Assumption 2, for every $\mathbf{x}\in\mathbb{R}^{d_x}$ and $\mathbf{y},\mathbf{y}'\in\mathbb{R}^{d_y}$, we have*

$$\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})^{\top}\left(\mathbf{y}-\mathbf{y}'\right)+\frac{1}{2L}\left\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})\right\|^2+\frac{\mu}{2}\left\|\mathbf{y}-\mathbf{y}'\right\|^2\leq f\left(\mathbf{x},\mathbf{y}^+\right)-f\left(\mathbf{x},\mathbf{y}'\right)$$

where $\mathbf{y}^+=\mathbf{y}-\frac{1}{L}\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})$.

Lemma 14 *For a set of arbitrary vectors $a_1, \ldots, a_n$ such that $a_i \in \mathbb{R}^{d_x}$, we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} a_i \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} \|a_i\|^2$$

Lemma 15 (Young's + Cauchy-Schwarz Inequality) *For any vectors $a, b \in \mathbb{R}^{d_x}$ and $\alpha > 0$ we have*

$$2\langle a, b \rangle \leq \alpha \|a\|^2 + \frac{1}{\alpha} \|b\|^2$$

and

$$\|a + b\|^2 \leq (1 + \alpha)\|a\|^2 + (1 + \frac{1}{\alpha})\|b\|^2$$