SCAFF-PD: Communication Efficient Fair and Robust Federated Learning

Anonymous Authors[1]

Abstract

We present SCAFF-PD, a fast and communication-efficient algorithm for distributionally robust federated learning. Our approach improves fairness by optimizing a family of distributionally robust objectives tailored to heterogeneous clients. We leverage the special structure of these objectives, and design an accelerated primal dual (APD) algorithm which uses bias corrected local steps (as in SCAFFOLD) to achieve significant gains in communication efficiency and convergence speed. We evaluate SCAFF-PD on several benchmark datasets and demonstrate its effectiveness in improving fairness and robustness while maintaining competitive accuracy. Our results suggest that SCAFF-PD is a promising approach for federated learning in resource-constrained and heterogeneous settings.

1. Introduction

Federated learning is a popular approach for training machine learning models on decentralized data, where data privacy concerns or other constraints prevent centralized data aggregation (McMahan et al., 2017; Kairouz et al., 2021). In federated learning, model updates are computed locally on each device (the *client*) and then aggregated to train a global model at the center (the *server*). This approach has gained traction due to its ability to leverage data from multiple sources while preserving privacy, security, and autonomy, and has the potential to make machine learning more participatory in a range of interesting problem domains (Kulynych et al., 2020; Jones and Tonetti, 2020; Pentland et al., 2021).

Federated learning is naturally most attractive when the participating clients have access to different data, leading to data heterogeneity (du Terrail et al., 2022). This heterogeneity can lead to significant fairness issues, where the performance of the global model can be biased towards the data distribution of some clients over others (Dwork et al., 2012; Li et al., 2019; Abay et al., 2020). Heterogeneity can also hurt the generalization of the global model (Quinonero-Candela et al., 2008; Mohri et al., 2019). Specifically, if some clients have a disproportionate influence on the global model, the resulting model is neither fair nor will it generalize well to new clients. Such disparities are especially prevalent and detrimental in medical research, and have resulted in misdiagnosis and suboptimal treatment (Graham, 2015; Albain et al., 2009; Nana-Sinkam et al., 2021).

To address these challenges, distributionally robust objectives (DRO) explicitly account for the heterogeneity across clients and seek to optimize performance under the worst-case data distribution across clients, rather than just the average performance (Rahimian and Mehrotra, 2019). This approach can lead to more robust models that are less biased towards specific clients and more likely to generalize to new clients (Mohri et al., 2019; Duchi et al., 2023). However, such robust objectives are significantly harder to optimize. Current algorithms have very slow convergence, potentially to the point of being impractical (Ro et al., 2021). This leads to the central question of our work:

Can we design federated optimization techniques for the DRO problem with convergence rates that match their average objective counterparts?

1.1. Our Contributions

We summarize our contributions below.

Framework. We present a general formulation for the cross-silo federated DRO problem:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}\in\Lambda} \left\{ F(\boldsymbol{x},\boldsymbol{\lambda}) := \sum_{i=1}^{N} \lambda_i \cdot f_i(\boldsymbol{x}) - \psi(\boldsymbol{\lambda}) \right\}, \quad (1.1)$$

where $f_i(\boldsymbol{x})$ is the loss suffered by client i. Instead of minimizing a simple average of the client losses, equation (1.1) incorporates weights using $\boldsymbol{\lambda} \in \mathbb{R}^N$. The choice of $\boldsymbol{\lambda}$ is made in a *worst-case* manner, while being subject to the constraint set Λ and regularized with $\psi(\boldsymbol{\lambda})$. As we will show, this formulation is a generalization of several specific fair objectives that have been proposed in the federated

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

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

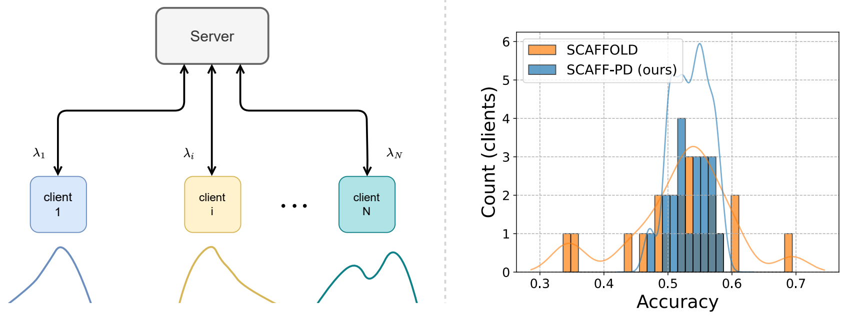


Figure 1: (**left**) In federated learning, the data distribution across individual clients differ significantly from one another. (**right**) When directly applying SOTA federated optimization algorithm (SCAFFOLD), the learned global model is biased toward certain clients, leading to noticeably worse performance when applied to a subset of participating clients. Our proposed algorithm—SCAFF-PD—largely mitigates this bias via learning a distributionally robust global model, which significantly enhances the performance of the most challenging subset of clients, specifically the worst 20%.

learning literature (Mohri et al., 2019; Li et al., 2019; 2020a; Zhang et al., 2022a; Pillutla et al., 2021).

Algorithm. The objective defined in equation (1.1) is a min-max problem and can be directly optimized using well-established algorithms such as gradient descent ascent (GDA). However, such approaches ignore the unique structure of our formulation, particularly the linearity of the interaction term between $\boldsymbol{\lambda}$ and $\boldsymbol{x}$. We leverage this to design an accelerated primal-dual (APD) algorithm (Hamedani and Aybat, 2021). Additionally, we propose to use control variates (à la SCAFFOLD) to correct the bias caused by local steps, making optimal use of local client computation (Karimireddy et al., 2020). Our proposed method, SCAFF-PD, combines these ideas to provide an efficient and practical algorithm, compatible with secure aggregation.

Convergence. We provide strong convergence guarantees for SCAFF-PD when f_i are strongly convex. If $\psi(\boldsymbol{\lambda})$ is a generic convex function, we achieve an accelerated $O\left(1/T^2\right)$ rate of convergence. Furthermore, if ψ is strongly convex, SCAFF-PD converges linearly at a rate of $\exp\left(-O(T)\right)$. This represents the first federated approach for the DRO problem that achieves *linear* convergence, let alone an *accelerated* rate. Finally, we extend our analysis to the stochastic setting, where we obtain an optimal rate of $O\left(1/T\right)$, and improve over the previous $O(1/\sqrt{T})$ rate.

Practical Evaluation. We conducted comprehensive simulations and demonstrate accelerated convergence, robustness to data heterogeneity, and the ability to leverage local computations.

For deep learning models, we avail ourselves of a two-stage Train-Convexify-Train method (Yu et al., 2022). First, we train a deep learning model using conventional federated learning methods, such as FedAvg. Then, we apply SCAFF-PD to fine tune a convex approximation. To evaluate our algorithms, we use several real-world datasets with various distributionally robust objectives, and we study the trade-off between the mean and tail accuracy of these methods.

Outline. In consideration of the page constraints, we introduce the related work in Section A and establish the problem setup in Section B. Our proposed algorithms are presented in Section 2, accompanied by convergence analysis in Section 3. The results of our experiments are summarized in Section C.

2. SCAFF-PD: Accelerated Primal-Dual Federated Algorithm with Bias Corrected Local Steps

In this section, we describe our proposed algorithm SCAFF-PD (Stochastic Controlled Averaging with Primal-Dual updates) for solving the federated DRO problem (1.1). We present the pseudo-code for SCAFF-PD in Algorithm 1 and algorithm used for local updates in Algorithm 2.

As described in Algorithm 1, SCAFF-PD comprises three main steps that are executed at each communication round r: (1). Collecting loss vector $[L_1^r, \dots, L_N^r]^\top$ and gradients $\{g_i(\boldsymbol{x}^r)\}_{i=1}^N$ (for bias correction); (2). Update to the dual variable; (3). Local updates to each client model, and aggregating the updates by using the updated dual variable. We provide the pseudo-code for local updates in Algorithm 2.

Extrapolated Dual Update. Based on the computed loss vector $\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) = [L_1^r, \dots, L_N^r]^\top$ in the first step, we update the weight vector $\boldsymbol{\lambda}$. Importantly, when $\theta_r > 0$, we use both the dual gradient from the current round ($\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r)$) as well as the past round ($\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r-1})$) to obtain the extrapolated gradient $\boldsymbol{s}^r$. The gradient extrapolation step is widely used in primal-dual hybrid gradient (PDHG) methods (Chambolle and Pock, 2016) for solving convex-concave saddle-point problems, and it provides the key component in our algorithm for

Algorithm 1 SCAFF-PD($\boldsymbol{x}^0, \boldsymbol{\lambda}^0$)

for $r = 1, 2, \ldots, R$ **do**
  #(1). Collect gradient and loss vector
  Set parameters $\{\tau_r, \sigma_r, \gamma_r, \theta_r\}$
  for $i = 1, 2, \ldots, N$ **do**
    $L_i^r = f_i(\boldsymbol{x}^r)$, $\boldsymbol{c}_i^r = g_i(\boldsymbol{x}^r)$,
    Communicate $(L_i^r, \boldsymbol{c}_i^r)$ to center
  end for
  #(2). Update dual $\boldsymbol{\lambda}$

$$\boldsymbol{s}^r = (1+\theta_r)\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) - \theta_r \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r-1})$$

$$\boldsymbol{\lambda}^{r+1} = \mathrm{argmin}_{\boldsymbol{\lambda}\in\Lambda}\left\{\psi(\boldsymbol{\lambda}) - \langle \boldsymbol{s}^r, \boldsymbol{\lambda}\rangle + \frac{1}{\sigma_r}\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r)\right\}$$

  # (3).Update primal $\boldsymbol{x}$
  $\boldsymbol{c}^r = \sum_{i=1}^N \lambda_i^{r+1}\boldsymbol{c}_i^r$, Communicate $\boldsymbol{c}^r$ to each client
  for $i = 1, 2, \ldots, N$ **do**
    $\Delta \boldsymbol{u}_i^r \leftarrow$ LOCAL-UPDATE($\boldsymbol{x}^r, \boldsymbol{c}_i^r, \boldsymbol{c}^r$),
    Communicate $\Delta \boldsymbol{u}_i^r$ to the center
  end for
  Aggregate updates from different client via the weight vector $\boldsymbol{\lambda}^{r+1}$

$$\boldsymbol{x}^{r+1} = \mathrm{argmin}_{\boldsymbol{x}}\left\{\langle \sum_{i=1}^N \lambda_i^{r+1}\Delta \boldsymbol{u}_i^r, \boldsymbol{x}\rangle + \frac{1}{\tau_r}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r)\right\}$$

end for
Return: $(\boldsymbol{x}^{R+1}, \boldsymbol{\lambda}^{R+1})$

achieving acceleration. The extrapolation step used in dual update is to Nesterov's acceleration (Nesterov, 2003), which can lead to faster convergence rate and has been widely utilized for achieving acceleration in solving various optimization problems. (Chambolle and Pock, 2011; 2016; Zhang and Lin, 2015; Hamedani and Aybat, 2021).

Local Steps and Control Variates $\boldsymbol{c}_i$. Supposing that communication is not a limiting factor, each client can compute its local gradient and transmit it to the server without any local steps. In this case, the update to the primal variable $\boldsymbol{x}$ becomes

$$\Delta \boldsymbol{u}_i^r = g_i(\boldsymbol{x}^r),$$
$$\mathrm{argmin}_{\boldsymbol{x}}\Big\{\langle \sum_{i=1}^N \lambda_i^{r+1} g_i(\boldsymbol{x}^r), \boldsymbol{x}\rangle + \frac{1}{\tau_r}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r)\Big\}. \tag{2.1}$$

This update performs the primal update with the unbiased gradient $\nabla_{\boldsymbol{x}} F(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1})$, which is equivalent to the standard primal update in primal-dual-based algorithms (Chambolle and Pock, 2016; Hamedani and Aybat, 2021; Zhang et al., 2022b). However, such an update does not effectively utilize the local computational resources available on each client. Hence, we would like to perform multiple local update steps. The catch is that performing multiple

Algorithm 2 LOCAL-UPDATE($\boldsymbol{x}, \boldsymbol{c}_i, \boldsymbol{c}$)

Input: optimization parameters (η_ℓ, J), model parameters $(\boldsymbol{c}_i, \boldsymbol{c}, \boldsymbol{x})$
$\boldsymbol{u}_{i,0} = \boldsymbol{x}$
for $j = 1, 2, \ldots, J$ **do**
  $\boldsymbol{u}_{i,j} = \boldsymbol{u}_{i,j-1} - \eta_\ell \cdot (g_i(\boldsymbol{u}_{i,j-1}) - \boldsymbol{c}_i + \boldsymbol{c})$
end for
$\Delta \boldsymbol{u}_i = (\boldsymbol{x} - \boldsymbol{u}_{i,J})/(\eta_\ell J)$
Return: $\Delta \boldsymbol{u}_i$

local steps is known to lead to biased updates and "client-drift" (Karimireddy et al., 2020; Woodworth et al., 2020; Wang et al., 2020). We explicitly correct for this bias using control variates $\{\boldsymbol{c}_i\}_{i\in[N]}$ similar to SCAFFOLD. As we will demonstrate in the subsequent theoretical analysis, this correction allows SCAFF-PD to converge to the saddle-point solution of the DRO problem regardless of the data heterogeneity.

While we use local updates on the primal variable, we do not perform any on the dual variable. This is unlike general federated min-max optimization algorithms (Hou et al., 2021; Beznosikov et al., 2022). This design aligns well with the federated DRO formulation since it is impractical for each client to update the weight vector at each local step due to their lack of knowledge regarding the loss values of other clients. The aggregation of SCAFF-PD on the server resembles federated algorithms used for solving minimization problems, with the key difference being the utilization of the updated weight vector for primal aggregation.

3. Theoretical Analysis

We now present the convergence results for SCAFF-PD in solving the min-max optimization problem described in Eq. (1.1). Firstly, in Section 3.1, we introduce the results for the strongly-convex-concave setting. Subsequently, in Section 3.2, we present the results for the strongly-convex-convex setting.

3.1. Strongly-convex-concave Setting

We first introduce how to choice the parameters for SCAFF-PD in when ψ is convex and $\{f_i\}_{i\in[N]}$ are strongly convex in Condition 3.1.

Condition 3.1. *The parameters of Algorithm 1 are defined as $\sigma_{-1} = \gamma_0\bar{\tau}$, $\sigma_r = \gamma_r\tau_r$, $\theta_r = \sigma_{r-1}/\sigma_r$, $\gamma_{r+1} = \gamma_r(1+\mu_{\boldsymbol{x}}\tau_r)$.*

Next we present our convergence results in this setting.

Theorem 3.1. *Supppose $\{f_i\}_{i\in[N]}$ are $\mu_{\boldsymbol{x}}$-strongly convex. If Assumption B.3 and AssumptionB.4 hold, and we let the parameters $\{\tau_r, \sigma_r, \gamma_r, \theta_r\}$ of Algorithm 1 satisfy Condi-*

tion 3.1, then the R-th iterate $(\boldsymbol{x}^R, \boldsymbol{\lambda}^R)$ *satisfies*

$$\mathbb{E}\left[\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2\right] \le \frac{C_1}{R^2} D_0 + \frac{C_2}{R}\zeta^2, \tag{3.1}$$

where $C_1, C_2 \ge 0$ *are non-negative constants, and* $D_0 = \|\boldsymbol{x}^\star - \boldsymbol{x}^0\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2$.

Corollary 3.2. *Under the assumptions in Theorem 3.1,*

- *(deterministic local gradient): If the local gradient satisfies* $g_i(\boldsymbol{x}) = \nabla f_i(\boldsymbol{x})$ *for* $i \in [N]$, *then after* $O\left(\frac{\|\boldsymbol{x}^\star - \boldsymbol{x}^0\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2}{\sqrt{\varepsilon}}\right)$ *rounds, we have* $\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2 \le \varepsilon$.

- *(stochastic local gradient): If the local gradient satisfies Assumption B.4 with* $\sigma > 0$, *then after* $O\left(\frac{\|\boldsymbol{x}^\star - \boldsymbol{x}^0\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2}{\sqrt{\varepsilon}} + \frac{\zeta^2}{\varepsilon}\right)$ *rounds, we have* $\mathbb{E}\left[\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2\right] \le \varepsilon$.

Remark 3.3. *As suggested by the Corollary 3.2, in the deterministic setting* ($\zeta = 0$, *when applying* SCAFF-PD *for solving the min-max problems in the vanilla AFL and the super-quantile approach,* SCAFF-PD *achieves the convergence rate of* $O(1/R^2)$. *The rate of* SCAFF-PD *is faster than existing algorithms – the convergence rate is* $O(1/R)$ *in both Mohri et al. (2019); Pillutla et al. (2021). In addition, the algorithm with theoretical convergence guarantees introduced in Mohri et al. (2019) does not apply local steps (i.e., number of local updates* $J = 1$*), resulting in inferior performance in practical applications.*

Remark 3.4. SCAFF-PD *matches the rates* ($O(1/R^2)$) *of the centralized accelerated primal-dual algorithm (Hamedani and Aybat, 2021) when* $\zeta = 0$. *Meanwhile, our proposed algorithm converges faster compared to directly applying centralized gradient descent ascent (GDA) and extra-gradient method (EG) for solving Eq. (1.1), which achieve a rate of* $O(1/R)$.

3.2. Strongly-convex-strongly-concave Setting

We next present results for the strongly-convex-strongly-concave setting. Differing from the strongly-convex-concave setting, the parameters of Algorithm 1 are fixed across different rounds, as follows.

Condition 3.2. *The parameters of Algorithm 1 are defined as* $\mu_{\boldsymbol{x}}\tau = O\left(\frac{1-\theta}{\theta}\right)$, $\mu_{\boldsymbol{\lambda}}\sigma = O\left(\frac{1-\theta}{\theta}\right)$, $\frac{1}{1-\theta} = O\left(\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right) \vee \frac{\zeta^2}{\mu_{\boldsymbol{x}}\varepsilon}\right)$.

Theorem 3.5. *Suppose* $\{f_i\}_{i\in[N]}$ *are* $\mu_{\boldsymbol{x}}$*-strongly convex and* ψ *is* $\mu_{\boldsymbol{y}}$*-strongly convex. If Assumption B.3 and AssumptionB.4 hold, and we let the parameters* $\{\tau, \sigma, \theta\}$ *of Algorithm 1 satisfy Condition 3.2, then the R-th iterate* $(\boldsymbol{x}^R, \boldsymbol{\lambda}^R)$ *satisfies*

$$\mathbb{E}\left[\mu_{\boldsymbol{x}}\|\boldsymbol{x}^r - \boldsymbol{x}^\star\|^2\right] \le C_1 D_0 \theta^R + C_2(1-\theta)\frac{\zeta^2}{\mu_{\boldsymbol{x}}}, \tag{3.2}$$

where $C_1, C_2 \ge 0$ *are non-negative constants and* $D_0 = \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2$.

Corollary 3.6. *Under the assumptions in Theorem 3.5,*

- *(deterministic local gradient): If the local gradient satisfies* $g_i(\boldsymbol{x}) = \nabla f_i(\boldsymbol{x})$ *for* $i \in [N]$, *then after* $\widetilde{O}\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right)$ *rounds,* $\mu_{\boldsymbol{x}}\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2 \le \varepsilon$.

- *(stochastic local gradient): If the local gradient satisfies Assumption B.4 with* $\zeta > 0$, *then after* $\widetilde{O}\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}} + \frac{\zeta^2}{\mu_{\boldsymbol{x}}\varepsilon}\right)$ *rounds,* $\mathbb{E}\left[\mu_{\boldsymbol{x}}\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2\right] \le \varepsilon$.

Remark 3.7. *Our algorithm converges linearly to the global saddle point when each client applies a noiseless gradient for local updates (i.e.,* $\zeta = 0$*) in the presence of data heterogeneity and client-drift in federated learning. In contrast, previous approaches exhibit only sub-linear convergence. In the strongly-convex-strongly-concave setting, DRFA (Deng et al., 2020) converges to the saddle-point solution with rate* $O(1/R)$ *when there is no data heterogeneity and* $\zeta = 0$.

Remark 3.8. *By applying bias correction in local updates, the convergence rates of our algorithm match those of the centralized accelerated primal-dual algorithm (Zhang et al., 2021) in both deterministic and stochastic settings.*

Remark 3.9. *Compared to the standard minimization in federated learning, the DRO objective results in a slightly worse condition number in terms of convergence rate. In comparison to the standard minimization objective in federated learning, the DRO objective yields a slightly worse condition number. Solving DRO with* SCAFF-PD *requires* $(\sqrt{L_{\boldsymbol{xx}}/\mu_{\boldsymbol{x}}} + \sqrt{L_{\boldsymbol{\lambda x}}^2/(L_{\boldsymbol{xx}}\mu_{\boldsymbol{\lambda}})})$ *times more communication rounds compared to solving minimization problems with ProxSkip (Mishchenko et al., 2022).*

4. Conclusions

We have demonstrated the ability of SCAFF-PD to address challenges of fairness and robustness in federated learning. Theoretically, we obtained accelerated convergence rates for solving a wide class of federated DRO problems. Experimentally, we demonstrated strong empirical performance of SCAFF-PD on real-world datasets, improving upon existing approaches in both communication efficiency and model performance. An interesting future direction is the integration of DRO and privacy-preserving techniques in the context of federated learning, making SCAFF-PD applicable for a wider range of real-world applications. Another exciting direction is to explicitly integrate SCAFF-PD with game-theoretic mechanisms. Finally, studying the interplay between distributional robustness and personalization is an important open problem.

References

A. Abay, Y. Zhou, N. Baracaldo, S. Rajamoni, E. Chuba, and H. Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

K. S. Albain, J. M. Unger, J. J. Crowley, C. A. Coltman, and D. L. Hershman. Racial disparities in cancer survival among randomized clinical trials patients of the southwest oncology group. *JNCI: Journal of the National Cancer Institute*, 101(14):984–992, 2009.

A. Beznosikov, P. Dvurechenskii, A. Koloskova, V. Samokhin, S. U. Stich, and A. Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35:38116–38133, 2022.

A. Blum, N. Haghtalab, R. L. Phillips, and H. Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014. PMLR, 2021.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122, 2020.

J. O. du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic settings. 2022.

J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664, 2023.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

G. Graham. Disparities in cardiovascular disease risk in the united states. *Current Cardiology Reviews*, 11(3): 238–245, 2015.

E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

C. Hou, K. K. Thekumparampil, G. Fanti, and S. Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022.

C. I. Jones and C. Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58, 2020.

P. Kairouz, H. B. McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

S. P. Karimireddy, W. Guo, and M. I. Jordan. Mechanisms that incentivize data sharing in federated learning. *arXiv preprint arXiv:2207.04557*, 2022.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

B. Kulynych, D. Madras, S. Milli, I. D. Raji, A. Zhou, and R. Zemel. Participatory approaches to machine learning. International Conference on Machine Learning Workshop, 2020.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

B. Li, M. N. Schmidt, T. S. Alstrøm, and S. U. Stich. Partial variance reduction improves non-convex federated learning on heterogeneous data. *arXiv preprint arXiv:2212.02191*, 2022a.

Q. Li, Y. Diao, Q. Chen, and B. He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978. IEEE, 2022b.

T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

T. Li, A. Beirami, M. Sanjabi, and V. Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020a.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

N. L. Martinez, M. A. Bertran, A. Papadaki, M. Rodrigues, and G. Sapiro. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*, pages 7492–7501. PMLR, 2021.

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

K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! *arXiv preprint arXiv:2202.09357*, 2022.

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

P. Nana-Sinkam, J. Kraschnewski, R. Sacco, J. Chavez, M. Fouad, T. Gal, M. AuYoung, A. Namoos, R. Winn, V. Sheppard, et al. Health disparities and equity in the era of covid-19. *Journal of Clinical and Translational Science*, 5(1):e99, 2021.

J. F. Nash Jr. The bargaining problem. *Econometrica*, pages 155–162, 1950.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

A. Pentland, A. Lipton, and T. Hardjono. *Building the New Economy: Data as Capital*. MIT Press, 2021.

K. Pillutla, Y. Laguel, J. Malick, and Z. Harchaoui. Federated learning with superquantile aggregation for heterogeneous data. *arXiv e-prints*, pages arXiv–2112, 2021.

V. Piratla, P. Netrapalli, and S. Sarawagi. Focus on the common good: Group distributional robustness follows. *arXiv preprint arXiv:2110.02619*, 2021.

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008.

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

J. Ro, M. Chen, R. Mathews, M. Mohri, and A. T. Suresh. Communication-efficient agnostic federated averaging. *arXiv preprint arXiv:2104.02748*, 2021.

S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

R. H. L. Sim, Y. Zhang, M. C. Chan, and B. K. H. Low. Collaborative machine learning with incentive-aware model rewards. In *International Conference on Machine Learning*, pages 8927–8936. PMLR, 2020.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. Continual learning in medical devices: Fda's action plan and beyond. *The Lancet Digital Health*, 3(6):e337–e338, 2021.

J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33:7611–7623, 2020.

J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6): 1358–1376, 2014.

B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020.

X. Xu, L. Lyu, X. Ma, C. Miao, C. S. Foo, and B. K. H. Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021.

T. Yu, E. Bagdasaryan, and V. Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Y. Yu, A. Wei, S. P. Karimireddy, Y. Ma, and M. Jordan. Tct: Convexifying federated learning using bootstrapped neural tangent kernels. *Advances in Neural Information Processing Systems*, 35:30882–30897, 2022.

M. Zecchin, M. Kountouris, and D. Gesbert. Communication-efficient distributionally robust decentralized learning. *arXiv preprint arXiv:2205.15614*, 2022.

G. Zhang, S. Malekmohammadi, X. Chen, and Y. Yu. Proportional fairness in federated learning. *arXiv preprint arXiv:2202.01666*, 2022a.

X. Zhang, N. S. Aybat, and M. Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *arXiv preprint arXiv:2111.12743*, 2021.

X. Zhang, N. Aybat, and M. Gurbuzbalaban. SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=GiUpEVQmNx8.

Y. Zhang and X. Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, pages 353–361. PMLR, 2015.

A. Related Work

Cross-silo FL. Federated learning (FL) is a distributed machine learning paradigm that enables model training without exchanging raw data. In cross-silo FL (which is our focus), valuable data is split across different organizations, and each organization is either protected by privacy regulations or unwilling to share their raw data. Such organizations are referred to as "data islands" and can be found in hospital networks, financial institutions, autonomous-vehicle companies, etc. Thus, cross-silo FL involves a few highly reliable clients who potentially have extremely diverse data.

The most widely used federated optimization algorithm is Federated Averaging (FedAvg) (McMahan et al., 2017), which averages the local model updates to produce a global model. However, FedAvg is known to suffer from poor convergence when the local datasets are heterogeneous (Hsieh et al., 2020; Li et al., 2020b; Karimireddy et al., 2020; Reddi et al., 2021; Wang et al., 2021; du Terrail et al., 2022). Scaffold (Karimireddy et al., 2020) corrects for this heterogeneity, leading to more accurate updates and faster convergence (Mishchenko et al., 2022; Li et al., 2022a; Yu et al., 2022). However, all of these methods are restricted to optimizing the average of the client objectives.

Distributionally Robust Optimization. DRO is a framework for optimization under uncertainty, where the goal is to optimize the worst-case performance over a set of probability distributions. See Rahimian and Mehrotra (2019) for a review and its history in risk management, economics, and finance. Fast centralized optimization methods have been developed when uncertainity is represented by f-divergences (Wiesemann et al., 2014; Namkoong and Duchi, 2016; Levy et al., 2020) or Wasserstein distances (Mohajerin Esfahani and Kuhn, 2018; Gao and Kleywegt, 2022). The former approach accounts for changing proportions of subpopulations, relating it to notions of subpopulation fairness (Duchi et al., 2023; Santurkar et al., 2020; Piratla et al., 2021; Martinez et al., 2021). Our work also implicitly focuses on f-divergences. Deng et al. (2020) and Zecchin et al. (2022) adapt the gradient-descent-ascent (GDA) algorithm to solve the federated and decentralized DRO problems respectively. However, these methods inherit the slowness of both the GDA and FedAvg algorithms, making their performance trail the state of the art for the average objective (Mishchenko et al., 2022).

Fairness in FL. While fairness is an extremely multi-faceted concept, here we are concerned with the distribution of model performance across clients. Mohri et al. (2019) noted that minimizing the average of the client losses may lead to unfair distribution of errors, and instead proposed an *agnostic FL* (AFL) framework which minimizes a worst-case mixture of the client losses. Alternatives and extensions to AFL have also been proposed subsequently Li et al. (2019; 2020a); Pillutla et al. (2021). Again, the convergence of optimization methods for these losses (when analyzed) is significantly slower than their centralized counterparts.

While all of these works demand equitable performance across all clients, others propose to scale a client's accuracy in proportion to their contribution (Sim et al., 2020; Blum et al., 2021; Xu et al., 2021; Zhang et al., 2022a; Karimireddy et al., 2022). These methods are motivated by game-theoretic considerations to incentivize clients and improve the quality of the data contributions. Our framework (1.1) can be applied to such mechanisms by an appropriate choice of $\{f_i\}, \Lambda,$ and ψ. For example, Zhang et al. (2022a) show how to set these to recover the Nash bargaining solution (Nash Jr, 1950). Thus, our work can be seen as a practical optimization algorithm to implement many of the mechanisms studied in FL.

Finally, personalization—serving a separate model to each client—has also been proposed as a method to improve the distribution of client performance (Yu et al., 2020). However, personalized models are sometimes not feasible either due to regulations (Vokinger et al., 2021) or because the client may not have additional data. Further, personalization does not remove the differences in performance (though it does reduce it) (Yu et al., 2020), nor does it solve the game-theoretic considerations described above. Extending our work to this setting is an important question we leave for future work.

B. Problem Setup

We consider the min-max optimization problem in the context of federated learning, where the objective function, defined in Eq. (1.1), is distributed among N clients. Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local function on the i-th client, where $f_i(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[f(\boldsymbol{x}, \xi)]$ and $\mathcal{D}_i$ is the data distribution of the i-th client. For example, we can define $\mathcal{D}_i$ as the uniform distribution over the training dataset present on the i-th client.

Notation. We use the notation $\boldsymbol{x}^r \in \mathbb{R}^d$ to denote the global iterate at the r-th round, and use $\boldsymbol{u}_{i,j}^r \in \mathbb{R}^d$ to denote the local iterate at the j-th step on the i-th client (at the r-th round). We apply $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_N]^\top \in \mathbb{R}^N$ to denote the weight vector, where λ_i is the weight for client i. We let $[N]$ denote the set $\{1, \dots, N\}$. To facilitate clarity and presentation, we let

$\Phi(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i \cdot f_i(\boldsymbol{x})$. For local gradients, we let $g_i(\boldsymbol{u}_{i,j-1})$ denote the stochastic gradient of f_i at iterate $\boldsymbol{u}_{i,j-1}$:

$$g_i(\boldsymbol{u}^r_{i,j-1}) = \nabla f_i(\boldsymbol{u}^r_{i,j-1}, \xi^r_{i,j-1}). \tag{B.1}$$

Choosing ψ and Λ. We let $\psi : \mathbb{R}^N \to \mathbb{R}$ denote the regularization on the weight vector $\boldsymbol{\lambda}$. The χ^2 penalty (Levy et al., 2020) involves setting

$$\psi(\boldsymbol{\lambda}) = \frac{\rho}{2N} \sum_{i=1}^{N} (N\lambda_i - 1)^2, \text{ and } \Lambda = \Delta^N . \tag{B.2}$$

When regularization is set to zero with $\rho = 0$, the DRO formulation (1.1) recovers the agnostic federated learning (AFL) of Mohri et al. (2019). A non-zero value of ρ can be used to trade off the worst-case loss against the average loss. In particular, setting $\rho \to \infty$ recovers the standard average FL objective. While we will primarily focus on (B.2) in this work, other choices are also possible. The DRO objective becomes the α-Conditional Value at Risk (CVaR) loss (Duchi and Namkoong, 2021), also known as super-quantile loss (Pillutla et al., 2021) by setting

$$\psi(\boldsymbol{\lambda}) = 0, \text{ and } \Lambda = \{\boldsymbol{\lambda} \in \Delta, \lambda_i \leq 1/(\alpha N)\} .$$

Finally, we can recover the Q-FL loss of Li et al. (2019) by setting

$$\psi(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|^{1+\frac{1}{q}}, \text{ and } \Lambda = \mathbb{R}^N .$$

Definitions and assumptions. In the convergence analysis of our proposed algorithms, we rely on the following definitions and assumptions regarding the local functions and the regularization term ψ:

Definition B.1 (Smoothness)**.** *$f(\cdot)$ is convex and differentiable, and there exists $L \geq 0$ such that for any $\boldsymbol{x}_1, \boldsymbol{x}_2$ in the domain of $f_i(\cdot)$,*

$$\|\nabla f_i(\boldsymbol{x}_1) - \nabla f_i(\boldsymbol{x}_2)\| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{B.3}$$

Definition B.2 (Strong convexity)**.** *$f(\cdot)$ is μ-strongly convex, i.e.,*

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \langle \nabla f(\boldsymbol{x}_1), \boldsymbol{x}_2 - \boldsymbol{x}_1 \rangle + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|^2.$$

Assumption B.3 (Smoothness w.r.t. Φ)**.** *$\Phi(\boldsymbol{x}, \cdot)$ is concave and differentiable, and there exists $L_{\boldsymbol{\lambda x}} \geq 0$ such that for any $\boldsymbol{x}_1, \boldsymbol{x}_2$ in the domain of $\Phi(\cdot, \boldsymbol{\lambda})$ and $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ in the domain of $\Phi(\boldsymbol{x}, \cdot)$,*

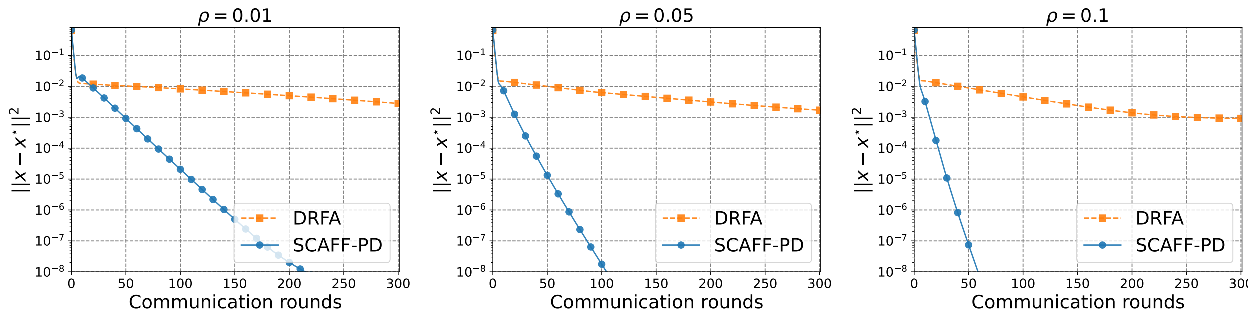
$$\|\nabla_{\boldsymbol{\lambda}} \Phi(\boldsymbol{x}_1, \boldsymbol{\lambda}_1) - \nabla_{\boldsymbol{\lambda}} \Phi(\boldsymbol{x}_2, \boldsymbol{\lambda}_2)\| \leq L_{\boldsymbol{\lambda x}}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

Assumption B.4 (Bounded noise)**.** *There exist $\zeta \geq 0$ such that for all $i \in [N]$, the local gradient $g_i(\boldsymbol{x})$ defined in Eq. (B.1) satisfies*

$$\mathbb{E}\left[\|g_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2\right] \leq \zeta^2, \quad \mathbb{E}\left[g_i(\boldsymbol{x})\right] = \nabla f_i(\boldsymbol{x}).$$

C. Experiments

We now study the performance of SCAFF-PD for solving federated DRO problems on both synthetic datasets and real-world datasets. Our primary objective when working with synthetic datasets is to validate the convergence analysis of SCAFF-PD. On real-world datasets, we compare with existing federated optimization algorithms for learning robust and fair models (DRFA (Deng et al., 2020), AFL (Mohri et al., 2019), and q-FFL (Li et al., 2019)) as well as widely used federated algorithms for solving minimization problems including FedAvg (McMahan et al., 2017) and SCAFFOLD (Karimireddy et al., 2020). After conducting thorough evaluations, we have observed that our proposed accelerated algorithms achieve fast convergence rates and strong empirical performance on real-world datasets. We have provided supplementary experimental results in Appendix F, which includes additional baseline methods, ablations on our algorithm, and other relevant findings.



(a) $\rho = 0.01$. (b) $\rho = 0.05$. (c) $\rho = 0.1$.

Figure 2: We compare our proposed algorithm with the existing method DFRA (Deng et al., 2020) on synthetic datasets. ρ is the strength of regularization ψ (defined in Eq. (B.2)). X-axis represents the number of communication rounds, and Y-axis represents the distance to optimal solution.

C.1. Results on Synthetic Datasets

To construct the synthetic datasets, we follow the setup described in Eq. (1.1) and consider a simple robust regression problem. Specifically, for the i-th client, the local function f_i is defined as $f_i(\boldsymbol{x}) = \frac{1}{m_i}\sum_{j=1}^{m_i}(\langle \boldsymbol{a}_i^j, \boldsymbol{x}\rangle - y_i^j)^2 + \frac{\mu_{\boldsymbol{x}}}{2}\|\boldsymbol{x}\|^2$, where j is sample index on this client and there are m_i training samples on client-i. We apply the χ^2 penalty for regularizing the weight vector $\boldsymbol{\lambda}$. To generate the data, each input $\boldsymbol{a}_i^j$ is sampled from a Gaussian distribution $\boldsymbol{a}_i^j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d\times d})$. Then we random generate $\widehat{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, c^2\boldsymbol{I}_{d\times d})$, and $\delta_i^{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_{d\times d})$. Based on $(\widehat{\boldsymbol{x}}, \delta_i^{\boldsymbol{x}})$, we generate y_i^i as $y_i^i = \langle \boldsymbol{a}_i^j, \widehat{\boldsymbol{x}} + \delta_i^{\boldsymbol{x}}\rangle$. Therefore, there exist distribution shifts across different clients (i.e., concept shifts). We set $N = 5$, $d = 10$, and $m_i = 100$ for $i \in [N]$. To measure the algorithm performance, we evaluate the distance between $\boldsymbol{x}^R$ and the optimal solution $\boldsymbol{x}^\star$: $\|\boldsymbol{x}^R - \boldsymbol{x}^\star\|^2$.

We compare SCAFF-PD with DRFA (Deng et al., 2020) on this synthetic dataset. The regularization parameter ρ for ψ is varied from 0.01 to 0.1. For both algorithms, we set the number of local steps to be 100 and select the algorithm parameters through grid search. The comparison results are summarized in Fig 2. As shown in Fig 2, we observe that our proposed algorithm SCAFF-PD achieves linear convergence rates in all three settings. In contrast, DRFA converges much more slowly compared to SCAFF-PD. We have included more experimental results under this synthetic setup in Appendix F, including results on the effect of local steps and data heterogeneity.

C.2. Results on Real-world Datasets

Dataset setup. We evaluate the performance of various federated learning algorithms on CIFAR100 (Krizhevsky et al., 2009) and TinyImageNet (Le and Yang, 2015). We follow the setup used in Li et al. (2022b): we consider different degrees of data heterogeneity by applying Dirichlet allocation, denoted by Dir(α), to partition the dataset into different clients. Smaller α values in Dir(α) leads to higher data heterogeneity. Additionally, after the data partition through the Dirichlet allocation, we randomly sample 30% of the clients and remove 70% training samples from those clients. Such a sub-sampling procedure can better model real-world data-imbalance scenarios. We consider the number of clients $N = 20$ for both datasets. Results on larger number of clients and other real-world datasets can be found in Appendix F.

Model setup. We consider learning a linear classifier by using representations extracted from pre-trained deep neural networks. Previous studies have demonstrated the efficacy of this approach, particularly in the context of data heterogeneity (Yu et al., 2022) as well as sub-group robustness (Izmailov et al., 2022). For both datasets, we apply the ResNet-18 (He et al., 2016) pre-trained on ImageNet-1k (Deng et al., 2009) as the backbone for extracting feature representations of the image samples. To apply the pre-trained ResNet-18, we resize the images from CIFAR100 and TinyImageNet to 3×224×224.

Comparisons with existing approaches. We consider three data heterogeneity settings for both datasets. To measure the performance of different algorithms, beside the average classification accuracy across clients, we also evaluate the `worst-20%` accuracy[1] for comparing fairness and robustness of different federated learning algorithms. Previous studies

[1] First sort the clients by test accuracy, then select the lower 20% of clients and compute the mean from this subset.

Table 1: The average and worst-20% top-1 accuracy of our algorithm (SCAFF-PD) vs. state-of-the-art federated learning algorithms evaluated on CIFAR100 and Tiny-ImageNet. The highest top-1 accuracy in each setting is highlighted in **bold**.

Datasets	Methods	Non-i.i.d. degree					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		average	worst-20%	average	worst-20%	average	worst-20%
CIFAR-100	FedAvg	38.77	15.93	35.96	24.43	36.57	26.50
	SCAFFOLD	37.38	14.65	35.28	24.77	35.63	25.61
	q-FFL	26.39	5.43	29.60	18.62	30.38	21.98
	AFL	47.38	18.04	**44.73**	22.06	**44.89**	27.27
	DRFA	46.47	26.77	41.61	27.66	43.20	32.04
	SCAFF-PD	**49.03**	**29.30**	42.06	**28.37**	43.69	**32.77**
		average	worst-20%	average	worst-20%	average	worst-20%
TinyImageNet	FedAvg	33.66	18.18	31.53	23.46	35.08	27.61
	SCAFFOLD	31.79	15.85	30.43	22.57	34.58	27.33
	q-FFL	25.50	9.70	27.45	19.38	32.90	26.24
	AFL	**45.32**	18.65	**45.54**	28.02	**46.11**	29.50
	DRFA	36.80	22.32	37.39	28.38	37.39	28.38
	SCAFF-PD	41.26	**25.32**	39.32	**30.27**	41.23	**29.78**

have employed this metric for comparing different model in federated learning (Li et al., 2019). The comparative results are summarized in Table 1. We find that our proposed algorithm outperforms existing methods in most settings, especially under higher heterogeneity. For example, when the level of data heterogeneity is low ($\alpha = 0.1$), applying SCAFF-PD does not yield very large improvements compared to the existing algorithms. In the case of high data heterogeneity ($\alpha = 0.01$), our proposed algorithm largely improves the worst-20% accuracy performance on both datasets.

Effect of ρ in DRO. To gain a better understanding of the empirical performance of our algorithm, we investigate the role of ρ in DRO when applying our algorithm. We consider $\rho \in \{0.1, 0.2, 0.5\}$ and measure both the average and worst-20% accuracy during training. We present the results in Fig 3. We find that when ρ is small, SCAFF-PD can achieve better fairness/robustness—the worst-20% accuracy significantly improves when we decrease the ρ in SCAFF-PD. Meanwhile, the experimental results suggest that smaller ρ leads to faster convergence w.r.t. worst-20% accuracy for our algorithm. On the other hand, when applying smaller ρ, the condition number of the min-max optimization problem becomes worse. Fortunately, our algorithm is guaranteed to achieve accelerated rates, making it particularly beneficial in scenarios where $\mu_{\boldsymbol{\lambda}}$ is small. As we have demonstrated in Fig 2, our proposed algorithm still converges relatively fast when ρ is small.

In addition, we study the trade-off between average accuracy vs. worst-20% accuracy vs. best-20% accuracy for different algorithms. The results are summarized in Fig 4 (in Appendix F). Without sacrificing much on average accuracy and best-20% accuracy, our algorithm largely improves the worst-20% accuracy.

D. Technical Lemmas

This section is dedicated to presenting several lemmas that serve as building blocks in proving the convergence of our proposed algorithms.

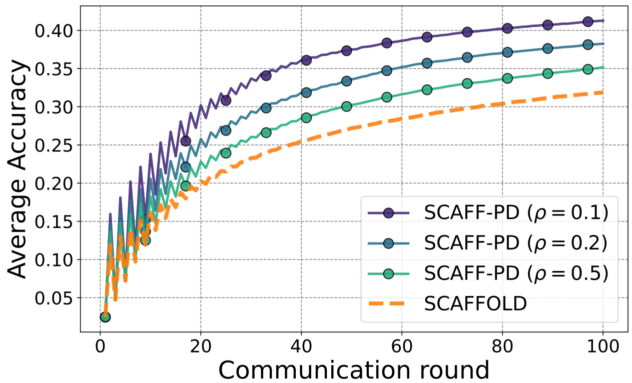
Lemma D.1 (Perturbed strong convexity, Karimireddy et al. (2020))**.** *Suppose the function $f(\cdot) : \mathcal{X} \to \mathbb{R}$ is L-smooth and μ-strongly convex, then for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}$,*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{y} \rangle \geq f(\boldsymbol{z}) - f(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y} - \boldsymbol{z}\|^2 - L\|\boldsymbol{z} - \boldsymbol{z}\|^2. \tag{D.1}$$

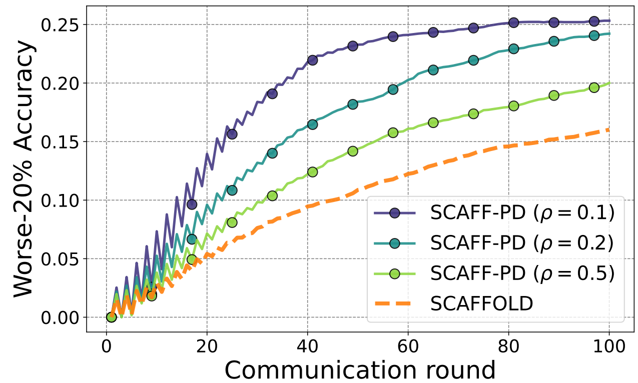
We now present the lemma for analyzing the drift term.

Lemma D.2 (Bounded drift)**.** *Suppose $\tau_r = J\,\eta_\ell\,\eta_g$, and $\eta_g \geq 1$, then we have*

$$\mathcal{E}_r \leq \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1})\right\|^2\right] + \frac{12\tau^2}{\eta_g^2}\left(1+\chi\right)\zeta^2 + \frac{3\tau^2}{\eta_g^2 J}\zeta^2, \tag{D.2}$$



(a) Average accuracy.



(b) Worst-20% accuracy.

Figure 3: We study the effect of regularization term ρ in our proposed algorithm SCAFF-PD. We measure both the average test accuracy (**a**) and worst-20% accuracy (**b**) during training. In addition, we include SCAFFOLD (orange dashed lines) as a baseline method for comparison.

where $\mathcal{E}_r$ is defined as

$$\mathcal{E}_r = \frac{1}{J} \sum_{i=1}^{N} \sum_{j=1}^{J} \lambda_i \mathbb{E}\left[\|\boldsymbol{u}_{i,j} - \boldsymbol{x}\|^2\right], \tag{D.3}$$

and χ is defined as

$$\chi = \max_{\boldsymbol{\lambda} \in \Lambda} \sum_{i=1}^{N} \lambda_i^2. \tag{D.4}$$

Proof. We omit the r superscript in the following proof. Recall that the definition of $\boldsymbol{u}_{i,j}$ in Algorithm 1, i.e.,

$$\boldsymbol{u}_{i,j} = \boldsymbol{u}_{i,j-1} - \eta_\ell \left(g_i(\boldsymbol{u}_{i,j-1}) - \hat{\boldsymbol{c}}_i + \hat{\boldsymbol{c}}\right),$$

and we have

$$\begin{aligned} \mathbb{E}\left[g_i(\boldsymbol{u}_{i,j-1})\right] &= \nabla f_i(\boldsymbol{u}_{i,j-1}), \\ \mathbb{E}\left[\hat{\boldsymbol{c}}_i\right] &= \nabla f_i(\boldsymbol{x}) = \boldsymbol{c}_i, \\ \mathbb{E}\left[\hat{\boldsymbol{c}}\right] &= \sum_{i=1}^{N} \lambda_i \nabla f_i(\boldsymbol{x}) = \boldsymbol{c}. \end{aligned} \tag{D.5}$$

Then we can upper bound $\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right]$ as follows,

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right] \\
=&\,\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}-\eta_\ell(g_i(\boldsymbol{u}_{i,j-1})-\hat{\boldsymbol{c}}_i+\hat{\boldsymbol{c}})\|^2\right] \\
=&\,\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}-\eta_\ell(\nabla f_i(\boldsymbol{u}_{i,j-1})-\hat{\boldsymbol{c}}_i+\hat{\boldsymbol{c}})\|^2\right]+\eta_\ell^2\mathbb{E}\left[\|g_i(\boldsymbol{u}_{i,j-1})-\nabla f_i(\boldsymbol{u}_{i,j-1})\|^2\right] \\
\leq&\,\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}-\eta_\ell(\nabla f_i(\boldsymbol{u}_{i,j-1})-\hat{\boldsymbol{c}}_i+\hat{\boldsymbol{c}})\|^2\right]+\eta_\ell^2\zeta^2 \\
\leq&\left(1+\frac{1}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+J\eta_\ell^2\mathbb{E}\left[\|\nabla f_i(\boldsymbol{u}_{i,j-1})-\hat{\boldsymbol{c}}_i+\hat{\boldsymbol{c}}\|^2\right]+\eta_\ell^2\zeta^2 \\
=&\left(1+\frac{1}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\nabla f_i(\boldsymbol{u}_{i,j-1})-\hat{\boldsymbol{c}}_i+\hat{\boldsymbol{c}}\|^2\right]+\eta_\ell^2\zeta^2 \\
\leq&\left(1+\frac{1}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{2\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\nabla f_i(\boldsymbol{u}_{i,j-1})-\boldsymbol{c}_i+\boldsymbol{c}\|^2\right] \\
&\quad+\frac{2\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\hat{\boldsymbol{c}}_i-\boldsymbol{c}_i+\boldsymbol{c}-\hat{\boldsymbol{c}}\|^2\right]+\eta_\ell^2\zeta^2 \\
\leq&\left(1+\frac{1}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{2\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\nabla f_i(\boldsymbol{u}_{i,j-1})-\boldsymbol{c}_i+\boldsymbol{c}\|^2\right] \\
&\quad+\underbrace{\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\hat{\boldsymbol{c}}_i-\boldsymbol{c}_i\|^2\right]+\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\hat{\boldsymbol{c}}-\boldsymbol{c}\|^2\right]+\eta_\ell^2\zeta^2}_{\Gamma},
\end{aligned}
\tag{D.6}
$$

where $\Gamma=0$ if the local gradients are deterministic. Next, we could first upper bound the term $\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right]$ as

$$
\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right] \\
\leq&\left(1+\frac{1}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\nabla f_i(\boldsymbol{u}_{i,j-1})-\boldsymbol{c}_i\|^2\right]+\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+\Gamma \\
\leq&\left(1+\frac{1}{J-1}+\frac{4\tau^2 L_{\boldsymbol{xx}}^2}{\eta_g^2 J}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+\Gamma \\
\leq&\left(1+\frac{2}{J-1}\right)\mathbb{E}\left[\|\boldsymbol{u}_{i,j-1}-\boldsymbol{x}\|^2\right]+\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+\Gamma,
\end{aligned}
\tag{D.7}
$$

where we apply the condition that $\frac{4\tau^2 L_{\boldsymbol{xx}}^2}{\eta_g^2 J}\leq\frac{1}{J-1}$. Then we have

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right]&\leq\sum_{i=1}^{j-1}\left(1+\frac{2}{J-1}\right)^i\left(\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+\Gamma\right) \\
&\leq 3J\left(\frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+\Gamma\right)=\frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right]+3J\,\Gamma.
\end{aligned}
\tag{D.8}
$$

Then the drift error $\mathcal{E}_r$ can be upper bounded as follows,

$$\begin{aligned}
\mathcal{E}_r &= \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i\mathbb{E}\left[\|\boldsymbol{u}_{i,j}-\boldsymbol{x}\|^2\right] \\
&\leq \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i\frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\|\boldsymbol{c}\|^2\right] + \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}3\lambda_i J\,\Gamma \\
&= \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\Big\|\sum_{i=1}^{N}\lambda_i\nabla f_i(\boldsymbol{x})\Big\|^2\right] + 3J\cdot\left(\frac{4\tau^2}{\eta_g^2 J}\sum_{i=1}^{N}\lambda_i\mathbb{E}\left[\|\hat{\boldsymbol{c}}_i-\boldsymbol{c}_i\|^2\right] + \frac{4\tau^2}{\eta_g^2 J}\mathbb{E}\left[\|\hat{\boldsymbol{c}}-\boldsymbol{c}\|^2\right] + \eta_\ell^2\zeta^2\right) \\
&\leq \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\Big\|\sum_{i=1}^{N}\lambda_i\nabla f_i(\boldsymbol{x})\Big\|^2\right] + \left(\frac{12\tau^2}{\eta_g^2}\sum_{i=1}^{N}\lambda_i\mathbb{E}\left[\|\hat{\boldsymbol{c}}_i-\boldsymbol{c}_i\|^2\right] + \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\|\hat{\boldsymbol{c}}-\boldsymbol{c}\|^2\right] + \frac{3\tau^2}{\eta_g^2 J}\zeta^2\right) \\
&\leq \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x},\boldsymbol{\lambda})\|^2\right] + \frac{12\tau^2}{\eta_g^2}\zeta^2 + \frac{12\tau^2}{\eta_g^2}\underbrace{\left(\sum_{i=1}^{N}\lambda_i^2\right)}_{\leq\chi}\zeta^2 + \frac{3\tau^2}{\eta_g^2 J}\zeta^2 \\
&\leq \frac{12\tau^2}{\eta_g^2}\mathbb{E}\left[\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x},\boldsymbol{\lambda})\|^2\right] + \frac{12\tau^2}{\eta_g^2}\left(1+\chi\right)\zeta^2 + \frac{3\tau^2}{\eta_g^2 J}\zeta^2,
\end{aligned}$$

which completes the proof. □

The next lemma is useful in effectively controlling the drift term in our later analysis.

Lemma D.3. *Suppose $\tau_r = J\,\eta_\ell\,\eta_g$, and $\eta_g \geq 1$, then we have*

$$\frac{1}{\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right] \geq -\tau_r L_{\boldsymbol{xx}}^2\mathcal{E}_r + \frac{\tau_r}{2}\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})\right\|^2, \tag{D.9}$$

where $\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})$ is defined as

$$\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1}) = \sum_{i=1}^{N}\lambda_i^{r+1}\nabla f_i(\boldsymbol{x}^r). \tag{D.10}$$

Proof. In start with, we analyze $\frac{1}{4\tau_r}\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2$ based on the local updates, i.e.,

$$\begin{aligned}
&\frac{1}{\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right] \\
&= \tau_r\mathbb{E}\left[\Big\|\frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}g_i(\boldsymbol{u}_{i,j-1}^r)\Big\|^2\right] \\
&\geq \tau_r\mathbb{E}\left[\Big\|\frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\nabla f_i(\boldsymbol{u}_{i,j-1}^r)\Big\|^2\right] \\
&\geq -\tau_r\mathbb{E}\left[\Big\|\frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\nabla f_i(\boldsymbol{u}_{i,j-1}^r) - \sum_{i=1}^{N}\lambda_i^{r+1}\nabla f_i(\boldsymbol{x}^r)\Big\|^2\right] + \frac{\tau_r}{2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})\right\|^2\right] \\
&\geq -\tau_r\frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\mathbb{E}\left[\left\|\nabla f_i(\boldsymbol{u}_{i,j-1}^r)-\nabla f_i(\boldsymbol{x}^r)\right\|^2\right] + \frac{\tau_r}{2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})\right\|^2\right] \\
&\geq -\tau_r L_{\boldsymbol{xx}}^2\underbrace{\frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\mathbb{E}\left[\left\|\boldsymbol{u}_{i,j-1}^r-\boldsymbol{x}^r\right\|^2\right]}_{\mathcal{E}_r} + \frac{\tau_r}{2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})\right\|^2\right] \\
&= -\tau_r L_{\boldsymbol{xx}}^2\mathcal{E}_r + \frac{\tau_r}{2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^{r+1})\right\|^2\right],
\end{aligned}$$

which completes the proof. □

The next two lemmas focus on the primal update and dual update, respectively.

Lemma D.4. *Suppose $\tau_r = J\,\eta_\ell\,\eta_g$, and $\eta_g \geq 1$, then we have*

$$\begin{aligned}&\psi(\boldsymbol{\lambda}^{r+1}) - \langle \boldsymbol{s}^r, \boldsymbol{\lambda}^{r+1}\rangle \\ \leq\, &\psi(\boldsymbol{\lambda}) - \langle \boldsymbol{s}^r, \boldsymbol{\lambda}\rangle + \frac{1}{\sigma_r}\left[\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r) - \mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) - \mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^r)\right] - \frac{\mu_{\boldsymbol{\lambda}}}{2}\|\boldsymbol{\lambda}_{r+1} - \boldsymbol{\lambda}\|^2.\end{aligned} \tag{D.11}$$

Proof. Based on Property 1 in Tseng (2008), and $\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|^2/2$. □

Lemma D.5. *Suppose $\tau_r = J\,\eta_\ell\,\eta_g$, and $\eta_g \geq 1$, then we have*

$$\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \leq \frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r) - \mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r+1}) - \mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right], \tag{D.12}$$

and

$$\begin{aligned}&\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \\ =\, &\underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r} - \boldsymbol{x}\right\rangle\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[\left\langle \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_2} + \underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r - \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_3},\end{aligned} \tag{D.13}$$

where

$$\begin{aligned}&\mathcal{T}_1 \geq \mathbb{E}\left[\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] + \frac{\mu_{\boldsymbol{x}}}{4}\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] - L_{\boldsymbol{xx}}\mathcal{E}_r, \\ &\mathcal{T}_2 \geq \mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1})\right] - 2L_{\boldsymbol{xx}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] - 2L_{\boldsymbol{xx}}\mathcal{E}_r, \\ &\mathcal{T}_3 \geq -\frac{2\chi\,\tau_r}{J}\zeta^2 - \frac{1}{4\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right],\end{aligned} \tag{D.14}$$

and $\Delta\boldsymbol{x}^r$ and $\widetilde{\Delta}\boldsymbol{x}^r$ are defined as

$$\Delta\boldsymbol{x}^r = \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}g_i(\boldsymbol{u}_{i,j-1}^r), \quad \widetilde{\Delta}\boldsymbol{x}^r = \mathbb{E}\left[\Delta\boldsymbol{x}^r\right] = \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\mathbb{E}\left[\nabla f_i(\boldsymbol{u}_{i,j-1}^r)\right]. \tag{D.15}$$

Proof. Based on Property 1 in Tseng (2008), for the update step of $\boldsymbol{x}^{r+1}$, we have

$$\mathbb{E}\left[\left\langle \sum_{i=1}^{N}\lambda_i^{r+1}\Delta\boldsymbol{u}_i^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \leq \frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r) - \mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r+1}) - \mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right], \tag{D.16}$$

then we analyze the term $\Delta\boldsymbol{x}^r = \sum_{i=1}^{N}\lambda_i^{r+1}\Delta\boldsymbol{u}_i^r$, i.e.,

$$\begin{aligned}\sum_{i=1}^{N}\lambda_i^{r+1}\Delta\boldsymbol{u}_i^r &= \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\left(g_i(\boldsymbol{u}_{i,j-1}^r) - \hat{\boldsymbol{c}}_i^r + \hat{\boldsymbol{c}}^r\right) \\ &= \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}g_i(\boldsymbol{u}_{i,j-1}^r) - \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\hat{\boldsymbol{c}}_i^r + \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\hat{\boldsymbol{c}}^r \\ &= \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}g_i(\boldsymbol{u}_{i,j-1}^r).\end{aligned} \tag{D.17}$$

Next we decompose $\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right]$ as follows,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \\ =\, &\underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r} - \boldsymbol{x}\right\rangle\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[\left\langle \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_2} + \underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r - \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_3}.\end{aligned} \tag{D.18}$$

We then analyze the upper bound for $|\mathcal{T}_3|$, i.e.,

$$
\begin{aligned}
|\mathcal{T}_3| &= \mathbb{E}\left[\left|\langle \Delta \boldsymbol{x}^r - \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\rangle\right|\right] \\
&\leq \tau_r \mathbb{E}\left[\|\Delta \boldsymbol{x}^r - \widetilde{\Delta}\boldsymbol{x}^r\|^2\right] + \frac{1}{4\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] \\
&\leq \frac{2\chi\tau_r}{J}\zeta^2 + \frac{1}{4\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right].
\end{aligned} \tag{D.19}
$$

Next, we analyze term $\mathcal{T}_1$, i.e.,

$$
\begin{aligned}
\mathcal{T}_1 &= \mathbb{E}\left[\left\langle \Delta \boldsymbol{x}^r, \boldsymbol{x}^r - \boldsymbol{x}\right\rangle\right] = \mathbb{E}\left[\left\langle \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^r - \boldsymbol{x}\right\rangle\right] \\
&= \mathbb{E}\Big[\Big\langle \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\nabla f_i(\boldsymbol{u}_{i,j-1}^r), \boldsymbol{x}^r - \boldsymbol{x}\Big\rangle\Big] \\
&\geq \frac{1}{J}\sum_{i,j}\lambda_i^{r+1}\mathbb{E}\left[f_i(\boldsymbol{x}^r) - f_i(\boldsymbol{x}) + \frac{\mu_{\boldsymbol{x}}}{4}\|\boldsymbol{x}^r - \boldsymbol{x}\|^2 - L_{\boldsymbol{x}\boldsymbol{x}}\|\boldsymbol{u}_{i,j-1}^r - \boldsymbol{x}^r\|^2\right] \\
&= \mathbb{E}\left[\sum_i \lambda_i^{r+1} f_i(\boldsymbol{x}^r) - \sum_i \lambda_i^{r+1} f_i(\boldsymbol{x}) + \frac{\mu_{\boldsymbol{x}}}{4}\|\boldsymbol{x}^r - \boldsymbol{x}\|^2 - L_{\boldsymbol{x}\boldsymbol{x}}\frac{1}{J}\sum_{i,j}\lambda_i^{r+1}\|\boldsymbol{u}_{i,j-1}^r - \boldsymbol{x}^r\|^2\right] \\
&= \mathbb{E}\left[\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] + \frac{\mu_{\boldsymbol{x}}}{4}\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] - L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r,
\end{aligned}
$$

where we apply the perturbed strong convexity lemma (Lemma D.1) for the first inequality.

We then analyze term $\mathcal{T}_2$,

$$
\begin{aligned}
\mathcal{T}_2 &= \mathbb{E}\left[\left\langle \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right] \\
&= \mathbb{E}\Big[\Big\langle \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{J}\lambda_i^{r+1}\nabla f_i(\boldsymbol{u}_{i,j-1}^r), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\Big\rangle\Big] \\
&\geq \mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1}) + \frac{\mu_{\boldsymbol{x}}}{4}\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 - \frac{L_{\boldsymbol{x}\boldsymbol{x}}}{J}\sum_{i,j}\lambda_i^{r+1}\|\boldsymbol{u}_{i,j-1}^r - \boldsymbol{x}^{r+1}\|^2\right] \\
&\geq \mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1}) + \left(\frac{\mu_{\boldsymbol{x}}}{4} - 2L_{\boldsymbol{x}\boldsymbol{x}}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] - 2L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r \\
&\geq \mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1})\right] - 2L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] - 2L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r,
\end{aligned}
$$

where we apply the perturbed strong convexity lemma (Lemma D.1) for the first inequality and apply $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2$ for the second inequality. This completes our proof. □

E. Convergence of SCAFF-PD

In this section, we present the missing proofs in Section 3. Specifically, Section E.1 contains the proofs for the strongly convex-concave setting (Section 3.1), while Section E.2 includes the proofs for the strongly convex-strongly concave setting (Section 3.2).

E.1. Proofs – strongly-convex-concave (SC-C) setting

In this subsection, we first present the technical lemma in Section E.1.1. Next, we analyze how to set the step size related parameters in Section E.1.2. Finally, we prove Theorem 3.1 in Section E.1.3.

E.1.1. Technical Lemma

Lemma E.1. *If we set the step size in Algorithm 1 as $\tau_r \cdot L_{\boldsymbol{x}\boldsymbol{x}} \leq 1$, and the parameters of Algorithm 1 satisfy Condition 3.1, then for any $\boldsymbol{x}, \boldsymbol{\lambda}$ we have*

$$\mathbb{E}\left[F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] \leq -Z_{r+1} + V_r + \Delta_r + C\tau_r\zeta^2, \tag{E.1}$$

where Z_{r+1}, V_r, Δ_r are defined as

$$\begin{aligned}
Z_{r+1} &= \mathbb{E}\left[\langle \boldsymbol{q}^{r+1}, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle + \frac{1}{2\sigma_r}\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\|^2 + \left(\frac{1}{2\tau_r} + \frac{\mu_{\boldsymbol{x}}}{8}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}\|^2 + \frac{1}{2\alpha_{r+1}}\|\boldsymbol{q}^{r+1}\|^2\right], \\
V_r &= \mathbb{E}\left[\theta_r\langle \boldsymbol{q}^r, \boldsymbol{\lambda}^r - \boldsymbol{\lambda}\rangle + \frac{1}{2\sigma_r}\|\boldsymbol{\lambda}^r - \boldsymbol{\lambda}\|^2 + \frac{1}{2\tau_r}\|\boldsymbol{x}^r - \boldsymbol{x}\|^2 + \frac{\theta_r}{2\alpha_r}\|\boldsymbol{q}^r\|^2\right], \\
\Delta_r &= \mathbb{E}\left[\left(\frac{\alpha_r\theta_r}{2} - \frac{1}{2\sigma_r}\right)\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2 + \left(\frac{L_{\boldsymbol{\lambda}\boldsymbol{x}}^2}{2\alpha_{r+1}} + 3L_{\boldsymbol{x}\boldsymbol{x}} - \frac{1}{4\tau_r}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right],
\end{aligned} \tag{E.2}$$

$\boldsymbol{q}^r$ is defined as

$$\boldsymbol{q}^r = \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) - \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r-1}), \tag{E.3}$$

and $C \geq 0$ is a constant.

Proof. To start with, by applying Lemma D.4, we have

$$\psi(\boldsymbol{\lambda}^{r+1}) - \langle \boldsymbol{s}^r, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle \leq \psi(\boldsymbol{\lambda}) + \underbrace{\frac{1}{\sigma_r}\left[\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r) - \mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) - \mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^r)\right]}_{B_r}, \tag{E.4}$$

where we define B_r as

$$B_r = \frac{1}{\sigma_r}\left[\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r) - \mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) - \mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^r)\right]. \tag{E.5}$$

Then by applying Lemma D.5, for the update step of $\boldsymbol{x}^{r+1}$, we have

$$\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \leq \frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r) - \mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r+1}) - \mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right], \tag{E.6}$$

where $\Delta\boldsymbol{x}^r$ is defined in Eq. (D.15). Then we decompose the term $\mathbb{E}[\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\rangle]$ as follows,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}\right\rangle\right] \\
&= \underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r, \boldsymbol{x}^r - \boldsymbol{x}\right\rangle\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[\left\langle \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_2} + \underbrace{\mathbb{E}\left[\left\langle \Delta\boldsymbol{x}^r - \widetilde{\Delta}\boldsymbol{x}^r, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\right\rangle\right]}_{\mathcal{T}_3} \\
&= \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3,
\end{aligned} \tag{E.7}$$

and by Lemma D.5,

$$\begin{aligned}
\mathcal{T}_1 &\geq \mathbb{E}\left[\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] + \frac{\mu_{\boldsymbol{x}}}{4}\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] - L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r, \\
\mathcal{T}_2 &\geq \mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^{r+1}) - \Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^{r+1})\right] - 2L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] - 2L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r, \\
\mathcal{T}_3 &\geq -\frac{2\chi\tau_r}{J}\zeta^2 - \frac{1}{4\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right].
\end{aligned} \tag{E.8}$$

Therefore, by combining Eq. (E.6) and Eq. (E.8), we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x},\boldsymbol{\lambda}^{r+1})\right]\\
\leq&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1})+\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1})-\Phi(\boldsymbol{x}^{r},\boldsymbol{\lambda}^{r+1})\right]-\mathcal{T}_2-\mathcal{T}_3\\
&+\frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^r)-\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^{r+1})-\mathrm{D}(\boldsymbol{x}^{r+1},\boldsymbol{x}^r)\right]-\frac{\mu_{\boldsymbol{x}}}{4}\mathbb{E}\left[\|\boldsymbol{x}^r-\boldsymbol{x}\|^2\right]+L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r\\
\leq&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1})\right]+\frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^r)-\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^{r+1})-\frac{1}{2}\mathrm{D}(\boldsymbol{x}^{r+1},\boldsymbol{x}^r)\right]\\
&-\frac{\mu_{\boldsymbol{x}}}{4}\mathbb{E}\left[\|\boldsymbol{x}^r-\boldsymbol{x}\|^2\right]+2L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+\frac{2\chi\tau_r}{J}\zeta^2\\
\leq&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1})\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+\frac{2\chi\tau_r}{J}\zeta^2\\
&+\underbrace{\frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^r)-\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^{r+1})-\frac{1}{2}\mathrm{D}(\boldsymbol{x}^{r+1},\boldsymbol{x}^r)\right]-\frac{\mu_{\boldsymbol{x}}}{8}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}\|^2\right]}_{A_r}\\
=&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1})\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+\frac{2\chi\tau_r}{J}\zeta^2+A_r,
\end{aligned}\tag{E.9}$$

where we apply the lower bound of $\mathcal{T}_2$ and $\mathcal{T}_3$ (Eq. (E.8)) for the second inequality, and apply $-\|\boldsymbol{x}\|^2\leq\|\boldsymbol{y}\|^2-\frac{1}{2}\|\boldsymbol{x}+\boldsymbol{y}\|^2$ for the third inequality, and apply $\mu_{\boldsymbol{x}}\leq L_{\boldsymbol{x}\boldsymbol{x}}$ for the last inequality. We define A_r as

$$A_r=\frac{1}{\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^r)-\mathrm{D}(\boldsymbol{x},\boldsymbol{x}^{r+1})-\frac{1}{2}\mathrm{D}(\boldsymbol{x}^{r+1},\boldsymbol{x}^r)\right]-\frac{\mu_{\boldsymbol{x}}}{8}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}\|^2\right].\tag{E.10}$$

Next, we apply the concavity of $\Phi(\boldsymbol{x}^{r+1},\cdot)$ and combining the above two steps, for the $\boldsymbol{x}$-update we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\Phi(\boldsymbol{x},\boldsymbol{\lambda}^{r+1})\right]\\
\leq&\mathbb{E}[\langle\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1}),\boldsymbol{\lambda}-\boldsymbol{\lambda}^{r+1}\rangle]+A_r+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+\frac{2\chi\tau_r}{J}\zeta^2,
\end{aligned}\tag{E.11}$$

By combining the inequality of $\boldsymbol{\lambda}$-update (Eq. (E.4)) and $\boldsymbol{x}$-update (Eq. (E.11)), we can get

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-F(\boldsymbol{x},\boldsymbol{\lambda}^{r+1})\right]\\
=&\mathbb{E}\left[\left(\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda})-\psi(\boldsymbol{\lambda})\right)-\left(\Phi(\boldsymbol{x},\boldsymbol{\lambda}^{r+1})-\psi(\boldsymbol{\lambda}^{r+1})\right)\right]\\
\leq&\mathbb{E}\left[\langle\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r+1},\boldsymbol{\lambda}^{r+1}),\boldsymbol{\lambda}-\boldsymbol{\lambda}^{r+1}\rangle\right]+\langle\boldsymbol{s}^r,\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}\rangle+A_r+B_r\\
&+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+\frac{2\chi\tau_r}{J}\zeta^2\\
=&-\mathbb{E}\left[\langle\boldsymbol{q}^{r+1},\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}\rangle\right]+\theta_r\mathbb{E}\left[\langle\boldsymbol{q}^r,\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}\rangle\right]+A_r+B_r\\
&+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+\frac{2\chi\tau_r}{J}\zeta^2\\
=&-\langle\mathbb{E}\left[\boldsymbol{q}^{r+1},\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}\rangle\right]+\theta_r\mathbb{E}\left[\langle\boldsymbol{q}^r,\boldsymbol{\lambda}^r-\boldsymbol{\lambda}\rangle\right]+\theta_r\mathbb{E}\left[\langle\boldsymbol{q}^r,\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\rangle\right]\\
&+A_r+B_r+3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r+3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\right]+\frac{2\chi\tau_r}{J}\zeta^2,
\end{aligned}\tag{E.12}$$

where we apply Eq. (E.11) for the first inequality, and the definition of $\boldsymbol{q}^r$ for the second equality, and $\boldsymbol{q}^r$ is defined as

$$\boldsymbol{q}^r=\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^r)-\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1},\boldsymbol{\lambda}^{r-1}).\tag{E.13}$$

The term $\theta^r\langle\boldsymbol{q}^r,\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\rangle$ can be upper bounded as

$$\begin{aligned}
\theta_r\langle\boldsymbol{q}^r,\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\rangle&=\theta_r\langle\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^r)-\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1},\boldsymbol{\lambda}^{r-1}),\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\rangle\\
&=\theta_r\langle\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^r)-\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1},\boldsymbol{\lambda}^r),\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\rangle\\
&\leq\frac{\theta_r}{2\alpha_r}\|\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r,\boldsymbol{\lambda}^r)-\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1},\boldsymbol{\lambda}^r)\|^2+\frac{\alpha_r\theta_r}{2}\|\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^r\|^2
\end{aligned}\tag{E.14}$$

where we apply $\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r}) = \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r-1})$ (because the Φ is linear in $\boldsymbol{\lambda}$) for the second equality, and apply the smoothness assumption

$$\|\boldsymbol{q}^{r}\| = \|\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r}, \boldsymbol{\lambda}^{r}) - \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r})\| \le L_{\boldsymbol{\lambda}\boldsymbol{x}}\|\boldsymbol{x}^{r} - \boldsymbol{x}^{r-1}\|$$

in the second inequality. Then by combining Eq. (E.14) and Eq. (E.12), we have

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] \\
\le& -\underbrace{\mathbb{E}\left[\langle \boldsymbol{q}^{r+1}, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle + \frac{1}{\sigma_r}\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) + \frac{1}{\tau_r}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r+1}) + \frac{\mu_{\boldsymbol{x}}}{8}\|\boldsymbol{x}^{r+1} - \boldsymbol{x}\|^2 + \frac{1}{2\alpha_{r+1}}\|\boldsymbol{q}^{r+1}\|^2\right]}_{Z_{r+1}} \\
&+ \underbrace{\mathbb{E}\left[\theta_r\langle \boldsymbol{q}^{r}, \boldsymbol{\lambda}^{r} - \boldsymbol{\lambda}\rangle + \frac{1}{\sigma_r}\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r}) + \frac{1}{\tau_r}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r}) + \frac{\theta_r}{2\alpha_r}\|\boldsymbol{q}^{r}\|^2\right]}_{V_r} + \frac{\alpha_r\theta_r}{2}\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^{r}\|^2\right] \\
&+ \mathbb{E}\left[\frac{1}{2\alpha_{r+1}}\|\boldsymbol{q}^{r+1}\|^2 + 3L_{\boldsymbol{x}\boldsymbol{x}}\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2 + 3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{2\tau_r}\mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^{r}) - \frac{1}{\sigma_r}\mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^{r})\right] + \frac{2\chi\tau_r}{J}\zeta^2 \\
\le& -Z_{r+1} + V_r + \frac{\alpha_r\theta_r}{2}\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^{r}\|^2\right] + \frac{L_{\boldsymbol{\lambda}\boldsymbol{x}}^2}{2\alpha_{r+1}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right] + 3L_{\boldsymbol{x}\boldsymbol{x}}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right] \\
&+ 3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{2\tau_r}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^{r})\right] - \frac{1}{\sigma_r}\left[\mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^{r})\right] + \frac{2\chi\tau_r}{J}\zeta^2 \\
=& -Z_{r+1} + V_r + \left(\frac{\alpha_r\theta_r}{2} - \frac{1}{2\sigma_r}\right)\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^{r}\|^2\right] + \left(\frac{L_{\boldsymbol{\lambda}\boldsymbol{x}}^2}{2\alpha_{r+1}} + 3L_{\boldsymbol{x}\boldsymbol{x}} - \frac{1}{4\tau_r}\right)\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right] \\
&+ \frac{2\chi\tau_r}{J}\zeta^2 + \underbrace{3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{8\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right]}_{\mathcal{T}_4}.
\end{aligned} \tag{E.15}$$

Next, to get the upper bound of $\mathcal{T}_4$, we apply Lemma D.3 to analyze the term $\frac{1}{8\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right]$,

$$\frac{1}{8\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right] \ge -\frac{\tau_r L_{\boldsymbol{x}\boldsymbol{x}}^2}{8}\mathcal{E}_r + \frac{\tau_r}{16}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^{r}, \boldsymbol{\lambda}^{r+1})\right\|^2\right]. \tag{E.16}$$

By applying Lemma D.2, we can upper bound the drift error as follows,

$$\mathcal{E}_r \le \frac{12\tau_r^2}{\eta_g^2}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^{r}, \boldsymbol{\lambda}^{r+1})\right\|^2\right] + \left[\frac{8\tau_r^2}{\eta_g^2}(1+\chi) + \frac{3\tau_r^2}{\eta_g^2 J}\right]\zeta^2, \tag{E.17}$$

Then if we set the effective step size as $\tau_r = O(1/L_{\boldsymbol{x}\boldsymbol{x}}))$, the term $\mathcal{T}_4$ can be upper bounded as

$$\begin{aligned}
\mathcal{T}_4 &= 3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{8\tau_r}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2\right] \\
&\le \left(3L_{\boldsymbol{x}\boldsymbol{x}} + \frac{\tau_r L_{\boldsymbol{x}\boldsymbol{x}}^2}{8}\right)\mathcal{E}_r - \frac{\tau_r}{16}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^{r}, \boldsymbol{\lambda}^{r+1})\right\|^2\right] \\
&\le \underbrace{\left(\frac{36\tau_r^2 L_{\boldsymbol{x}\boldsymbol{x}}}{\eta_g^2} + \frac{2\tau_r^3 L_{\boldsymbol{x}\boldsymbol{x}}^2}{\eta_g^2} - \frac{\tau_r}{16}\right)}_{\le 0}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\Phi(\boldsymbol{x}^{r}, \boldsymbol{\lambda}^{r+1})\right\|^2\right] \\
&\quad + \left(3L_{\boldsymbol{x}\boldsymbol{x}} + \frac{\tau_r L_{\boldsymbol{x}\boldsymbol{x}}^2}{8}\right)\left[\frac{8\tau_r^2}{\eta_g^2}(1+\chi) + \frac{3\tau_r^2}{\eta_g^2 J}\right]\zeta^2 \\
&\le 12\tau_r\left(3(1+\chi) + \frac{1}{J}\right)\zeta^2 \le C\tau_r\zeta^2,
\end{aligned} \tag{E.18}$$

where $C \geq 0$ is a non-negative constant. Then by combining Eq. (E.18) and Eq. (E.15), we have

$$\begin{aligned}
&\mathbb{E}\left[L(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - L(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] \\
\leq& - Z_{r+1} + V_r + \frac{2\chi\,\tau_r}{J}\zeta^2 + C\tau_r\zeta^2 \\
&+ \underbrace{\left(\frac{\alpha_r\theta_r}{2} - \frac{1}{2\sigma_r}\right)\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2\right] + \left(\frac{L_{\boldsymbol{\lambda x}}^2}{2\alpha_{r+1}} + 3L_{\boldsymbol{xx}} - \frac{1}{4\tau_r}\right)\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right]}_{\Delta_r} \\
=& - Z_{r+1} + V_r + \Delta_r + C\tau_r\zeta^2,
\end{aligned} \tag{E.19}$$

which completes the proof. □

E.1.2. How to Set Parameters in Strongly-convex-concave (SC-C) Setting?

Next we study how to set the parameters of Algorithm 1 in the strongly-convex-concave setting.

Lemma E.2. *In Algorithm 1, if we set the parameters as*

$$\sigma_{-1} = \gamma_0\bar{\tau}, \quad \sigma_r = \gamma_r\tau_r, \quad \theta_r = \sigma_{r-1}/\sigma_r, \quad \gamma_{r+1} = \gamma_r(1 + \mu_{\boldsymbol{x}}\tau_r), \tag{E.20}$$

and we set t_r as

$$t_r = \sigma_r/\sigma_0, \tag{E.21}$$

then we have

$$t_r\left(\frac{1}{\tau_r} + \mu_{\boldsymbol{x}}\right) \geq \frac{t_{r+1}}{\tau_{r+1}}, \quad \frac{t_r}{\sigma_r} \geq \frac{t_{r+1}}{\sigma_{r+1}}, \quad \frac{t_r}{t_{r+1}} = \theta_{r+1}. \tag{E.22}$$

Proof. Because we have $t_r = \sigma_r/\sigma_0$, then $t_r\left(\frac{1}{\tau_r} + \mu_{\boldsymbol{x}}\right) \geq \frac{t_{r+1}}{\tau_{r+1}}$ can be written as

$$(1 + \tau_r\mu_{\boldsymbol{x}}) \geq \frac{\tau_r}{\tau_{r+1}}\frac{t_{r+1}}{t_r} = \frac{\tau_r}{\tau_{r+1}}\frac{\sigma_{r+1}}{\sigma_r}, \tag{E.23}$$

then due to the updates of γ_r ($\gamma_{r+1} = \gamma_r(1 + \tau_r\mu_{\boldsymbol{x}})$) and update of σ_r ($\sigma_r = \gamma_r\tau_r$), we have

$$(1 + \tau_r\mu_{\boldsymbol{x}}) = \frac{\gamma_{r+1}}{\gamma_r} = \frac{\sigma_{r+1}}{\tau_{r+1}}\frac{\tau_r}{\sigma_r}, \tag{E.24}$$

therefore, the three inequalities in Eq. (E.22) are satisfied. □

Lemma E.3. *For Algorithm 1, we have*

$$\frac{\tau_r}{\sigma_r} = \frac{1}{\gamma_r} = O\left(\frac{1}{r^2}\right), \quad \gamma_r = O\left(r^2\right), \quad \sigma_r = O(r), \quad \tau_r\sigma_r = \tau_0^2\gamma_0. \tag{E.25}$$

Proof. Since $\tau_{r+1} = \tau_r\sqrt{\gamma_r/\gamma_{r+1}}$, then we have $\tau_r = \tau_0\sqrt{\gamma_0/\gamma_r}$, then based on the update rule for γ_r ($\gamma_{r+1} = \gamma_r(1 + \mu_{\boldsymbol{x}}\tau_r)$), we have

$$\gamma_{r+1} = \gamma_r(1 + \mu_{\boldsymbol{x}}\tau_r) = \gamma_r + \mu_{\boldsymbol{x}}\tau_0\sqrt{\gamma_0\gamma_r}. \tag{E.26}$$

Then we apply induction to prove that

$$\gamma_r \geq \frac{\mu_{\boldsymbol{x}}^2\tau_0^2\gamma_0}{9}r^2. \tag{E.27}$$

Therefore, for σ_r, we have

$$\sigma_r = \gamma_r\tau_r = \frac{\gamma_{r+1} - \gamma_r}{\mu_{\boldsymbol{x}}} \geq \tau_0\sqrt{\gamma_0\gamma_r} \geq \frac{\mu\tau_0^2\gamma_0}{3}r, \tag{E.28}$$

and

$$\tau_r\sigma_r = \frac{\sigma_r^2}{\gamma_r} = \frac{(\gamma_{r+1} - \gamma_r)^2}{\mu_{\boldsymbol{x}}^2\gamma_r} = \tau_0^2\gamma_0 = \text{constant}, \tag{E.29}$$

furthermore, we have

$$\frac{\tau_r}{\sigma_r} = \frac{1}{\gamma_r} = O\left(\frac{1}{r^2}\right). \tag{E.30}$$

□

Remark E.4. *For the sake of simplicity, we establish the validity of the aforementioned two lemmas by considering the case where the parameter* $(1/\tau_r + \mu_{\boldsymbol{x}})$ *is used. It is worth noting that in subsequent proofs (Theorem E.6), it suffices to substitute a smaller value of* $\mu_{\boldsymbol{x}}$*, such as* $(1/\tau_r + \mu_{\boldsymbol{x}}/4)$*.*

Proposition E.5. *If we first set* $\theta_0 = 1$*, then we set* $\tau_r, \sigma_r, \theta_r$ *such that*

$$\frac{1-\delta}{2\tau_r} \geq 6L_{\boldsymbol{xx}} + \frac{L^2_{\boldsymbol{\lambda x}}}{\alpha_{r+1}}, \quad \frac{1-\delta}{\sigma_r} \geq \theta_r \alpha_r, \tag{E.31}$$

where $\delta \in (0,1)$*. Then* $\Delta_r \leq 0$ *for* $r = 1, \ldots, R$*.*

E.1.3. CONVERGENCE ANALYSIS

Finally, we prove the convergence of Algorithm 1 in the strongly-convex-concave setting.

Theorem E.6. *Under the assumptions of Theorem 3.1, Algorithm 1 will converge to* $\boldsymbol{x}^\star$*, and*

$$\mathbb{E}\left[\|\boldsymbol{x}^{R+1} - \boldsymbol{x}^\star\|^2\right] \leq \frac{C_1}{R^2}\left[\|\boldsymbol{x}^\star - \boldsymbol{x}^0\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2\right] + \frac{C_2}{R}\zeta^2, \tag{E.32}$$

where $C_1, C_2 > 0$ *are constants.*

Proof. For θ_r, we have

$$\theta_{r+1} = \frac{\sigma_r}{\sigma_{r+1}} = \frac{\tau_r\gamma_r}{\tau_{r+1}\gamma_{r+1}} = \sqrt{\frac{\gamma_r}{\gamma_{r+1}}} = \frac{1}{\sqrt{1+\mu_{\boldsymbol{x}}\tau_r}}, \quad \tau_{r+1} = \tau_r\sqrt{\frac{\gamma_r}{\gamma_{r+1}}} = \theta_{r+1}\tau_r, \tag{E.33}$$

where we apply the fact that $\tau_{r+1} = \tau_r\sqrt{\gamma_r/\gamma_{r+1}}$. Next we set t_r, α_r as

$$t_r = \sigma_r/\sigma_0, \quad \alpha_r = c_\alpha/\sigma_{r-1}, \tag{E.34}$$

where $c_\alpha \in (0,1)$ is a constant. then Eq. (E.31) can be written as

$$\frac{1-\delta}{\tau_r} \geq 12L_{\boldsymbol{xx}} + \frac{2L^2_{\boldsymbol{\lambda x}}\sigma_r}{c_\alpha}, \quad 1-(\delta + c_\alpha) \geq 0, \tag{E.35}$$

the second one can be easily satisfied, the first one we apply induction to prove it,

$$\frac{1-\delta}{\tau_{r+1}} = \frac{1-\delta}{\tau_r}\sqrt{\frac{\gamma_{r+1}}{\gamma_r}} \geq \left(12L_{\boldsymbol{xx}} + \frac{2L^2_{\boldsymbol{\lambda x}}\sigma_r}{c_\alpha}\right)\sqrt{\frac{\gamma_{r+1}}{\gamma_r}} \geq \left(12L_{\boldsymbol{xx}} + \frac{2L^2_{\boldsymbol{\lambda x}}\sigma_{r+1}}{c_\alpha}\right), \tag{E.36}$$

where we apply the fact that

$$\gamma_{r+1}/\gamma_r \geq 1, \quad \sigma_{r+1} = \sigma_r\sqrt{\gamma_{k+1}/\gamma_k}. \tag{E.37}$$

Therefore, by Eq. (E.35), we can prove that

$$\Delta_r = \mathbb{E}\left[\left(\frac{\alpha_r\theta_r}{2} - \frac{1}{2\sigma_r}\right)\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2 + \left(\frac{L^2_{\boldsymbol{\lambda x}}}{2\alpha_{r+1}} + 3L_{\boldsymbol{xx}} - \frac{1}{4\tau_r}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] \leq 0.$$

Meanwhile, given the parameters of Algorithm 1 satisfy Condition 3.1, by Lemma E.2, we have

$$t_{r+1}V_{r+1} \leq t_r Z_{r+1}. \tag{E.38}$$

Then, by multiplying Eq. (E.1) and summing up from $r = 0, \cdots, R$, we have

$$\begin{aligned}&\left[\sum_{r=0}^{R} t_r\right] \mathbb{E}\left[F(\bar{\boldsymbol{x}}^{R+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \bar{\boldsymbol{\lambda}}^{R+1})\right] + \frac{t_R}{\tau_R} \cdot \mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{R+1}, \boldsymbol{x}^\star)\right] \\ \leq& \frac{t_0}{\tau_0}\mathrm{D}(\boldsymbol{x}^\star, \boldsymbol{x}^0) + \frac{t_0}{\sigma_0}\mathrm{D}(\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}^0) + \sum_{r=0}^{R}(t_r \cdot C\tau_r\zeta^2),\end{aligned} \tag{E.39}$$

where we defined $\bar{\boldsymbol{x}}^{R+1}, \bar{\boldsymbol{\lambda}}^{R+1}$ as

$$\bar{\boldsymbol{x}}^{R+1} = \frac{1}{\sum_{r=0}^{R} t_r}\sum_{r=0}^{R} t_r \boldsymbol{x}^r, \quad \bar{\boldsymbol{\lambda}}^{R+1} = \frac{1}{\sum_{r=0}^{R} t_r}\sum_{r=0}^{R} t_r \boldsymbol{\lambda}^r, \tag{E.40}$$

because by Lemma E.3, we have

$$\sigma_R/\tau_R = O(R^2), \quad \sum_{r=0}^{R} t_r = O(R^2), \quad t_R = \sigma_R/\sigma_0, \quad t_r\tau_r = \tau_0^2\gamma_0, \tag{E.41}$$

then we have

$$\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{R+1}, \boldsymbol{x}^\star)\right] \leq \frac{C_1}{R^2}\left[\frac{t_0}{\tau_0}\mathrm{D}(\boldsymbol{x}^\star, \boldsymbol{x}^0) + \frac{t_0}{\sigma_0}\mathrm{D}(\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}^0)\right] + \frac{C_2}{R}\zeta^2, \tag{E.42}$$

where we apply the fact that

$$F(\bar{\boldsymbol{x}}^{R+1}, \boldsymbol{\lambda}^\star) - F(\boldsymbol{x}^\star, \bar{\boldsymbol{\lambda}}^{R+1}) \geq 0. \tag{E.43}$$

This completes our proof. □

E.2. Proofs – strongly-convex-strongly-concave (SC-SC) setting

In this subsection, we first present the technical lemma in Section E.2.1. Next, we analyze how to set the step size related parameters in Section E.2.2. Finally, we prove Theorem 3.5 in Section E.2.3.

E.2.1. TECHNICAL LEMMAS

Lemma E.7. *If we set the step size in Alg 1 as $\tau \cdot L_{\boldsymbol{x}\boldsymbol{x}} \leq 1$, then for any $\boldsymbol{x}, \boldsymbol{\lambda}$ we have*

$$\mathbb{E}\left[F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] \leq -Z_{r+1} + V_r + \Delta_r + C\tau\zeta^2, \tag{E.44}$$

where Z_r, V_r, Δ_r are defined as

$$\begin{aligned}
Z_{r+1} &= \mathbb{E}\left[\langle \boldsymbol{q}^{r+1}, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle + \left(\frac{1}{2\sigma} + \frac{\mu_{\boldsymbol{\lambda}}}{2}\right)\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\|^2 + \left(\frac{1}{2\tau} + \frac{\mu_{\boldsymbol{x}}}{8}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}\|^2\right], \\
V_r &= \mathbb{E}\left[\theta^r\langle \boldsymbol{q}^r, \boldsymbol{\lambda}^r - \boldsymbol{\lambda}\rangle + \frac{1}{2\sigma}\|\boldsymbol{\lambda}^r - \boldsymbol{\lambda}\|^2 + \frac{1}{2\tau}\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right], \\
\Delta_r &= \mathbb{E}\left[\left(\frac{\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2\pi} - \frac{1}{2\sigma}\right)\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2 + \left(\frac{\pi\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2} + 3L_{\boldsymbol{x}\boldsymbol{x}} - \frac{1}{4\tau}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right],
\end{aligned} \tag{E.45}$$

where $\pi > 0$ is a parameter and $C \geq 0$ is a constant.

Proof. Most of the steps are the same as in the Lemma E.1. To start with, based on the condition that $\psi(\boldsymbol{\lambda})$ is strongly convex in $\boldsymbol{\lambda}$, we apply Lemma D.4,

$$\begin{aligned}
&\psi(\boldsymbol{\lambda}^{r+1}) - \langle \boldsymbol{s}^r, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle \\
\leq &\psi(\boldsymbol{\lambda}^r) + \frac{1}{\sigma}\left[\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r) - \mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) - \mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^r)\right] - \frac{\mu_{\boldsymbol{\lambda}}}{2}\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\|^2.
\end{aligned} \tag{E.46}$$

Next, we change the way we upper bound $\theta\langle \boldsymbol{q}^r, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\rangle$ in the strongly-convex-concave setting, and we upper bound this term as follows,

$$\begin{aligned}
\theta\langle \boldsymbol{q}^r, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\rangle &= \theta\langle \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) - \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^{r-1}), \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\rangle \\
&= \theta\langle \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) - \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^r), \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\rangle \\
&\leq \theta\|\nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) - \nabla_{\boldsymbol{\lambda}}\Phi(\boldsymbol{x}^{r-1}, \boldsymbol{\lambda}^r)\|\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\| \\
&\leq \frac{\pi\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2}\|\boldsymbol{x}^r - \boldsymbol{x}^{r-1}\|^2 + \frac{\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2\pi}\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2,
\end{aligned} \tag{E.47}$$

where $\pi > 0$ is a constant. Then we have

$$\begin{aligned}
&F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1}) \\
\leq &-\underbrace{\mathbb{E}\left[\langle \boldsymbol{q}^{r+1}, \boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\rangle + \frac{1}{\tau}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^{r+1}) + \frac{\mu_{\boldsymbol{x}}}{8}\|\boldsymbol{x}^{r+1} - \boldsymbol{x}\|^2 + \frac{1}{\sigma}\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{r+1}) + \frac{\mu_{\boldsymbol{\lambda}}}{2}\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}\|^2\right]}_{Z_{r+1}} \\
&+\underbrace{\mathbb{E}\left[\theta\langle \boldsymbol{q}^r, \boldsymbol{\lambda}^r - \boldsymbol{\lambda}\rangle + \frac{1}{\sigma}\mathrm{D}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^r) + \frac{1}{\tau}\mathrm{D}(\boldsymbol{x}, \boldsymbol{x}^r)\right]}_{V_r} + \frac{\pi\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2}\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}^{r-1}\|^2\right] + \frac{\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2\pi}\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2\right] \\
&+ 3L_{\boldsymbol{x}\boldsymbol{x}}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] + 3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{2\tau}\mathbb{E}\left[\mathrm{D}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r)\right] - \frac{1}{\sigma}\mathbb{E}\left[\mathrm{D}(\boldsymbol{\lambda}^{r+1}, \boldsymbol{\lambda}^r)\right] + \frac{2\chi\tau}{J}\zeta^2 \\
= &-Z_{r+1} + V_r + \left(\frac{\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2\pi} - \frac{1}{2\sigma}\right)\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2\right] + \left(\frac{\pi\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2} + 3L_{\boldsymbol{x}\boldsymbol{x}} - \frac{1}{4\tau}\right)\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] \\
&+ 3L_{\boldsymbol{x}\boldsymbol{x}}\mathcal{E}_r - \frac{1}{8\tau}\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] + \frac{2\chi\tau}{J}\zeta^2 \\
\geq &-Z_{r+1} + V_r + \underbrace{\left(\frac{\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2\pi} - \frac{1}{2\sigma}\right)\mathbb{E}\left[\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2\right] + \left(\frac{\pi\theta L_{\boldsymbol{\lambda}\boldsymbol{x}}}{2} + 3L_{\boldsymbol{x}\boldsymbol{x}} - \frac{1}{4\tau}\right)\mathbb{E}\left[\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right]}_{\Delta_r} \\
&+ C\tau\zeta^2,
\end{aligned}$$

where the last inequality is because Eq. (E.18). This completes our proof. □

E.2.2. HOW TO SET PARAMETERS IN STRONGLY-CONVEX-STRONGLY-CONCAVE (SC-SC) SETTING?

Lemma E.8. *For Algorithm 1, if we set the parameters as*

$$\mu_{\boldsymbol{x}}\tau = O\left(\frac{1-\theta}{\theta}\right), \quad \mu_{\boldsymbol{\lambda}}\sigma = O\left(\frac{1-\theta}{\theta}\right), \quad \frac{1}{1-\theta} = O\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right), \tag{E.48}$$

then we have

$$\frac{1}{2\tau} + \frac{\mu_{\boldsymbol{x}}}{8} \geq \frac{1}{2\tau\theta}, \quad \frac{1}{2\sigma} + \frac{\mu_{\boldsymbol{\lambda}}}{2} \geq \frac{1}{2\sigma\theta}, \quad \frac{1}{\tau} \geq 12L_{\boldsymbol{xx}} + 2\pi\theta L_{\boldsymbol{\lambda x}}, \quad \frac{1}{\sigma} \geq \frac{\theta L_{\boldsymbol{\lambda x}}}{\pi}. \tag{E.49}$$

Proof. The conditions in Eq. (E.49) can be reformulated as follows,

$$\begin{aligned}
\frac{1}{2\tau} + \frac{\mu_{\boldsymbol{x}}}{8} \geq \frac{1}{2\tau\theta} \quad &\Leftrightarrow \quad \mu_{\boldsymbol{x}}\tau \geq 4\frac{1-\theta}{\theta}, \\
\frac{1}{2\sigma} + \frac{\mu_{\boldsymbol{\lambda}}}{2} \geq \frac{1}{2\sigma\theta} \quad &\Leftrightarrow \quad \mu_{\boldsymbol{\lambda}}\sigma \geq \frac{1-\theta}{\theta}, \\
\frac{1}{\tau} \geq 12L_{\boldsymbol{xx}} + 2\pi\theta L_{\boldsymbol{\lambda x}} \quad &\Leftarrow \quad \frac{1}{\tau} \geq 12L_{\boldsymbol{xx}} + 2\pi L_{\boldsymbol{\lambda x}}, \\
\frac{1}{\sigma} \geq \frac{\theta L_{\boldsymbol{\lambda x}}}{\pi} \quad &\Leftarrow \quad \frac{c}{\sigma} \geq \frac{\theta L_{\boldsymbol{\lambda x}}}{\pi},
\end{aligned} \tag{E.50}$$

where $c \in (0, 1]$.

Next we study how to set $\{\tau, \sigma, \theta\}$ such that Eq. (E.50) holds, we could set

$$\begin{aligned}
\tau \geq \frac{4}{\mu_{\boldsymbol{x}}}\frac{1-\theta}{\theta}, &\quad \sigma \geq \frac{1}{\mu_{\boldsymbol{\lambda}}}\frac{1-\theta}{\theta}, \\
\pi = &\frac{\theta\sigma L_{\boldsymbol{\lambda x}}}{c}, \\
\frac{\mu_{\boldsymbol{x}}\theta}{4(1-\theta)} - 12L_{\boldsymbol{xx}} \geq &\frac{2\theta\sigma L_{\boldsymbol{\lambda x}}^2}{c} \geq (1-\theta)\frac{2L_{\boldsymbol{\lambda x}}^2}{c\mu_{\boldsymbol{\lambda}}},
\end{aligned} \tag{E.51}$$

therefore, once θ satisfy the following condition

$$\frac{\mu_{\boldsymbol{x}}\theta}{4(1-\theta)} - 12L_{\boldsymbol{xx}} \geq (1-\theta)\frac{2L_{\boldsymbol{\lambda x}}^2}{c\mu_{\boldsymbol{\lambda}}}, \tag{E.52}$$

and then we can set τ and σ based on the value of θ according to Eq. (E.51). Then if we let

$$\omega = \frac{1}{1-\theta}, \tag{E.53}$$

therefore, based on Eq. (E.52), by setting $c = 1$, we have

$$\begin{aligned}
&\frac{\omega-1}{\omega}\frac{\omega\mu_{\boldsymbol{x}}}{4} - 12L_{\boldsymbol{xx}} \geq \frac{1}{\omega}\frac{2L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{\lambda}}}, \\
\Leftrightarrow \quad &\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}\omega^2 - (\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}} + 48\mu_{\boldsymbol{\lambda}}L_{\boldsymbol{xx}})\omega - 8L_{\boldsymbol{\lambda x}}^2 \geq 0, \\
\Leftarrow \quad &\omega = C_\omega\frac{(\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}} + 48\mu_{\boldsymbol{\lambda}}L_{\boldsymbol{xx}}) + \sqrt{(\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}} + 48\mu_{\boldsymbol{\lambda}}L_{\boldsymbol{xx}})^2 + 32\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}L_{\boldsymbol{\lambda x}}^2}}{2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}, \\
\Leftrightarrow \quad &\omega = C_\omega\left(\frac{1}{2} + \frac{24L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\left(\frac{1}{2} + \frac{24L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}}\right)^2 + \frac{16L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right), \\
\Leftrightarrow \quad &\omega = O\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right),
\end{aligned} \tag{E.54}$$

where $C_\omega \geq 1$ is a constant. This completes our proof.

□

E.2.3. CONVERGENCE ANALYSIS

Theorem E.9. *Under the assumptions in Theorem 3.5, Algorithm 1 will converge to $\boldsymbol{x}^\star$, and*

$$\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}^\star\|^2\right] \le C_1\theta^R\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^\star\|^2\right] + C_2(1-\theta)\frac{\zeta^2}{\mu_{\boldsymbol{x}}^2}, \tag{E.55}$$

where $C_1, C_2 \ge 0$ are non-negative constants.

Proof. The last two conditions in Eq. (E.49) ensure

$$\Delta_r = \mathbb{E}\left[\left(\frac{\theta L_{\boldsymbol{\lambda x}}}{2\pi} - \frac{1}{2\sigma}\right)\|\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r\|^2 + \left(\frac{\pi\theta L_{\boldsymbol{\lambda x}}}{2} + 3L_{\boldsymbol{xx}} - \frac{1}{4\tau}\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2\right] \le 0,$$

for $r = 0, \ldots, R$. The first two conditions in Eq. (E.49) ensure

$$Z_{r+1} \ge \frac{1}{\theta}V_{r+1}.$$

Therefore, by applying Lemma E.7, we have

$$\mathbb{E}\left[F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}) - F(\boldsymbol{x}, \boldsymbol{\lambda}^{r+1})\right] + \frac{1}{\theta}V_{r+1} \le V_r + \Delta_r + C\tau\zeta^2, \tag{E.56}$$

Then we plug $\boldsymbol{x} = \boldsymbol{x}^\star, \boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$ in Eq. (E.44), and we have $F(\boldsymbol{x}^{r+1}, \boldsymbol{\lambda}^\star) - F(\boldsymbol{x}^\star, \boldsymbol{\lambda}^{r+1}) \ge 0$, then we have

$$V_{r+1} \le \theta V_r + \theta\Delta_r + C\theta\tau\zeta^2, \tag{E.57}$$

therefore, we can derive that

$$V_R \le \theta^R V_0 + \theta\Delta_R + \frac{C\tau\theta\zeta^2}{1-\theta}, \tag{E.58}$$

meanwhile, we can set the parameters $\{\tau, \sigma, \theta\}$ (according to Eq. (E.50)) such that

$$\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] \le 4\tau\theta^R V_0 + \frac{C\tau^2\theta\zeta^2}{1-\theta} \tag{E.59}$$

since $\tau = 2(1-\theta)/(\theta\mu_{\boldsymbol{x}})$,

$$\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] \le 4\tau\theta^R V_0 + \frac{4C(1-\theta)\zeta^2}{\theta\mu_{\boldsymbol{x}}^2} \tag{E.60}$$

We need to run at least N_ε rounds such that $4\tau\theta^R V_0 + \frac{4C(1-\theta)\zeta^2}{\theta\mu_{\boldsymbol{x}}^2} \le 2\varepsilon$.

Suppose N_ε satisfies

$$N_\varepsilon = O\left(\ln\left(\frac{V_0}{\varepsilon}\right)\Big/\ln\left(\frac{1}{\theta}\right)\right), \tag{E.61}$$

then we have $4\tau\theta^R V_0 \le \varepsilon$. Because $\ln(1/\theta)$ is convex in $\theta \in \mathbb{R}_+$, then we have

$$\ln\left(\frac{1}{\theta}\right) \le \frac{1}{1-\theta}, \quad \theta \in (0,1), \tag{E.62}$$

therefore, to get an upper bound for N_ε, we only need to get the upper bound for $\frac{1}{1-\theta}$. Then if we set $\omega = \frac{1}{1-\theta}$, then based on Eq. (E.54), we have

$$\omega = O\left(\frac{L_{\boldsymbol{xx}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L_{\boldsymbol{\lambda x}}^2}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}}}\right), \tag{E.63}$$

meanwhile, we need to ensure $\frac{4C(1-\theta)\zeta^2}{\theta\mu_{\boldsymbol{x}}^2}$ is small, i.e.,

$$\frac{4C(1-\theta)\zeta^2}{\theta\mu_{\boldsymbol{x}}^2} = \varepsilon \quad \Leftrightarrow \quad \frac{1}{1-\theta} = \frac{4C\zeta^2}{\theta\mu_{\boldsymbol{x}}^2\varepsilon}. \tag{E.64}$$

therefore, in ensure $\mathbb{E}\left[\|\boldsymbol{x}^r - \boldsymbol{x}\|^2\right] \leq 2\varepsilon$, the number of communication rounds satisfies

$$N_\varepsilon = \widetilde{O}\left(\frac{L_{\boldsymbol{x}\boldsymbol{x}}}{\mu_{\boldsymbol{x}}} + \sqrt{\frac{L^2_{\boldsymbol{\lambda}\boldsymbol{x}}}{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{\lambda}}} + \frac{\zeta^2}{\mu^2_{\boldsymbol{x}}\varepsilon}}\right), \tag{E.65}$$

which completes our proof. □

F. Additional Implementation Details and Experimental Results

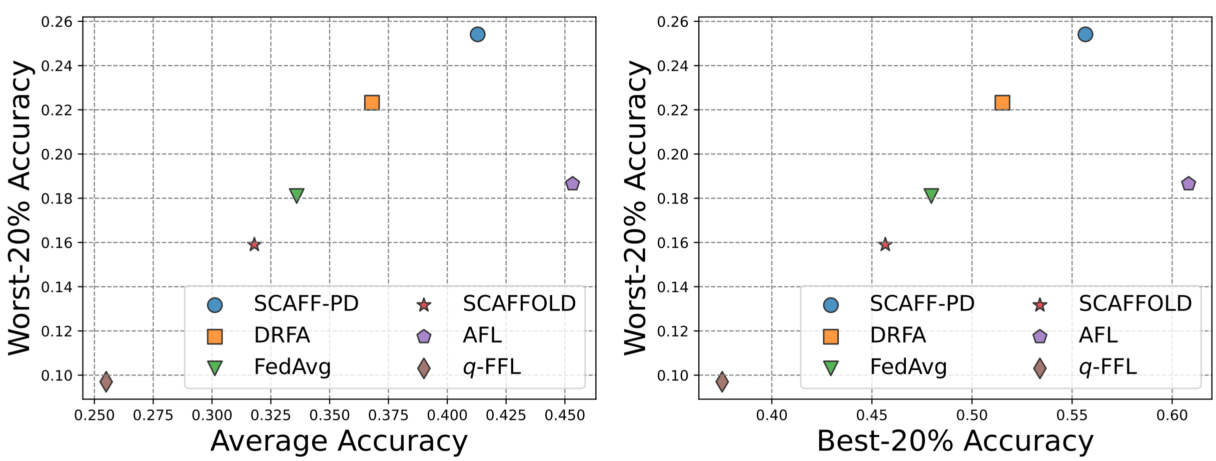
In this section, we provide further details for algorithm implementations (Section F.1) as well as additional experimental results – trade-off between worst-20% and average accuracy (Section F.2), convergence performance on synthetic datasets (Section F.3), and comparison with existing methods (Section F.4).

F.1. Additional Experimental Details

In order to enhance the performance of baseline methods, we incorporate local steps into the AFL (Mohri et al., 2019) method. We find that employing local steps yields significantly better performance compared to taking a single gradient step. To ensure a fair comparison, we employ identical feature extraction procedures across all methods. Following the setup outlined in Yu et al. (2022), we first compute the empirical neural tangent kernel (eNTK) representations of the input samples. Then, we randomly select 50,000 features from the eNTK representation through subsampling. For the (local) objective function, we utilize the mean squared error (MSE) loss, which has been used for classification tasks as described in Yu et al. (2022). To calculate the average accuracy, we begin by computing the test accuracy of each client. Then, we compute the average accuracy by averaging the results from all clients.

F.2. Trade-off between Worst-20% Accuracy and Average Accuracy

We present the trade-off between worst-20% accuracy and average/best-20% accuracy through a scatter plot, as illustrated in Figure 4. We consider the TinyImageNet dataset with the Non-i.i.d. degree parameter $\alpha = 0.01$. Our proposed algorithm, as illustrated in Figure 4, showcases a compelling trade-off between accuracy in the worst-20% and the average/best-20% scenarios.



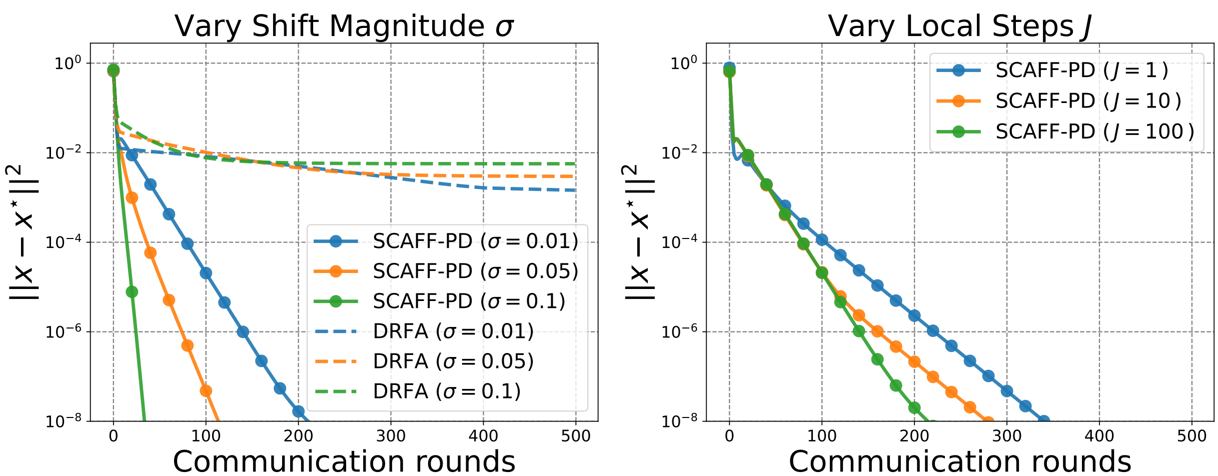
(a) Worst-20% v.s. Average accuracy.

(b) Worst-20% v.s. Best-20% accuracy.

Figure 4: Compare the average/worst-20%/best-20% accuracy of different algorithms on TinyImageNet with $\alpha = 0.01$.

F.3. Additional Experiments on Synthetic Datasets

We vary the level of data heterogeneity by changing the parameter σ from 0.01 to 0.1, where σ is used for generating $\delta_i^{\boldsymbol{x}}$ ($\delta_i^{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{d\times d})$). Figure 5a illustrates the fast convergence of SCAFF-PD to the optimal solution across various data heterogeneity settings. We also explore the effect of varying the number of local steps. Figure 5b demonstrates that increasing the number of local steps results in faster convergence towards the optimal solution.



(a) Vary the level of data heterogeneity.

(b) Effect of local steps.

Figure 5: (**left**) Compare SCAFF-PD and DRFA under different levels of data heterogeneity. (**right**) Study the effect of local steps for our proposed algorithm on the synthetic dataset.

F.4. Additional Experiments on Comparison with Existing Methods

More clients. On the CIFAR100 dataset, we conduct a comparison of different algorithms in the 50 clients setting, following the configuration outlined in Table 1. The summarized results are presented in Table 2. Consistent with our previous findings, SCAFF-PD exhibits superior robustness when compared to existing methods.

Table 2: The average and worst-20% top-1 accuracy of our algorithm (SCAFF-PD) vs. state-of-the-art federated learning algorithms evaluated on CIFAR100 with 50 clients. The highest top-1 accuracy in each setting is highlighted in **bold**.

Datasets	Methods	Non-i.i.d. degree	
		$\alpha = 0.01$	
		average	worst-20%
CIFAR-100	FedAvg	45.45	20.64
	SCAFFOLD	43.73	18.33
	q-FFL	33.42	8.13
	AFL	49.93	31.87
	DRFA	**51.07**	31.23
	SCAFF-PD	50.43	**33.03**

Additional dataset. We consider another dataset – CIFAR10 dataset, the setup mostly follows the configuration outlined in Table 1. We summarize the results in Table 3. We observe that SCAFF-PD outperforms existing methods.

Table 3: The average and worst-20% top-1 accuracy of our algorithm (SCAFF-PD) vs. state-of-the-art federated learning algorithms evaluated on CIFAR10 with 20 clients and $\alpha = 0.05$. The highest top-1 accuracy in each setting is highlighted in **bold**.

Datasets	Methods	Non-i.i.d. degree	
		$\alpha = 0.05$	
		`average`	`worst-20%`
CIFAR-100	FedAvg	77.42	60.63
	SCAFFOLD	77.75	62.89
	q-FFL	68.52	41.26
	AFL	78.89	65.07
	DRFA	79.04	65.02
	SCAFF-PD	**79.71**	**69.59**

Additional baselines. In addition to the baseline methods listed in Table 1, we include Δ-FL (Pillutla et al., 2021) and FedProx (Li et al., 2020b) in our evaluation. We adopt a similar setup as presented in Table 1 to assess the performance of these two methods. The summarized results are presented in Table 4, indicating that our proposed algorithm surpasses both Δ-FL and FedProx in terms of worst-20% accuracy and average accuracy.

Table 4: The average and worst-20% top-1 accuracy of our algorithm (SCAFF-PD) vs. state-of-the-art federated learning algorithms evaluated on CIFAR100 and Tiny-ImageNet. The highest top-1 accuracy in each setting is highlighted in **bold**.

Datasets	Methods	Non-i.i.d. degree					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		`average`	`worst-20%`	`average`	`worst-20%`	`average`	`worst-20%`
CIFAR-100	FedProx	38.76	15.58	35.91	24.57	36.49	26.45
	Δ-FL	30.09	7.26	33.18	15.82	31.69	16.63
	SCAFF-PD	**49.03**	**29.30**	**42.06**	**28.37**	**43.69**	**32.77**
		`average`	`worst-20%`	`average`	`worst-20%`	`average`	`worst-20%`
TinyImageNet	FedProx	33.65	18.09	31.52	23.62	34.98	27.59
	Δ-FL	29.06	11.94	36.77	22.24	36.47	20.13
	SCAFF-PD	**41.26**	**25.32**	**39.32**	**30.27**	**41.23**	**29.78**