AN ADAPTIVE PENALTY METHOD FOR HETEROGENEOUS FEDERATED LEARNING WITH NONLINEAR CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) enables collaborative model training without sharing raw data and is typically formulated as unconstrained minimization of a global loss. To the best of our knowledge, no existing methods provide provable convergence guarantees for nonlinearly constrained FL while explicitly mitigating client drift. We propose an adaptive-penalty Moreau-smoothed stochastic controlled federated averaging method that extends SCAFFOLD to FL with both local and global nonlinear constraints. The penalty parameter and stepsize are updated at each communication round according to analytically derived rules. In the convex setting, we achieve $\widetilde{O}(\epsilon^{-2})$ communication complexity for both global suboptimality and constraint violation. In the strongly convex setting, the complexity for both quantities improves to $\widetilde{O}(\epsilon^{-1})$. Some of the unique features of our method include the following: (i) In both regimes, the global suboptimality complexity matches the ϵ-dependence of SCAFFOLD and FedAvg in the unconstrained setting; (ii) unlike limited existing constrained FL analyses that rely on Lagrangian duality and constraint qualifications, our penalty-based analysis establishes upper bounds on both global suboptimality and constraint violation without requiring constraint qualifications. We show that a constraint qualification is needed only to establish the lower bound on global suboptimality; and (iii) under Slater's condition, we show that the expected suboptimality and infeasibility errors converge asymptotically to zero for any arbitrary Moreau smoothing parameter. Numerical experiments on MNIST, EMNIST, and CIFAR-10 show that the proposed method attains competitive performance while consistently reducing constraint violation.

1 INTRODUCTION

Federated Learning (FL) enables communication-efficient training across distributed, heterogeneous, and privacy-sensitive clients, with deployments in healthcare, mobile systems, and industry platforms (Sheller et al., 2020; Hard et al., 2018; Kairouz et al., 2021). Recent work has developed methods with convergence guarantees under non-IID data and characterized client drift (McMahan et al., 2017; Karimireddy et al., 2020; Li et al., 2020; Mishchenko et al., 2022). Among these, SCAFFOLD (Karimireddy et al., 2020) provably mitigates client drift via control variates in heterogeneous settings. Related approaches include FedProx, which stabilizes local updates via proximal regularization, and a growing line of personalized FL methods—such as pFedMe (T Dinh et al., 2020), Per-FedAvg (Fallah et al., 2020), Ditto (Li et al., 2021), PFL (Tan et al., 2022), and FedRep (Collins et al., 2021)—that learn client-specific models.

Despite these advances, federated learning under explicit local or global functional constraints remains largely underexplored. In many applications, learned models must satisfy feasibility or structural constraints imposed directly on the optimization variables, such as norm bounds, sparsity, or general nonlinear constraint sets. However, most existing convergence and personalization frameworks are developed for unconstrained FL objectives, and extending theoretical guarantees to constrained federated optimization remains largely open, particularly under non-IID data. Motivated by this lacuna, in this paper we consider a federated optimization problem under explicit nonlinear local and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

global constraints, of the form

$$\begin{aligned} \min_{x} \quad & f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\tilde{f}_i(x,\xi_i)\right] \\ \text{s.t.} \quad & h_i(x) \leq 0, \quad i = 1,\ldots,m, \\ & h_0(x) \leq 0. \end{aligned} \tag{1}$$

Here, a group of m clients collaboratively seeks to minimize a global objective, where client i is associated with a local dataset $\mathcal{D}_i$, a stochastic loss function $\tilde{f}_i : \mathbb{R}^n \times \mathcal{D}_i \to \mathbb{R}$, and a local constraint function h_i. In addition, all clients are subject to a shared global constraint h_0.

Motivation. The need for FL methods with provable guarantees for solving equation 1 is particularly pronounced in heterogeneous and constrained regimes, where feasibility is closely tied to client participation. In such settings, clients may be unwilling to participate unless their local performance requirements are guaranteed, as global updates can otherwise degrade local utility under data heterogeneity. For instance, in cross-silo FL among hospitals, all sites may train a shared model while hospital i requires $\mathbb{E}_{\zeta_i \sim \mathcal{D}_i^{\text{test}}}\left[\tilde{g}_i(x,\zeta_i)\right] \leq u_i$ on its local patient population, with heterogeneous distributions and acceptable risk thresholds across hospitals. These considerations naturally lead to FL formulations with nonlinear functional constraints at both the client and system levels.

Background and gaps. Penalty methods for convex optimization with inequality constraints replace feasibility with a quadratic ReLU penalty of the form $\frac{1}{2}\sum_j(\max\{0, h_j(x)\})^2$. Classical penalty and exact penalty frameworks (Bertsekas, 1999; Nocedal & Wright, 2006; Rockafellar, 1997) guarantee convergence as the penalty parameter $\rho \to \infty$, but provide limited guidance for selecting or updating ρ, which can lead to either persistent infeasibility or poor conditioning in practice.

Adaptive penalty schemes in convex optimization (Lan & Monteiro, 2013; Chambolle & Pock, 2011; He et al., 2000) update ρ_k based on constraint violation, improving empirical performance. However, they typically provide only asymptotic feasibility guarantees and do not explicitly characterize how the violation metric should govern penalty adaptation or yield sharp non-asymptotic convergence rates. Lagrange multiplier–based methods, including ADMM and primal–dual schemes, have been used for constrained federated learning, e.g., (Gong et al., 2022) and related primal–dual formulations (Wang et al., 2022). These approaches primarily leverage Lagrangian multiplier theory to handle the consensus constraints inherent in FL and do not address the presence of nonlinear local or global functional constraints that may arise in more general FL formulations. Consequently, existing work provides limited understanding of constraint-violation dynamics and tight non-asymptotic feasibility guarantees for federated learning problems with nonlinear constraints.

Contributions. Our goal in this work is to address the problem in equation 1 in non-IID settings and obtain sharp communication complexity bounds for both the suboptimality and infeasibility errors. We summarize our contributions in the following.

(I) A single-loop FL method with adaptive penalty and stepsize. We propose an adaptive penalty stochastic control variate federated averaging method, named `AdaP-SCAFFOLD`. This method extends SCAFFOLD (Karimireddy et al., 2020) to handle FL with local and global nonlinear constraints. Both the penalty parameter and the stepsize are updated at each communication round according to analytically derived update rules. Each client i at round r uses the local penalty gradient term $\rho_r \min\{\max\{0, \eta^{-1}h_i(x)\}, 1\}\nabla h_i(x)$. In contrast to the classical quadratic ReLU penalty, we measure infeasibility via an η-smoothed Moreau envelope of ReLU. We establish key smoothness and growth properties of this penalty, which are central to the analysis. In particular, the smoothed penalty grows linearly in the infeasible region while remaining smooth near the boundary.

(II) Non-asymptotic guarantees under weak assumptions. The proposed penalty scheme yields sharp non-asymptotic rates, with the growth of the penalty parameter ρ_r playing a central role. In particular, there is an inherent tradeoff: faster growth of ρ_r leads to a more rapid reduction in constraint violation, but slows convergence to the optimal solution. In the convex setting, we achieve $\widetilde{O}(\epsilon^{-2})$ communication complexity for both global suboptimality and constraint violation. In the strongly convex setting, the complexity for both quantities improves to $\widetilde{O}(\epsilon^{-1})$. In both regimes, the global suboptimality complexity matches the ϵ-dependence of SCAFFOLD and FedAvg in the unconstrained setting. Unlike limited existing constrained FL analyses that rely on Lagrangian duality and constraint qualifications, our penalty-based analysis establishes upper bounds on both global

suboptimality and constraint violation without requiring constraint qualifications. We show that a constraint qualification is needed only to establish the lower bound on global suboptimality.

(III) Asymptotic guarantees for the original constrained problem. If Slater's condition holds, we show that, for any fixed smoothing parameter $\eta > 0$, the expected smoothed infeasibility and absolute suboptimality errors converge to zero as the number of communication rounds tends to infinity. Importantly, we show that the Moreau smoothing scheme employed in the construction of the penalty term preserves the original feasible set and does not introduce bias.

Related work. Among the works most closely related to ours are Federated Learning with Convex Global and Local Constraints (He et al., 2024), FedDA (Li et al., 2024), and FedSGM (Upadhyay et al., 2026). He et al. consider both local and global nonlinear functional constraints. However, their approach relies on a two-loop proximal augmented-Lagrangian framework with full-batch local gradients, requires full client participation, does not explicitly correct for client drift, and requires constraint qualifications for its convergence guarantees. FedDA addresses federated constrained optimization over a global convex constraint set, with feasibility enforced through projection/proximal mappings, rather than considering client-specific nonlinear functional constraints. FedSGM supports partial participation and nonlinear functional constraints, but enforces an aggregated constraint across clients rather than requiring each client-specific constraint to be satisfied individually. Furthermore, its analysis relies on constraint qualifications. These works leave open the problem of jointly handling client-specific and shared functional constraints while explicitly controlling client drift under partial participation. Table 1 summarizes the key distinctions between our proposed method and the most closely related approaches for constrained FL.

Table 1: Comparison with closely related constrained federated learning methods.

Method	Local functional constraints	Shared/global constraint	Partial participation	Stochastic gradients	Drift correction
He et al. (He et al., 2024)	✓	✓	–	–	–
FedDA (Li et al., 2024)	–	Shared set $\mathcal{X}$	–	✓	–
FedSGM (Upadhyay et al., 2026)	–	Aggregated functional constraint	✓	✓	–
`AdaP-SCAFFOLD` **(our work)**	✓	✓	✓	✓	✓

2 Algorithm Outline

In this section, we introduce our method, where we impose the following assumptions.

Assumption 2.1. Consider the problem in equation 1. Let the following hold.

(i) The function $f_i(x) \triangleq \mathbb{E}_{\xi_i}[\tilde{f}_i(x, \xi_i)]$ is L_i^f-smooth and convex, for all $i \in [m] \triangleq \{1, \ldots, m\}$.

(ii) For any $i \in [m]$, client i has access to a stochastic gradient oracle $\nabla \tilde{f}_i(x, \xi_i)$ such that, for any random x, the following holds almost surely: (ii-a) $\mathbb{E}[\nabla \tilde{f}_i(x, \xi_i) \mid x] = \nabla f_i(x)$. (ii-b) There exists $\sigma > 0$ such that $\mathbb{E}[\|\nabla \tilde{f}_i(x, \xi_i) - \nabla f_i(x)\|^2 \mid x] \leq \sigma^2$.

(iii) The function h_i is L_i^h-smooth and convex, for all $i \in \{0\} \cup [m]$.

(iv) The set $\{x \in \mathbb{R}^n \mid h_i(x) \leq 0\}$ is bounded, for all $i \in \{0\} \cup [m]$.

(v) The global objective $f(x)$ is bounded below on $\mathbb{R}^n$, i.e., there exists $f_{\text{low}} := \inf_x f(x) > -\infty$.

While some of the main non-asymptotic guarantees for our method do not rely on constraint qualifications, we invoke the following assumption for a subset of our results.

Assumption 2.2. Slater's condition holds, i.e., there exists $\tilde{x}$ such that $h_i(\tilde{x}) < 0$, for all $i \in \{0\} \cup [m]$.

Remark 2.3. (i) The boundedness of the feasible set of the problem in equation 1 holds under some standard assumptions. For instance, if each function h_i is either coercive or strongly convex, the sublevel set $\{x : h_i(x) \leq 0\}$ is bounded (Bertsekas et al., 2003, Proposition 3.2.1). Furthermore, the closedness of the feasible set is guaranteed in view of the convexity of the function h_i for all $i \in [m]$ (Bertsekas et al., 2003, Proposition 1.1.2).

(ii) Slater's condition is a widely used assumption in constrained optimization and ensures the existence of at least one point that strictly satisfies all client-level and global constraints. For example, consider the case where client i requires a bounded expected loss on its local test distribution, $\mathbb{E}_{\zeta_i \sim \mathcal{D}_i^{\text{test}}}\left[\tilde{g}_i(x, \zeta_i)\right] \leq u_i$. Slater's condition is met, e.g., if u_i is estimated using a standard first-order method as $u_i \approx \delta_i + \min_x \mathbb{E}_{\zeta_i \sim \mathcal{D}_i^{\text{test}}}\left[\tilde{g}_i(x, \zeta_i)\right]$, where $\delta_i > 0$ is a client-specific hyperparameter

introduced to provide slack and help avoid overfitting. In ill-conditioned settings where $\delta_i := 0$, Slater's condition may fail to hold. Our theory allows us to address the constrained FL problem both when Slater's condition holds and when it may fail to hold.

To address the constrained FL problem in equation 1, we propose `AdaP-SCAFFOLD` given by Algorithm 1. At the beginning of each communication round r, the server samples a subset of clients

Algorithm 1 Adaptive penalty stochastic controlled federated averaging (`AdaP-SCAFFOLD`)

input: Server chooses $\hat{x}_1 \in \mathbb{R}^n$, initial control variates $\hat{c}_1 = 0$ and $c_i^1 = 0$ for $i \in [m]$. Local stepsize $\{\gamma_{l,r}\}$, global stepsize γ_g, penalty parameter $\{\rho_r\}$, Moreau smoothing parameter $\eta > 0$, and synchronization indices T_r for $r = 1, \ldots, R+1$ where $T_{r+1} - T_r = H$.
for $r = 1, \ldots, R$ **do**
  Server samples a subset $\mathcal{S}_r \subseteq [m]$ with $|\mathcal{S}_r| = s$, and broadcasts $\hat{x}_r$ and $\hat{c}_r$ to all clients in $\mathcal{S}_r$.
  For $i \in \mathcal{S}_r$, initialize local state: $x_{i,T_r} := \hat{x}_r$.
  for $t := T_r, \ldots, T_{r+1} - 1$ **in parallel do**
    Client i samples a mini-batch $\xi_{i,t} \sim \mathcal{D}_i$
    $g_{i,t} := \nabla \tilde{f}_i(x_{i,t}, \xi_{i,t}) + \rho_r \sum_{j \in \{0,i\}} \min\left\{\max\{0, \eta^{-1} h_j(x_{i,t})\}, 1\right\} \nabla h_j(x_{i,t})$
    $x_{i,t+1} := x_{i,t} - \gamma_{l,r}(g_{i,t} - c_i^r + \hat{c}_r)$
  end for
  For each $i \in \mathcal{S}_r$, compute:
    $c_i^{r+1} := c_i^r - \hat{c}_r + \frac{1}{H\gamma_{l,r}}(\hat{x}_r - x_{i,T_{r+1}})$
    $\Delta x_{i,r} := x_{i,T_{r+1}} - \hat{x}_r, \quad \Delta c_i^r := c_i^{r+1} - c_i^r$
  Server aggregates:
    $\hat{x}_{r+1} := \hat{x}_r + \frac{\gamma_g}{s} \sum_{i \in \mathcal{S}_r} \Delta x_{i,r}, \quad \hat{c}_{r+1} := \hat{c}_r + \frac{1}{m} \sum_{i \in \mathcal{S}_r} \Delta c_i^r$
end for
output: Final global model $\hat{x}_{R+1}$

$\mathcal{S}_r$, and broadcasts the current global model $\hat{x}_r$ and global control variate $\hat{c}_r$ to the selected clients. Each selected client initializes its local model at $\hat{x}_r$ and performs H local updates using the smoothed adaptive penalty stochastic gradient $g_{i,t}$. The correction term $-c_i^r + \hat{c}_r$ follows the SCAFFOLD control-variate technique used to reduce client drift caused by data heterogeneity (Karimireddy et al., 2020). After completing the local updates, each selected client updates its local control variate c_i^{r+1}, and sends both the model update $\Delta x_{i,r}$ and the control-variate update Δc_i^r back to the server. The server then aggregates the received model updates to obtain $\hat{x}_{r+1}$, and updates the global control variate $\hat{c}_{r+1}$. Clients that are not selected keep their local control variates unchanged.

A key novelty lies in an adaptive penalty-driven update that handles both local and global constraints through a single gradient term $\rho_r \sum_{j \in \{0,i\}} \min\{\max\{0, \eta^{-1} h_j(x_{i,t})\}, 1\} \nabla h_j(x_{i,t})$, with a round-varying penalty parameter ρ_r. This allows both local ($j = i$) and global ($j = 0$) constraints to be enforced within the same update. Unlike SCAFFOLD, which uses a fixed stepsize, we introduce a round-varying local stepsize $\gamma_{l,r}$ to analytically balance optimization and constraint satisfaction under heterogeneity. Rather than treating penalty growth and stepsize independently, we design them jointly, and derive explicit update rules for ρ_r and $\gamma_{l,r}$ that ensure both asymptotic convergence and sharp non-asymptotic rates.

3 Moreau Smoothed ReLU Penalty Function

Given a functional constraint $h_i(x) \leq 0$, a standard penalty used in augmented Lagrangian methods (Bertsekas, 1999; Nocedal & Wright, 2006; Rockafellar, 1997) is the quadratic ReLU penalty $p_i(x) \triangleq \frac{1}{2}(\max\{0, h_i(x)\})^2$. If h_i is convex, then p_i is convex and continuously differentiable with $\nabla p_i(x) = \max\{0, h_i(x)\} \nabla h_i(x)$. However, even when h_i is smooth (i.e., has Lipschitz continuous gradients), ∇p_i need not be globally Lipschitz, due to the potentially unbounded growth of the product $h_i(x) \nabla h_i(x)$. This violates standard L-smoothness assumptions commonly required in the analysis of federated optimization methods. To circumvent this, next we consider an η-Moreau smoothed version of the ReLU penalty function as follows. Figure 1 provides a comparison of the ReLU, quadratic ReLU, and its Moreau-smoothed variant.

Definition 3.1. Let $r(\tau) \triangleq \max\{0, \tau\}$ for $\tau \in \mathbb{R}$. The closed form of η-Moreau envelope of r is given as

$$r_\eta(\tau) \triangleq \frac{1}{2\eta}[\min\{\eta, \max\{0, \tau\}\}]^2 + \max\{0, \tau - \eta\}.$$

For each $i \in \{0, \dots, m\}$, we define the smoothed penalty function $g_{i,\eta}(x) \triangleq r_\eta(h_i(x))$. The global aggregate smoothed penalty function is defined as $\bar{g}_\eta(x) \triangleq g_{0,\eta}(x) + \frac{1}{m}\sum_{i=1}^m g_{i,\eta}(x)$.

Our goal in the remainder of this section is to provide detailed properties of the penalty function $g_{i,\eta}(x)$. To this end, we use the following terms.

Definition 3.2. For a function $h : \mathbb{R}^n \to \mathbb{R}$ and $\alpha \in \mathbb{R}$, we define the sublevel and superlevel sets as

$$\mathcal{L}_\alpha^{\leq}(h) \triangleq \{x \in \mathbb{R}^n \mid h(x) \leq \alpha\},$$
$$\mathcal{L}_\alpha^{\geq}(h) \triangleq \{x \in \mathbb{R}^n \mid h(x) \geq \alpha\}.$$

Similarly, $\mathcal{L}_\alpha^{<}(h)$ consists of all points x such that $h(x)$ is strictly less than α, and $\mathcal{L}_\alpha^{>}(h)$ consists of all points x such that $h(x)$ is strictly greater than α.

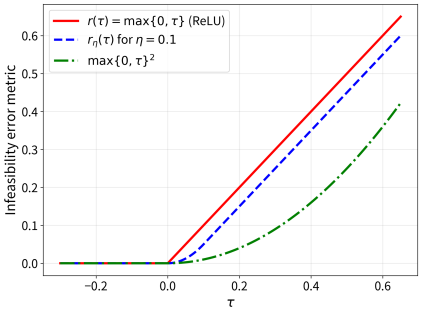


Figure 1: $r_\eta(\tau)$ smoothly approximates ReLU while preserving its minimizers ($\tau \leq 0$). It grows linearly for $\tau > \eta$, unlike the quadratic ReLU $\max\{0, \tau\}^2$.

Next, for any $i \in \{0, 1, \dots, m\}$, we partition $\mathbb{R}^n$ into three local sets as $\mathcal{X}_i \triangleq \mathcal{L}_0^{\leq}(h_i)$, $\mathcal{X}_i^{(\eta,\infty)} \triangleq \mathcal{L}_\eta^{>}(h_i)$, and $\mathcal{X}_i^{(0,\eta]} \triangleq \mathcal{L}_\eta^{\leq}(h_i) \cap \mathcal{L}_0^{>}(h_i)$. We further split $\mathcal{X}_i^{(\eta,\infty)}$ into $\mathcal{X}_i^{(\eta,\eta+1]} \triangleq \mathcal{L}_{\eta+1}^{\leq}(h_i) \cap \mathcal{L}_\eta^{>}(h_i)$ and $\mathcal{X}_i^{(\eta+1,\infty)} \triangleq \mathcal{L}_{\eta+1}^{>}(h_i)$.

Notably, the feasible set in equation 1 can be characterized as the sublevel intersection $\cap_{i=0}^m \mathcal{X}_i$. Definition 3.2 separates the behavior of each constraint according to the infeasibility level. Points in $\mathcal{X}_i$ are feasible for client i, points in $\mathcal{X}_i^{(0,\eta]}$ lie in the smoothing region near the constraint boundary, and points in $\mathcal{X}_i^{(\eta,\infty)}$ are infeasible. Next, we provide a compactness property followed by a definition.

Lemma 3.3. Suppose Assumption 2.1 holds. Consider Definition 3.2. The closures of the sets $\mathcal{X}_i^{(0,\eta]}$ and $\mathcal{X}_i^{(\eta,\eta+1]}$ are compact sets.

Definition 3.4. For any $i \in \{0, \dots, m\}$, we define $L_{\mathcal{X}_i^{(0,\eta]}}$ and $L_{\mathcal{X}_i^{(\eta,\infty)}}$ as the Lipschitz constants of h_i within the sets $\mathcal{X}_i^{(0,\eta]}$ and $\mathcal{X}_i^{(\eta,\eta+1]}$, respectively.

$$L_{\mathcal{X}_i^{(0,\eta]}} \triangleq \sup_{x \in \mathcal{X}_i^{(0,\eta]}} \|\nabla h_i(x)\|, \quad L_{\mathcal{X}_i^{(\eta,\eta+1]}} \triangleq \sup_{x \in \mathcal{X}_i^{(\eta,\eta+1]}} \|\nabla h_i(x)\|,$$

where $L_{\mathcal{X}_i^{(0,\eta]}} := 0$ if $\mathcal{X}_i^{(0,\eta]} = \emptyset$. Similarly, $L_{\mathcal{X}_i^{(\eta,\eta+1]}} := 0$ if $\mathcal{X}_i^{(\eta,\eta+1]} = \emptyset$. Furthermore, we define the following strictly positive scalars given as

$$\delta_i \triangleq \inf_{x \in \mathcal{X}_i^{(0,\eta]},\, y \in \mathcal{X}_i^{(\eta+1,\infty)}} \|x - y\|, \quad \nu_i \triangleq \inf_{x \in \mathcal{X}_i,\, y \in \mathcal{X}_i^{(\eta,\infty)}} \|x - y\|,$$

where $\delta_i := \infty$ if $\mathcal{X}_i^{(0,\eta]} = \emptyset$ or $\mathcal{X}_i^{(\eta,\eta+1]} = \emptyset$. Further, $\nu_i := \infty$ if $\mathcal{X}_i^{(\eta,\infty)} = \emptyset$.

Remark 3.5. The existence of finite scalars $L_{\mathcal{X}_i^{(0,\eta]}}$ and $L_{\mathcal{X}_i^{(\eta,\infty)}}$ in Definition 3.4 is guaranteed by the smoothness of h_i and Lemma 3.3. Since h_i is continuously differentiable, the gradient norm $\|\nabla h_i(x)\|$ is continuous. Consequently, on any bounded domain, the gradient norm is bounded (Rudin, 1976, Theorem 4.15). This upper bound serves as the local Lipschitz constant for h_i on the respective sets.

Proposition 3.6 (Moreau smoothed ReLU penalty)**.** Let Assumption 2.1 hold and suppose $\eta > 0$ is an arbitrary scalar. Then, the local and global penalty functions satisfy the following properties:

(I) Basic properties without constraint qualifications.
(I-i) A vector x is feasible for the problem in equation 1 if and only if $g_{i,\eta}(x) = 0$ for all $i \in \{0, \dots, m\}$.

(I-ii) For all $i \in \{0, \dots, m\}$, the function $g_{i,\eta}(x)$ is continuously differentiable and

$$\nabla g_{i,\eta}(x) = \min\left\{\max\{0, \eta^{-1} h_i(x)\}, 1\right\} \nabla h_i(x).$$

Further, $g_{i,\eta}(x)$ is convex and L_i^g-smooth, where

$$L_i^g \triangleq L_i^h + L_{\mathcal{X}_i^{(0,\eta]}} \max\{L_{\mathcal{X}_i^{(0,\eta]}}/\eta,\, L_{\mathcal{X}_i^{(\eta,\infty)}}/\eta,\, 1/\delta_i,\, 1/\nu_i\}.$$

(II) Error bound under Slater's condition. Additionally, let Assumption 2.2 hold. Then, for any $i \in \{0, \dots, m\}$, there exists $\hat{y}_i \in \mathbb{R}^n$ such that $h_i(\hat{y}_i) = 0$ and $\|\nabla h_i(\hat{y}_i)\| > 0$. Further:

(II-i) For all $x \in \mathcal{X}_i$, $g_{i,\eta}(x) = 0$. (II-ii) For all $x \in \mathcal{X}_i^{(0,\eta]}$, $\frac{\|\nabla h_i(\hat{y}_i)\|^2}{2\eta} \operatorname{dist}^2(x, \mathcal{X}_i) \le g_{i,\eta}(x) \le \frac{\eta}{2}$. (II-iii) For all $x \in \mathcal{X}_i^{(\eta,\infty)}$, $\frac{\|\nabla h_i(\hat{y}_i)\|}{2} \operatorname{dist}(x, \mathcal{X}_i) \le g_{i,\eta}(x)$.

Proposition 3.6 provides insights on why the smoothed penalty is suitable for our algorithm and analysis. Part (I) establishes its basic properties without requiring any constraint qualification conditions. In (I-i), we show that the smoothed penalty preserves the exact notion of feasibility: zero penalty is equivalent to satisfying the original constraints. Part (I-ii) shows that smoothing turns the nonsmooth violation function into a differentiable, convex, and smooth function, so it can be incorporated into gradient-based local updates and the SCAFFOLD correction. Under Slater's condition, part (II) provides the error-bound relation between the smoothed penalty value and the distance to the feasible set. This relates the smoothed infeasibility metrics to the distance from the original local and global feasibility sets.

4 Main Convergence Guarantees

In this section, we present the main convergence guarantees for the proposed adaptive-penalty method in both convex and strongly convex settings. While the analysis is deferred to the Supplementary Material, we highlight key technical challenges and novel ideas for addressing them.

Technical challenges. Naturally, the analysis of adaptive penalty optimization schemes is substantially more challenging than their constant-penalty counterparts, mainly because as the penalty parameter increases, the smoothness parameter of the penalty function increases, often enforcing the need for an adaptive stepsize rule. In our proposed FL method, this issue exacerbates, because under client heterogeneity and partial participation, increasing ρ_r strengthens feasibility enforcement but also increases the L-smooth parameter of the dynamic (round-varying) penalty function, and thus amplifies penalty-dependent errors in the client-drift and control-lag estimates. Another key challenge lies in decoupling suboptimality and infeasibility error metrics and establishing separate bounds on these metrics. This is non-trivial because the local update in our adaptive-penalty scheme couples the gradient of the local objective with those of the local and global constraints through the penalty term. Furthermore, $f(\hat{x}_r) - f^*$ may be negative at infeasible iterates, so convergence of the penalized objective does not directly yield separate suboptimality and infeasibility guarantees.

We address these challenges by analyzing and balancing penalty growth with stepsize decay and decoupling the two errors through suitably defined weighted averaged models. The analysis derives separate bounds for suboptimality and infeasibility by (i) exploiting the nonnegativity of the smoothed infeasibility error and (ii) using a lower bound on the suboptimality error. Both resulting bounds depend explicitly on the penalty and stepsize sequences. This allows us to jointly choose the penalty and stepsize updates so that the penalty grows while the resulting updates remain controlled despite client heterogeneity and partial participation. We then obtain separate upper bounds on suboptimality and infeasibility without requiring constraint qualifications. Under Slater's condition, we further invoke the penalty–distance relation in Proposition 3.6 and obtain a lower bound on the suboptimality error. We further utilize this bound to establish the asymptotic convergence properties.

Theorem 4.1 (Convex Setting)**.** Let Assumption 2.1 hold. Choose $\gamma_g = \sqrt{s}$, $\gamma_{l,r} = \frac{c_l}{H\sqrt{s}\sqrt{r}(1+\log r)}$, and $\rho_r = c_\rho \sqrt{r}(1 + \log r)$ for scalars $c_l, c_\rho > 0$ satisfying conditions in Lemma A.18. Define the weighted output $\bar{x}_R := \frac{\sum_{r=1}^R w_{r+1}\tilde{\gamma}_r \hat{x}_r}{\sum_{r=1}^R w_{r+1}\tilde{\gamma}_r}$, where $\{w_r\}$ is recursively defined by $w_1 = 1$ and $w_{r+1}(1 + B_r) = w_r$ for any $r \ge 1$. Then, the following hold (see Definition A.9 for notation).

(i) [Suboptimality-Upper Bound and Lower Estimate]

$$-G_*\mathbb{E}[\mathrm{dist}(\bar{x}_R,\mathcal{X})] \le \mathbb{E}\left[f(\bar{x}_R)-f^*\right] \le \frac{\exp(448c_l^2L_f^2)(1+\log R)}{\sqrt{R}}\left[\frac{D^2}{2c_l}+20c_lG_*^2+\frac{8c_l\sigma^2}{sH}+\frac{52c_l\sigma^2}{H}\right].$$

(ii) [Infeasibility]

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \le \frac{\exp(448c_l^2L_f^2)(1+\log R)}{c_\rho\sqrt{R}}\left[\frac{D^2}{c_l}+40c_lG_*^2+\frac{16c_l\sigma^2}{sH}+\frac{104c_l\sigma^2}{H}+4(f^*-f_{\mathrm{low}})\right].$$

(iii) [Guarantees under Slater's condition] Suppose, in addition, that Assumption 2.2 holds.
(iii-1) [Lower bound on suboptimality]

$$\mathbb{E}[f(\bar{x}_R)-f^*] \ge -\frac{2mG_*}{\kappa_{\min}}\frac{\exp(448c_l^2L_f^2)(1+\log R)}{c_\rho\sqrt{R}}\left[\frac{D^2}{c_l}+40c_lG_*^2+\frac{16c_l\sigma^2}{sH}+\frac{104c_l\sigma^2}{H}+4(f^*-f_{\mathrm{low}})\right]-\mathcal{O}(\eta).$$

(iii-2) [Asymptotic convergence] For any fixed $\eta>0$, as $R\to\infty$, the expected smoothed infeasibility $\mathbb{E}[\bar{g}_\eta(\bar{x}_R)]$ and the expected absolute suboptimality error $\mathbb{E}[|f(\bar{x}_R)-f^*|]$ converge to zero.

Remark 4.2. Theorem 4.1 establishes convergence rate guarantees for the problem in equation 1. Compared to SCAFFOLD, a main challenge arises because the penalized objective changes across rounds, which introduces additional penalty-dependent error terms. The theorem shows that these error terms can be effectively controlled by combining control variates with a suitably increasing penalty parameter, relying on the infeasibility growth property established in Proposition 3.6. Consequently, under the proposed stepsize and penalty update rules, the method achieves an upper bound of $\widetilde{O}(\epsilon^{-2})$ on both suboptimality and infeasibility, as well as an $\widetilde{O}(\epsilon^{-2})$ bound on the η-*approximate* suboptimality lower bound.

Theorem 4.3 (Strongly convex setting)**.** Let Assumption 2.1 hold. Suppose that each local objective f_i is μ-strongly convex for some $\mu>0$. Let x^* denote the unique optimal solution in equation 1. Choose $\gamma_g=\sqrt{s}$, $\gamma_{l,r}=\frac{c_l}{H\sqrt{s}\left(1+\frac{\mu c_l}{4}(r-1)\right)}$, and $\rho_r=c_\rho\left(1+\frac{\mu c_l}{4}(r-1)\right)$, where $c_l,c_\rho>0$ satisfy the conditions in Lemma A.22. Define the weighted output $\bar{x}_R^{\mathrm{sc}}:=\frac{\sum_{r=1}^R w_{r+1}\tilde{\gamma}_r\hat{x}_r}{\sum_{r=1}^R w_{r+1}\tilde{\gamma}_r}$, such that $w_1=1$ and $w_{r+1}\left(1-\frac{\mu\tilde{\gamma}_r}{2}+B_r\right)=w_r$. Then the following hold (see Definition A.9 for notation).
(i) [Suboptimality-Upper Bound]

$$\frac{\mu}{2}\mathbb{E}\left[\|\bar{x}_R^{\mathrm{sc}}-x^*\|^2\right]-G_*\,\mathbb{E}[\mathrm{dist}(\bar{x}_R^{\mathrm{sc}},\mathcal{X})] \le \mathbb{E}[f(\bar{x}_R^{\mathrm{sc}})-f^*] \le \frac{4\exp\left(\frac{2240c_lL_f^2}{\mu}\right)}{\mu(R+1)}\left[\frac{D^2}{c_l^2R}+20G_*^2+\frac{8\sigma^2}{sH}+\frac{52\sigma^2}{H}\right].$$

(ii) [Infeasibility]

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \le \frac{8\exp\left(\frac{2240c_lL_f^2}{\mu}\right)}{\mu^2c_lc_\rho(R+1)}\left[\frac{\mu D^2}{c_lR}+G_*^2+\frac{1+\log R}{R}\left(80G_*^2+\frac{32\sigma^2}{sH}+\frac{208\sigma^2}{H}\right)\right].$$

(iii) [Guarantees under Slater's condition] Suppose additionally that Assumption 2.2 holds.
(iii-1) [η-approximate lower bound on suboptimality]

$$\begin{aligned}\mathbb{E}[f(\bar{x}_R^{\mathrm{sc}})-f^*] \ge{}& \frac{\mu}{2}\mathbb{E}\left[\|\bar{x}_R^{\mathrm{sc}}-x^*\|^2\right]-\mathcal{O}(\eta)\\ &-\frac{2mG_*}{\kappa_{\min}}\frac{8\exp\left(\frac{2240c_lL_f^2}{\mu}\right)}{\mu^2c_lc_\rho(R+1)}\left[\frac{\mu D^2}{c_lR}+G_*^2+\frac{1+\log R}{R}\left(80G_*^2+\frac{32\sigma^2}{sH}+\frac{208\sigma^2}{H}\right)\right].\end{aligned}$$

(iii-2) [Asymptotic convergence] For any fixed $\eta>0$, as $R\to\infty$, the expected smoothed infeasibility error $\mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})]$, the expected absolute suboptimality error $\mathbb{E}[|f(\bar{x}_R^{\mathrm{sc}})-f^*|]$, and $\mathbb{E}\left[\|\bar{x}_R^{\mathrm{sc}}-x^*\|^2\right]$ converge to zero.

Remark 4.4. Theorem 4.3 shows that strong convexity improves the communication complexity of `AdaP-SCAFFOLD`. In the strongly convex setting, the additional contraction term $-\mu\tilde{\gamma}_r/2$ in the Lyapunov recursion yields a stronger weighted telescoping argument than in the convex case. This stronger contraction leads to sharper control of the accumulated optimization and infeasibility errors, resulting in faster convergence rates. Consequently, both the suboptimality upper bound and infeasibility decrease at a faster rate. As a result, `AdaP-SCAFFOLD` achieves $\widetilde{O}(\epsilon^{-1})$ communication complexity for both suboptimality and infeasibility in the strongly convex setting, improving upon the $\widetilde{O}(\epsilon^{-2})$ complexity obtained in the convex setting.

Remark 4.5. Theorems A.16 and A.20 in the supplementary material establish vanishing expected smoothed infeasibility and absolute suboptimality error for any fixed smoothing parameter $\eta > 0$. Since Proposition 3.6 shows that $g_{i,\eta}(x) = 0$ if and only if $h_i(x) \leq 0$, the Moreau smoothing preserves the feasible set and does not introduce asymptotic feasibility bias.

5 EXPERIMENTS

In this section, we evaluate `AdaP-SCAFFOLD` on MNIST, EMNIST, and CIFAR-10 under heterogeneous data distributions. Our experiments consist of two parts: comparison with existing constrained FL methods, and comparison with standard and personalized FL baselines. For each client, the constraint threshold u_i is calibrated using a warmup procedure. Detailed experimental settings are provided in Appendix A.5.

5.1 COMPARISON WITH PROXAL UNDER DATA HETEROGENEITY

We compare `AdaP-SCAFFOLD` with ProxAL (He et al., 2024), a closely related constrained FL method for local and global functional constraints. We vary the Dirichlet parameter α to study performance under different levels of client heterogeneity and report the results in Table 2.

Table 2: Performance under varying data heterogeneity.

Dataset	Heterogeneity	Method	Acc↑	Loss↓	Violation↓	Feasible Ratio↑
CIFAR-10	$\alpha = 10$	ProxAL	0.3992 ± 0.0009	1.7453 ± 0.0016	0.0969 ± 0.0111	$\mathbf{0.4000 \pm 0.0000}$
		AdaP-SCAFFOLD	$\mathbf{0.4026 \pm 0.0006}$	$\mathbf{1.7361 \pm 0.0031}$	$\mathbf{0.0949 \pm 0.0139}$	$\mathbf{0.4000 \pm 0.0000}$
	$\alpha = 1$	ProxAL	0.3888 ± 0.0051	1.7639 ± 0.0185	$\mathbf{0.0816 \pm 0.0103}$	0.2667 ± 0.1155
		AdaP-SCAFFOLD	$\mathbf{0.4028 \pm 0.0038}$	$\mathbf{1.7376 \pm 0.0154}$	0.0884 ± 0.0210	$\mathbf{0.4000 \pm 0.0000}$
	$\alpha = 0.1$	ProxAL	0.4038 ± 0.0177	1.7445 ± 0.0330	$\mathbf{0.0426 \pm 0.0274}$	0.1333 ± 0.2309
		AdaP-SCAFFOLD	$\mathbf{0.4301 \pm 0.0053}$	$\mathbf{1.6955 \pm 0.0214}$	0.0833 ± 0.0454	$\mathbf{0.4000 \pm 0.2000}$
EMNIST	$\alpha = 10$	ProxAL	0.6408 ± 0.0002	1.3444 ± 0.0002	0.0950 ± 0.0002	0.0000 ± 0.0000
		AdaP-SCAFFOLD	$\mathbf{0.6870 \pm 0.0004}$	$\mathbf{1.1703 \pm 0.0005}$	$\mathbf{0.0083 \pm 0.0022}$	$\mathbf{0.6667 \pm 0.1155}$
	$\alpha = 1$	ProxAL	0.6407 ± 0.0008	1.3437 ± 0.0029	0.0943 ± 0.0029	0.0000 ± 0.0000
		AdaP-SCAFFOLD	$\mathbf{0.6864 \pm 0.0013}$	$\mathbf{1.1710 \pm 0.0030}$	$\mathbf{0.0067 \pm 0.0057}$	$\mathbf{0.7333 \pm 0.1155}$
	$\alpha = 0.1$	ProxAL	0.6417 ± 0.0044	1.3430 ± 0.0090	0.0933 ± 0.0091	0.0000 ± 0.0000
		AdaP-SCAFFOLD	$\mathbf{0.6880 \pm 0.0041}$	$\mathbf{1.1662 \pm 0.0138}$	$\mathbf{0.0198 \pm 0.0084}$	$\mathbf{0.6667 \pm 0.1155}$
MNIST	$\alpha = 10$	ProxAL	0.9045 ± 0.0012	0.3605 ± 0.0007	0.0374 ± 0.0008	0.2667 ± 0.1155
		AdaP-SCAFFOLD	$\mathbf{0.9219 \pm 0.0003}$	$\mathbf{0.2783 \pm 0.0018}$	$\mathbf{0.0092 \pm 0.0005}$	$\mathbf{0.6000 \pm 0.0000}$
	$\alpha = 1$	ProxAL	0.9024 ± 0.0045	0.3647 ± 0.0051	0.0407 ± 0.0070	0.4000 ± 0.2000
		AdaP-SCAFFOLD	$\mathbf{0.9221 \pm 0.0021}$	$\mathbf{0.2762 \pm 0.0035}$	$\mathbf{0.0126 \pm 0.0056}$	$\mathbf{0.7333 \pm 0.1155}$
	$\alpha = 0.1$	ProxAL	0.9007 ± 0.0008	0.3674 ± 0.0053	0.0416 ± 0.0132	0.3333 ± 0.1155
		AdaP-SCAFFOLD	$\mathbf{0.9241 \pm 0.0049}$	$\mathbf{0.2703 \pm 0.0116}$	$\mathbf{0.0188 \pm 0.0077}$	$\mathbf{0.5333 \pm 0.1155}$

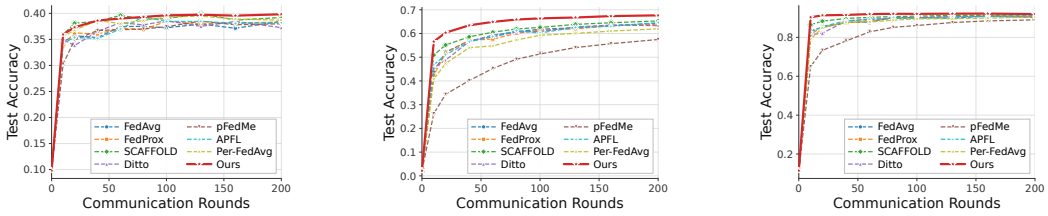
ProxAL assumes **full client participation** and **deterministic full gradients**, whereas AdaP-SCAFFOLD supports partial participation and stochastic gradients. For a controlled comparison, we evaluate both methods under full participation with full-gradient updates and match their computational budgets at 2,000 equivalent full-gradient evaluations. Although this setting removes sampling noise, deterministic client drift remains under heterogeneous local objectives, which AdaP-SCAFFOLD mitigates through its control variates. In some CIFAR-10 settings, ProxAL achieves a smaller average violation because its augmented-Lagrangian subproblems place greater emphasis on reducing the magnitude of active constraint residuals, potentially at the expense of predictive performance. Moreover, average violation and feasible ratio capture different aspects of feasibility: ProxAL may produce more clients with small positive violations, whereas AdaP-SCAFFOLD makes more clients feasible but leaves larger residuals on a few infeasible clients.

5.2 COMPARISON WITH EXISTING FL METHODS

In the heterogeneous setting with $\alpha = 0.3$, Table 4 in the Appendix and Figure 2 show that `AdaP-SCAFFOLD` maintains competitive predictive performance while consistently improving constraint satisfaction, achieving lower average violation and higher feasible ratios than the unconstrained FL baselines.

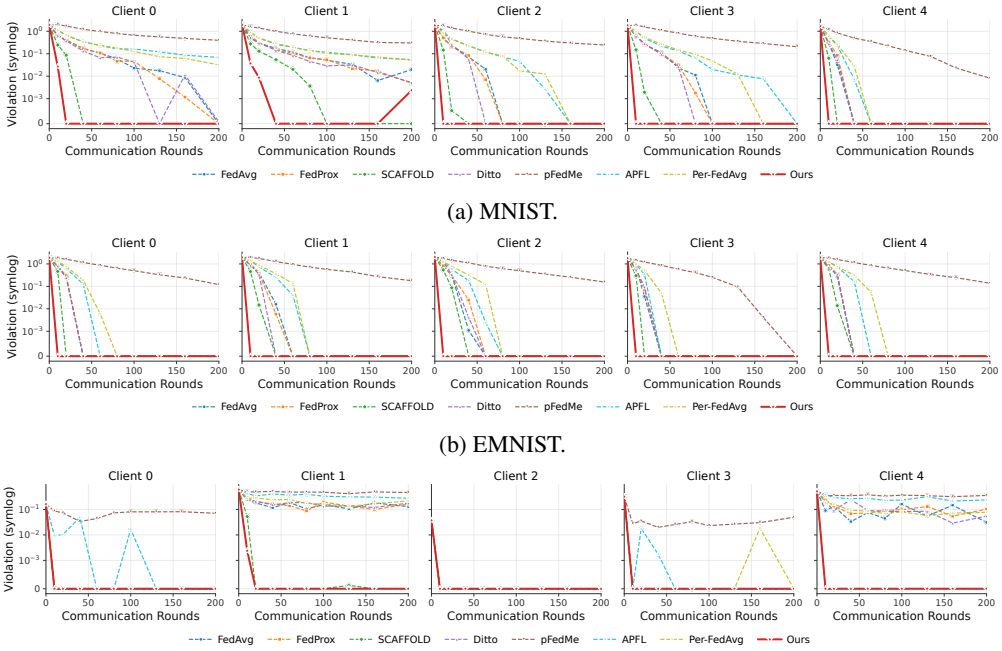
Unlike SCAFFOLD, which minimizes only the unconstrained training objective, AdaP-SCAFFOLD enforces client-wise constraints during training. Therefore, improved constraint satisfaction on the

Figure 2: Convergence comparison with personalized FL methods.



held-out test data is expected when the learned feasibility generalizes beyond the training sample, which indicates that AdaP-SCAFFOLD improves feasibility without sacrificing predictive performance. Figure 3 reports per-client constraint violations across the three datasets. `AdaP-SCAFFOLD` generally drives violations toward the feasibility floor more rapidly and consistently across clients than the unconstrained FL baselines. These results complement the aggregate metrics in Table 4, indicating that the improvement in constraint satisfaction is broadly distributed across clients rather than driven by a small subset.

Figure 3: Per-client constraint violation over training rounds.



(a) MNIST.

(b) EMNIST.

(c) CIFAR-10.

6 Conclusion

In this paper, we studied federated learning with explicit local and global nonlinear constraints under heterogeneous data distributions. To address this setting, we proposed `AdaP-SCAFFOLD`, an adaptive penalty extension of SCAFFOLD that combines control variates with analytically designed penalty and stepsize updates. Our analysis established sharp non-asymptotic guarantees for both suboptimality and constraint violation in convex and strongly convex settings, while preserving the convergence rates of SCAFFOLD for the objective error. Under Slater's condition, we also showed that the proposed adaptive penalty mechanism yields asymptotic convergence to an optimal solution of the original constrained problem. Experimental results on MNIST, EMNIST, and CIFAR-10 demonstrated that `AdaP-SCAFFOLD` consistently improves feasibility performance while maintaining competitive predictive accuracy under heterogeneous client data. Future work includes extending the framework to the general nonconvex setting, designing adaptive penalty update rules for this setting, and evaluating the method in more realistic large-scale federated learning applications.

Reproducibility Statement

Detailed experimental settings, including the datasets, model architecture, client partitioning, training configuration, constraint construction, baseline implementations, and evaluation metrics, are described in the numerical experiments in the main text and in Appendix A.5. Complete statements and proofs of the theoretical results are provided in the supplementary material.

AI Use Statement

In this work, we used generative AI tools for language editing and polishing. We have not used generative AI tools for developing ideas or mathematical proof sketches, and remaining are not applicable to this work. Additionally, we used generative AI tools for review code for syntax errors. We have reviewed all AI-assisted work. Any AI-assisted suggestions regarding syntax were independently checked by the authors, and the relevant code was executed and verified for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Heinz H Bauschke, Jonathan M Borwein, and Wu Li. Strong conical hull intersection property, bounded linear regularity, jameson's property (g), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

Yonghai Gong, Yichuan Li, and Nikolaos M. Freris. Fedadmm: A robust federated deep learning framework with adaptivity to system heterogeneity. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 2575–2587, 2022. doi: 10.1109/ICDE53745.2022.00238.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Bing-Sheng He, Hai Yang, and SL Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications*, 106(2):337–356, 2000.

Chuan He, Le Peng, and Ju Sun. Federated learning with convex global and local constraints. *Transactions on machine learning research*, 2024.

Peter Kairouz, Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Guanghui Lan and Renato DC Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1):115–139, 2013.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Junyi Li, Feihu Huang, and Heng Huang. Fedda: Faster adaptive gradient methods for federated constrained optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International conference on machine learning*, pp. 15750–15769. PMLR, 2022.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN 9780070856134. URL `https://books.google.com/books?id=kwqzPAAACAAJ`.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Antesh Upadhyay, Sang Bin Moon, and Abolfazl Hashemi. Fedsgm: A unified framework for constraint aware, bidirectionally compressed, multi-step federated optimization. *arXiv preprint arXiv:2601.16897*, 2026.

Han Wang, Siddartha Marella, and James Anderson. Fedadmm: A federated primal-dual algorithm allowing partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 287–294. IEEE, 2022.

A Technical Appendices and Supplementary Material

A.1 Proof of Lemma 3.3

Proof. First, we show that the sets are bounded. By Assumption 2.1 (iv), the feasible set $\{x \mid h_i(x) \leq 0\}$ is compact (and thus bounded). Since h_i is a continuous function (Assumption 2.1 (iii)), the boundedness of one nonempty level set of $h_i(x)$ implies that $\{x \mid h_i(x) \leq \alpha\}$ is bounded for any $\alpha \in \mathbb{R}$ (Rockafellar, 1997, Corollary 8.7.1). Observe that both $\mathcal{X}_i^{(0,\eta]}$ and $\mathcal{X}_i^{(\eta,\eta+1]}$ are subsets of the sublevel set $\{x \mid h_i(x) \leq \eta+1\}$. Since this sublevel set is bounded, the closures $\mathcal{X}_i^{(0,\eta]}$ and $\mathcal{X}_i^{(\eta,\eta+1]}$ are necessarily bounded. Recall that a convex function over $\mathbb{R}^n$ is continuous. The closedness of the two sets follows from the continuity of $h_i(x)$ (Bertsekas et al., 2003). □

A.2 Proof of Proposition 3.6

Proof. (I-i) In the first step, let x_0 be a feasible vector to the problem in equation 1. Then, $h_i(x_0) \leq 0$ for all $i \in \{0, \ldots, m\}$. This implies that

$$g_{i,\eta}(x_0) = r_\eta(h_i(x_0)) = \frac{1}{2\eta}[\min\{\eta, \max\{0, h_i(x_0)\}\}]^2 + \max\{0, h_i(x_0) - \eta\} = 0.$$

In the second step, let x_0 be a vector such that $g_{i,\eta}(x_0) = 0$ for all $i \in \{0, \ldots, m\}$. This implies that $\frac{1}{2\eta}[\min(\eta, \max(0, h_i(x_0)))]^2 = 0$. Thus, $\max(0, h_i(x_0)) = 0$, implying that $h_i(x_0) \leq 0$, for all i. This completes the proof of (I-i).

(I-ii) From Definition 3.1 and the differentiability of $r_\eta(\cdot)$, $g_{i,\eta}(\cdot)$ is continuously differentiable and

$$\nabla g_{i,\eta}(x) = \begin{cases} 0, & h_i(x) < 0, \\ \frac{h_i(x)\nabla h_i(x)}{\eta}, & 0 \leq h_i(x) < \eta, \\ \nabla h_i(x), & h_i(x) \geq \eta. \end{cases}$$

This can be compactly written as $\nabla g_{i,\eta}(x) = \min\left\{\max\{0, \eta^{-1} h_i(x)\}, 1\right\} \nabla h_i(x)$.

Next, we show the smoothness property. Consider the following six cases:

Case 1. If either $x, y \in \mathcal{X}_i$ or $x, y \in \mathcal{X}_i^{(\eta,\infty)}$, then in view of Assumption 2.1, we have

$$\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| \leq L_i^h \|x - y\|.$$

Case 2. If $x, y \in \mathcal{X}_i^{(0,\eta]}$, then we have

$$\begin{aligned} \|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| &\leq \frac{1}{\eta}\|h_i(x)\nabla h_i(x) - h_i(y)\nabla h_i(y)\| \\ &= \frac{1}{\eta}\|h_i(x)(\nabla h_i(x) - \nabla h_i(y)) + \nabla h_i(y)(h_i(x) - h_i(y)\| \\ &\leq L_i^h \|x - y\| + \frac{L_{\mathcal{X}_i^{(0,\eta]}}}{\eta}\|\nabla h_i(y)\|\|x - y\|. \end{aligned}$$

In view of Definition 3.4, we have that $\|\nabla h_i(y)\| \leq L_{\mathcal{X}_i^{(0,\eta]}}$. We obtain

$$\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| \leq \left(L_i^h + \frac{L^2_{\mathcal{X}_i^{(0,\eta]}}}{\eta}\right)\|x - y\|.$$

Case 3. If $x \in \mathcal{X}_i^{(0,\eta]}$ and $y \in \mathcal{X}_i$, then we obtain

$$\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| \leq \frac{1}{\eta}|h_i(x)|\|\nabla h_i(x)\|.$$

From $y \in \mathcal{X}_i$, we have $h_i(y) \leq 0$. Also, from $x \in \mathcal{X}_i^{(0,\eta]}$, we have $h_i(x) \geq 0$. Thus,

$$h(x) \leq h(x) - h(y) = |h(x) - h(y)| \leq L_{\mathcal{X}_i^{(0,\eta]}}\|x - y\|.$$

From the two preceding relations, we obtain

$$\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| \leq \frac{1}{\eta}|h_i(x)|\|\nabla h_i(x)\| \leq \frac{1}{\eta}L_{\mathcal{X}_i^{(0,\eta]}}\|x-y\|L_{\mathcal{X}_i^{(0,\eta]}} = \frac{L^2_{\mathcal{X}_i^{(0,\eta]}}}{\eta}\|x-y\|$$

Case 4. If $x \in \mathcal{X}_i^{(0,\eta]}$ and $y \in \mathcal{X}_i^{(\eta,\infty)}$, consider two settings as follows. First, suppose $y \in \mathcal{X}_i^{(\eta,\eta+1]}$. Since ∇h is a continuous mapping and $\mathcal{X}_i^{(\eta,\eta+1]}$ is a compact set, we have $|h_i(x) - h_i(y)| \leq L_{\mathcal{X}_i^{(\eta,\eta+1]}}\|x-y\|$. We may write

$$\begin{aligned}
\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| &= \|\frac{1}{\eta}h_i(x)\nabla h_i(x) - \nabla h_i(y)\| \\
&\leq \|\frac{1}{\eta}h_i(x)\nabla h_i(x) - \nabla h_i(x)\| + \|\nabla h_i(x) - \nabla h_i(y)\| \\
&\leq \frac{\eta - h_i(x)}{\eta}\|\nabla h_i(x)\| + L_i^h\|x-y\| \\
&\leq \frac{|h_i(y) - h_i(x)|}{\eta}L_{\mathcal{X}_i^{(0,\eta]}} + L_i^h\|x-y\| \\
&\leq \frac{L_{\mathcal{X}_i^{(\eta,\infty)}}L_{\mathcal{X}_i^{(0,\eta]}}}{\eta}\|x-y\| + L_i^h\|x-y\| \\
&= \left(\frac{L_{\mathcal{X}_i^{(\eta,\infty)}}L_{\mathcal{X}_i^{(0,\eta]}}}{\eta} + L_i^h\right)\|x-y\|.
\end{aligned}$$

Suppose now that $y \in \mathcal{X}_i^{(\eta+1,\infty)}$. This implies that $\|x-y\| \geq \delta_i$. We obtain

$$\begin{aligned}
\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| &= \|\frac{1}{\eta}h_i(x)\nabla h_i(x) - \nabla h_i(y)\| \\
&\leq \|\frac{1}{\eta}h_i(x))\nabla h_i(x) - \nabla h_i(x)\| + \|\nabla h_i(x) - \nabla h_i(y)\| \\
&\leq \frac{\eta - h_i(x)}{\eta}\|\nabla h_i(x)\| + L_i^h\|x-y\| \\
&\leq \|\nabla h_i(x)\| + L_i^h\|x-y\| \\
&\leq \frac{1}{\delta_i}\|x-y\|L_{\mathcal{X}_i^{(0,\eta]}} + L_i^h\|x-y\| \\
&= \left(\frac{L_{\mathcal{X}_i^{(0,\eta]}}}{\delta_i} + L_i^h\right)\|x-y\|.
\end{aligned}$$

Case 5. If $x \in \mathcal{X}_i^{(\eta,\infty)}$ and $y \in \mathcal{X}_i$, then $h_i(x) > \eta$ while $h_i(y) \leq 0$. From the continuity of the function h_i, there exits a scalar $\alpha \in [0,1]$ such that for $z \triangleq \alpha x + (1-\alpha)y$ we have $h_i(z) = \eta$. This implies that $z \in \mathcal{X}_i^{(0,\eta]}$. We have

$$\begin{aligned}
\|\nabla g_{i,\eta}(x) - \nabla g_{i,\eta}(y)\| &= \|\nabla h_i(x)\| \\
&\leq \|\nabla h_i(x) - \nabla h_i(z)\| + \|\nabla h_i(z)\| \\
&\leq L_i^h\|x-z\| + L_{\mathcal{X}_i^{(0,\eta]}} \\
&\leq L_i^h\|x - \alpha x - (1-\alpha)y\| + L_{\mathcal{X}_i^{(0,\eta]}} \\
&\leq L_i^h\|x-y\| + L_{\mathcal{X}_i^{(0,\eta]}} \\
&\leq L_i^h\|x-y\| + \frac{1}{\nu_i}L_{\mathcal{X}_i^{(0,\eta]}}\|x-y\| \\
&= \left(L_i^h + \frac{L_{\mathcal{X}_i^{(0,\eta]}}}{\nu_i}\right)\|x-y\|,
\end{aligned}$$

where we used $\|x - y\| \geq \nu_i$.

Case 6. If $x \in \mathcal{X}_i$ and $y \in \mathcal{X}_i^{(\eta,\infty)}$, then following similar steps in Case 5, we obtain an identical bound.

Combining the results in Cases 1–6, we arrive at the smoothness result in (i).

(II) Recall that $g_{i,\eta}(x) \triangleq r_\eta(h_i(x))$. We have

$$g_{i,\eta}(x) = \begin{cases} 0, & h_i(x) \leq 0, \\ \frac{(h_i(x))^2}{2\eta}, & 0 < h_i(x) \leq \eta, \\ h_i(x) - \frac{\eta}{2}, & h_i(x) > \eta. \end{cases}$$

Let $i \in \{0, \ldots, m\}$ be given. We consider the following cases.

(II-i). If $x \in \mathcal{X}_i$, then $g_{i,\eta}(x) = 0$ and $\text{dist}(x, \mathcal{X}_i) = 0$ and so $g_{i,\eta}(x) \geq \alpha_{i,1} \,\text{dist}^2(x, \mathcal{X}_i)$ for all $x \in \mathcal{X}_i$.

(II-ii). If $x \in \mathcal{X}_i^{(0,\eta]}$, let us define $y(x) \triangleq \Pi_{\mathcal{X}_i}[x]$. Consider the projection problem given as

$$\begin{aligned} y(x) = \arg\min_z \ & \frac{1}{2}\|z - x\|^2 \\ \text{s.t.} \quad & h_i(z) \leq 0 \end{aligned}$$

The KKT optimality condition implies that if there exists a Lagrange multiplier $v \geq 0$ such that:

$$y(x) - x + \nu \nabla h_i(y(x)) = 0 \quad \Longrightarrow \quad x - y(x) = \nu \nabla h_i(y(x)),$$

and by the complementary slackness, we have

$$\nu h_i(y(x)) = 0.$$

Since $x \notin \mathcal{X}_i$ and $y(x) \in \mathcal{X}_i$, then $x - y(x) \neq 0$, implying that the Lagrangian multiplier is strictly positive, i.e., $\nu > 0$, and thus $h_i(y(x)) = 0$. We have

$$\langle \nabla h_i(y(x)), x - y(x) \rangle = \|\nabla h_i(y(x))\| \|x - y(x)\|.$$

Recall that h_i is convex. We obtain, for $x \in \mathcal{X}_i^{(0,\eta]}$,

$$h_i(x) \geq h_i(y(x)) + \langle \nabla h_i(y(x)), x - y(x) \rangle = \|\nabla h_i(y(x))\| \|x - y(x)\|. \tag{2}$$

Next, we show that

$$\inf \left\{ \|\nabla h_i(y(x))\| \;\middle|\; x \in \mathcal{X}_i^{(0,\eta]},\ h_i(y(x)) = 0 \right\} > 0. \tag{3}$$

It suffices to show that $\inf \{\|\nabla h_i(y)\| \mid y \in \mathbb{R}^n, h_i(y) = 0\} > 0$. Let $\mathcal{B} \triangleq \{u \in \mathbb{R}^n \mid h_i(u) = 0\}$. The continuity of h_i implies that $\mathcal{B}$ is closed. Further, in view of $\mathcal{B} \subset \mathcal{X}_i$ and Lemma 3.3, the set $\mathcal{B}$ is bounded. Thus, it is a compact set. Invoking the Weierstrass theorem, there exists $\hat{y}_i \in \mathcal{B}$ such that $\|\nabla h_i(\hat{y}_i)\| = \inf \{\|\nabla h_i(y)\| \mid y \in \mathcal{B}\}$. Suppose, for the sake of contradiction, $\|\nabla h_i(\hat{y}_i)\| = 0$. Since h_i is convex and differentiable, this implies that $\hat{y}_i$ is a global minimizer of h_i, i.e., $h(x) \geq h(\hat{y}_i) = 0$. However, this contradicts the Slater's condition. We conclude that $\inf \left\{ \|\nabla h_i(y(x))\| \;\middle|\; x \in \mathcal{X}_i^{(0,\eta]},\ h_i(y(x)) = 0 \right\} > 0$. Next, from equation 2 and equation 3 and the definition of $g_{i,\eta}$, we have for $x \in \mathcal{X}_i^{(0,\eta]}$,

$$g_{i,\eta}(x) = \frac{(h_i(x))^2}{2\eta} \geq \frac{1}{2\eta}\|\nabla h_i(y(x))\|^2 \|x - y(x)\|^2 \geq \frac{\|\nabla h_i(\hat{y}_i)\|^2}{2\eta} \text{dist}^2(x, \mathcal{X}_i).$$

Further, from $x \in \mathcal{X}_i^{(0,\eta]}$, we obtain $g_{i,\eta}(x) = \frac{(h_i(x))^2}{2\eta} \leq \frac{\eta}{2}$.

(II-iii) If $x \in \mathcal{X}_i^{(\eta,\infty)}$, then a similar fashion to the proof in (II-ii), the KKT optimality conditions hold and thus, for $x \in \mathcal{X}_i^{(\eta,\infty)}$, the inequality equation 2 holds and $h_i(y(x)) = 0$. In a similar vein to the proof in Case B, we can show that

$$\inf \left\{ \|\nabla h_i(y(x))\| \;\middle|\; x \in \mathcal{X}_i^{(\eta,\infty)},\ h_i(y(x)) = 0 \right\} \geq \|\nabla h_i(\hat{y}_i)\| > 0. \tag{4}$$

This can be shown following similar steps in (II-ii). Since $x \in \mathcal{X}_i^{(\eta,\infty)}$, we have $h_i(x) > \eta$. From the convexity of h_i, we obtain

$$g_{i,\eta}(x) = h_i(x) - \frac{\eta}{2} \geq \frac{h_i(x)}{2} \geq \frac{1}{2}\left(0 + \|\nabla h_i(y(x))\|\|x - y(x)\|\right) \geq \frac{\|\nabla h_i(\hat{y}_i)\|}{2}\text{dist}(x, \mathcal{X}_i).$$

□

A.3 Proof of Theorem 4.1

A.3.1 Additional lemmas

Lemma A.1 ((Karimireddy et al., 2020, Lemma 5))**.** Let $h : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex. Then, for any $x, y, z \in \mathbb{R}^n$,

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

Lemma A.2. Let $\{X_i\}_{i=1}^n$ be a sequence of random vectors in $\mathbb{R}^d$ adapted to a filtration $\{\mathcal{F}_i\}_{i=0}^n$. Suppose that

$$\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0, \qquad \forall i \in [n].$$

Then, for any $1 \leq i < j \leq n$,

$$\mathbb{E}\langle X_i, X_j\rangle = 0.$$

Consequently,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|^2\right] = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|^2\right].$$

Proof. For any $1 \leq i < j \leq n$, since X_i is $\mathcal{F}_{j-1}$-measurable, we have

$$\mathbb{E}\langle X_i, X_j\rangle = \mathbb{E}\left[\mathbb{E}\left[\langle X_i, X_j\rangle \mid \mathcal{F}_{j-1}\right]\right].$$

Using the $\mathcal{F}_{j-1}$-measurability of X_i,

$$\mathbb{E}\left[\langle X_i, X_j\rangle \mid \mathcal{F}_{j-1}\right] = \langle X_i, \mathbb{E}[X_j \mid \mathcal{F}_{j-1}]\rangle = 0.$$

Therefore,

$$\mathbb{E}\langle X_i, X_j\rangle = 0.$$

Expanding the squared norm gives

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|^2\right] = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2] + 2\sum_{1\leq i<j\leq n} \mathbb{E}\langle X_i, X_j\rangle.$$

Since all cross terms vanish, we obtain

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|^2\right] = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2].$$

□

Lemma A.3. Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth. Then, for any $x, y \in \mathbb{R}^n$ and any $\gamma \in [0, 2/L]$, we have

$$\left\|x - y - \gamma\big(\nabla F(x) - \nabla F(y)\big)\right\|^2 \leq \|x - y\|^2.$$

Proof. Expanding the squared norm gives

$$\begin{aligned}&\left\|x - y - \gamma\big(\nabla F(x) - \nabla F(y)\big)\right\|^2 \\ &= \|x - y\|^2 - 2\gamma\left\langle x - y, \nabla F(x) - \nabla F(y)\right\rangle + \gamma^2\left\|\nabla F(x) - \nabla F(y)\right\|^2.\end{aligned}$$

Since F is convex and L-smooth, its gradient is $1/L$-cocoercive, i.e.,

$$\langle x - y, \nabla F(x) - \nabla F(y) \rangle \geq \frac{1}{L} \|\nabla F(x) - \nabla F(y)\|^2 .$$

Therefore,

$$\begin{aligned} &\big\|x - y - \gamma\big(\nabla F(x) - \nabla F(y)\big)\big\|^2 \\ &\leq \|x - y\|^2 - \left(\frac{2\gamma}{L} - \gamma^2\right) \|\nabla F(x) - \nabla F(y)\|^2 . \end{aligned}$$

When $\gamma \in [0, 2/L]$, we have

$$\frac{2\gamma}{L} - \gamma^2 \geq 0.$$

Hence,

$$\big\|x - y - \gamma\big(\nabla F(x) - \nabla F(y)\big)\big\|^2 \leq \|x - y\|^2.$$

□

Lemma A.4. Let

$$\mathcal{X}_i := \{x \in \mathbb{R}^n : \, h_i(x) \leq 0\}, \qquad i \in \{0, \ldots, m\},$$

and define

$$\mathcal{X} := \bigcap_{i=0}^{m} \mathcal{X}_i.$$

Suppose Assumption 2.1 2.2 holds. Then there exists a constant $C_{\mathcal{X}} > 0$ such that, for every $x \in \mathbb{R}^n$,

$$\operatorname{dist}(x, \mathcal{X}) \leq C_{\mathcal{X}} \sum_{i=0}^{m} \operatorname{dist}(x, \mathcal{X}_i).$$

Proof. Under Assumption, each set $\mathcal{X}_i$ is closed and convex, and Slater's condition implies

$$\bigcap_{i=0}^{m} \operatorname{ri}(\mathcal{X}_i) \neq \emptyset.$$

Hence, the collection $\{\mathcal{X}_i\}_{i=0}^m$ is boundedly linearly regular (Bauschke et al., 1999). In particular, for every bounded set $\mathcal{B} \subseteq \mathbb{R}^n$, there exists a constant $\kappa_{\mathcal{B}} > 0$ such that

$$\operatorname{dist}\left(y, \bigcap_{i=0}^{m} \mathcal{X}_i\right) \leq \kappa_{\mathcal{B}} \max_{0 \leq i \leq m} \operatorname{dist}(y, \mathcal{X}_i), \qquad y \in \mathcal{B}.$$

Fix $x \in \mathbb{R}^n$ and let

$$q := \Pi_{\mathcal{X}_0}(x).$$

Since $\mathcal{X}_0$ is bounded, there exists a bounded set $\mathcal{B}$ containing $\mathcal{X}_0$, and thus $q \in \mathcal{B}$. Let

$$d_i := \operatorname{dist}(x, \mathcal{X}_i), \qquad i \in \{0, \ldots, m\}.$$

By the triangle inequality,

$$\begin{aligned} \operatorname{dist}(x, \mathcal{X}) &\leq \|x - q\| + \operatorname{dist}(q, \mathcal{X}) \\ &\leq d_0 + \kappa_{\mathcal{B}} \max_{0 \leq i \leq m} \operatorname{dist}(q, \mathcal{X}_i). \end{aligned}$$

Moreover, for every $i \in \{0, \ldots, m\}$,

$$\operatorname{dist}(q, \mathcal{X}_i) \leq \|q - x\| + \operatorname{dist}(x, \mathcal{X}_i) = d_0 + d_i.$$

Therefore,

$$\max_{0 \leq i \leq m} \operatorname{dist}(q, \mathcal{X}_i) \leq d_0 + \max_{0 \leq i \leq m} d_i \leq 2 \sum_{i=0}^{m} d_i.$$

Combining the preceding inequalities gives

$$\operatorname{dist}(x, \mathcal{X}) \leq (1 + 2\kappa_{\mathcal{B}}) \sum_{i=0}^{m} \operatorname{dist}(x, \mathcal{X}_i).$$

The result follows by setting

$$C_{\mathcal{X}} := 1 + 2\kappa_{\mathcal{B}}.$$

□

Lemma A.5 (Harmonic series bounds)**.** For every integer $R \geq 1$ and scalar $a \geq 0$,

$$\sum_{r=1}^{R} \frac{1}{r^a} \leq \begin{cases} \dfrac{R^{1-a}}{1-a}, & 0 \leq a < 1, \\ 1 + \log R, & a = 1, \\ \dfrac{a}{a-1}, & a > 1. \end{cases}$$

Proof. Since $t \mapsto t^{-a}$ is nonincreasing,

$$\sum_{r=1}^{R} r^{-a} \leq 1 + \int_1^R t^{-a}\, dt.$$

Evaluating the integral gives the first two bounds. For $a > 1$, extending the integral to infinity gives $1 + \int_1^\infty t^{-a}\, dt = a/(a-1)$. □

A.3.2 Convergence analysis of AdaP-SCAFFOLD

Definition A.6 (History of the method)**.** Let us define the history of the method in Algorithm 1 using the filtration defined as

$$\mathcal{F}_t \triangleq \begin{cases} \mathcal{S}_r \cup \mathcal{F}_{t-1} \cup \{\xi_{i,t-1}\}_{i \in \mathcal{S}_{r-1}}, & \text{if } t = T_r \text{ and } r \geq 2, \\ \mathcal{F}_{t-1} \cup \{\xi_{i,t-1}\}_{i \in \mathcal{S}_r}, & \text{if } T_r < t \leq T_{r+1} - 1 \text{ and } r \geq 1, \end{cases}$$

where $T_1 = 0$ and $\mathcal{F}_0 \triangleq S_1 \cup \{\hat{x}_1\}$. The set $\mathcal{F}_t$ contains all random variables generated up to the iteration t.

Definition A.7 (Local and global penalized objectives)**.** For a given round $r \geq 1$ and $i \in [m]$, we define the stochastic local penalty objective for client i as

$$\tilde{F}_{i,r}(x, \xi_i) \triangleq \tilde{f}_i(x, \xi_i) + \rho_r \left(g_{i,\eta}(x) + g_{0,\eta}(x) \right),$$

the deterministic counterpart as $F_{i,r}(x) \triangleq \mathbb{E}[\tilde{F}_{i,r}(x, \xi_i)]$, and the corresponding global objective as

$$F_r(x) \triangleq f(x) + \rho_r \bar{g}_\eta(x) = \frac{1}{m} \sum_{i=1}^{m} F_{i,r}(x).$$

Next, we define the smoothness and variance parameters as follows:

- **Local smoothness.** For each client $i \in [m]$, the local functions f_i and for each $i \in \{0\} \cup [m]$, $g_{i,\eta}$ are L_i^f-smooth and L_i^g-smooth.
- **Global smoothness.** Let us define $L_f \triangleq \max_{i \in [m]} L_i^f$, $L_{g,\eta} \triangleq \max_{i \in \{0\} \cup [m]} L_i^g$, and $L_{r,\eta} \triangleq L_f + \rho_r L_{g,\eta}$. Thus, both the local global penalized objectives, i.e., $F_{i,r}$ and F_r, are $L_{r,\eta}$-smooth.
- **Stochastic variance.** The variance of the local stochastic gradients evaluated on $F_{i,r}$ is bounded by $\sigma_r^2 \triangleq \sigma^2 + \rho_r^2 \sigma_g^2$. Since $g_{i,\eta}(x)$ doesn't have any stochastic term, then $\sigma_g^2 = 0$, which implies $\sigma_r^2 = \sigma^2$.

Definition A.8. Consider Algorithm 1. Given $r \geq 1$, we define the effective stepsize as

$$\tilde{\gamma}_r \triangleq H \gamma_{l,r} \gamma_g,$$

define the client drift as

$$e_r \triangleq \frac{1}{Hm} \sum_{t=T_r}^{T_{r+1}-1} \sum_{i=1}^{m} \mathbb{E}\left[\|x_{i,t} - \hat{x}_r\|^2 \right],$$

and define the control lag as

$$\mathcal{C}_r \triangleq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\|c_i^r - \nabla F_{i,r}(x^*)\|^2 \right],$$

where x^* is an optimal solution of the problem in equation 1 and H denotes the number of local steps in Algorithm 1.

Definition A.9. Define $\mathcal{X} := \bigcap_{i=0}^{m}\{x \in \mathbb{R}^n : h_i(x) \le 0\} \neq \emptyset$. Define the optimal set $\mathcal{X}^* \triangleq \arg\min_{x\in\mathcal{X}} f(x)$, and

$$
\begin{aligned}
A_r &\triangleq 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2,\\
B_r &\triangleq 200\tilde{\gamma}_r^2 L_f^2 + 24H^2\gamma_{l,r}^2 L_f^2 A_r,\\
\Phi_r &\triangleq \mathbb{E}\big[\|\hat{x}_r - x^*\|^2\big] + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_r,\\
\mathcal{R}_r &\triangleq \left(8\tilde{\gamma}_r^2 + 12H^2\gamma_{l,r}^2 A_r\right)\|\nabla f(x^*)\|^2 + \left(\frac{8\tilde{\gamma}_r^2}{sH} + \frac{48\tilde{\gamma}_r^2}{H} + 4H\gamma_{l,r}^2 A_r\right)\sigma^2,\\
G_* &\triangleq \|\nabla f(x^*)\|,\\
D^2 &\triangleq \Phi_1 := \mathbb{E}\big[\|\hat{x}_1 - x^*\|^2\big] + \frac{24m\tilde{\gamma}_1^2}{s}\mathcal{C}_1.
\end{aligned}
$$

Under Assumption 2.2, define $\kappa_{\min} := \min_{i\in\{0,\dots,m\}}\|\nabla h_i(\hat{y}_i)\| > 0$, where $\hat{y}_i$ is defined in Proposition 3.6(iii).

Definition A.10. Given $x \in \mathbb{R}^n$, we define the index sets

$$
\begin{aligned}
\mathcal{I}^{(\eta,\infty)}(x) &= \{i \in \{0,\dots,m\} \mid h_i(x) > \eta\},\\
\mathcal{I}^{(-\infty,\eta]}(x) &= \{i \in \{0,\dots,m\} \mid h_i(x) \le \eta\}.
\end{aligned}
$$

We also define

$$
G(x) \triangleq \sum_{i\in\mathcal{I}^{(\eta,\infty)}(x)} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(x) + \sum_{i\in\mathcal{I}^{(-\infty,\eta]}(x)} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|}\sqrt{g_{i,\eta}(x)},
$$

Lemma A.11. Let

$$
\mathcal{X}_i := \{x \in \mathbb{R}^n : h_i(x) \le 0\}, \qquad \mathcal{X} := \bigcap_{i=0}^{m}\mathcal{X}_i,
$$

and let

$$
x := \Pi_{\mathcal{X}}[\hat{x}_r].
$$

From Lemma A.4

$$
\operatorname{dist}(z,\mathcal{X}) \le \sum_{i=0}^{m}\operatorname{dist}(z,\mathcal{X}_i), \qquad z \in \mathbb{R}^n.
$$

Then

$$
\begin{aligned}
f(\hat{x}_r) - f^* \ge\ & -\|\nabla f(x^*)\|\Bigg(\sum_{i\in\mathcal{I}^{(\eta,\infty)}(\hat{x}_r)} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\hat{x}_r)\\
&+ \sum_{i\in\mathcal{I}^{(-\infty,\eta]}(\hat{x}_r)} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|}\sqrt{g_{i,\eta}(\hat{x}_r)}\Bigg) + \frac{\mu_f}{2}\|\hat{x}_r - x^*\|^2. \qquad (5)
\end{aligned}
$$

Moreover,

$$
\operatorname{dist}(\hat{x}_r,\mathcal{X}) \le \sum_{i\in\mathcal{I}^{(\eta,\infty)}(\hat{x}_r)} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\hat{x}_r) + \sum_{i\in\mathcal{I}^{(-\infty,\eta]}(\hat{x}_r)} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|}\sqrt{g_{i,\eta}(\hat{x}_r)}. \qquad (6)
$$

Proof. By convexity and μ_f-strong convexity of f,

$$
f(\hat{x}_r) - f^* \ge \langle \nabla f(x^*), \hat{x}_r - x^*\rangle + \frac{\mu_f}{2}\|\hat{x}_r - x^*\|^2.
$$

Since x^* is optimal over the convex set $\mathcal{X}$,

$$
\langle \nabla f(x^*), x - x^*\rangle \ge 0.
$$

Therefore,

$$
\langle \nabla f(x^*), \hat{x}_r - x^*\rangle = \langle \nabla f(x^*), \hat{x}_r - x\rangle + \langle \nabla f(x^*), x - x^*\rangle
$$

$$\geq -\|\nabla f(x^*)\|\|\hat{x}_r - x\|$$
$$= -\|\nabla f(x^*)\| \operatorname{dist}(\hat{x}_r, \mathcal{X}).$$

This proves the first inequality once equation 6 is established.

For each i, Proposition 3.6 gives

$$\operatorname{dist}(\hat{x}_r, \mathcal{X}_i) \leq \begin{cases} \dfrac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\hat{x}_r), & h_i(\hat{x}_r) > \eta, \\ \dfrac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|} \sqrt{g_{i,\eta}(\hat{x}_r)}, & 0 < h_i(\hat{x}_r) \leq \eta, \\ 0, & h_i(\hat{x}_r) \leq 0. \end{cases}$$

Using

$$\operatorname{dist}(\hat{x}_r, \mathcal{X}) \leq C_{\mathcal{X}} \sum_{i=0}^{m} \operatorname{dist}(\hat{x}_r, \mathcal{X}_i)$$

and summing the three cases yields equation 6. □

Lemma A.12. Consider Algorithm 1 and let Assumption 2.1 hold. Let $x^* \in \mathbb{R}^n$ denote an optimal solution of the problem in equation 1. Then, for any $r \geq 1$, we have

$$\mathbb{E}[\|\hat{x}_{r+1} - \hat{x}_r\|^2] \leq 8\tilde{\gamma}_r^2 L_{r,\eta}^2 e_r + 12\tilde{\gamma}_r^2 \mathcal{C}_r + 8\tilde{\gamma}_r^2 \|\nabla f(x^*)\|^2 + \frac{8\tilde{\gamma}_r^2 \sigma^2}{sH}$$
$$+ 8\tilde{\gamma}_r^2 L_f^2 \mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + 32\tilde{\gamma}_r^2 \rho_r^2 L_{g,\eta} \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right].$$

Proof. Using the triangle inequality, we may write

$$\mathbb{E}[\|\hat{x}_{r+1} - \hat{x}_r\|^2] = \mathbb{E}\left[\left\| -\frac{\tilde{\gamma}_r}{sH} \sum_{i \in \mathcal{S}_r} \sum_{t=T_r}^{T_{r+1}-1} \left(\nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - c_i^r + \hat{c}_r\right) \right\|^2\right]$$
$$\leq \underbrace{4\tilde{\gamma}_r^2 \mathbb{E}\left[\left\| \frac{1}{sH} \sum_{i \in \mathcal{S}_r} \sum_{t=T_r}^{T_{r+1}-1} \left(\nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla F_{i,r}(\hat{x}_r)\right) \right\|^2\right]}_{T_1}$$
$$+ \underbrace{4\tilde{\gamma}_r^2 \mathbb{E}[\|\hat{c}_r\|^2]}_{T_2} + \underbrace{4\tilde{\gamma}_r^2 \mathbb{E}\left[\| \frac{1}{s} \sum_{i \in \mathcal{S}_r} (\nabla F_{i,r}(x^*) - c_i^r) \|^2\right]}_{T_3}$$
$$+ \underbrace{4\tilde{\gamma}_r^2 \mathbb{E}\left[\left\| \frac{1}{sH} \sum_{i \in \mathcal{S}_r} \sum_{t=T_r}^{T_{r+1}-1} (\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*)) \right\|^2\right]}_{T_4}.$$

Let us define the indicator function, for $i \in [m]$, as

$$I_{i,r} = \begin{cases} 1, & i \in \mathcal{S}_r, \\ 0, & i \notin \mathcal{S}_r. \end{cases}$$

For T_1, by adding and subtracting $\nabla F_{i,r}(x_{i,t})$, we have

$$T_1 = 4\tilde{\gamma}_r^2 \mathbb{E}\left[\left\| \frac{1}{sH} \sum_{i \in \mathcal{S}_r} \sum_{t=T_r}^{T_{r+1}-1} \left(\nabla \widetilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla F_{i,r}(\hat{x}_r)\right) \right\|^2\right]$$

$$\leq \underbrace{8\tilde{\gamma}_r^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}(\nabla F_{i,r}(x_{i,t})-\nabla F_{i,r}(\hat{x}_r))\right\|^2\right]}_{T_{1,1}}$$
$$+\underbrace{8\tilde{\gamma}_r^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\left(\nabla\widetilde{F}_{i,r}(x_{i,t},\xi_{i,t})-\nabla F_{i,r}(x_{i,t})\right)\right\|^2\right]}_{T_{1,2}}.$$

For $T_{1,1}$, applying Jensen's inequality and the law of total expectation, we obtain

$$\begin{aligned}
&8\tilde{\gamma_r}^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}(\nabla F_{i,r}(x_{i,t})-\nabla F_{i,r}(\hat{x}_r))\right\|^2\right]\\
&=8\tilde{\gamma_r}^2\mathbb{E}_{\mathcal{S}_r}\left[\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}(\nabla F_{i,r}(x_{i,t})-\nabla F_{i,r}(\hat{x}_r))\right\|^2\mid\mathcal{S}_r\right]\right]\\
&\leq 8\tilde{\gamma_r}^2\mathbb{E}\left[\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\|(\nabla F_{i,r}(x_{i,t})-\nabla F_{i,r}(\hat{x}_r))\|^2\right]\\
&=\frac{8\tilde{\gamma_r}^2}{mH}\sum_{i=1}^{m}\sum_{t=T_r}^{T_{r+1}-1}\mathbb{E}\left[\|\nabla F_{i,r}(x_{i,t})-\nabla F_{i,r}(\hat{x}_r)\|^2\right]\\
&\leq\frac{8\tilde{\gamma_r}^2L_{r,\eta}^2}{mH}\sum_{i=1}^{m}\sum_{t=T_r}^{T_{r+1}-1}\mathbb{E}\left[\|x_{i,t}-\hat{x}_r\|^2\right]\\
&=8\tilde{\gamma_r}^2L_{r,\eta}^2e_r,
\end{aligned}$$

Fot $T_{1,2}$, let

$$\varepsilon_{i,t}^r\triangleq\nabla\widetilde{F}_{i,r}(x_{i,t},\xi_{i,t})-\nabla F_{i,r}(x_{i,t}).$$

By Definition A.6, for every $t\in\{T_r,\dots,T_{r+1}-1\}$, the client set $\mathcal{S}_r$ and the local iterates $\{x_{i,t}\}_{i\in\mathcal{S}_r}$ are $\mathcal{F}_t$-measurable, whereas the stochastic oracle samples $\{\xi_{i,t}\}_{i\in\mathcal{S}_r}$ are drawn after $\mathcal{F}_t$ is generated. Therefore, by Assumption 2.1,

$$\mathbb{E}\left[\varepsilon_{i,t}^r\mid\mathcal{F}_t\right]=0,\qquad\mathbb{E}\left[\left\|\varepsilon_{i,t}^r\right\|^2\mid\mathcal{F}_t\right]\leq\sigma^2.$$

Moreover, conditional on $\mathcal{F}_t$, the stochastic oracle samples are independent across clients. Hence, for $i\neq j$,

$$\begin{aligned}
\mathbb{E}\left[\left\langle\varepsilon_{i,t}^r,\varepsilon_{j,t}^r\right\rangle\middle|\mathcal{F}_t\right]&=\left\langle\mathbb{E}\left[\varepsilon_{i,t}^r\mid\mathcal{F}_t\right],\mathbb{E}\left[\varepsilon_{j,t}^r\mid\mathcal{F}_t\right]\right\rangle\\
&=0.
\end{aligned}$$

For two different local iterations $\tau<t$, $\varepsilon_{j,\tau}^r$ is $\mathcal{F}_t$-measurable by Definition A.6. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left\langle\varepsilon_{j,\tau}^r,\varepsilon_{i,t}^r\right\rangle\right]&=\mathbb{E}\left[\mathbb{E}\left[\left\langle\varepsilon_{j,\tau}^r,\varepsilon_{i,t}^r\right\rangle\middle|\mathcal{F}_t\right]\right]\\
&=\mathbb{E}\left[\left\langle\varepsilon_{j,\tau}^r,\mathbb{E}\left[\varepsilon_{i,t}^r\mid\mathcal{F}_t\right]\right\rangle\right]\\
&=0.
\end{aligned}$$

Therefore, all cross terms vanish. Consequently,

$$T_{1,2}=8\tilde{\gamma}_r^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\varepsilon_{i,t}^r\right\|^2\right]$$

$$
\begin{aligned}
&= \frac{8\tilde{\gamma}_r^2}{s^2H^2}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\left\|\varepsilon_{i,t}^r\right\|^2\right] \\
&= \frac{8\tilde{\gamma}_r^2}{s^2H^2}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\mathbb{E}\left[\left\|\varepsilon_{i,t}^r\right\|^2\middle|\mathcal{F}_t\right]\right] \\
&\leq \frac{8\tilde{\gamma}_r^2}{s^2H^2}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\sigma^2\right] \\
&= \frac{8\tilde{\gamma}_r^2}{s^2H^2}sH\sigma^2 \\
&= \frac{8\tilde{\gamma}_r^2\sigma^2}{sH}.
\end{aligned}
$$

Combining the above two bounds yields

$$
T_1 \leq 8\tilde{\gamma}_r^2 L_{r,\eta}^2 e_r + \frac{8\tilde{\gamma}_r^2\sigma^2}{sH}.
$$

For $T_2 = 4\tilde{\gamma}_r^2\mathbb{E}[\|\hat{c}_r\|^2]$.

$$
\begin{aligned}
4\tilde{\gamma}_r^2\mathbb{E}[\|\hat{c}_r\|^2] &= 4\tilde{\gamma}_r^2\mathbb{E}[\|\frac{1}{m}\sum_{i=1}^m c_i^r\|^2] \\
&= 4\tilde{\gamma}_r^2\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m(c_i^r - \nabla F_{i,r}(x^*)) + \frac{1}{m}\sum_{i=1}^m\nabla F_{i,r}(x^*)\|^2\right] \\
&\leq 8\tilde{\gamma}_r^2\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m(c_i^r - \nabla F_{i,r}(x^*))\|^2\right] + 8\tilde{\gamma}_r^2\|\nabla f(x^*)\|^2 \\
&\leq \frac{8\tilde{\gamma}_r^2}{m}\sum_{i=1}^m\mathbb{E}\left[\|c_i^r - \nabla F_{i,r}(x^*)\|^2\right] + 8\tilde{\gamma}_r^2\|\nabla f(x^*)\|^2 \\
&= 8\tilde{\gamma}_r^2\mathcal{C}_r + 8\tilde{\gamma}_r^2\|\nabla f(x^*)\|^2
\end{aligned}
$$

In bounding T_3

$$
\begin{aligned}
T_3 &= 4\tilde{\gamma}_r^2\mathbb{E}\left[\|\frac{1}{s}\sum_{i\in\mathcal{S}_r}(\nabla F_{i,r}(x^*) - c_i^r)\|^2\right] \\
&\leq \frac{4\tilde{\gamma}_r^2}{s}\mathbb{E}\sum_{i\in\mathcal{S}_r}\left[\|(\nabla F_{i,r}(x^*) - c_i^r\|^2\right] \\
&= \frac{4\tilde{\gamma}_r^2}{s}\frac{s}{m}\sum_{i=1}^m\mathbb{E}\left[\|\nabla F_{i,r}(x^*) - c_i^r\|^2\right] \\
&= 4\tilde{\gamma}_r^2\mathcal{C}_r
\end{aligned}
$$

Next, we bound T_4 as follows.

$$
\begin{aligned}
T_4 &= 4\tilde{\gamma}_r^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}(\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*))\right\|^2\right] \\
&\leq \underbrace{8\tilde{\gamma}_r^2\mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}(\nabla f_i(\hat{x}_r) - \nabla f_i(x^*))\right\|^2\right]}_{T_{4,1}}
\end{aligned}
$$

$$+ \underbrace{8\tilde{\gamma}_r^2 \mathbb{E}\left[\left\|\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\rho_r\left(\nabla g_{i,\eta}(\hat{x}_r)+\nabla g_{0,\eta}(\hat{x}_r)-\nabla g_{i,\eta}(x^*)-\nabla g_{0,\eta}(x^*)\right)\right\|^2\right]}_{T_{4,2}}.$$

For the first term $T_{4,1}$, since the terms $\nabla f_i(\hat{x}_r)$ and $\nabla f_i(x^*)$ are independent of the local step t, we can simplify the summation over t and apply Jensen's inequality:

$$\begin{aligned}
T_{4,1} &= 8\tilde{\gamma}_r^2 \mathbb{E}\left[\left\|\frac{1}{s}\sum_{i\in\mathcal{S}_r}\left(\nabla f_i(\hat{x}_r)-\nabla f_i(x^*)\right)\right\|^2\right] \\
&\leq 8\tilde{\gamma}_r^2 \mathbb{E}\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|\nabla f_i(\hat{x}_r)-\nabla f_i(x^*)\|^2\right] \\
&= 8\tilde{\gamma}_r^2 \mathbb{E}\left[\frac{1}{s}\sum_{i=1}^{m} I_{i,r}\cdot\|\nabla f_i(\hat{x}_r)-\nabla f_i(x^*)\|^2\right] \\
&= 8\tilde{\gamma}_r^2 \mathbb{E}\left[\frac{1}{s}\sum_{i=1}^{m} \mathbb{E}[I_{i,r}]\cdot\|\nabla f_i(\hat{x}_r)-\nabla f_i(x^*)\|^2\right] \\
&= \frac{8\tilde{\gamma}_r^2}{m}\sum_{i=1}^{m}\mathbb{E}\left[\|\nabla f_i(\hat{x}_r)-\nabla f_i(x^*)\|^2\right] \\
&\leq 8\tilde{\gamma}_r^2 L_f^2 \mathbb{E}\|\hat{x}_r-x^*\|^2.
\end{aligned}$$

For the second term $T_{4,2}$, since $\nabla g_{i,\eta}(\hat{x}_r)$ and $\nabla g_{i,\eta}(x^*)$ are independent of the local step t, we can simplify the inner summation. We obtain

$$\begin{aligned}
T_{4,2} &= 8\tilde{\gamma}_r^2\rho_r^2 \mathbb{E}\left[\left\|\frac{1}{s}\sum_{i\in\mathcal{S}_r}\left(\nabla g_{i,\eta}(\hat{x}_r)+\nabla g_{0,\eta}(\hat{x}_r)-\nabla g_{i,\eta}(x^*)-\nabla g_{0,\eta}(x^*)\right)\right\|^2\right] \\
&\leq 8\tilde{\gamma}_r^2\rho_r^2 \mathbb{E}\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|(\nabla g_{i,\eta}(\hat{x}_r)+\nabla g_{0,\eta}(\hat{x}_r)-\nabla g_{i,\eta}(x^*)-\nabla g_{0,\eta}(x^*))\|^2\right] \\
&= 8\tilde{\gamma}_r^2\rho_r^2 \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\|(\nabla g_{i,\eta}(\hat{x}_r)+\nabla g_{0,\eta}(\hat{x}_r)-\nabla g_{i,\eta}(x^*)-\nabla g_{0,\eta}(x^*))\|^2\right] \\
&\leq 16\tilde{\gamma}_r^2\rho_r^2 \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\|(\nabla g_{i,\eta}(\hat{x}_r)-\nabla g_{i,\eta}(x^*))\|^2\right] \\
&+ 16\tilde{\gamma}_r^2\rho_r^2 \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\|(\nabla g_{0,\eta}(\hat{x}_r)-\nabla g_{0,\eta}(x^*))\|^2\right] \\
&\leq 32\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[g_{i,\eta}(\hat{x}_r)-g_{i,\eta}(x^*)-\langle\nabla g_{i,\eta}(x^*),\hat{x}_r-x^*\rangle\right] \\
&+ 32\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta}\,\mathbb{E}\left[g_{0,\eta}(\hat{x}_r)-g_{0,\eta}(x^*)-\langle\nabla g_{0,\eta}(x^*),\hat{x}_r-x^*\rangle\right] \\
&= 32\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right],
\end{aligned}$$

where we used $\nabla g_{i,\eta}(x^*)=0$ and $g_{i,\eta}(x^*)=0$ for all i in view of Proposition 3.6 and that x^* is an optimal and thus, a feasible solution. □

Lemma A.13. Consider Algorithm 1 and let Assumption 2.1 hold. Let $x^*\in\mathbb{R}^n$ denote an optimal solution of the problem in equation 1. Then, for any $r\geq 1$,

$$\mathcal{C}_{r+1}\leq\left(1-\frac{s}{m}\right)\mathcal{C}_r+\frac{2s\sigma^2}{mH}+\frac{4sL_{r,\eta}^2}{m}e_r+\frac{8sL_f^2}{m}\mathbb{E}\left[\|\hat{x}_r-x^*\|^2\right]+\frac{32s\rho_r^2L_{g,\eta}}{m}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right].$$

Proof. Define selection indicator

$$I_{i,r} = \begin{cases} 1, & i \in \mathcal{S}_r, \\ 0, & i \notin \mathcal{S}_r. \end{cases}$$

Note that

$$c_i^{r+1} = \begin{cases} c_i^r, & i \notin \mathcal{S}_r, \\ \frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}), & i \in \mathcal{S}_r. \end{cases}$$

Therefore

$$c_i^{r+1} = (1 - I_{i,r})c_i^r + I_{i,r}\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}),$$

Then, since $I_{i,r}(1 - I_{i,r}) = 0$, expand the norm

$$\begin{aligned} \|c_i^{r+1} - \nabla F_{i,r}(x^*)\|^2 &= \|c_i^{r+1} - \nabla f_i(x^*)\|^2 \\ &= \|(1 - I_{i,r})(c_i^r - \nabla f_i(x^*)) + I_{i,r}(\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla f_i(x^*))\|^2 \\ &= (1 - I_{i,r})\|c_i^r - \nabla f_i(x^*)\|^2 + I_{i,r}\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla f_i(x^*)\|^2 \end{aligned}$$

Taking expectation from both sides

$$\begin{aligned} \mathbb{E}\left[\|c_i^{r+1} - \nabla F_{i,r}(x^*)\|^2\right] = (1 - \frac{s}{m})\mathbb{E}[\|c_i^r - \nabla f_i(x^*)\|^2] + \\ \frac{s}{m}\mathbb{E}\left[\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla f_i(x^*)\|^2\right] \end{aligned}$$

Consider $\mathcal{C}_r$ given by Definition A.8. We have

$$\begin{aligned} \mathcal{C}_{r+1} &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r+1} - \nabla F_{i,r}(x^*)\|^2\right] \\ &= (1 - \frac{s}{m})\mathcal{C}_r + \frac{s}{m^2}\sum_{i=1}^m \mathbb{E}\left[\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1} \nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla f_i(x^*)\|^2\right] \\ &= (1 - \frac{s}{m})\mathcal{C}_r \\ &\quad + \frac{s}{m^2}\sum_{i=1}^m \mathbb{E}\left[\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1}\left(\nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla F_{i,r}(x_{i,t})\right) + \frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1}\left(\nabla F_{i,r}(x_{i,t}) - \nabla f_i(x^*)\right)\|^2\right] \\ &\leq (1 - \frac{s}{m})\mathcal{C}_r + \frac{2s}{m^2}\sum_{i=1}^m \mathbb{E}\left[\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1}\left(\nabla \tilde{F}_{i,r}(x_{i,t}, \xi_{i,t}) - \nabla F_{i,r}(x_{i,t})\right)\|^2\right] \\ &\quad + \frac{2s}{m^2}\sum_{i=1}^m \mathbb{E}\left[\|\frac{1}{H}\sum_{t=T_r}^{T_{r+1}-1}\left(\nabla F_{i,r}(x_{i,t}) - \nabla f_i(x^*)\right)\|^2\right] \\ &\leq (1 - \frac{s}{m})\mathcal{C}_r + \frac{2s\sigma^2}{mH} + \frac{4s}{m^2H}\sum_{i=1}^m \sum_{t=T_r}^{T_{r+1}-1} \mathbb{E}\left[\|\left(\nabla F_{i,r}(x_{i,t}) - \nabla F_{i,r}(\hat{x}_r)\right)\|^2\right] \\ &\quad + \frac{4s}{m^2H}\sum_{i=1}^m \sum_{t=T_r}^{T_{r+1}-1} \mathbb{E}\left[\|\left(\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*)\right)\|^2\right] \end{aligned}$$

$$\leq (1-\frac{s}{m})\mathcal{C}_r + \frac{2s\sigma^2}{mH} + \frac{4sL_{r,\eta}^2}{m}e_r + \frac{4s}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\|\left(\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*)\right)\|^2\right]$$

$$\leq (1-\frac{s}{m})\mathcal{C}_r + \frac{2s\sigma^2}{mH} + \frac{4sL_{r,\eta}^2}{m}e_r + \frac{8s}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\|\nabla f_i(\hat{x}_r) - \nabla f_i(x^*)\|^2\right]$$

$$+ \frac{8s\rho_r^2}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\|\nabla g_{i,\eta}(\hat{x}_r) + \nabla g_{0,\eta}(\hat{x}_r) - \nabla g_{i,\eta}(x^*) - \nabla g_{0,\eta}(x^*)\|^2\right]$$

Taking similar steps in the proof of Lemma A.12, we obtain

$$\mathcal{C}_{r+1} \leq (1-\frac{s}{m})\mathcal{C}_r + \frac{2s\sigma^2}{mH} + \frac{4sL_{r,\eta}^2}{m}e_r + \frac{8sL_f^2}{m}\mathbb{E}[\|\hat{x}_r - x^*\|^2] + \frac{32s\rho_r^2 L_{g,\eta}}{m}\mathbb{E}[\bar{g}_\eta(\hat{x}_r)].$$

□

Lemma A.14. Consider Algorithm 1 and assume that the local stepsize satisfies $\gamma_{l,r} \leq \frac{2}{H(L_f+\rho_r L_{g,\eta})}$ for $r \geq 1$. Let Assumption 2.1 hold. Let $x^* \in \mathbb{R}^n$ denote an optimal solution of the problem in equation 1. Then, for any $r \geq 1$,

$$e_r \leq 12H^2\gamma_{l,r}^2\mathcal{C}_r + 12H^2\gamma_{l,r}^2\|\nabla f(x^*)\|^2 + 24H^2\gamma_{l,r}^2 L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right]$$
$$+ 96H^2\gamma_{l,r}^2\rho_r^2 L_{g,\eta}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + 4H\gamma_{l,r}^2\sigma^2.$$

Proof. Using the local update rules in the algorithm, for any $T_r + 1 \leq t < T_{r+1}$, for any $r \geq 1$, we have

$$\mathbb{E}\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|x_{i,t} - \hat{x}_r\|^2\right] = \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\left\|x_{i,t-1} - \gamma_{l,r}\nabla\tilde{F}_{i,r}(x_{i,t-1},\xi_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\right\|^2\right]$$

$$= \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\right.$$
$$\left.-\gamma_{l,r}\left(\nabla\tilde{F}_{i,r}(x_{i,t-1},\xi_{i,t-1}) - \nabla F_{i,r}(x_{i,t-1})\right)\right\|^2\Bigg]$$

$$= \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\|^2\right]$$
$$+ \frac{\gamma_{l,r}^2}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\left\|\nabla\tilde{F}_{i,r}(x_{i,t-1},\xi_{i,t-1}) - \nabla F_{i,r}(x_{i,t-1})\right\|^2\right]$$
$$- \frac{2\gamma_{l,r}}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\langle x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r,\right.$$
$$\left.\nabla\tilde{F}_{i,r}(x_{i,t-1},\xi_{i,t-1}) - \nabla F_{i,r}(x_{i,t-1})\right\rangle\Bigg]$$

$$\leq \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\|^2\right] + \gamma_{l,r}^2\sigma^2.$$

Invoking the triangle inequality, for any $a > 0$ we have

$$\frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t} - \hat{x}_r\|^2\right] = \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla\tilde{F}_{i,r}(x_{i,t-1},\xi_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\|^2\right]$$

$$\leq \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) - \gamma_{l,r}(\hat{c}_r - c_i^r) - \hat{x}_r\|^2\right] + \gamma_{l,r}^2\sigma^2$$

$$\leq (1+a)\underbrace{\mathbb{E}\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) + \gamma_{l,r}\nabla F_{i,r}(\hat{x}_r) - \hat{x}_r\|^2\right]}_{T_5}$$

$$+ (1+\frac{1}{a})\gamma_{l,r}^2 \mathbb{E}\underbrace{\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|\hat{c}_r - c_i^r + \nabla F_{i,r}(\hat{x}_r)\|^2\right]}_{T_6} + \gamma_{l,r}^2\sigma^2.$$

Since $F_{i,r}$ is convex and smooth. For $\gamma_{l,r} \leq 2/(L_{r,\eta})$, applying Lemma A.3, we obtain

$$T_5 = \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \gamma_{l,r}\nabla F_{i,r}(x_{i,t-1}) + \gamma_{l,r}\nabla F_{i,r}(\hat{x}_r) - \hat{x}_r\|^2\right] \leq \frac{1}{s}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\|x_{i,t-1} - \hat{x}_r\|^2\right].$$

Using the relaxed triangle inequality to expand T_6, we obtain

$$\begin{aligned}
T_6 &= \mathbb{E}\left[\frac{1}{s}\sum_{i\in\mathcal{S}_r}\|\hat{c}_r - c_i^r + \nabla F_{i,r}(\hat{x}_r)\|^2\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|\hat{c}_r - c_i^r + \nabla F_{i,r}(\hat{x}_r)\|^2\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\Bigg[\Bigg\|\left(\hat{c}_r - \frac{1}{m}\sum_{j=1}^m \nabla F_{j,r}(x^*)\right) - (c_i^r - \nabla F_{i,r}(x^*)) \\
&\qquad\qquad + \frac{1}{m}\sum_{j=1}^m \nabla F_{j,r}(x^*) + (\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*))\Bigg\|^2\Bigg] \\
&\leq \frac{3}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\left(\hat{c}_r - \frac{1}{m}\sum_{j=1}^m \nabla F_{j,r}(x^*)\right) - (c_i^r - \nabla F_{i,r}(x^*))\right\|^2\right] \\
&\quad + 3\left\|\frac{1}{m}\sum_{j=1}^m \nabla F_{j,r}(x^*)\right\|^2 \\
&\quad + \frac{3}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*)\|^2\right].
\end{aligned}$$

where we use the identity that for any vector $z_1, \ldots, z_m$, let $\bar{z} = \frac{1}{m}\sum_{i=1}^m z_i$, then

$$\frac{1}{m}\sum_{i=1}^m \|\bar{z} - z_i\|^2 = \frac{1}{m}\sum_{i=1}^m \|z_i\|^2 - \|\bar{z}\|^2$$

Therefore,

$$T_6 \leq 3\mathcal{C}_r + 3\|\nabla f(x^*)\|^2 + \frac{3}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla F_{i,r}(\hat{x}_r) - \nabla F_{i,r}(x^*)\|^2\right].$$

Using

$$F_{i,r}(x) = f_i(x) + \rho_r\big(g_{i,\eta}(x) + g_{0,\eta}(x)\big),$$

we have

$$\begin{aligned}
T_6 \leq\ & 3\mathcal{C}_r + 3\|\nabla f(x^*)\|^2 + \frac{6}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla f_i(\hat{x}_r) - \nabla f_i(x^*)\|^2\right] \\
& + \frac{6\rho_r^2}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla g_{i,\eta}(\hat{x}_r) + \nabla g_{0,\eta}(\hat{x}_r) - \nabla g_{i,\eta}(x^*) - \nabla g_{0,\eta}(x^*)\|^2\right].
\end{aligned}$$

Taking similar steps as in the proof of Lemma A.12, we obtain

$$T_6 \leq 3\mathcal{C}_r + 3\|\nabla f(x^*)\|^2 + 6L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + 24\rho_r^2 L_{g,\eta}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right].$$

Note that when $H = 1$, $e_r = 0$ since $x_{i,T_r} = \hat{x}_r$ for all $i \in [m]$ and that $\mathcal{C}_r$ and the right-hand side are both nonnegative. Thus the lemma holds true for $H = 1$. Suppose $H > 1$. Combining the bounds on T_5 and T_6 and choosing $a := \frac{1}{H-1}$ yields

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\|x_{i,t}-\hat{x}_r\|^2\right] \le \left(1+\frac{1}{H-1}\right)\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\|x_{i,t-1}-\hat{x}_r\|^2\right] + \gamma_{l,r}^2\sigma^2 \\ + H\gamma_{l,r}^2\Bigg(3\mathcal{C}_r + 3\|\nabla f(x^*)\|^2 + 6L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] \\ + 24\rho_r^2 L_{g,\eta}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right]\Bigg).\end{aligned}$$

Unrolling the preceding recursion and noting that $x_{i,T_r} = \hat{x}_r$, we obtain the following for any $k \in \{1,\dots,H\}$.

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\|x_{i,T_r+k}-\hat{x}_r\|^2\right] &\le \Bigg(3H\gamma_{l,r}^2\mathcal{C}_r + 3H\gamma_{l,r}^2\|\nabla f(x^*)\|^2 + 6H\gamma_{l,r}^2L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] \\ &\quad + 24H\gamma_{l,r}^2\rho_r^2L_{g,\eta}\mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + \gamma_{l,r}^2\sigma^2\Bigg)\left(\sum_{\tau=0}^{k-1}\left(1+\frac{1}{H-1}\right)^\tau\right) \\ &\le 4H\Bigg(3H\gamma_{l,r}^2\mathcal{C}_r + 3H\gamma_{l,r}^2\|\nabla f(x^*)\|^2 + 6H\gamma_{l,r}^2L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] \\ &\quad + 24H\gamma_{l,r}^2\rho_r^2L_{g,\eta}\mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + \gamma_{l,r}^2\sigma^2\Bigg).\end{aligned}$$

where we used the identity $(H-1)\left(\left(1+\frac{1}{H-1}\right)^H - 1\right) \le 4H$ for $H > 1$. Averaging over k yields

$$\begin{aligned}e_r \le 12H^2\gamma_{l,r}^2\mathcal{C}_r + 12H^2\gamma_{l,r}^2\|\nabla f(x^*)\|^2 + 24H^2\gamma_{l,r}^2L_f^2\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] \\ + 96H^2\gamma_{l,r}^2\rho_r^2L_{g,\eta}\mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + 4H\gamma_{l,r}^2\sigma^2.\end{aligned}$$

This completes the proof. □

Lemma A.15. Consider Algorithm 1 and let Assumption 2.1 hold. Let $x^* \in \mathbb{R}^n$ denote an optimal solution of the problem in equation 1. Suppose $\tilde{\gamma}_r$ is nonincreasing, ρ_r is nondecreasing, and satisfy

$$\frac{\mu m\gamma_g^2\tilde{\gamma}_r}{s} + A_r \le \gamma_g^2, \tag{7}$$

$$\tilde{\gamma}_r\rho_r L_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) \le 1, \tag{8}$$

For $\mu > 0$ and $0 < \mu\tilde{\gamma}_r < 2$. In convex case,

$$\Phi_{r+1} \le (1+B_r)\Phi_r - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + \mathcal{R}_r. \tag{9}$$

In the strongly convex case,

$$\Phi_{r+1} \le (1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r)\Phi_r - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + \mathcal{R}_r. \tag{10}$$

Proof. We may write

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}_{r+1} - x^*\|^2\right] &= \mathbb{E}\left[\|\hat{x}_r + (\hat{x}_{r+1} - \hat{x}_r) - x^*\|^2\right] \\ &= \mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + 2\mathbb{E}\left[\langle\hat{x}_{r+1} - \hat{x}_r, \hat{x}_r - x^*\rangle\right] + \mathbb{E}\left[\|\hat{x}_{r+1} - \hat{x}_r\|^2\right].\end{aligned}$$

Denote the cross term by

$$T_{\text{cross}} \triangleq 2\mathbb{E}\left[\langle\hat{x}_{r+1} - \hat{x}_r, \hat{x}_r - x^*\rangle\right].$$

Invoking Lemma A.1, we have

$$\begin{aligned}
T_{\text{cross}} &= 2\mathbb{E}\left[\langle \hat{x}_{r+1} - \hat{x}_r, \hat{x}_r - x^* \rangle\right] \\
&= -2\tilde{\gamma}_r \mathbb{E}\left[\frac{1}{sH}\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\langle \nabla F_{i,r}(x_{i,t}), \hat{x}_r - x^*\rangle\right] \\
&\leq \frac{1}{sH}\mathbb{E}\left[\sum_{i\in\mathcal{S}_r}\sum_{t=T_r}^{T_{r+1}-1}\left(-2\tilde{\gamma}_r\left(F_{i,r}(\hat{x}_r) - F_{i,r}(x^*)\right) - \frac{\mu\tilde{\gamma}_r}{2}\|\hat{x}_r - x^*\|^2 + 2\tilde{\gamma}_r L_{r,\eta}\|\hat{x}_r - x_{i,t}\|^2\right)\right] \\
&\leq -\frac{\mu\tilde{\gamma}_r}{2}\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - 2\tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + 2\tilde{\gamma}_r L_{r,\eta} e_r.
\end{aligned}$$

Applying Lemma A.12, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}_{r+1} - x^*\|^2\right] \leq\ & \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + 8\tilde{\gamma}_r^2 L_f^2\right)\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] \\
&- (2\tilde{\gamma}_r\rho_r - 32\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta})\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + \left(A_r - 96\tilde{\gamma}_r^2 L_{r,\eta}^2\right)e_r + 12\tilde{\gamma}_r^2\mathcal{C}_r \\
&+ 8\tilde{\gamma}_r^2\|\nabla f(x^*)\|^2 + \frac{8\tilde{\gamma}_r^2\sigma^2}{sH}. \qquad (11)
\end{aligned}$$

Multiplying the inequality in Lemma A.13 by $\frac{24m\tilde{\gamma}_r^2}{s}$, we have

$$\begin{aligned}
\frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_{r+1} \leq\ & \left(1 - \frac{s}{m}\right)\frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_r + 96\tilde{\gamma}_r^2 L_{r,\eta}^2 e_r + 768\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta}\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \\
&+ 192\tilde{\gamma}_r^2 L_f^2\mathbb{E}[\|\hat{x}_r - x*\|^2] + \frac{48\tilde{\gamma}_r^2\sigma^2}{H}. \qquad (12)
\end{aligned}$$

Adding equation 11 and equation 12, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}_{r+1} - x^*\|^2 + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_{r+1}\right] \leq\ & \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + 200\tilde{\gamma}_r^2 L_f^2\right)\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + \left[\left(1 - \frac{s}{m}\right)\frac{24m\tilde{\gamma}_r^2}{s} + 12\tilde{\gamma}_r^2\right]\mathcal{C}_r \\
&- 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \left(2\tilde{\gamma}_r\rho_r - 800\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta}\right)\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + A_r e_r \\
&+ 8\tilde{\gamma}_r^2\|\nabla f(x^*)\|^2 + \left[\frac{8\tilde{\gamma}_r^2}{sH} + \frac{48\tilde{\gamma}_r^2}{H}\right]\sigma^2. \qquad (13)
\end{aligned}$$

Substituting Lemma A.14 into equation 13:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}_{r+1} - x^*\|^2\right] + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_{r+1} \leq\ & \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right)\mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + \left[\frac{24m\tilde{\gamma}_r^2}{s} - 12\tilde{\gamma}_r^2 + 12H^2\gamma_{l,r}^2 A_r\right]\mathcal{C}_r \\
&- 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \left[2\tilde{\gamma}_r\rho_r - 800\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta} - 96H^2\gamma_{l,r}^2\rho_r^2 L_{g,\eta} A_r\right]\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \\
&+ \mathcal{R}_r. \qquad (14)
\end{aligned}$$

First, note that

$$\begin{aligned}
\frac{24m\tilde{\gamma}_r^2}{s} - 12\tilde{\gamma}_r^2 + 12H^2\gamma_{l,r}^2 A_r &= \frac{24m\tilde{\gamma}_r^2}{s}[1 - \frac{s}{2m}(1 - \frac{A_r}{\gamma_g^2})] \\
&\leq \frac{24m\tilde{\gamma}_r^2}{s}(1 - \frac{\mu\tilde{\gamma}_r}{2}).
\end{aligned}$$

By equation 8,

$$2\tilde{\gamma}_r\rho_r - 800\tilde{\gamma}_r^2\rho_r^2 L_{g,\eta} - 96H^2\gamma_{l,r}^2\rho_r^2 L_{g,\eta} A_r \geq \tilde{\gamma}_r\rho_r.$$

Since $B_r, \mathcal{C}_r \geq 0$, we obtain

$$\begin{aligned}
\Phi_{r+1} &\leq \mathbb{E}\left[\|\hat{x}_{r+1} - x^*\|^2\right] + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_{r+1} \\
&\leq \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right)\mathbb{E}[\|\hat{x}_r - x^*\|^2] + \frac{24m\tilde{\gamma}_r^2}{s}\left(1 - \frac{\mu\tilde{\gamma}_r}{2}\right)\mathcal{C}_r - 2\tilde{\gamma}_r\mathbb{E}[f(\hat{x}_r) - f^*]
\end{aligned}$$

$$
\begin{aligned}
&- \tilde{\gamma}_r \rho_r \mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + \mathcal{R}_r \\
&\leq \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right) \mathbb{E}[\|\hat{x}_r - x^*\|^2] + \frac{24m\tilde{\gamma}_r^2}{s} \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right) \mathcal{C}_r - 2\tilde{\gamma}_r \mathbb{E}[f(\hat{x}_r) - f^*] \\
&- \tilde{\gamma}_r \rho_r \mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + \mathcal{R}_r \\
&= \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right) \Phi_r - 2\tilde{\gamma}_r \mathbb{E}[f(\hat{x}_r) - f^*] - \tilde{\gamma}_r \rho_r \mathbb{E}[\bar{g}_\eta(\hat{x}_r)] + \mathcal{R}_r.
\end{aligned}
$$

□

Table 3: Initialization of Key Parameters

Quantity	Initialization
$\hat{x}_1$	arbitrary in $\mathbb{R}^n$
c_i^1	0
$\hat{c}^1$	0
$\mathcal{C}_1$	$\frac{1}{m}\sum_i \|\nabla f_i(x^*)\|^2$
ρ_1	> 0, determined by schedule
$\gamma_{l,1}$	> 0, determined by schedule
$\tilde{\gamma}_1$	> 0
T_1	0

Theorem A.16 (Convex setting)**.** Consider the general convex setting, i.e., $\mu = 0$. Let Assumption 2.1 hold, and let x^* be an optimal solution of the problem in equation 1. Assume that the effective stepsizes $\{\tilde{\gamma}_r\}_{r=1}^{R+1}$ are nonincreasing, and the penalty parameters $\{\rho_r\}_{r=1}^{R+1}$ are nondecreasing with $\rho_r \geq \rho_1 > 0$. Suppose that the stepsize and penalty parameters satisfy the conditions in Lemma A.15, so that equation 9 holds. Let

$$
\begin{aligned}
A_r &\triangleq 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2, \\
B_r &\triangleq 200\tilde{\gamma}_r^2 L_f^2 + 24H^2\gamma_{l,r}^2 L_f^2 A_r, \\
\Phi_r &\triangleq \mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_r,
\end{aligned}
$$

and define

$$
\mathcal{R}_r \triangleq \left(8\tilde{\gamma}_r^2 + 12H^2\gamma_{l,r}^2 A_r\right)\|\nabla f(x^*)\|^2 + \left(\frac{8\tilde{\gamma}_r^2}{sH} + \frac{48\tilde{\gamma}_r^2}{H} + 4H\gamma_{l,r}^2 A_r\right)\sigma^2.
$$

Define the positive weights by

$$
w_1 = 1, \qquad w_{r+1}(1 + B_r) = w_r, \qquad r = 1, \ldots, R,
$$

and let

$$
S_R \triangleq \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r, \qquad \bar{x}_R \triangleq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \hat{x}_r}{S_R}.
$$

Then the following error bounds hold.

(i) Upper bound on suboptimality error.

$$
\mathbb{E}\left[f(\bar{x}_R) - f^*\right] \leq \mathcal{E}_R \triangleq \frac{\Phi_1 + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r}{2S_R}.
$$

(ii) Upper and lower bounds on infeasibility error.

$$
0 \leq \mathbb{E}\left[\bar{g}_\eta(\bar{x}_R)\right] \leq U_R,
$$

where

$$
U_R \triangleq \frac{1}{S_R}\left[\frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R}\frac{w_{r+1}\mathcal{R}_r}{\rho_r} + 2(f^* - f_{\text{low}})\sum_{r=1}^{R}\frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}\right].
$$

(iii) Lower bound on suboptimality error. Suppose that Assumption 2.2 holds, Then

$$\mathbb{E}\left[f(\bar{x}_R) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}} U_R - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}} \sqrt{U_R}.$$

Proof. Starting from equation 9, multiplying both side by w_{r+1}, we get

$$2w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[f(\hat{x}_r) - f^*\right] + w_{r+1}\tilde{\gamma}_r\rho_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \leq w_r\Phi_r - w_{r+1}\Phi_{r+1} + w_{r+1}\mathcal{R}_r. \tag{15}$$

Summing yields

$$2\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[f(\hat{x}_r) - f^*\right] + \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\rho_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \leq \Phi_1 - w_{R+1}\Phi_{R+1} + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r.$$

(i) Upper bound on suboptimality error. Since $\bar{g}_\eta(\cdot) \geq 0$ and $\rho_r > 0$, we can drop the second term on the left-hand side. By the convexity of f,

$$f(\bar{x}_R) = f\left(\frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\hat{x}_r}{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r}\right) \leq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r f(\hat{x}_r)}{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r}.$$

Taking expectations and using the preceding inequality yields

$$\begin{aligned}
\mathbb{E}\left[f(\bar{x}_R) - f^*\right] &\leq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[f(\hat{x}_r) - f^*\right]}{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r} \\
&\leq \frac{\Phi_1 - w_{R+1}\Phi_{R+1} + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r}{2\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r} \\
&\leq \frac{\Phi_1 + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r}{2\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r},
\end{aligned}$$

where the last inequality follows from $w_{R+1} > 0$ and $\Phi_{R+1} \geq 0$.

(ii) Upper and lower bounds on infeasibility error. By the lower bound on f, we have

$$\mathbb{E}\left[f(\hat{x}_r) - f^*\right] \geq f_{\text{low}} - f^*.$$

Multiplying the one-step recursion by w_{r+1}/ρ_r and using the definition of the weights, we obtain

$$\begin{aligned}
w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] &\leq \frac{w_r\Phi_r - w_{r+1}\Phi_{r+1}}{\rho_r} - \frac{2w_{r+1}\tilde{\gamma}_r}{\rho_r}\mathbb{E}\left[f(\hat{x}_r) - f^*\right] + \frac{w_{r+1}\mathcal{R}_r}{\rho_r} \\
&\leq \frac{w_r\Phi_r - w_{r+1}\Phi_{r+1}}{\rho_r} + 2(f^* - f_{\text{low}})\frac{w_{r+1}\tilde{\gamma}_r}{\rho_r} + \frac{w_{r+1}\mathcal{R}_r}{\rho_r}.
\end{aligned}$$

Since $\{\rho_r\}$ is positive and nondecreasing, $w_r\Phi_r \geq 0$, and $w_1 = 1$, we have

$$\begin{aligned}
\sum_{r=1}^{R} \frac{w_r\Phi_r - w_{r+1}\Phi_{r+1}}{\rho_r} &= \frac{\Phi_1}{\rho_1} + \sum_{r=2}^{R} w_r\Phi_r\left(\frac{1}{\rho_r} - \frac{1}{\rho_{r-1}}\right) - \frac{w_{R+1}\Phi_{R+1}}{\rho_R} \\
&\leq \frac{\Phi_1}{\rho_1}.
\end{aligned}$$

Consequently, summing over $r = 1, \ldots, R$ yields

$$\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \leq \frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R} \frac{w_{r+1}\mathcal{R}_r}{\rho_r} + 2(f^* - f_{\text{low}})\sum_{r=1}^{R} \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}.$$

By the convexity and nonnegativity of $\bar{g}_\eta$,

$$\begin{aligned}
0 &\leq \mathbb{E}\left[\bar{g}_\eta(\bar{x}_R)\right] \\
&\leq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right]}{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r}
\end{aligned}$$

$$\leq \frac{\frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R} \frac{w_{r+1}\mathcal{R}_r}{\rho_r} + 2(f^* - f_{\text{low}}) \sum_{r=1}^{R} \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}}{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r}.$$

(iii) Lower bound on suboptimality error. Starting from Lemma A.11,

$$f(\bar{x}_R) - f^* \geq -\|\nabla f(x^*)\| \Bigg(\sum_{i \in \mathcal{I}^{[\eta,\infty)}(\bar{x}_R)} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\bar{x}_R) + \sum_{i \in \mathcal{I}^{(-\infty,\eta]}(\bar{x}_R)} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|} \sqrt{g_{i,\eta}(\bar{x}_R)} \Bigg).$$

Using $\|\nabla h_i(\hat{y}_i)\| \geq \kappa_{\min}$, we obtain

$$f(\bar{x}_R) - f^* \geq -\frac{2\|\nabla f(x^*)\|}{\kappa_{\min}} \sum_{i \in \mathcal{I}^{[\eta,\infty)}(\bar{x}_R)} g_{i,\eta}(\bar{x}_R) - \frac{\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}} \sum_{i \in \mathcal{I}^{(-\infty,\eta]}(\bar{x}_R)} \sqrt{g_{i,\eta}(\bar{x}_R)}.$$

Since

$$\sum_{i \in \mathcal{I}^{[\eta,\infty)}(\bar{x}_R)} g_{i,\eta}(\bar{x}_R) \leq \sum_{i=0}^{m} g_{i,\eta}(\bar{x}_R) = m\bar{g}_\eta - (m-1)g_{0,\eta} \leq m\bar{g}_\eta(\bar{x}_R),$$

and

$$\sum_{i \in \mathcal{I}^{(-\infty,\eta]}(\bar{x}_R)} \sqrt{g_{i,\eta}(\bar{x}_R)} \leq \sum_{i=0}^{m} \sqrt{g_{i,\eta}(\bar{x}_R)} \leq m\sqrt{\bar{g}_\eta(\bar{x}_R)},$$

we further get

$$f(\bar{x}_R) - f^* \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}} \bar{g}_\eta(\bar{x}_R) - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}} \sqrt{\bar{g}_\eta(\bar{x}_R)}.$$

Taking expectations and using Jensen's inequality for the concave square-root function yield

$$\mathbb{E}\left[f(\bar{x}_R) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}} \mathbb{E}\left[\bar{g}_\eta(\bar{x}_R)\right] - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}} \sqrt{\mathbb{E}\left[\bar{g}_\eta(\bar{x}_R)\right]}.$$

Substituting the upper bound on $\mathbb{E}[\bar{g}_\eta(\bar{x}_R)]$, we obtain

$$\mathbb{E}\left[f(\bar{x}_R) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}} U_R - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}} \sqrt{U_R}.$$

This completes the proof. □

Lemma A.17. Suppose that the convex recursion equation 9 holds, and use the notation of Theorem A.16. Let H, s, and $\eta > 0$ be fixed, and choose

$$\gamma_g = \sqrt{s}, \qquad \gamma_{l,r} = \frac{c_l}{H\sqrt{s}\sqrt{r}(1+\log r)}, \qquad \rho_r = c_\rho \sqrt{r}(1+\log r),$$

where $c_l, c_\rho > 0$ satisfy the conditions in Lemma A.18. Then

$$\mathcal{E}_R \leq \frac{e^{448c_l^2 L_f^2}(1+\log R)}{\sqrt{R}} \left[\frac{\Phi_1}{2c_l} + 20c_l\|\nabla f(x^*)\|^2 + \frac{8c_l\sigma^2}{sH} + \frac{52c_l\sigma^2}{H} \right].$$

and

$$U_R \leq \frac{e^{448c_l^2 L_f^2}(1+\log R)}{c_\rho\sqrt{R}} \left[\frac{\Phi_1}{c_l} + 40c_l\|\nabla f(x^*)\|^2 + \frac{16c_l\sigma^2}{sH} + \frac{104c_l\sigma^2}{H} + 4(f^* - f_{\text{low}}) \right].$$

Proof. By Lemma A.18, $0 \le A_r \le s$. Using $H^2\gamma_{l,r}^2 = \tilde{\gamma}_r^2/s$, the definitions of B_r and $\mathcal{R}_r$ give

$$0 \le B_r \le 224\tilde{\gamma}_r^2 L_f^2 = \frac{224c_l^2 L_f^2}{r(1+\log r)^2},$$

and

$$0 \le \mathcal{R}_r \le \frac{c_l^2}{r(1+\log r)^2}\left[20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

Since

$$\sum_{r=1}^{R} \frac{1}{r(1+\log r)^2} \le 1 + \int_1^R \frac{dt}{t(1+\log t)^2} = 2 - \frac{1}{1+\log R} \le 2,$$

we have $\sum_{r=1}^R B_r \le 448c_l^2 L_f^2$. Therefore, using $w_1 = 1$, $w_{r+1}(1+B_r) = w_r$, and $\log(1+B_r) \le B_r$,

$$1 \ge w_{r+1} = \exp\left(-\sum_{j=1}^{r} \log(1+B_j)\right) \ge e^{-448c_l^2 L_f^2}.$$

Consequently,

$$\begin{aligned} S_R &= \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \\ &\ge c_l e^{-448c_l^2 L_f^2} \sum_{r=1}^{R} \frac{1}{\sqrt{r}(1+\log r)} \\ &\ge c_l e^{-448c_l^2 L_f^2} \frac{\sqrt{R}}{1+\log R}. \end{aligned}$$

Also,

$$\sum_{r=1}^{R} w_{r+1}\mathcal{R}_r \le 2c_l^2\left[20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

Substituting these estimates into

$$\mathcal{E}_R = \frac{\Phi_1 + \sum_{r=1}^R w_{r+1}\mathcal{R}_r}{2S_R}$$

proves the suboptimality upper bound.

For the infeasibility bound, note that $\rho_r \ge c_\rho$ and

$$\sum_{r=1}^{R} \frac{w_{r+1}\mathcal{R}_r}{\rho_r} \le \frac{1}{c_\rho}\sum_{r=1}^{R} w_{r+1}\mathcal{R}_r,$$

while

$$\sum_{r=1}^{R} \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r} \le \frac{c_l}{c_\rho}\sum_{r=1}^{\infty} \frac{1}{r(1+\log r)^2} \le \frac{2c_l}{c_\rho}.$$

Substituting these estimates and the lower bound on S_R into

$$U_R = \frac{1}{S_R}\left[\frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R} \frac{w_{r+1}\mathcal{R}_r}{\rho_r} + 2(f^* - f_{\text{low}})\sum_{r=1}^{R} \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}\right],$$

with $\rho_1 = c_\rho$, yields the claimed bound. □

Proof of Theorem 4.1

(I) SUBOPTIMALITY UPPER AND LOWER BOUNDS. The upper bound follows directly from Theorem A.16(i) and Lemma A.17, which give

$$\mathbb{E}[f(\bar{x}_R) - f^*] \le \mathcal{E}_R \le \frac{\exp(448c_l^2 L_f^2)(1+\log R)}{\sqrt{R}} \left[\frac{D^2}{2c_l} + 20c_l G_*^2 + \frac{8c_l\sigma^2}{sH} + \frac{52c_l\sigma^2}{H}\right].$$

For the lower bound, let $x = \Pi_{\mathcal{X}}[\bar{x}_R]$. By convexity of f,

$$f(\bar{x}_R) - f^* \ge \langle \nabla f(x^*), \bar{x}_R - x^* \rangle.$$

Since x^* minimizes f over the convex set $\mathcal{X}$,

$$\langle \nabla f(x^*), x - x^* \rangle \ge 0.$$

Hence

$$\begin{aligned} f(\bar{x}_R) - f^* &\ge \langle \nabla f(x^*), \bar{x}_R - x \rangle \\ &\ge -G_* \|\bar{x}_R - x\| \\ &= -G_* \operatorname{dist}(\bar{x}_R, \mathcal{X}). \end{aligned}$$

Taking expectations gives the claimed lower bound.

(II) INFEASIBILITY. By Theorem A.16(ii),

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \le U_R.$$

Applying the explicit estimate for U_R in Lemma A.17 gives

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \le \frac{\exp(448c_l^2 L_f^2)(1+\log R)}{c_\rho\sqrt{R}} \left[\frac{D^2}{c_l} + 40c_l G_*^2 + \frac{16c_l\sigma^2}{sH} + \frac{104c_l\sigma^2}{H} + 4(f^* - f_{\text{low}})\right],$$

which is the stated infeasibility bound.

(III-1) η-APPROXIMATE LOWER BOUND ON SUBOPTIMALITY. Suppose Assumption 2.2 holds. From Lemma A.11,

$$f(\bar{x}_R) - f^* \ge -G_* \left(\sum_{i \in \mathcal{I}^{[\eta,\infty)}} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\bar{x}_R) + \sum_{i \in \mathcal{I}^{(-\infty,\eta]}} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|} \sqrt{g_{i,\eta}(\bar{x}_R)} \right).$$

Using $\|\nabla h_i(\hat{y}_i)\| \ge \kappa_{\min}$ and

$$\sum_{i=0}^{m} g_{i,\eta}(x) \le m\bar{g}_\eta(x),$$

the first term is bounded by

$$\frac{2mG_*}{\kappa_{\min}} \bar{g}_\eta(\bar{x}_R).$$

Moreover, for every $i \in \mathcal{I}^{(-\infty,\eta]}$, Proposition 3.6(iii) gives either $g_{i,\eta}(\bar{x}_R) = 0$ or $g_{i,\eta}(\bar{x}_R) \le \eta/2$. Therefore,

$$\sqrt{2\eta}\sqrt{g_{i,\eta}(\bar{x}_R)} \le \eta,$$

and the second term is $\mathcal{O}(\eta)$. Hence

$$f(\bar{x}_R) - f^* \ge -\frac{2mG_*}{\kappa_{\min}} \bar{g}_\eta(\bar{x}_R) - \mathcal{O}(\eta).$$

(III-2) ASYMPTOTIC CONVERGENCE. For the two assertions highlighted in the theorem, Lemma A.17 implies

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \le U_R = \mathcal{O}\left(\frac{1+\log R}{\sqrt{R}}\right),$$

and therefore

$$\mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \longrightarrow 0.$$

From the pointwise lower bound above, we have

$$\max\{f^* - f(\bar{x}_R), 0\} \le \frac{2mG_*}{\kappa_{\min}}\bar{g}_\eta(\bar{x}_R) + \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{\bar{g}_\eta(\bar{x}_R)}.$$

Using

$$|f(\bar{x}_R) - f^*| = f(\bar{x}_R) - f^* + 2\max\{f^* - f(\bar{x}_R), 0\},$$

taking expectations, and applying Jensen's inequality to the square-root term, we obtain

$$\mathbb{E}[|f(\bar{x}_R) - f^*|] \le \mathbb{E}[f(\bar{x}_R) - f^*] + \frac{4mG_*}{\kappa_{\min}}\mathbb{E}[\bar{g}_\eta(\bar{x}_R)] + \frac{2m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{\mathbb{E}[\bar{g}_\eta(\bar{x}_R)]}.$$

By Theorem A.16(i)–(ii),

$$\mathbb{E}[f(\bar{x}_R) - f^*] \le \mathcal{E}_R, \qquad \mathbb{E}[\bar{g}_\eta(\bar{x}_R)] \le U_R.$$

Therefore,

$$0 \le \mathbb{E}[|f(\bar{x}_R) - f^*|] \le \mathcal{E}_R + \frac{4mG_*}{\kappa_{\min}}U_R + \frac{2m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{U_R}.$$

For any fixed $\eta > 0$, Lemma A.17 gives $\mathcal{E}_R \to 0$ and $U_R \to 0$ as $R \to \infty$. Hence,

$$\mathbb{E}[|f(\bar{x}_R) - f^*|] \longrightarrow 0.$$

(III-3) EXACT LOWER BOUND ON SUBOPTIMALITY. Theorem A.16(iii) gives

$$\mathbb{E}[f(\bar{x}_R) - f^*] \ge -\frac{2mG_*}{\kappa_{\min}}U_R - \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{U_R}.$$

Substituting the explicit upper bound on U_R from Lemma A.17 gives Taking expectations and applying part (i) gives

$$\begin{aligned}\mathbb{E}[f(\bar{x}_R) - f^*] \ge &-\frac{2mG_*}{\kappa_{\min}}\frac{\exp(448c_l^2L_f^2)(1+\log R)}{c_\rho\sqrt{R}}\left[\frac{D^2}{c_l} + 40c_lG_*^2 + \frac{16c_l\sigma^2}{sH} + \frac{104c_l\sigma^2}{H} + 4(f^* - f_{\text{low}})\right] \\ &-\frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\left\{\frac{\exp(448c_l^2L_f^2)(1+\log R)}{c_\rho\sqrt{R}}\left[\frac{D^2}{c_l} + 40c_lG_*^2 + \frac{16c_l\sigma^2}{sH} + \frac{104c_l\sigma^2}{H} + 4(f^* - f_{\text{low}})\right]\right\}^{1/2}.\end{aligned}$$

Lemma A.18 (Feasible parameter choice for the convex setting)**.** Suppose $\mu = 0$ and $\gamma_g = \sqrt{s}$. Let $c_\rho > 0$ be arbitrary and fixed, and choose

$$\gamma_{l,r} = \frac{c_l}{H\sqrt{s}\sqrt{r}(1+\log r)}, \qquad \rho_r = c_\rho\sqrt{r}(1+\log r), \qquad r \ge 1.$$

With $L_{1,\eta}$ denoting $L_{r,\eta}$ at $r = 1$, let

$$A_1 := 2c_lL_{1,\eta} + 104c_l^2L_{1,\eta}^2.$$

If $c_l > 0$ satisfies

$$A_1 \le s, \qquad c_lc_\rho L_{g,\eta}(800s + 96A_1) \le s,$$

then the parameter conditions required for the revised convex recursion equation 9 hold for all $r \ge 1$. Specifically,

$$\begin{aligned}&\gamma_{l,r} \le \frac{2}{HL_{r,\eta}}, \\ &A_r \le \gamma_g^2, \\ &\tilde{\gamma}_r\rho_rL_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) \le 1.\end{aligned}$$

Moreover, $\tilde{\gamma}_r$ is positive and nonincreasing, and ρ_r is positive and nondecreasing. For every fixed $c_\rho > 0$, such a $c_l > 0$ exists.

Proof. The effective stepsize satisfies

$$\tilde{\gamma}_r = H\gamma_{l,r}\gamma_g = \frac{c_l}{\sqrt{r}(1+\log r)}, \qquad \tilde{\gamma}_r\rho_r = c_l c_\rho.$$

The stated positivity and monotonicity follow immediately. Using the definition of $L_{r,\eta}$, we obtain

$$\tilde{\gamma}_r L_{r,\eta} \le c_l L_{1,\eta}.$$

Consequently,

$$\begin{aligned} A_r &= 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2 \\ &\le 2c_l L_{1,\eta} + 104 c_l^2 L_{1,\eta}^2 = A_1 \le s = \gamma_g^2. \end{aligned}$$

Since $104c_l^2 L_{1,\eta}^2 \le A_1 \le s$,

$$H\gamma_{l,r}L_{r,\eta} = \frac{\tilde{\gamma}_r L_{r,\eta}}{\sqrt{s}} \le \frac{c_l L_{1,\eta}}{\sqrt{s}} \le \frac{1}{\sqrt{104}} < 2.$$

This verifies the local stepsize condition. The penalty condition follows from

$$\begin{aligned} \tilde{\gamma}_r\rho_r L_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) &\le c_l c_\rho L_{g,\eta}\left(800 + \frac{96A_1}{s}\right) \\ &\le 1. \end{aligned}$$

Finally, for fixed $c_\rho > 0$, $L_{1,\eta}$ is fixed, and both A_1 and $c_l c_\rho L_{g,\eta}(800s + 96A_1)$ tend to zero as $c_l \downarrow 0$. Thus the conditions hold for sufficiently small $c_l > 0$. □

Corollary A.19. Suppose $\mu = 0$, $1 \le s \le m$, and $\gamma_g = \sqrt{s}$. Fix any $c_\rho > 0$, and let

$$L_{1,\eta} := L_f + c_\rho L_{g,\eta}.$$

Define

$$\bar{c}_l^{\mathrm{cvx}} := \min\left\{\frac{\sqrt{1+104s}-1}{104L_{1,\eta}}, \frac{1}{896 c_\rho L_{g,\eta}}\right\}.$$

Choose any

$$0 < c_l \le \bar{c}_l^{\mathrm{cvx}},$$

and set

$$\gamma_{l,r} = \frac{c_l}{H\sqrt{s}\sqrt{r}(1+\log r)}, \qquad \rho_r = c_\rho\sqrt{r}(1+\log r), \qquad r \ge 1.$$

Then all parameter conditions required for the convex recursion equation 9 hold for every $r \ge 1$. In particular,

$$\begin{aligned} &\gamma_{l,r} \le \frac{2}{HL_{r,\eta}}, \\ &A_r \le \gamma_g^2, \\ &\tilde{\gamma}_r\rho_r L_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) \le 1. \end{aligned}$$

Moreover, $\tilde{\gamma}_r$ is positive and nonincreasing, while ρ_r is positive and nondecreasing.

Proof. Let

$$z := c_l L_{1,\eta}.$$

By the choice of c_l,

$$z \le \frac{\sqrt{1+104s}-1}{104}.$$

The right-hand side is the positive root of

$$104z^2 + 2z - s = 0.$$

Hence,

$$2c_l L_{1,\eta} + 104 c_l^2 L_{1,\eta}^2 \le s.$$

Therefore,

$$A_1 \le s.$$

Furthermore, since $A_1 \le s$,

$$800s + 96A_1 \le 896s.$$

Using

$$c_l \le \frac{1}{896 c_\rho L_{g,\eta}},$$

we obtain

$$c_l c_\rho L_{g,\eta}(800s + 96A_1) \le 896s\, c_l c_\rho L_{g,\eta} \le s.$$

Thus both conditions of Lemma A.18 are satisfied, and the claimed parameter properties follow directly from that lemma. □

A.4 Proof of Theorem 4.3

Theorem A.20 (Strongly convex setting)**.** Let Assumption 2.1 hold, and suppose that each local objective f_i is μ-strongly convex, where $\mu > 0$. Let x^* be the optimal solution of the problem in equation 1. Assume the smoothness and sampling conditions used in the one-round analysis hold. Let the positive deterministic effective stepsizes $\{\tilde{\gamma}_r\}_{r=1}^{R+1}$ be nonincreasing and satisfy $0 < \mu\tilde{\gamma}_r < 2$. Let the deterministic penalty parameters $\{\rho_r\}_{r=1}^{R+1}$ be nondecreasing with $\rho_r \ge \rho_1 > 0$.

Define

$$\begin{aligned}
A_r &\triangleq 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2,\\
B_r &\triangleq 200\tilde{\gamma}_r^2 L_f^2 + 24H^2\gamma_{l,r}^2 L_f^2 A_r,\\
\Phi_r &\triangleq \mathbb{E}\left[\|\hat{x}_r - x^*\|^2\right] + \frac{24m\tilde{\gamma}_r^2}{s}\mathcal{C}_r,
\end{aligned}$$

and

$$\mathcal{R}_r \triangleq \left(8\tilde{\gamma}_r^2 + 12H^2\gamma_{l,r}^2 A_r\right)\|\nabla f(x^*)\|^2 + \left(\frac{8\tilde{\gamma}_r^2}{sH} + \frac{48\tilde{\gamma}_r^2}{H} + 4H\gamma_{l,r}^2 A_r\right)\sigma^2.$$

Suppose that, for $r = 1, \dots, R$,

$$\begin{aligned}
&\gamma_{l,r} \le \frac{2}{HL_{r,\eta}},\\
&A_r + \frac{\mu m \gamma_g^2 \tilde{\gamma}_r}{s} \le \gamma_g^2,\\
&\tilde{\gamma}_r \rho_r L_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) \le 1.
\end{aligned}$$

Define the positive weights by

$$w_1 = 1, \qquad w_{r+1}\left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right) = w_r, \qquad r = 1, \dots, R.$$

Let

$$T_R^{\mathrm{sc}} \triangleq \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r, \qquad \bar{x}_R^{\mathrm{sc}} \triangleq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\hat{x}_r}{T_R^{\mathrm{sc}}}.$$

Then the following bounds hold.

(i) Upper bound on suboptimality error.

$$\mathbb{E}\left[f(\bar{x}_R^{\mathrm{sc}}) - f^*\right] \le B_R^{\mathrm{sc}} \triangleq \frac{\Phi_1 + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r}{2T_R^{\mathrm{sc}}}.$$

(ii) Upper and lower bounds on infeasibility error.

$$0 \le \mathbb{E}\left[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})\right] \le U_R^{\mathrm{sc}},$$

where

$$U_R^{\mathrm{sc}} \triangleq \frac{1}{T_R^{\mathrm{sc}}}\left[\frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R}\frac{w_{r+1}\mathcal{R}_r}{\rho_r} + \frac{\|\nabla f(x^*)\|^2}{\mu}\sum_{r=1}^{R}\frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}\right].$$

(iii) Lower bound on suboptimality error. Suppose that Assumption 2.2 holds, Then

$$\mathbb{E}\left[f(\bar{x}_R^{\mathrm{sc}}) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}}U_R^{\mathrm{sc}} - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}}\sqrt{U_R^{\mathrm{sc}}}.$$

Proof. Starting from

$$\begin{aligned}\Phi_{r+1} \leq \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right)\mathbb{E}\|\hat{x}_r - x^*\|^2 + \frac{24m\tilde{\gamma}_r^2}{s}\left(1 - \frac{\mu\tilde{\gamma}_r}{2}\right)\mathcal{C}_r \\ - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + \mathcal{R}_r. \end{aligned} \tag{16}$$

Since $B_r \geq 0$ and $\mathcal{C}_r \geq 0$, it follows that

$$\Phi_{r+1} \leq \left(1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r\right)\Phi_r - 2\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] - \tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] + \mathcal{R}_r. \tag{17}$$

The coefficient $1 - \mu\tilde{\gamma}_r/2 + B_r$ is positive, so the weights are well defined and positive. Multiplying equation 17 by w_{r+1} gives

$$\begin{aligned}2w_{r+1}\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] + w_{r+1}\tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \\ \leq w_r\Phi_r - w_{r+1}\Phi_{r+1} + w_{r+1}\mathcal{R}_r.\end{aligned}$$

(i) Upper bound on suboptimality error. Summing the preceding inequality over $r = 1, \ldots, R$ yields

$$\begin{aligned}2\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right] + \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\rho_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \\ \leq \Phi_1 - w_{R+1}\Phi_{R+1} + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r.\end{aligned}$$

Using the nonnegativity of $\bar{g}_\eta$ and Φ_{R+1}, and applying Jensen's inequality to f, we obtain

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}_R^{\mathrm{sc}}) - f^*\right] &\leq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r\mathbb{E}\left[f(\hat{x}_r) - f^*\right]}{T_R^{\mathrm{sc}}} \\ &\leq \frac{\Phi_1 + \sum_{r=1}^{R} w_{r+1}\mathcal{R}_r}{2T_R^{\mathrm{sc}}} = B_R^{\mathrm{sc}}.\end{aligned}$$

(ii) Upper and lower bounds on infeasibility error. By strong convexity, for every $x \in \mathbb{R}^n$,

$$\begin{aligned}f(x) - f^* &\geq \langle\nabla f(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|^2 \\ &= \frac{\mu}{2}\left\|x - x^* + \frac{\nabla f(x^*)}{\mu}\right\|^2 - \frac{\|\nabla f(x^*)\|^2}{2\mu} \\ &\geq -\frac{\|\nabla f(x^*)\|^2}{2\mu}.\end{aligned}$$

Multiplying equation 17 by w_{r+1}/ρ_r and using this lower bound gives

$$\begin{aligned}w_{r+1}\tilde{\gamma}_r\mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right] \leq \frac{w_r\Phi_r - w_{r+1}\Phi_{r+1}}{\rho_r} + \frac{w_{r+1}\mathcal{R}_r}{\rho_r} \\ + \frac{\|\nabla f(x^*)\|^2}{\mu}\frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}.\end{aligned}$$

Since ρ_r is nondecreasing and $w_r\Phi_r \geq 0$,

$$\sum_{r=1}^{R} \frac{w_r\Phi_r - w_{r+1}\Phi_{r+1}}{\rho_r} = \frac{\Phi_1}{\rho_1} + \sum_{r=2}^{R} w_r\Phi_r \left(\frac{1}{\rho_r} - \frac{1}{\rho_{r-1}}\right) - \frac{w_{R+1}\Phi_{R+1}}{\rho_R}$$
$$\leq \frac{\Phi_1}{\rho_1}.$$

Summing and applying the convexity of $\bar{g}_\eta$ yields

$$0 \leq \mathbb{E}\left[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})\right]$$
$$\leq \frac{\sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r \mathbb{E}\left[\bar{g}_\eta(\hat{x}_r)\right]}{T_R^{\mathrm{sc}}}$$
$$\leq \frac{1}{T_R^{\mathrm{sc}}}\left[\frac{\Phi_1}{\rho_1} + \sum_{r=1}^{R} \frac{w_{r+1}\mathcal{R}_r}{\rho_r} + \frac{\|\nabla f(x^*)\|^2}{\mu}\sum_{r=1}^{R} \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}\right] = U_R^{\mathrm{sc}}.$$

(iii) Lower bound on suboptimality error. Using the geometric lower bound from the convex analysis, for every $x \in \mathbb{R}^n$,

$$f(x) - f^* \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}}\bar{g}_\eta(x) - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}}\sqrt{\bar{g}_\eta(x)}.$$

Applying this inequality to $x = \bar{x}_R^{\mathrm{sc}}$, taking expectations, and using Jensen's inequality for the concave square-root function gives

$$\mathbb{E}\left[f(\bar{x}_R^{\mathrm{sc}}) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}}\mathbb{E}\left[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})\right]$$
$$- \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}}\sqrt{\mathbb{E}\left[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})\right]}.$$

Substituting the upper bound U_R^{sc} from part (ii) yields

$$\mathbb{E}\left[f(\bar{x}_R^{\mathrm{sc}}) - f^*\right] \geq -\frac{2m\|\nabla f(x^*)\|}{\kappa_{\min}}U_R^{\mathrm{sc}} - \frac{m\sqrt{2\eta}\|\nabla f(x^*)\|}{\kappa_{\min}}\sqrt{U_R^{\mathrm{sc}}}.$$

□

Lemma A.21. Suppose that the strongly convex recursion equation 10 holds, and use the notation of Theorem A.20. Let $1 \leq s \leq m$, and let H and $\eta > 0$ be fixed. Choose

$$\gamma_g = \sqrt{s}, \qquad \gamma_{l,r} = \frac{c_l}{H\sqrt{s}\left(1 + \frac{\mu c_l}{4}(r-1)\right)}, \qquad \rho_r = c_\rho\left(1 + \frac{\mu c_l}{4}(r-1)\right),$$

where $c_l, c_\rho > 0$ satisfy the parameter conditions of Theorem A.20 Then for every $R \geq 1$, the suboptimality upper bound satisfies

$$B_R^{\mathrm{sc}} \leq \frac{4e^{2240c_l L_f^2/\mu}}{\mu(R+1)}\left[\frac{\Phi_1}{c_l^2 R} + 20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

The infeasibility upper bound satisfies

$$U_R^{\mathrm{sc}} \leq \frac{8e^{2240c_l L_f^2/\mu}}{\mu^2 c_l c_\rho(R+1)}\left[\frac{\mu\Phi_1}{c_l R} + \|\nabla f(x^*)\|^2 + \frac{1+\log R}{R}\left(80\|\nabla f(x^*)\|^2 + \frac{32\sigma^2}{sH} + \frac{208\sigma^2}{H}\right)\right].$$

Proof. The parameter conditions imply $A_r \leq s$ and

$$\mu m c_l + A_1 \leq s \leq m, \qquad \text{hence} \qquad 0 < \mu c_l \leq 1.$$

We first estimate the weights. Factoring each term in the weight recursion and substituting the prescribed stepsize give

$$w_{r+1} = \prod_{k=1}^{r}\left(1 - \frac{\mu\tilde{\gamma}_k}{2} + B_k\right)^{-1}$$

$$= \left[\prod_{k=1}^{r}\left(1-\frac{\mu\tilde{\gamma}_k}{2}\right)^{-1}\right]\prod_{k=1}^{r}\left(1+\frac{B_k}{1-\mu\tilde{\gamma}_k/2}\right)^{-1}$$
$$= \left[\prod_{k=1}^{r}\frac{1+\frac{\mu c_l}{4}(k-1)}{1+\frac{\mu c_l}{4}(k-3)}\right]\prod_{k=1}^{r}\left(1+\frac{B_k}{1-\mu\tilde{\gamma}_k/2}\right)^{-1}$$
$$= \frac{\left(1+\frac{\mu c_l}{4}(r-2)\right)\left(1+\frac{\mu c_l}{4}(r-1)\right)}{\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)}\prod_{k=1}^{r}\left(1+\frac{B_k}{1-\mu\tilde{\gamma}_k/2}\right)^{-1}.$$

To bound the remaining product, note that $H^2\gamma_{l,k}^2 = \tilde{\gamma}_k^2/s$ and $A_k \le s$ imply

$$0 \le B_k \le 224\tilde{\gamma}_k^2 L_f^2.$$

Moreover, integral comparison yields

$$\sum_{k=1}^{r}\tilde{\gamma}_k^2 \le c_l^2\left[1+\int_0^{r-1}\frac{dt}{(1+\frac{\mu c_l}{4}t)^2}\right]$$
$$\le c_l^2 + \frac{4c_l}{\mu} \le \frac{5c_l}{\mu}.$$

Since $1-\mu\tilde{\gamma}_k/2 \ge 1/2$, we obtain

$$\sum_{k=1}^{r}\frac{B_k}{1-\mu\tilde{\gamma}_k/2} \le 448L_f^2\sum_{k=1}^{r}\tilde{\gamma}_k^2 \le \frac{2240c_lL_f^2}{\mu}.$$

Thus, by $\log(1+t)\le t$,

$$e^{-2240c_lL_f^2/\mu} \le \prod_{k=1}^{r}\left(1+\frac{B_k}{1-\mu\tilde{\gamma}_k/2}\right)^{-1} \le 1.$$

We next bound the denominator. Multiplying the lower bound on w_{r+1} by $\tilde{\gamma}_r$ cancels the factor $1+\frac{\mu c_l}{4}(r-1)$, giving

$$T_R^{\mathrm{sc}} = \sum_{r=1}^{R} w_{r+1}\tilde{\gamma}_r$$
$$\ge \frac{c_l e^{-2240c_lL_f^2/\mu}}{\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)}\sum_{r=1}^{R}\left(1+\frac{\mu c_l}{4}(r-2)\right)$$
$$\ge \frac{e^{-2240c_lL_f^2/\mu}\mu c_l^2 R(R+1)}{8\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)},$$

where the last inequality uses $1+\frac{\mu c_l}{4}(r-2)\ge\frac{\mu c_l}{4}r$.

For the suboptimality bound, the residual satisfies

$$\mathcal{R}_r \le \tilde{\gamma}_r^2\left[20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

The upper bound on the weights gives

$$w_{r+1}\tilde{\gamma}_r^2 \le \frac{c_l^2}{\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)}\frac{1+\frac{\mu c_l}{4}(r-2)}{1+\frac{\mu c_l}{4}(r-1)} \le \frac{c_l^2}{\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)}.$$

Consequently,

$$\sum_{r=1}^{R} w_{r+1}\mathcal{R}_r \le \frac{c_l^2 R}{\left(1-\frac{\mu c_l}{4}\right)\left(1-\frac{\mu c_l}{2}\right)}\left[20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

Combining this estimate with the lower bound on T_R^{sc} and using

$$0 < \left(1 - \frac{\mu c_l}{4}\right)\left(1 - \frac{\mu c_l}{2}\right) \le 1,$$

we obtain

$$B_R^{\mathrm{sc}} = \frac{\Phi_1 + \sum_{r=1}^R w_{r+1}\mathcal{R}_r}{2T_R^{\mathrm{sc}}} \le \frac{4e^{2240 c_l L_f^2/\mu}}{\mu(R+1)}\left[\frac{\Phi_1}{c_l^2 R} + 20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

For the infeasibility bound. Since

$$\rho_r = c_\rho\left(1 + \frac{\mu c_l}{4}(r-1)\right), \qquad 1 + \frac{\mu c_l}{4}(r-1) \ge \frac{\mu c_l}{4} r,$$

the preceding bound on $w_{r+1}\tilde{\gamma}_r^2$ and $\sum_{r=1}^R r^{-1} \le 1 + \log R$ yield

$$\sum_{r=1}^R \frac{w_{r+1}\mathcal{R}_r}{\rho_r} \le \frac{4c_l(1+\log R)}{\mu c_\rho\left(1 - \frac{\mu c_l}{4}\right)\left(1 - \frac{\mu c_l}{2}\right)}\left[20\|\nabla f(x^*)\|^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

Also, $\tilde{\gamma}_r \rho_r = c_l c_\rho$, so

$$\sum_{r=1}^R \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r} = \frac{1}{c_l c_\rho}\sum_{r=1}^R w_{r+1}\tilde{\gamma}_r^2 \le \frac{c_l R}{c_\rho\left(1 - \frac{\mu c_l}{4}\right)\left(1 - \frac{\mu c_l}{2}\right)}.$$

Substituting these two estimates and $\rho_1 = c_\rho$ into

$$U_R^{\mathrm{sc}} = \frac{1}{T_R^{\mathrm{sc}}}\left[\frac{\Phi_1}{c_\rho} + \sum_{r=1}^R \frac{w_{r+1}\mathcal{R}_r}{\rho_r} + \frac{\|\nabla f(x^*)\|^2}{\mu}\sum_{r=1}^R \frac{w_{r+1}\tilde{\gamma}_r}{\rho_r}\right],$$

and applying the same lower bound on T_R^{sc}, we obtain

$$U_R^{\mathrm{sc}} \le \frac{8e^{2240 c_l L_f^2/\mu}}{\mu^2 c_l c_\rho (R+1)}\left[\frac{\mu\Phi_1}{c_l R} + \|\nabla f(x^*)\|^2 + \frac{1+\log R}{R}\left(80\|\nabla f(x^*)\|^2 + \frac{32\sigma^2}{sH} + \frac{208\sigma^2}{H}\right)\right].$$

This completes the proof. □

Proof of Theorem 4.3

(I) SUBOPTIMALITY UPPER AND LOWER BOUNDS. The upper bound follows directly from Theorem A.20(i) and Lemma A.21, which give

$$\mathbb{E}[f(\bar{x}_R^{\mathrm{sc}}) - f^*] \le B_R^{\mathrm{sc}} \le \frac{4\exp\left(\frac{2240 c_l L_f^2}{\mu}\right)}{\mu(R+1)}\left[\frac{D^2}{c_l^2 R} + 20G_*^2 + \frac{8\sigma^2}{sH} + \frac{52\sigma^2}{H}\right].$$

For the lower bound, let $x = \Pi_{\mathcal{X}}[\bar{x}_R^{\mathrm{sc}}]$. By the μ-strong convexity of f,

$$f(\bar{x}_R^{\mathrm{sc}}) - f^* \ge \langle \nabla f(x^*), \bar{x}_R^{\mathrm{sc}} - x^*\rangle + \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2.$$

Since x^* minimizes f over the convex set $\mathcal{X}$,

$$\langle \nabla f(x^*), x - x^*\rangle \ge 0.$$

Therefore,

$$\begin{aligned} f(\bar{x}_R^{\mathrm{sc}}) - f^* &\ge \langle \nabla f(x^*), \bar{x}_R^{\mathrm{sc}} - x\rangle + \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2 \\ &\ge -G_* \operatorname{dist}(\bar{x}_R^{\mathrm{sc}}, \mathcal{X}) + \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2. \end{aligned}$$

Taking expectations gives the claimed lower bound.

(II) INFEASIBILITY. By Theorem A.20(ii),

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \le U_R^{\mathrm{sc}}.$$

Applying Lemma A.21 gives

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \le \frac{8\exp\left(\frac{2240c_l L_f^2}{\mu}\right)}{\mu^2 c_l c_\rho (R+1)} \times \left[\frac{\mu D^2}{c_l R} + G_*^2 + \frac{1+\log R}{R}\left(80G_*^2 + \frac{32\sigma^2}{sH} + \frac{208\sigma^2}{H}\right)\right],$$

which proves the stated infeasibility bound.

(III-1) η-APPROXIMATE LOWER BOUND ON SUBOPTIMALITY. Suppose Assumption 2.2 holds. By Lemma A.11,

$$f(\bar{x}_R^{\mathrm{sc}}) - f^* \ge \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2 - G_*\left(\sum_{i\in\mathcal{I}^{[\eta,\infty)}} \frac{2}{\|\nabla h_i(\hat{y}_i)\|} g_{i,\eta}(\bar{x}_R^{\mathrm{sc}}) + \sum_{i\in\mathcal{I}^{(-\infty,\eta]}} \frac{\sqrt{2\eta}}{\|\nabla h_i(\hat{y}_i)\|}\sqrt{g_{i,\eta}(\bar{x}_R^{\mathrm{sc}})}\right).$$

Using $\|\nabla h_i(\hat{y}_i)\| \ge \kappa_{\min}$ and

$$\sum_{i\in\mathcal{I}^{[\eta,\infty)}} g_{i,\eta}(\bar{x}_R^{\mathrm{sc}}) \le m\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}}),$$

the first penalty term is bounded by

$$\frac{2mG_*}{\kappa_{\min}}\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}}).$$

For $i \in \mathcal{I}^{(-\infty,\eta]}$, Proposition 3.6(iii) gives either $g_{i,\eta}(\bar{x}_R^{\mathrm{sc}}) = 0$ or $g_{i,\eta}(\bar{x}_R^{\mathrm{sc}}) \le \eta/2$. Thus the second penalty term is $\mathcal{O}(\eta)$, and hence

$$f(\bar{x}_R^{\mathrm{sc}}) - f^* \ge \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2 - \frac{2mG_*}{\kappa_{\min}}\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}}) - \mathcal{O}(\eta).$$

Taking expectations and using $\mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \le U_R^{\mathrm{sc}}$ gives

$$\mathbb{E}[f(\bar{x}_R^{\mathrm{sc}}) - f^*] \ge \frac{\mu}{2}\mathbb{E}\left[\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2\right] - \frac{2mG_*}{\kappa_{\min}}U_R^{\mathrm{sc}} - \mathcal{O}(\eta).$$

Substituting the bound on U_R^{sc} from Lemma A.21 proves the stated result.

(III-2) ASYMPTOTIC CONVERGENCE. For the two assertions highlighted in the theorem, Lemma A.21 gives

$$0 \le \mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \le U_R^{\mathrm{sc}} \longrightarrow 0$$

for every fixed $\eta > 0$. Hence,

$$\mathbb{E}[\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})] \longrightarrow 0.$$

It remains to establish convergence of the expected absolute suboptimality error. From the pointwise lower bound used above, after dropping the nonnegative strongly convex term,

$$\max\{f^* - f(\bar{x}_R^{\mathrm{sc}}), 0\} \le \frac{2mG_*}{\kappa_{\min}}\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}}) + \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})}.$$

Using

$$|f(\bar{x}_R^{\mathrm{sc}}) - f^*| = f(\bar{x}_R^{\mathrm{sc}}) - f^* + 2\max\{f^* - f(\bar{x}_R^{\mathrm{sc}}), 0\},$$

taking expectations, and applying Jensen's inequality, we obtain

$$\mathbb{E}[|f(\bar{x}_R^{\mathrm{sc}}) - f^*|] \le B_R^{\mathrm{sc}} + \frac{4mG_*}{\kappa_{\min}}U_R^{\mathrm{sc}} + \frac{2m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{U_R^{\mathrm{sc}}}.$$

For any fixed $\eta > 0$, Lemma A.21 gives

$$B_R^{\mathrm{sc}} \longrightarrow 0, \qquad U_R^{\mathrm{sc}} \longrightarrow 0.$$

Therefore,

$$\mathbb{E}[|f(\bar{x}_R^{\mathrm{sc}}) - f^*|] \longrightarrow 0.$$

(III-3) EXACT LOWER BOUND ON SUBOPTIMALITY. Applying Lemma A.11 and the same aggregation argument as in the proof of Theorem A.20(iii), we obtain

$$f(\bar{x}_R^{\mathrm{sc}}) - f^* \geq \frac{\mu}{2}\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2 - \frac{2mG_*}{\kappa_{\min}}\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}}) - \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{\bar{g}_\eta(\bar{x}_R^{\mathrm{sc}})}.$$

Taking expectations and applying Jensen's inequality to the square-root term gives

$$\mathbb{E}[f(\bar{x}_R^{\mathrm{sc}}) - f^*] \geq \frac{\mu}{2}\mathbb{E}\big[\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2\big] - \frac{2mG_*}{\kappa_{\min}}U_R^{\mathrm{sc}} - \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\sqrt{U_R^{\mathrm{sc}}}.$$

Substituting the explicit estimate for U_R^{sc} from Lemma A.21 gives

$$\begin{aligned}\mathbb{E}[f(\bar{x}_R^{\mathrm{sc}}) - f^*] \geq\ & \frac{\mu}{2}\mathbb{E}\big[\|\bar{x}_R^{\mathrm{sc}} - x^*\|^2\big] \\ & - \frac{2mG_*}{\kappa_{\min}}\frac{8\exp\!\left(\frac{2240c_l L_f^2}{\mu}\right)}{\mu^2 c_l c_\rho (R+1)}\left[\frac{\mu D^2}{c_l R} + G_*^2 + \frac{1+\log R}{R}\left(80G_*^2 + \frac{32\sigma^2}{sH} + \frac{208\sigma^2}{H}\right)\right] \\ & - \frac{m\sqrt{2\eta}\,G_*}{\kappa_{\min}}\left\{\frac{8\exp\!\left(\frac{2240c_l L_f^2}{\mu}\right)}{\mu^2 c_l c_\rho (R+1)}\left[\frac{\mu D^2}{c_l R} + G_*^2 + \frac{1+\log R}{R}\left(80G_*^2 + \frac{32\sigma^2}{sH} + \frac{208\sigma^2}{H}\right)\right]\right\}^{1/2}.\end{aligned}$$

Lemma A.22 (Feasible parameter choice for the strongly convex setting)**.** Consider the parameter conditions in Theorem A.20. Suppose that $\mu > 0$, $1 \leq s \leq m$, and $\gamma_g = \sqrt{s}$. Let $c_\rho > 0$ be arbitrary and fixed, and choose

$$\gamma_{l,r} = \frac{c_l}{H\sqrt{s}\left(1 + \frac{\mu c_l}{4}(r-1)\right)}, \qquad \rho_r = c_\rho\left(1 + \frac{\mu c_l}{4}(r-1)\right), \qquad r \geq 1.$$

Let

$$L_{r,\eta} := L_f + \rho_r L_{g,\eta}, \qquad L_{1,\eta} = L_f + c_\rho L_{g,\eta},$$

and define

$$A_1 := 2c_l L_{1,\eta} + 104c_l^2 L_{1,\eta}^2.$$

Suppose that $c_l > 0$ satisfies

$$\mu m c_l + A_1 \leq s, \qquad c_l c_\rho L_{g,\eta}(800s + 96A_1) \leq s.$$

Then all parameter conditions in Theorem A.20 hold for every $r \geq 1$. Specifically, with

$$\tilde{\gamma}_r := H\gamma_{l,r}\gamma_g, \qquad A_r := 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2,$$

we have

$$\begin{aligned}&\gamma_{l,r} \leq \frac{2}{HL_{r,\eta}}, \\ &A_r + \frac{\mu m \gamma_g^2 \tilde{\gamma}_r}{s} \leq \gamma_g^2, \\ &\tilde{\gamma}_r \rho_r L_{g,\eta}\left(800 + \frac{96A_r}{\gamma_g^2}\right) \leq 1, \\ &0 < \mu\tilde{\gamma}_r < 2.\end{aligned}$$

Moreover, $\tilde{\gamma}_r$ is positive and nonincreasing, and ρ_r is positive and nondecreasing. Since $B_r \geq 0$,

$$0 < 1 - \frac{\mu\tilde{\gamma}_r}{2} < 1, \qquad 1 - \frac{\mu\tilde{\gamma}_r}{2} + B_r > 0,$$

so the weights in that theorem are well defined and positive. For every fixed $c_\rho > 0$, such a $c_l > 0$ exists.

Proof. Write

$$q_r := 1 + \frac{\mu c_l}{4}(r-1).$$

Then $q_r \geq 1$ is nondecreasing, and

$$\tilde{\gamma}_r = H\gamma_{l,r}\gamma_g = \frac{c_l}{q_r}, \qquad \rho_r = c_\rho q_r, \qquad \tilde{\gamma}_r\rho_r = c_l c_\rho.$$

Thus the stated positivity and monotonicity properties hold.

Using the definition of $L_{r,\eta}$ gives

$$\tilde{\gamma}_r L_{r,\eta} = \frac{c_l L_f}{q_r} + c_l c_\rho L_{g,\eta} \le c_l L_{1,\eta}.$$

Consequently,

$$A_r = 2\tilde{\gamma}_r L_{r,\eta} + 104\tilde{\gamma}_r^2 L_{r,\eta}^2 \le 2c_l L_{1,\eta} + 104 c_l^2 L_{1,\eta}^2 = A_1.$$

Since $104c_l^2 L_{1,\eta}^2 \le A_1 \le s$,

$$H\gamma_{l,r} L_{r,\eta} = \frac{\tilde{\gamma}_r L_{r,\eta}}{\sqrt{s}} \le \frac{c_l L_{1,\eta}}{\sqrt{s}} \le \frac{1}{\sqrt{104}} < 2.$$

This verifies the local stepsize condition.

Using $\gamma_g^2 = s$ and $\tilde{\gamma}_r \le c_l$, we obtain

$$A_r + \frac{\mu m \gamma_g^2 \tilde{\gamma}_r}{s} = A_r + \mu m \tilde{\gamma}_r \le A_1 + \mu m c_l \le s = \gamma_g^2.$$

The penalty condition follows from

$$\tilde{\gamma}_r \rho_r L_{g,\eta} \left(800 + \frac{96 A_r}{\gamma_g^2}\right) \le c_l c_\rho L_{g,\eta} \left(800 + \frac{96 A_1}{s}\right) \le 1.$$

Furthermore, $\mu m c_l + A_1 \le s$ and $s \le m$ imply

$$0 < \mu\tilde{\gamma}_r \le \mu c_l \le \frac{s}{m} \le 1 < 2.$$

Hence

$$\frac{1}{2} \le 1 - \frac{\mu\tilde{\gamma}_r}{2} < 1.$$

Together with $B_r \ge 0$, this establishes positivity of the coefficients defining the weights.

Finally, for fixed $c_\rho > 0$, $L_{1,\eta}$ is fixed. As $c_l \downarrow 0$,

$$A_1 \longrightarrow 0, \qquad \mu m c_l + A_1 \longrightarrow 0, \qquad c_l c_\rho L_{g,\eta}(800s + 96A_1) \longrightarrow 0.$$

Thus both required inequalities hold for sufficiently small $c_l > 0$. □

Corollary A.23. Suppose $\mu > 0$, $1 \le s \le m$, and $\gamma_g = \sqrt{s}$. Fix any $c_\rho > 0$, and let

$$L_{1,\eta} := L_f + c_\rho L_{g,\eta}.$$

Define

$$\bar{c}_l^{\mathrm{sc}} := \min\left\{ \frac{\sqrt{(2L_{1,\eta} + \mu m)^2 + 416 s L_{1,\eta}^2} - (2L_{1,\eta} + \mu m)}{208 L_{1,\eta}^2}, \frac{1}{896 c_\rho L_{g,\eta}} \right\}.$$

Choose any

$$0 < c_l \le \bar{c}_l^{\mathrm{sc}},$$

and set

$$\gamma_{l,r} = \frac{c_l}{H\sqrt{s}\left(1 + \frac{\mu c_l}{4}(r-1)\right)}, \qquad \rho_r = c_\rho \left(1 + \frac{\mu c_l}{4}(r-1)\right), \qquad r \ge 1.$$

Then all parameter conditions required in Theorem A.20 hold for every $r \ge 1$. In particular,

$$\gamma_{l,r} \le \frac{2}{H L_{r,\eta}},$$

$$A_r + \frac{\mu m \gamma_g^2 \tilde{\gamma}_r}{s} \le \gamma_g^2,$$

$$\tilde{\gamma}_r \rho_r L_{g,\eta} \left(800 + \frac{96 A_r}{\gamma_g^2}\right) \le 1,$$

$$0 < \mu\tilde{\gamma}_r < 2.$$

Moreover, $\tilde{\gamma}_r$ is positive and nonincreasing, ρ_r is positive and nondecreasing, and the weights in Theorem A.20 are well defined and positive.

Proof. By the choice of c_l,

$$104L_{1,\eta}^2 c_l^2 + (2L_{1,\eta} + \mu m)c_l \le s.$$

Therefore,

$$\mu m c_l + 2c_l L_{1,\eta} + 104 c_l^2 L_{1,\eta}^2 \le s.$$

Recalling

$$A_1 = 2c_l L_{1,\eta} + 104 c_l^2 L_{1,\eta}^2,$$

we obtain

$$\mu m c_l + A_1 \le s.$$

In particular, $A_1 \le s$, and hence

$$800s + 96A_1 \le 896s.$$

Using

$$c_l \le \frac{1}{896 c_\rho L_{g,\eta}},$$

we further obtain

$$c_l c_\rho L_{g,\eta}(800s + 96A_1) \le 896s\, c_l c_\rho L_{g,\eta} \le s.$$

Thus both conditions of Lemma A.22 are satisfied. The claimed parameter properties and positivity of the weights then follow directly from that lemma. □

A.5 Experiment Setup

Datasets, model and training protocol. We evaluate on three benchmark datasets under federated settings: MNIST (LeCun et al., 1998) EMNIST (Cohen et al., 2017), and CIFAR-10 (Krizhevsky & Hinton, 2009). For all experiments, we use a single-layer linear model to match our convex setting: $f(x) = \mathrm{softmax}(Wx + b)$ where $W \in \mathbb{R}^{d\times C}$, with d the input dimension and C the number of classes. We distribute data among $m = 5$ clients using a Dirichlet distribution $\mathrm{Dir}(\alpha)$ with $\alpha = 0.3$, which produces moderate-to-high statistical heterogeneity across clients. All methods share the following federated training configuration: $R = 200$ communication rounds, client participation fraction 0.7 (3 out of 5 clients per round), $E = 2$ local epochs per round, training batch size 64, and evaluation batch size 256. Results are averaged over 2 independent runs with different random seeds.

Constraint threshold u_i. Recall the client constraint $h_i(x) = f_i(x) + \frac{\gamma}{2}\|x\|^2 - u_i \le 0$. The threshold u_i is determined by a warmup phase: Before federated training begins, each client i runs unconstrained gradient descent on its local dataset for up to 1000 rounds (with early stopping when the loss stabilizes), recording the per-client loss value $h_{\mathrm{base},i}$. We then set $u_i = s \cdot \overline{h}_{\mathrm{base},i}$, where $\overline{h}_{\mathrm{base},i}$ is the time-averaged h_{base} and s is a dataset-specific scaling factor. This calibration ensures that u_i reflects each client's achievable loss level given its local data distribution, thereby producing heterogeneous constraints.

Evaluation metrics. We evaluate all methods on each client's test set at the end of training. We have the average test accuracy across clients, denoted as **Acc** and computed as $\frac{1}{m}\sum_{i=1}^m \mathrm{Acc}_i(\hat{x}_{R+1})$; and the average cross-entropy test loss across clients, denoted **Test Loss** and compute as $\frac{1}{m}\sum_{i=1}^m \mathrm{Loss}_i(\hat{x}_{R+1})$. To evaluate constraint satisfaction, we also report the average constraint **violation**, defined $\frac{1}{m}\sum_{i=1}^m \max\{h_i(\hat{x}_{R+1}), 0\}$, and the **feasible ratio**, defined as the fraction of clients satisfying the local constraint, $\frac{1}{m}\sum_{i=1}^m \mathbf{1}[h_i(\hat{x}_{R+1}) \le 0]$.

A.6 Additional Experiment

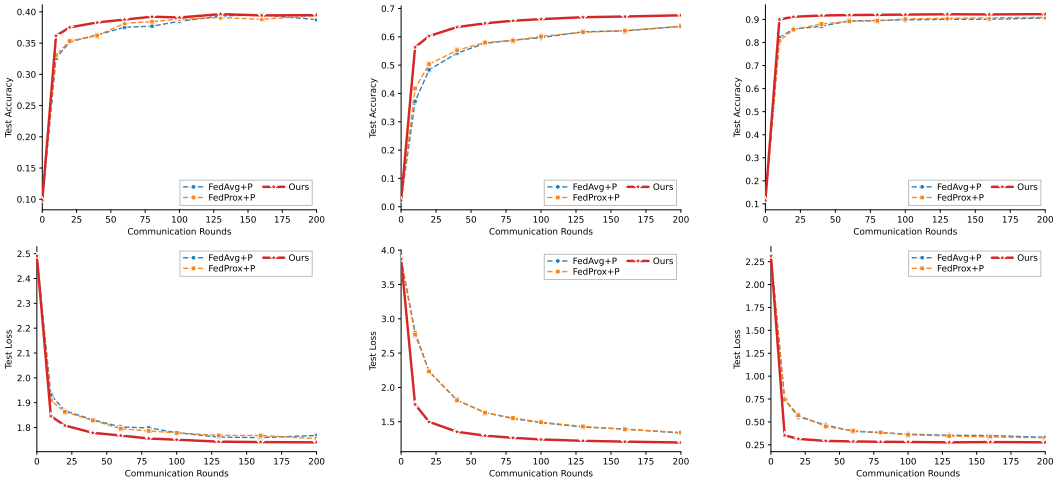
In this section, we conduct additional experiment compare to two federtaed method FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020) with penalty parameters.

A.6.1 Penalty FL

Penalty Methods We compare AdaP-SCAFFOLD with two penalty-augmented baselines (FedAvg+P, FedProx+P; best ρ chosen by grid search) and two constraint-FL methods (FedDA, ProxAL), all sharing the same warmup-calibrated u_i, $m = 5$ clients, Dirichlet α=0.3, K=20, 200 rounds. AdaP-SCAFFOLD attains the highest accuracy and lowest test loss on all three datasets. SCAFFOLD-corrected updates and the moving penalty schedule yield a more accurate model than methods enforcing the same constraint without variance reduction.

	Dataset	FedAvg	FedProx	SCAFFOLD	Ditto	APFL	Per-FedAvg	Ours
Acc.↑	MNIST	0.904	0.907	0.916	0.896	0.911	0.903	**0.923**
	CIFAR-10	0.377	0.391	0.392	0.365	0.377	**0.396**	0.393
	EMNIST	0.633	0.636	0.652	0.639	0.643	0.621	**0.677**
Loss↓	MNIST	0.334	0.335	0.304	0.350	0.317	0.354	**0.279**
	CIFAR-10	1.779	1.762	1.764	1.825	1.789	1.744	**1.743**
	EMNIST	1.342	1.338	1.290	1.338	1.320	1.430	**1.197**
Vio.↓	MNIST	0.286	0.292	0.210	0.291	0.408	0.392	**0.190**
	CIFAR-10	0.105	0.098	0.027	0.120	0.208	0.119	**0.026**
	EMNIST	0.294	0.302	0.236	0.304	0.398	0.416	**0.192**
Feas.↑	MNIST	0.40	0.44	0.72	0.38	0.20	0.24	**0.82**
	CIFAR-10	0.56	0.56	0.88	0.54	0.56	0.54	**0.90**
	EMNIST	0.66	0.70	0.74	0.68	0.58	0.52	**0.90**

Table 4: Average accuracy, loss, violation, and feasible ratio for all methods.



(a) CIFAR-10 (b) EMNIST (c) MNIST