DS FedProxGrad: Asymptotic Stationarity Without Noise Floor in Fair Federated Learning

Huzaifa Arif

Abstract

Recent work [1] introduced Federated Proximal Gradient (FedProxGrad) for solving non-convex composite optimization problems in group fair federated learning. However, the original analysis established convergence only to a *noise-dominated neighborhood of stationarity*, with explicit dependence on a variance-induced noise floor. In this work, we provide an improved asymptotic convergence analysis for a generalized FedProxGrad-type analytical framework with inexact local proximal solutions and explicit fairness regularization. We call this extended analytical framework **DS FedProxGrad** (Decay Step Size FedProxGrad). Under a Robbins-Monro step-size schedule [4] and a mild decay condition on local inexactness, we prove that $\liminf_{r\to\infty} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = 0$, i.e., the algorithm is asymptotically stationary and the convergence rate does not depend on a variance-induced noise floor.

1 Introduction

A recent TMLR paper [1] introduced FedProxGrad, an algorithm for composite optimization with smooth but non-convex objectives. The analysis showed convergence to a neighborhood of stationarity whose size depends on the stochastic gradient variance σ^2.

In this work, we provide an **improved asymptotic convergence analysis** for a generalized FedProxGrad-type analytical framework with inexact local proximal solutions and an explicit fairness regularizer. Under a Robbins–Monro step-size schedule and mild decay conditions on local inexactness, we establish

$$\liminf_{r\to\infty} \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] = 0. \tag{1}$$

1.1 Why Removing the Noise Floor Matters

From a theoretical perspective, previous FedProxGrad analyses [1] left open whether the method can achieve true stationarity under realistic assumptions, or whether the noise floor is intrinsic. We answer affirmatively: ***FedProxGrad*** *admits standard nonconvex SGD guarantees with decaying step sizes and controlled inexactness.*

For fairness and safety tuning, practitioners often extend training in late rounds to reduce fairness violations or residual bias. A noise-floor bound is pessimistic: it suggests a hard limit that training longer cannot overcome, potentially discouraging use of FedProxGrad for group fairness optimization. Our analysis shows that the expected gradient norm can be driven arbitrarily small, making extended training theoretically sound for reducing fairness loss or constraint violations.

Our analysis also provides a **unified view of stochasticity and inexactness**, cleanly separating the roles of stochastic gradient noise ($\mathbf{e}_r^{\text{stoch}}$) and deterministic local inexactness ($\mathbf{e}_{f,i}^r$). We show both can be made asymptotically harmless when their combined contribution is square-summable under the chosen schedule.

1.2 Relation to FedProxGrad

Our analysis follows FedProx [3]: each client approximately solves a proximal subproblem $f_i(\mathbf{w}) + \frac{1}{2\eta_r}\|\mathbf{w} - \mathbf{x}^r\|^2$ under bounded dissimilarity and inexact local solutions. Unlike [1], we adopt a SCAFFOLD-type viewpoint [2], rewriting the update as a gradient step on F perturbed by deterministic and stochastic errors. This yields a descent inequality

$$\mathbb{E}[F(\mathbf{x}^{r+1})] \leq \mathbb{E}[F(\mathbf{x}^r)] - \eta_r \Phi_r \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \mathcal{O}(\eta_r^2). \tag{2}$$

Combined with a Robbins–Monro stepsize and decaying inexactness budget $\gamma_r = \mathcal{O}(\eta_r)$, this yields $\liminf_r \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = 0$ without a noise floor and standard $\mathcal{O}(1/\sqrt{R})$ finite-time rates.

2 Algorithm and Update Dynamics

2.1 The DS `FedProxGrad` Analytical Framework

Our DS `FedProxGrad` analytical framework extends the original `FedProxGrad` framework [1] with explicit inexactness modeling and decaying step sizes. This is not a new algorithm, but rather an analytical approach for studying `FedProxGrad` under decay step size schedules. In round r, the algorithm proceeds as follows:

Step 1: Local Proximal Step. Each client i computes an approximate solution $\mathbf{y}_i^r$ to the proximal operator of their local loss f_i, centered at the global model $\mathbf{x}^r$ with step size η_r:

$$\mathbf{y}_i^r \approx \arg\min_{\mathbf{w}} \left\{ h_i(\mathbf{w}; \mathbf{x}^r) := f_i(\mathbf{w}) + \frac{1}{2\eta_r}\|\mathbf{w} - \mathbf{x}^r\|^2 \right\}. \tag{3}$$

Step 2: Aggregation. The server averages the local solutions:

$$\bar{\mathbf{y}}^r = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i^r. \tag{4}$$

Step 3: Global Update. The server applies the aggregated update and takes a gradient step on the fairness term $r(\mathbf{x})$:

$$\mathbf{x}^{r+1} = \bar{\mathbf{y}}^r - \eta_r \lambda \nabla r(\mathbf{x}^r) + \eta_r \mathbf{e}_r^{\text{stoch}}, \tag{5}$$

where $\mathbf{e}_r^{\text{stoch}}$ is the stochastic error in the fairness gradient estimation.

2.2 Abstract Error Decomposition

We rewrite the update in (5) as a gradient descent step on F perturbed by errors. Define the deterministic proximal error $\mathbf{e}_{f,i}^r$ such that:

$$\frac{\mathbf{y}_i^r - \mathbf{x}^r}{\eta_r} = -\nabla f_i(\mathbf{x}^r) + \mathbf{e}_{f,i}^r. \tag{6}$$

Rearranging and substituting into (5), the dynamics are governed by:

$$\mathbf{x}^{r+1} - \mathbf{x}^r = -\eta_r \nabla F(\mathbf{x}^r) + \eta_r \mathbf{e}_f^r + \eta_r \mathbf{e}_r^{\text{stoch}}, \tag{7}$$

where $\mathbf{e}_f^r = \frac{1}{N}\sum_{i=1}^N \mathbf{e}_{f,i}^r$.

Remark 1 (Sources of randomness)**.** *Unless stated otherwise, expectations are taken with respect to the randomness in the fairness gradient estimator* $\mathbf{e}_r^{stoch}$ *(and any randomness in the local solvers that produces* $\mathbf{e}_{f,i}^r$*). We condition on* $\mathbf{x}^r$ *whenever convenient and then take total expectations.*

3 Comparison with Original FedProxGrad

Table 1 provides a detailed comparison between the original `FedProxGrad` analysis in [1] and this work.

Table 1: Feature comparison: original `FedProxGrad` analysis vs. DS `FedProxGrad` (this work).

Feature	`FedProxGrad` **[1]**	**DS** `FedProxGrad` **(this work)**
Fairness-aware objective	✓	✓
Bounded fairness gradient	✗	✓
Inexactness schedule $\gamma_r \leq c_\gamma \eta_r$	✗	✓
Separate error modeling (stoch vs. prox)	✗	✓
Gradient noise floor	✓	✗
Asymptotic stationarity	✗	✓
Robbins–Monro stepsizes	✗	✓
Finite-horizon rate $\mathcal{O}(1/\sqrt{R})$	✗	✓

4 Assumptions

Assumption 1 (Regularity)**.**

(i) ***Smoothness:*** *Each local function* f_i *is* L_f*-smooth, and the regularizer* r *is* L_r*-smooth. The global objective* F *is* L_F*-smooth with* $L_F = L_f + \lambda L_r$*.*

(ii) ***Weak Convexity:*** *Each* f_i *is* ρ*-weakly convex (*$\rho \geq 0$*).*

Remark 2 (On weak convexity)**.** *The* ρ*-weak convexity requirement in Assumption 1(ii) guarantees that the local proximal objective is well-posed and that the strong convexity condition in Assumption 4(ii) can be enforced via a sufficiently small* η_r*. This requirement is identical to the original* `FedProxGrad` *analysis [1]. The subsequent descent analysis only uses smoothness, so the weak-convexity assumption does not otherwise appear in Lemmas 1–2.*

Assumption 2 (Variance and Unbiasedness). *There exist constants $G, B, \sigma > 0$ such that for all iterates $\mathbf{x}^r$,*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\mathbf{x})\|^2 \le G^2 + B^2\|\nabla f(\mathbf{x})\|^2, \tag{8}$$

$$\mathbb{E}[\mathbf{e}_r^{stoch} \mid \mathbf{x}^r] = 0, \qquad \mathbb{E}[\|\mathbf{e}_r^{stoch}\|^2 \mid \mathbf{x}^r] \le \sigma^2. \tag{9}$$

Assumption 3 (Fairness Gradient Control). *There exists a constant $G_r > 0$ such that $\|\nabla r(\mathbf{x})\| \le G_r$ for all $\mathbf{x}$.*

Remark 3 (Step-size–controlled local movement). *Lemma 3 in the Appendix shows that each local update satisfies*

$$\|\mathbf{y}_i^r - \mathbf{x}^r\| \le \eta_r(1+\gamma_r)\,\|\nabla f_i(\mathbf{x}^r)\|.$$

Assumption 4 (Step Size and Inexactness).

(i) ***Robbins-Monro:*** $\sum_{r=0}^{\infty}\eta_r = \infty$ *and* $\sum_{r=0}^{\infty}\eta_r^2 < \infty$.

(ii) ***Local Strong Convexity:*** $\eta_r \le \frac{1}{2\rho}$ *ensuring* $h_i(\cdot;\mathbf{x}^r)$ *is* μ_r*-strongly convex.*

(iii) ***Inexactness Condition:*** *The local solver output satisfies:*

$$\left\|\frac{\mathbf{y}_i^r - \mathbf{x}^r}{\eta_r} + \nabla f_i(\mathbf{x}^r)\right\| \le \gamma_r\|\nabla f_i(\mathbf{x}^r)\|, \tag{10}$$

where $0 \le \gamma_r \le c_\gamma\eta_r$.

(iv) ***Small Step Size:*** *For all* r, $\eta_r \le \frac{1}{6L_F}$.

5 Convergence Analysis

5.1 Bounding the Deterministic Error

Lemma 1 (Error Bound). *Under Assumption 4(iii) and 2, the aggregated deterministic error satisfies:*

$$\|\mathbf{e}_f^r\|^2 \le \gamma_r^2(G^2 + B^2\|\nabla f(\mathbf{x}^r)\|^2). \tag{11}$$

Proof. Using Jensen's inequality and (10):

$$\|\mathbf{e}_f^r\|^2 = \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_{f,i}^r \right\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{e}_{f,i}^r\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \gamma_r^2 \|\nabla f_i(\mathbf{x}^r)\|^2. \tag{12}$$

Applying Assumption 2 yields the result. □

Assumption 3 implies $\|\nabla r(\mathbf{x}^r)\| \leq G_r$; together with $\nabla F(\mathbf{x}^r) = \nabla f(\mathbf{x}^r) + \lambda \nabla r(\mathbf{x}^r)$ (where λ is the weight in (5)), we get

$$\|\nabla f(\mathbf{x}^r)\| \leq \|\nabla F(\mathbf{x}^r)\| + \lambda G_r.$$

Consequently,

$$\|\nabla f(\mathbf{x}^r)\|^2 \leq 2\|\nabla F(\mathbf{x}^r)\|^2 + 2\lambda^2 G_r^2. \tag{13}$$

This relation will be used to convert (11) into bounds that depend only on $\|\nabla F(\mathbf{x}^r)\|^2$.

5.2 One-Step Descent Lemma

Lemma 2 (Descent Inequality)**.** *Under Assumptions 1–4 and 3, there exist constants $\Phi_r \geq \frac{1}{4}$ and $\Xi_r = \mathcal{O}(\eta_r^2)$ such that*

$$\mathbb{E}[F(\mathbf{x}^{r+1})] - \mathbb{E}[F(\mathbf{x}^r)] \leq -\eta_r \Phi_r \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] + \Xi_r. \tag{14}$$

Proof. By L_F-smoothness, conditioning on $\mathbf{x}^r$ and taking expectations yields

$$\mathbb{E}[F(\mathbf{x}^{r+1})] \leq \mathbb{E}\big[F(\mathbf{x}^r) + \langle \nabla F(\mathbf{x}^r), \mathbf{x}^{r+1} - \mathbf{x}^r \rangle + \tfrac{L_F}{2} \|\mathbf{x}^{r+1} - \mathbf{x}^r\|^2\big]. \tag{15}$$

Let $\mathbf{g} = \nabla F(\mathbf{x}^r)$. Lemma 1 together with (13) yields

$$\|\mathbf{e}_f^r\|^2 \leq \gamma_r^2 \Big(G^2 + B^2 (2\|\mathbf{g}\|^2 + 2\lambda^2 G_r^2) \Big) \leq C_e \gamma_r^2 \big(\|\mathbf{g}\|^2 + K_e^2 \big), \tag{16}$$

with explicit constants $C_e = 2B^2$ and $K_e^2 = \frac{G^2}{2B^2} + \lambda^2 G_r^2$.

Analysis of the Linear Term. From (7),

$$\mathbb{E}[\langle \mathbf{g}, \mathbf{x}^{r+1} - \mathbf{x}^r \rangle] = \mathbb{E}[\langle \mathbf{g}, -\eta_r \mathbf{g} + \eta_r \mathbf{e}_f^r + \eta_r \mathbf{e}_r^{\text{stoch}} \rangle]. \tag{17}$$

By Assumption 2, $\mathbb{E}[\mathbf{e}_r^{\text{stoch}} \mid \mathbf{x}^r] = 0$, and using Young's inequality,

$$\mathbb{E}[\langle \mathbf{g}, -\eta_r \mathbf{g} + \eta_r \mathbf{e}_f^r \rangle] \leq -\eta_r \|\mathbf{g}\|^2 + \frac{\eta_r}{4} \|\mathbf{g}\|^2 + \eta_r \|\mathbf{e}_f^r\|^2. \tag{18}$$

Analysis of the Quadratic Term. Using (7),

$$\mathbb{E}[\|\mathbf{x}^{r+1} - \mathbf{x}^r\|^2] \leq 3\eta_r^2 \mathbb{E}[\|\mathbf{g}\|^2 + \|\mathbf{e}_f^r\|^2 + \|\mathbf{e}_r^{\text{stoch}}\|^2]. \tag{19}$$

Combining Terms and Absorption. Substituting into (15) yields the coefficient of $\|\mathbf{g}\|^2$ as

$$-\eta_r + \frac{\eta_r}{4} + \frac{3L_F \eta_r^2}{2} + \eta_r \left(1 + \frac{3L_F \eta_r}{2} \right) C_e \gamma_r^2. \tag{20}$$

Assumption 4(iv) implies $\frac{3L_F\eta_r^2}{2} \le \frac{\eta_r}{4}$, and with $\gamma_r \le c_\gamma \eta_r$ there exists r_0 so that for all $r \ge r_0$,

$$-\eta_r + \frac{\eta_r}{4} + \frac{3L_F\eta_r^2}{2} + \eta_r \left(1 + \frac{3L_F\eta_r}{2}\right) C_e\gamma_r^2 \le -\frac{\eta_r}{4}. \tag{21}$$

We therefore set $\Phi_r = \frac{1}{4}$ for $r \ge r_0$ and absorb the finite prefix $r < r_0$ into the constant on the right-hand side of the telescoping sum.

The remaining constant noise terms form

$$\Xi_r = \eta_r \left(1 + \frac{3L_F\eta_r}{2}\right) C_e\gamma_r^2 K_e^2 + \frac{3L_F\eta_r^2}{2}\sigma^2 = \mathcal{O}(\eta_r^3 + \eta_r^2) = \mathcal{O}(\eta_r^2). \tag{22}$$

Taking total expectations on both sides of (15) then gives (14). □

5.3 Main Result

Theorem 1 (Asymptotic Stationarity - Main Result)**.** *Under Assumptions 1–4 and 3, DS* `FedProxGrad` *achieves exact asymptotic stationarity:*

$$\liminf_{r\to\infty} \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] = 0. \tag{23}$$

This eliminates the $\mathcal{O}(\sigma^2)$ noise floor present in the original constant step-size `FedProxGrad` *analysis [1].*

Proof. Summing (14) over $r = 0, \ldots, R$ gives

$$\sum_{r=0}^{R} \eta_r \Phi_r \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] \le \mathbb{E}[F(\mathbf{x}^0)] - \mathbb{E}[F(\mathbf{x}^{R+1})] + \sum_{r=0}^{R} \Xi_r. \tag{24}$$

Letting $R \to \infty$, using $F(\mathbf{x}^{R+1}) \ge F^\star$ and $\sum_r \Xi_r < \infty$ yields

$$\sum_{r=0}^{\infty} \eta_r \Phi_r \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] < \infty. \tag{25}$$

Because $\sum_{r=0}^{\infty} \eta_r = \infty$ and $\Phi_r \ge \frac{1}{4}$, the only way for the weighted sum to remain finite is for $\liminf_{r\to\infty} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = 0$. □

Under additional standard conditions (e.g., via the Robbins–Siegmund lemma), one can often strengthen the result to an almost-sure liminf statement; here we focus on convergence in expectation.

Corollary 1 (Convergence Rates)**.**

(i) ***Infinite Horizon:*** *If $\eta_r = \frac{c}{r^\alpha}$ with $\alpha \in (0.5, 1]$, then $\liminf_{r\to\infty} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = 0$ by Theorem 1.*

(ii) ***Finite Horizon:*** *If $\eta_r = \frac{c}{\sqrt{R}}$ (constant over R rounds), then*

$$\min_{0\le r\le R} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{R}}\right). \tag{26}$$

Summing (14) over $r = 0, \ldots, R-1$ gives $\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = \mathcal{O}(1/\sqrt{R})$, and the minimum is bounded by this average. Note that this schedule does not satisfy $\sum_{r=0}^{\infty} \eta_r^2 < \infty$ as $R \to \infty$, hence it applies only to the finite-time minimum rate.

6 Conclusion

We introduced *DS* `FedProxGrad`, a fairness-aware analytical framework for `FedProxGrad` equipped with explicit modeling of inexact local proximal updates and a Robbins–Monro stepsize schedule.

Our analysis shows that, under mild regularity and inexactness conditions, DS `FedProxGrad` achieves *exact asymptotic stationarity* for the composite objective $F = f + \lambda r$:

$$\liminf_{r\to\infty} \mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] = 0,$$

thereby eliminating the $\mathcal{O}(\sigma^2)$ variance-induced noise floor that appears in the original constant stepsize `FedProxGrad` analysis [1].

Algorithmically, DS `FedProxGrad` stays within the FedProx/`FedProxGrad` family: clients approximately solve proximal subproblems around the global model, the server aggregates local updates, and a fairness-corrective step is applied on $r(\mathbf{x})$. Conceptually, however, our analytical framework is more realistic: we (i) optimize an explicit fairness-aware objective, (ii) separate stochastic gradient noise from deterministic inexactness, and (iii) couple both effects to a decaying stepsize so that their contribution becomes square-summable and asymptotically harmless. This yields standard nonconvex SGD-type guarantees and an $\mathcal{O}(1/\sqrt{R})$ finite-horizon rate for fair federated composite optimization.

From a practical standpoint, our results justify prolonged late-stage training of FedProx-style methods for fairness: under appropriate stepsize and inexactness schedules, the expected gradient norm can be driven arbitrarily small, rather than plateauing at a variance-determined floor. In this sense, DS `FedProxGrad` should be viewed as a theoretically grounded analytical framework for `FedProxGrad` in fair federated learning, particularly in regimes where long-horizon training and approximate local solves are unavoidable.

A Appendix

A.1 Iterate Boundedness

[3]

Lemma 3 (Local Update Boundedness)**.** *If the inexactness condition* (10) *holds, then*

$$\|\mathbf{y}_i^r - \mathbf{x}^r\| \leq \eta_r(1+\gamma_r)\|\nabla f_i(\mathbf{x}^r)\|. \tag{27}$$

Proof. From (6), we have

$$\frac{\mathbf{y}_i^r - \mathbf{x}^r}{\eta_r} = -\nabla f_i(\mathbf{x}^r) + \mathbf{e}_{f,i}^r, \tag{28}$$

where $\|\mathbf{e}_{f,i}^r\| \leq \gamma_r\|\nabla f_i(\mathbf{x}^r)\|$ by (10).

Taking norms on both sides:

$$\left\|\frac{\mathbf{y}_i^r - \mathbf{x}^r}{\eta_r}\right\| \leq \|\nabla f_i(\mathbf{x}^r)\| + \|\mathbf{e}_{f,i}^r\| \tag{29}$$

$$\leq \|\nabla f_i(\mathbf{x}^r)\| + \gamma_r\|\nabla f_i(\mathbf{x}^r)\| \tag{30}$$

$$= (1+\gamma_r)\|\nabla f_i(\mathbf{x}^r)\|. \tag{31}$$

Multiplying both sides by η_r completes the proof. □

References

[1] Huzaifa Arif, Pin-Yu Chen, Keerthiram Murugesan, and Alex Gittens. Group fair federated learning via stochastic kernel regularization. *Transactions on Machine Learning Research*, 2025.

[2] 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*, pages 5132–5143. PMLR, 2020.

[3] 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.

[4] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.