Tighter Performance Theory of FedExProx

Wojciech Anyszka Kaja Gruntkowska Alexander Tyurin Peter Richtárik

King Abdullah University of Science and Technology (KAUST)

arXiv:2410.15368

Optimization Problem

We consider the federated learning problem:

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{n}\sum_{i=1}^n f_i(x)\right\}$$

- $f_i(x)$ – local objective on client i
- **Goal:** find an ε–solution

Challenges in FL

Communication bottleneck: Communication often dominates computation

Partial participation: Only subset of clients active each round

Client drift: Local updates deviate from global objective

FedExProx

FedExProx performs extrapolated proximal updates:

$$x_{k+1}=x_k+\alpha_k\left(\frac{1}{n}\sum_{i=1}^n \mathrm{prox}_{\gamma f_i}(x_k)-x_k\right)$$

- Equivalent formulation via Moreau envelopes:

$$x_{k+1}=x_k-\alpha_k\gamma\frac{1}{n}\sum_{i=1}^n\nabla M_{f_i}^{\gamma}(x_k),$$

  where $M_f^{\gamma}(x):=\min\limits_{z\in\mathbb{R}^d}\left\{f(z)+\frac{1}{2\gamma}\|z-x\|^2\right\}$
- # of **GD** iterations to find $\mathrm{prox}_{\gamma f_i}(x)$ with accuracy ε: $\mathcal{O}\left((\gamma L_i+1)\log\frac{1}{\varepsilon}\right)$

Assumptions

A1: **Convexity**: f_i are proper, closed and convex and f attains a minimum at some (potentially non-unique) point x_*

A2: **Smoothness**: f is L–smooth, i.e.,

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$$

and each f_i is differentiable and L_i–smooth, with $L_{\max}:=\max_{i\in[n]}L_i$

A3: **Interpolation**: There exists $x_*\in\mathbb{R}^d$ such that $\nabla f_i(x_*)=0$ for all $i\in[n]$

A4: **Polyak-Łojasiewicz condition**:

$$\frac{1}{2}\|\nabla M^{\gamma}(x)\|^2\ge\mu_{\gamma}^{+}(M^{\gamma}(x)-M^{\gamma}(x_*))$$

A5: **Time model**: Time complexity $\pi(\gamma)$ of a **FedExProx** step is a non-decreasing function of γ.

Not Better than GD on Quadratics

Prior result: If $\alpha_k\equiv\alpha=\frac{1}{\gamma L_\gamma}>1$, **FedExProx** converges after $\mathcal{O}\left(\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\right)$ communication rounds, where L_γ is the smoothness constant of $M^{\gamma}(x):=\frac{1}{n}\sum_{i=1}^n M_{f_i}^{\gamma}(x)$ and $R^2:=\|x_*-x_0\|^2$.

Pessimistic Result

Let A3 and A5 hold. Consider solving a non-strongly convex quadratic optimization problem, where $f_i(x)=\frac{1}{2}x^\top\mathbf{A}_ix-b_i^\top x$, with $\mathbf{A}_i\in\mathrm{Sym}_+^d$ and $b_i\in\mathbb{R}^d$. Then the total time required by **FedExProx** to find $\bar{x}$ such that $\mathbb{E}[f(\bar{x})]-f(x_*)\le\varepsilon$ is

$$T(\gamma)=\pi(\gamma)\times\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\ge\pi(0)\times\frac{LR^2}{\varepsilon}$$

for all $\gamma>0$. Moreover, when $\gamma\to 0$, then $\pi(\gamma)\times\frac{L_\gamma(1+\gamma L_{\max})R^2}{\varepsilon}\to\pi(0)\times\frac{LR^2}{\varepsilon}$, and **FedExProx** effectively reduces to **GD**.

Contributions

1. **Tighter analysis for FedExProx.** We show that previous complexity bounds were overly pessimistic
2. **Linear convergence for quadratics.** We show that **FedExProx** can achieve

$$\mathcal{O}\left(\frac{L_\gamma}{\mu_\gamma^{+}}\log\frac{1}{\varepsilon}\right)$$

3. **Time complexity insight.** Optimal to choose $\gamma>0$ when communication dominates
4. **Extensions.** Partial participation, adaptive extrapolation, and PŁ condition

Tighter and More Optimistic Results

Fix any $\gamma>0$ and consider solving non-strongly convex quadratic optimization problem, where $f_i(x)=\frac{1}{2}x^\top\mathbf{A}_ix-b_i^\top x$, with $\mathbf{A}_i\in\mathrm{Sym}_+^d$ and $b_i\in\mathbb{R}^d$. Under A3, **FedExProx** with $\alpha=1/\gamma L_\gamma$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})]-f(x_*)\le\varepsilon$ after

$$\mathcal{O}\left(\frac{L_\gamma}{\mu_\gamma^{+}}\log\frac{1}{\varepsilon}\right)$$

iterations, where L_γ is the smoothness constant of M^{γ} and μ_γ^{+} is the smallest non-zero eigenvalue of the matrix $\nabla^2 M^{\gamma}$.

Time model

The time per one global iteration of **FedExProx** has two main sources:

1. **Local computation = $\widetilde{\mathcal{O}}\left(\tau\times(\gamma L_{\max}+1)\right)$ seconds:** Each step requires clients to compute $\mathrm{prox}_{\gamma f_i}(x_k)$ iteratively. Simple solvers return a solution of subproblem i after $\widetilde{\mathcal{O}}\left(\gamma L_i+1\right)$ local iterations. If each gradient calculation takes τ seconds, the total time required for all clients to calculate $\mathrm{prox}_{\gamma f_i}(x_k)$ is $\widetilde{\mathcal{O}}\left(\tau\times(\gamma L_{\max}+1)\right)$.
2. **Communication = μ seconds:** Once the local computations are completed, clients must communicate their results before the server can execute the global step.

Time Complexity

Total time complexity of **FedExProx**:

$$T_\mu(\gamma):=\widetilde{\mathcal{O}}\left((\mu+\tau\left(\gamma L_{\max}+1\right))\times\frac{L_\gamma}{\mu_\gamma^{+}}\right)$$

where μ = communication cost, τ = computation cost

Total time complexity of **GD**:

$$T_{\mathrm{GD}}:=T_\mu(0)=\widetilde{\mathcal{O}}\left((\mu+\tau)\times\frac{L}{\mu^{+}}\right)$$

Up to a constant factor, the time complexity is minimized by

$$\gamma\in\left[\frac{1}{\max_{i\in[n]}\lambda_{\max}(\mathbf{A}_i)},\min\left\{\frac{\frac{\mu}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\mathbf{A}_i)},\frac{1}{\min_{i\in[n]}\lambda_{\min}^{+}(\mathbf{A}_i)}\right\}\right]$$

if $\frac{\mu}{\tau}\ge 2$ and by

$$\gamma\in\left[0,\max\left\{0,\min\left\{\frac{\frac{\mu}{\tau}-1}{\max_{i\in[n]}\lambda_{\max}(\mathbf{A}_i)},\frac{1}{\min_{i\in[n]}\lambda_{\min}^{+}(\mathbf{A}_i)}\right\}\right\}\right]$$

if $\frac{\mu}{\tau}<2$, and

$$T_\mu(\gamma)\le T_{\mathrm{GD}}.$$

Key insight:

- If $\mu\gg\tau$, optimal $\gamma>0$
- **FedExProx** is provably better than **GD**

References

[1] Hanmin Li, Kirill Acharya, and Peter Richtarik. The power of extrapolation in federated learning. *NeurIPS*, 2024.

[2] 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*, 2020.

Partial Participation

Fix any $\gamma>0$ and consider solving non-strongly convex quadratic optimization problem, where $f_i(x)=\frac{1}{2}x^\top\mathbf{A}_ix-b_i^\top x$, with $\mathbf{A}_i\in\mathrm{Sym}_+^d$ and $b_i\in\mathbb{R}^d$. Under A3, **FedExProx** with $\alpha=1/\gamma L_{\gamma,S}$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})]-f(x_*)\le\varepsilon$ after

$$\mathcal{O}\left(\frac{L_{\gamma,S}}{\mu_\gamma^{+}}\log\frac{1}{\varepsilon}\right)$$

iterations, where $L_{\gamma,S}:=\frac{n-S}{S(n-1)}\frac{L_{\max}}{1+\gamma L_{\max}}+\frac{n(S-1)}{S(n-1)}L_\gamma$, L_γ is the smoothness constant of M^{γ} and μ_γ^{+} is the smallest non-zero eigenvalue of the matrix $\nabla^2 M^{\gamma}$.

Adaptive Strategies

Fix any $\gamma>0$ and consider solving non-strongly convex quadratic optimization problem, where $f_i(x)=\frac{1}{2}x^\top\mathbf{A}_ix-b_i^\top x$ for all $i\in[n]$, with $\mathbf{A}_i\in\mathrm{Sym}_+^d$ and $b_i\in\mathbb{R}^d$. Let A3 hold and consider two adaptive extrapolation strategies:

1. (**FedExProx-GraDS**) Set

$$\alpha_k=\alpha_k^{\mathrm{GraDS}}(x_k):=\frac{\frac{1}{n}\sum_{i=1}^n\left\|\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}{\left\|\frac{1}{n}\sum_{i=1}^n\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}\ge 1.$$

   Then, the iterates of **FedExProx** satisfy

$$\|x_K-\Pi(x_K)\|^2\le\left(1-\min_{k=0,\ldots,K-1}\alpha_k\gamma\frac{2+\gamma L_{\max}}{1+\gamma L_{\max}}\mu_\gamma^{+}\right)^K\|x_0-\Pi(x_0)\|^2.$$

2. (**FedExProx-StoPS**) Set

$$\alpha_k=\alpha_k^{\mathrm{StoPS}}(x_k):=\frac{\frac{1}{n}\sum_{i=1}^n\left(M_{f_i}^{\gamma}(x_k)-\inf M_{f_i}^{\gamma}\right)}{\gamma\left\|\frac{1}{n}\sum_{i=1}^n\nabla M_{f_i}^{\gamma}(x_k)\right\|^2}\ge\frac{1}{2\gamma L_\gamma}.$$

   Then, the iterates of **FedExProx** satisfy

$$\|x_K-\Pi(x_K)\|^2\le\left(1-\frac{3}{2}\min_{k=0,\ldots,K-1}\alpha_k\gamma\mu_\gamma^{+}\right)^K\|x_0-\Pi(x_0)\|^2.$$

Beyond Quadratics (PŁ)

Let A1, A2, A3, and A5 hold. For all $\gamma>0$, **FedExProx** with $\alpha=1/\gamma L_\gamma$ finds $\bar{x}$ such that $\mathbb{E}[f(\bar{x})]-f(x_*)\le\varepsilon$ in

$$\mathcal{O}\left(\frac{L_\gamma}{\mu_\gamma^{+}}\log\frac{1}{\varepsilon}\right)$$

iterations, where L_γ is a smoothness constant of M^{γ} and μ_γ^{+} is the PŁ constant.

Experiments

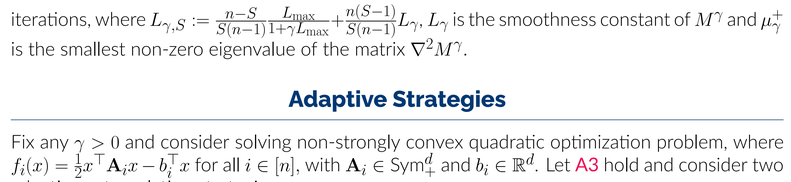

Figure 1. Empirical time complexities of FedExProx on a quadratic optimization task.

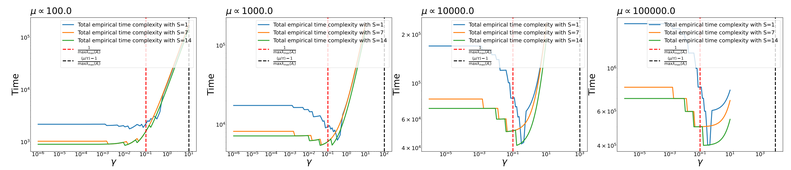

Figure 2. Empirical time complexities of FedExProx with partial client participation on a quadratic optimization task for $S\in\{1,7,14\}$ clients participating in each round.