FedProx: a Framework for Federated Optimization

**Global objective:** \( \min f(w) = \mathbb{E}_j [F_j(w)] \)

**Local objective on device:** \( \min_{W_i} F_i(W_i, X_i) \)

**Idea 1:** Allow for partial work to be performed on local devices based on systems constraints
- Generalization of the popular method FedAvg (FedAvg + allowing for variable local work + proximal term = FedProx)
- General: Can use any local solver; theory covers both convex and non-convex losses

**Idea 2:** At each round, each selected device solves a modified local subproblem:
- The proximal term incorporates noisy updates from variable local work; (2) explicitly limits the impact of local updates; (3) makes the method more amenable to theoretical analysis

**Proposed FedProx method**
- Until convergence:
  1. Server samples devices, and sends the current global model to all chosen devices
  2. Each device solves the following subproblem by performing variable local updates based on the underlying systems constraints
  3. Server aggregates local updates and forms a new global model

**Convergence Analysis**

**Characterize statistical heterogeneity:** B-dissimilarity

\[ B(w) = \sqrt{\mathbb{E}_j \left[ \left\| \frac{F_j(w)}{f_j(w)} \right\|_2^2 \right]} \]

**B quantifies statistical heterogeneity**

**Assumptions**

- **Assumption 1:** Bounded Dissimilarity
- **Assumption 2:** Modified Local subproblem is convex & smooth
- **Assumption 3:** Each local subproblem is solved inexactly to some optimality

**[Theorem]** Obtain suboptimality \( \epsilon \) after \( T \) iterations, with:

\[ T = \left( \frac{\mu (1 - \rho_k) - \mu^*}{\rho_k \mu^*} \right) \]

**Future Work**

- How to tune \( \mu \) automatically (hyper-parameter optimization for federated learning)?
- Can we quantify the statistical heterogeneity a priori and leverage it for improved performance?
- Better privacy metrics and mechanisms for federated learning?

**Evaluation**

**LEAF: A Benchmark for Learning in Federated Settings**

**Website:** leaf.cmu.edu

**Models:**
- FedAvg
- FedProx

**Effects of Idea 1 (partial work):**
- \( \mu > 0 \) leads to more stable convergence and enables otherwise divergent methods to converge

**Effects of Idea 2 (the proximal term):**
- \( \mu > 0 \) leads to worse convergence; Setting \( \mu = 0 \) can help to combat this