FeDis: Fast and Scalable Logistic Regression with Feature Distributed Stochastic Coordinate Descent - Theory Supplement

ABSTRACT

In this supplementary document, we give full theoretical proofs some of which are omitted from the main paper.

1. DETAILED PROOF

In this section, we give a theoretical convergence analysis of FeDis. Specifically, we prove that the output from FeDis converges to the solution of the logistic regression problem. Since the loss function is convex [2], it suffices to show that each iteration of FeDis decreases the loss function. Since FeDis randomly chooses coordinates, it is necessary to bound the expectation of the loss function where the expectation is over the random choices of the coordinates. Our main result is Theorem 1 which states that the expectation of the loss function decreases at each iteration when FeDis is run with a proper small step size. We first prove several lemmas, and use them to prove Theorem 1.

Without loss of generality, we assume that \( \text{diag}(\mathbf{X}^T \hat{\mathbf{X}}) = 1 \), following [1]. The Hessian of \( F(\hat{\theta}) \) is given by

\[
\frac{\partial^2 F(\hat{\theta})}{\partial \hat{\theta}_k \partial \hat{\theta}_k} = \sum_{i=1}^{n} \hat{X}_{ij} \hat{X}_{ik}(1 - \hat{p}_i) \hat{p}_i,
\]

where \( \hat{p}_i = 1/(1 + e^{-x_i \hat{\theta}}) \). Let \( \Delta \hat{\theta} \) be the change of \( \hat{\theta} \) at each iteration, and \( \Delta \hat{\theta}_j \) be the \( j \)-th coordinate of \( \Delta \hat{\theta} \) updated from machine \( k \). We first show the upper bound of \( F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta}) \).

**Lemma 1.** For any \( \hat{\theta}, F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta}) \leq (\Delta \hat{\theta})^T \nabla F(\hat{\theta}) + \frac{\beta}{2}(\Delta \hat{\theta})^T (\nabla^2 F(\theta')) \Delta \hat{\theta}, \) where \( \beta = \frac{1}{4} \) is a constant.

**Proof.** By Taylor’s theorem, there exists \( \theta' \) such that

\[
F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta}) = (\Delta \hat{\theta})^T \nabla F(\hat{\theta}) + \frac{1}{2} (\Delta \hat{\theta})^T (\nabla^2 F(\theta')) \Delta \hat{\theta}.
\]

Since \((1 - \hat{p}_i) \hat{p}_i \leq \frac{1}{4} = \beta \), it follows

\[
(\Delta \hat{\theta})^T (\nabla^2 F(\theta')) \Delta \hat{\theta} \leq \beta (\Delta \hat{\theta})^T \hat{X}^T \hat{X} \Delta \hat{\theta}.
\]

which proves the lemma. \( \square \)

Next, we give the relation between \( \nabla F(\hat{\theta}) \) and \( \nabla F_i(\hat{\theta}) \).

**Lemma 2.** \( \nabla F(\hat{\theta}) = \sum_{i=1}^{M} \nabla F_i(\hat{\theta}) \).

**Proof.**

\[
\frac{\partial F(\hat{\theta})}{\partial \hat{\theta}_j} = \sum_{i=1}^{n} (y_i \hat{X}_{ij}(\hat{p}_i - 1)) + \lambda = \sum_{i \in \mathcal{X}_k} (y_i \hat{X}_{ij}(\hat{p}_i - 1)) + M \cdot \frac{\lambda}{M} = \sum_{k=1}^{M} \frac{\partial F_i(\hat{\theta})}{\partial \hat{\theta}_j}.
\]

\( \square \)

Next, we give a loose bound of the difference between \( E[F(\hat{\theta} + \Delta \hat{\theta})] \) and \( E[F(\hat{\theta})] \).

**Lemma 3.** For \( M \geq 2 \), \( E[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \) is bounded by

\[
E[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \leq PE[\sum_{i=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta})_j] + \frac{\beta(1 + \epsilon)}{2} \sum_{i=1}^{M} (\Delta \hat{\theta}_j)^2,
\]

where \( \epsilon = \frac{\rho M - 1}{2d} \) and \( \rho \) is the spectral radius of \( \hat{X}^T \hat{X} \).

**Proof.** Let \( J \) be a set of sampled feature index of \( i \)-th iteration. Since we use random sampling without replacement, all elements of \( J \) are different from each other. Let \( E_{1k} [F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \) be the expected difference between \( F(\hat{\theta} + \Delta \hat{\theta}) \) and \( F(\hat{\theta}) \). Then by Lemma 1, the upper bound of \( E_{1k} [F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \) is given as follows:

\[
E_{1k} [F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \leq \sum_{j=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta})_j
\]

\[
+ E_{1k} [\frac{\beta}{2} \sum_{i,j \in J} \Delta \hat{\theta}_j (\hat{X}^T \hat{X})_{ij} \Delta \hat{\theta}_j]
\]

To separate the case of \( i = j \) from \( (i, j) \in J \), we re-express \( E_{1k} [\sum_{j=1}^{M} \Delta \hat{\theta}_j (\hat{X}^T \hat{X})_{ij} \Delta \hat{\theta}_j] \) as follows:

\[
E_{1k} [\frac{\beta}{2} \sum_{i \in J} \Delta \hat{\theta}_j (\hat{X}^T \hat{X})_{ij} \Delta \hat{\theta}_j] = E_{1k} [\frac{\beta}{2} \sum_{j \in J} \Delta \hat{\theta}_j^2]
\]

\[
+ E_{1k} [\sum_{i,j \in J \setminus \{i,j\}} \Delta \hat{\theta}_j (\hat{X}^T \hat{X})_{ij} \Delta \hat{\theta}_j]
\]

Now we compute the three parts in the upper bound of \( E_{1k} [F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \). The first term \( E_{1k} [\sum_{j=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta})_j] \) is given by

\[
E_{1k} [\sum_{j=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta})_j] = \frac{PM}{2d} \sum_{k=1}^{M} \left( \sum_{j=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta}_j) \right)
\]

\[
= \sum_{k=1}^{M} \left( \frac{P}{2d} \sum_{j=1}^{M} \Delta \hat{\theta}_j \nabla F(\hat{\theta})_j \right)
\]

\[
= \sum_{k=1}^{M} \frac{P}{2d} (\Delta \hat{\theta}_j)^2 \nabla F(\hat{\theta})_j)
\]
Here $\Delta_{y}^{j}$ is computed by $F_{k}$. Next we give $E_{k}[\frac{\beta}{2} \sum_{j \neq k} \Delta_{y}^{j}].$

$$E_{k}[\frac{\beta}{2} \sum_{j \neq k} \Delta_{y}^{j}] = \frac{\beta}{2} \frac{P M (PM - 1)}{2d(2d - 1)} \sum_{j=1}^{M} \left( \frac{1}{M} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{j} \right)$$

$$= \frac{\beta P}{2d} \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{j}.$$  

The third term $E_{k}[\sum_{l \neq j, k} \frac{\eta}{2} \Delta_{y}^{j} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l} ]$ is given by

$$E_{k}[\sum_{l \neq j, k} \frac{\eta}{2} \Delta_{y}^{j} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l}] = \frac{\beta}{2} \frac{P M (PM - 1)}{2d(2d - 1)} \sum_{l=1}^{M} \sum_{j=1}^{M} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l}.$$ 

Let $\sum_{k=1}^{M} \Delta_{y}^{j} = \sigma_{y}$. then we have the following equality:

$$\sum_{k=1}^{M} (\Delta_{y}^{j})^{T} \hat{X}_{y}^{T}X_{y} \Delta_{y}^{k} - (\sigma_{y})^{T} \hat{X}_{y}^{T}X_{y} \sigma_{y} = (\sigma_{y})^{T} \hat{X}_{y}^{T}X_{y} \sigma_{y} - (\sigma_{y})^{T} \sigma_{y}.$$ 

Summarizing the results up to now, we have the following:

$$E_{k}[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})]$$

$$\leq \sum_{k=1}^{M} \frac{P}{2d} \sum_{l=1}^{M} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l}$$

$$+ \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k} + \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k}$$

$$\leq \sum_{k=1}^{M} \frac{P}{2d} \sum_{l=1}^{M} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l}$$

$$+ \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k} + \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k}$$

$$= \sum_{k=1}^{M} \frac{P}{2d} \sum_{l=1}^{M} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l}$$

$$+ \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k}$$

Let $\epsilon = \frac{(PM - 1) \rho}{2d(2d - 1)}$. We finish the proof with the following:

$$E_{k}[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})]$$

$$\leq \sum_{k=1}^{M} \frac{P}{2d} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l} + \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k}$$

$$= PE_{k} \sum_{k=1}^{M} (\Delta_{y}^{j})^{T} (\hat{X}_{y}^{T}X_{y})_{l} \Delta_{y}^{l} + \beta \frac{P (PM - 1)}{2d(2d - 1)} \sum_{k=1}^{M} (\Delta_{y}^{k})^{T} \Delta_{y}^{k}$$

Let $\hat{\theta}$ be a diagonal matrix with $\hat{\theta}_{k, i} = \hat{\theta}_{j}$. And let $P_{i}(\hat{\theta})$ is a sub-matrix of $P(\hat{\theta})$ corresponding to $X_{k}$ and $y_{i}$ is a sub-vector of $y$ corresponding to $X_{k}$. The gradients of $F(\hat{\theta})$ and $F_{k}(\hat{\theta})$ are expressed by $P(\hat{\theta})$ and $P_{i}(\hat{\theta})$ as follows:

$$\frac{\partial F(\hat{\theta})}{\partial \hat{\theta}_{j}} = \langle \hat{X}_{y}^{T}(\hat{X}_{y}^{T} - I) y, \lambda \rangle$$

$$\frac{\partial F_{k}(\hat{\theta})}{\partial \hat{\theta}_{j}} = \langle \hat{X}_{y}^{i} (P_{i}(\hat{\theta}) - I), y_{i}, \lambda \rangle$$

We give an upper bound of $\sum_{k=1}^{M} (\nabla F_{k}(\hat{\theta}))^{2}$ in the following Lemma.

**Lemma 4.** For $\lambda \leq M$, $\sum_{k=1}^{M} (\nabla F_{k}(\hat{\theta}))^{2}$ has the following upper bound:

$$\sum_{k=1}^{M} (\nabla F_{k}(\hat{\theta}))^{2} \leq 2 \left( \| P(\hat{\theta}) - I \|_{y} y_{k} + \frac{\lambda}{M} \right)^{2}$$

**Proof.** We begin the proof with the matrix notation.

$$\nabla F_{k}(\hat{\theta}) = \left( \frac{\partial F_{k}(\hat{\theta})}{\partial \hat{\theta}_{j}} \right)^{2} = \left( \langle \hat{X}_{y}^{i} (P_{i}(\hat{\theta}) - I), y_{i}, \lambda \rangle \right)^{2}$$

Then, we have the following inequality:

$$\nabla F_{k}(\hat{\theta}) = \left( \langle \hat{X}_{y}^{i} (P_{i}(\hat{\theta}) - I), y_{i}, \lambda \rangle \right)^{2} \leq 2 \left( \| P_{i}(\hat{\theta}) - I \|_{y} y_{i} \right)^{2} + \frac{\lambda^{2}}{M^{2}}$$

$$\leq 2 \left( \| P_{i}(\hat{\theta}) - I \|_{y} y_{i} \right)^{2} + \frac{\lambda^{2}}{M^{2}}$$

Consequently, we get the following inequality:

$$\sum_{k=1}^{M} (\nabla F_{k}(\hat{\theta}))^{2} \leq \sum_{k=1}^{M} \left( \| P_{i}(\hat{\theta}) - I \|_{y} y_{i} \right)^{2} + \frac{\lambda^{2}}{M^{2}}$$

$$\leq 2 \left( \| P(\hat{\theta}) - I \|_{y} y_{i} \right)^{2} + \frac{\lambda^{2}}{M^{2}}$$

$$\leq 2 \left( \| P(\hat{\theta}) - I \|_{y} y_{i} \right)^{2} + \frac{\lambda^{2}}{M^{2}}$$

Now we provide the main theorem which proves that FedIS converges.

**Theorem 1.** In FedIS, for any iteration, feature $j$ and $\theta$, there exists a step size $\eta$ such that the expectation of the loss function of FedIS decreases: i.e.,

$$E[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] < 0$$
Proof: In FrDts, the $j$th coordinate of $\Delta \hat{\theta}$ updated from machine $k$ is given by $\Delta^k_{j} = \eta \cdot \text{max}(\hat{\theta}_j - \nabla F_k(\hat{\theta}))/\beta$. Since $\hat{\theta}_j$ is non-negative, there exists a non-negative constant $c$ satisfying $\text{max}(\hat{\theta}_j - \nabla F_k(\hat{\theta}))/\beta = -c \nabla F_k(\hat{\theta})/\beta$. Thus, $\Delta^k_{j} = -c \eta \nabla F_k(\hat{\theta})/\beta$.

Inserting $\Delta^k_{j}$ to the upper bound of $E[F(\hat{\theta} + \Delta \theta) - F(\hat{\theta})]$ in Lemma 3, we have the following inequality:

$$E[F(\hat{\theta} + \Delta \theta) - F(\hat{\theta})]$$

$$\leq PE[I \sum_{k=1}^{M} \Delta^k_{j} \nabla F_k(\hat{\theta})] + \frac{\beta (1 + \epsilon)}{2} \sum_{k=1}^{M} (\Delta^k_{j})^2$$

$$\leq \frac{P c \eta}{\beta} E[I \left( - \sum_{k=1}^{M} \nabla F_k(\hat{\theta}) \right) \nabla F(\hat{\theta}) + \frac{c \eta (1 + \epsilon)}{2} \sum_{k=1}^{M} (\nabla F_k(\hat{\theta}))^2]$$

$$= \frac{P c \eta}{\beta} E[I - (\nabla F(\hat{\theta}))^2 + \frac{c \eta (1 + \epsilon)}{2} \sum_{k=1}^{M} (\nabla F_k(\hat{\theta}))^2]$$

Let $\eta$ be a step size satisfying

$$\eta < \frac{(\nabla F(\hat{\theta}))^2}{c(1 + \epsilon) \left( \|P(\hat{\theta}) - I\|_2^2 + \lambda \right)}.$$ (1)

We want to show that the $\eta$ satisfies $-(\nabla F(\hat{\theta}))^2 + \frac{c \eta (1 + \epsilon)}{2} \sum_{k=1}^{M} (\nabla F_k(\hat{\theta}))^2 < 0$. By Lemma 4,

$$-(\nabla F(\hat{\theta}))^2 + \frac{c \eta (1 + \epsilon)}{2} \sum_{k=1}^{M} (\nabla F_k(\hat{\theta}))^2$$

$$\leq -(\nabla F(\hat{\theta}))^2 + c \eta (1 + \epsilon) \left( \|P(\hat{\theta}) - I\|_2^2 + \lambda \right)$$

Since $(\nabla F(\hat{\theta}))^2$ and $\left( \|P(\hat{\theta}) - I\|_2^2 + \lambda \right)$ are non-negative, we have following relation:

$$\frac{(\nabla F(\hat{\theta}))^2}{c(1 + \epsilon) \left( \|P(\hat{\theta}) - I\|_2^2 + \lambda \right)} > \eta$$

$$\iff -(\nabla F(\hat{\theta}))^2 + c \eta (1 + \epsilon) \left( \|P(\hat{\theta}) - I\|_2^2 + \lambda \right) < 0$$

which finishes the proof. □

2. REFERENCES