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 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 $diag(\hat{\mathbf{X}}^T\hat{\mathbf{X}}) = \mathbf{1}$, following [1]. The Hessian of $F(\hat{\boldsymbol{\theta}})$ is given by

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

where $\hat{p}_i = 1/(1+ext(-y_i\hat{\mathbf{x}}_i^T\hat{\boldsymbol{\theta}}))$. Let $\Delta\hat{\boldsymbol{\theta}}$ be the change of $\hat{\boldsymbol{\theta}}$ at each iteration, and $\Delta^k_{\hat{\theta}_j}$ be the jth coordinate of $\Delta\hat{\boldsymbol{\theta}}$ updated from machine k. We first show the upper bound of $F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})$.

Lemma 1. *For any* $\hat{\mathbf{X}}$, $F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}}) \le (\Delta\hat{\boldsymbol{\theta}})^T\nabla F(\hat{\boldsymbol{\theta}}) + \frac{\beta}{2}(\Delta\hat{\boldsymbol{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\Delta\hat{\boldsymbol{\theta}}$, *where* $\beta = \frac{1}{4}$ *is a constant.*

Proof. By Taylor's theorem, there exists $\boldsymbol{\theta}'$ such that

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

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

$$(\Delta\hat{\boldsymbol{\theta}})^T(\nabla^2 F(\boldsymbol{\theta}'))\Delta\hat{\boldsymbol{\theta}} \le \beta(\Delta\hat{\boldsymbol{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\Delta\hat{\boldsymbol{\theta}}.$$

which proves the lemma. □

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

Lemma 2. $\nabla F(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^{M}\nabla F_k(\hat{\boldsymbol{\theta}})$.

.

Proof.

$$\begin{aligned}\frac{\partial F(\hat{\boldsymbol{\theta}})}{\partial\hat{\theta}_j} &= \sum_{i=1}^{n}(y_i\hat{X}_{ij}(\hat{p}_i-1)) + \lambda \\ &= \sum_{i\in\cup\hat{\mathbf{X}}_k}^{n}(y_i\hat{X}_{ij}(\hat{p}_i-1)) + M\cdot\frac{\lambda}{M} \\ &= \sum_{k=1}^{M}\frac{\partial F_k(\hat{\boldsymbol{\theta}})}{\partial\hat{\theta}_j}.\end{aligned}$$

□

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

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

$$E[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})] \le PE_j[\sum_{k=1}^{M}\Delta^k_{\hat{\theta}_j}\nabla F(\hat{\boldsymbol{\theta}})_j + \frac{\beta(1+\epsilon)}{2}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}_j})^2],$$

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

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

$$\begin{aligned}E_{\mathbf{J}_t}\left[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})\right] \le& E_{\mathbf{J}_t}[\sum_{j\in\mathbf{J}_t}\Delta_{\hat{\theta}_j}\nabla F(\hat{\boldsymbol{\theta}})_j] \\ &+ E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{i,j\in\mathbf{J}_t}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}]\end{aligned}$$

To separate the case of $i = j$ from $\{i, j \in \mathbf{J}_t\}$, we re-express $E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{i,j\in\mathbf{J}_t}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}]$ as follows:

$$\begin{aligned}E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{i,j\in\mathbf{J}_t}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}] =& E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{j\in\mathbf{J}_t}\Delta^2_{\hat{\theta}_j}] \\ &+ E_{\mathbf{J}_t}[\sum_{i,j\in\mathbf{J}_t,i\neq j}\frac{\beta}{2}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}]\end{aligned}$$

Now we compute the three parts in the upper bound of $E_{\mathbf{J}_t}[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]$. The first term $E_{\mathbf{J}_t}[\sum_{j\in\mathbf{J}_t}\Delta_{\hat{\theta}_j}\nabla F(\hat{\boldsymbol{\theta}})_j]$ is given by

$$\begin{aligned}E_{\mathbf{J}_t}[\sum_{j\in\mathbf{J}_t}\Delta_{\hat{\theta}_j}\nabla F(\hat{\boldsymbol{\theta}})_j] &= \frac{PM}{2d}\sum_{j=1}^{2d}\left(\frac{1}{M}\sum_{k=1}^{M}\Delta^k_{\hat{\theta}_j}(\nabla F(\hat{\boldsymbol{\theta}})_j)\right) \\ &= \sum_{k=1}^{M}\left(\frac{P}{2d}\sum_{j=1}^{2d}\Delta^k_{\hat{\theta}_j}(\nabla F(\hat{\boldsymbol{\theta}})_j)\right) \\ &= \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}})).\end{aligned}$$

Here $\Delta^k_{\hat{\theta}}$ is computed by F_k. Next we give $E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{j\in\mathbf{J}_t}\Delta^2_{\hat{\theta}_j}]$.

$$E_{\mathbf{J}_t}[\frac{\beta}{2}\sum_{j\in\mathbf{J}_t}\Delta^2_{\hat{\theta}_j}] = \frac{\beta}{2}\frac{PM}{2d}\sum_{j=1}^{2d}\left(\frac{1}{M}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}_j})^2\right)$$
$$= \frac{\beta}{2}\frac{P}{2d}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}.$$

The third term $E_{\mathbf{J}_t}[\sum_{i,j\in\mathbf{J}_t,i\neq j}\frac{\beta}{2}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}]$ is given by

$$\begin{aligned}
&E_{\mathbf{J}_t}[\sum_{i,j\in\mathbf{J}_t,i\neq j}\frac{\beta}{2}\Delta_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta_{\hat{\theta}_j}]\\
&= \frac{\beta}{2}\frac{PM(PM-1)}{2d(2d-1)}\sum_{i,j=1,i\neq j}^{2d}\left(\frac{1}{M^2}\sum_{k,\ell=1}^{M}\Delta^k_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta^\ell_{\hat{\theta}_j}\right)\\
&= \frac{\beta}{2}\frac{P(PM-1)}{2d(2d-1)M}\sum_{k,\ell=1}^{M}\sum_{i,j=1,i\neq j}^{2d}\left(\Delta^k_{\hat{\theta}_i}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})_{i,j}\Delta^\ell_{\hat{\theta}_j}\right)\\
&= \frac{\beta}{2}\frac{P(PM-1)}{2d(2d-1)M}\sum_{k,\ell=1}^{M}\left((\Delta^k_{\hat{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\Delta^\ell_{\hat{\theta}} - (\Delta^k_{\hat{\theta}})^T\Delta^\ell_{\hat{\theta}}\right).
\end{aligned}$$

Let $\sum_{k=1}^{M}\Delta^k_{\hat{\theta}} = \sigma_{\hat{\theta}}$. then we have the following equality:

$$\sum_{k,\ell=1}^{M}\left((\Delta^k_{\hat{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\Delta^\ell_{\hat{\theta}} - (\Delta^k_{\hat{\theta}})^T\Delta^\ell_{\hat{\theta}}\right) = (\sigma_{\hat{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\sigma_{\hat{\theta}} - (\sigma_{\hat{\theta}})^T\sigma_{\hat{\theta}}.$$

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

$$\begin{aligned}
&E_{\mathbf{J}_t}[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]\\
&\le \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}})) + \frac{\beta}{2}\frac{P}{2d}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\\
&\quad + \frac{\beta}{2}\frac{P(PM-1)}{2d(2d-1)M}\left((\sigma_{\hat{\theta}})^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\sigma_{\hat{\theta}} - (\sigma_{\hat{\theta}})^T\sigma_{\hat{\theta}}\right).
\end{aligned}$$

Using the spectral radius ρ of $\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ with $\rho \ge 1$, we have the following inequalities:

$$\begin{aligned}
&E_{\mathbf{J}_t}[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]\\
&\le \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}})) + \frac{\beta}{2}\frac{P}{2d}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\\
&\quad + \frac{\beta}{2}\frac{P(PM-1)}{2d(2d-1)M}\left((\rho-1)(\sigma_{\hat{\theta}})^T\sigma_{\hat{\theta}}\right)\\
&\le \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}})) + \frac{\beta}{2}\frac{P}{2d}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\\
&\quad + \frac{\beta}{2}\frac{P(PM-1)}{2d(2d-1)}\frac{(\rho-1)}{M}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\\
&= \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}}))\\
&\quad + \frac{\beta}{2}\frac{P}{2d}\left(1 + \frac{(PM-1)(\rho-1)}{(2d-1)M}\right)\left(\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\right).
\end{aligned}$$

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

$$\begin{aligned}
&E_{\mathbf{J}_t}[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]\\
&\le \sum_{k=1}^{M}\frac{P}{2d}(\Delta^k_{\hat{\theta}})^T(\nabla F(\hat{\boldsymbol{\theta}})) + \frac{\beta}{2}\frac{P}{2d}(1+\epsilon)\left(\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}})^T\Delta^k_{\hat{\theta}}\right)\\
&= PE_j[\sum_{k=1}^{M}\Delta^k_{\hat{\theta}_j}\nabla F(\hat{\boldsymbol{\theta}})_j + \frac{\beta(1+\epsilon)}{2}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}_j})^2]
\end{aligned}$$

□

Let $\mathbf{P}(\hat{\boldsymbol{\theta}})$ be a diagonal matrix with $\mathbf{P}(\hat{\boldsymbol{\theta}})_{i,i} = \hat{p}_i$. And let $\mathbf{P}_k(\hat{\boldsymbol{\theta}})$ is a sub-matrix of $\mathbf{P}(\hat{\boldsymbol{\theta}})$ corresponding to $\hat{\mathbf{X}}_k$ and $\mathbf{y}_k$ is a sub-vector of $\mathbf{y}$ corresponding to $\hat{\mathbf{X}}_k$. The gradients of $F(\hat{\boldsymbol{\theta}})$ and $F_k(\hat{\boldsymbol{\theta}})$ are expressed by $\mathbf{P}(\hat{\boldsymbol{\theta}})$ and $\mathbf{P}_k(\hat{\boldsymbol{\theta}})$ as follows:

$$\begin{aligned}
\frac{\partial F(\hat{\boldsymbol{\theta}})}{\partial\hat{\theta}_j} &= (\hat{\mathbf{X}}^T(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y} + \lambda\mathbf{1})_j\\
\frac{\partial F_k(\hat{\boldsymbol{\theta}})}{\partial\hat{\theta}_j} &= (\hat{\mathbf{X}}_k^T(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k + \frac{\lambda}{M}\mathbf{1})_j\\
&= \left((\hat{\mathbf{X}}_k^T)_j(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k + \frac{\lambda}{M}\right)
\end{aligned}$$

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

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

$$\sum_{k=1}^{M}(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2 \le 2\left(\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \lambda\right).$$

Proof. We begin the proof with the matrix notation.

$$(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2 = \left(\frac{\partial F_k(\hat{\boldsymbol{\theta}})}{\partial\hat{\theta}_j}\right)^2 = \left((\hat{\mathbf{X}}_k^T)_j(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k + \frac{\lambda}{M}\right)^2$$

Then, we have the following inequality:

$$\begin{aligned}
(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2 &= \left((\hat{\mathbf{X}}_k^T)_j(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k + \frac{\lambda}{M}\right)^2\\
&\le 2\left((\hat{\mathbf{X}}_k^T)_j(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k\right)^2 + \frac{2\lambda^2}{M^2}\\
&\le 2\|(\hat{\mathbf{X}}_k^T)_j\|_2^2\|(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k\|_2^2 + \frac{2\lambda^2}{M^2}\\
&\le 2\|(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k\|_2^2 + \frac{2\lambda^2}{M^2}
\end{aligned}$$

Consequently, we get the following inequality:

$$\begin{aligned}
\sum_{k=1}^{M}(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2 &\le \sum_{k=1}^{M}\left(2\|(\mathbf{P}_k(\hat{\boldsymbol{\theta}}) - \mathbf{I}_k)\mathbf{y}_k\|_2^2 + \frac{2\lambda^2}{M^2}\right)\\
&= \left(2\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \frac{2\lambda^2}{M}\right)\\
&\le 2\left(\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \lambda\right)
\end{aligned}$$

□

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

Theorem 1. *In FeDis, for any iteration, feature j and $\hat{\boldsymbol{\theta}}$, there exists a step size η such that the expectation of the loss function of FeDis decreases: i.e.,*

$$E[F(\hat{\boldsymbol{\theta}}+\Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})] < 0$$

PROOF. In FEDIS, the jth coordinate of $\Delta\hat{\boldsymbol{\theta}}$ updated from machine k is given by $\Delta^k_{\hat{\theta}_j} = \eta \cdot max\{-\hat{\theta}_j, -(\nabla F_k(\hat{\boldsymbol{\theta}}))_j/\beta\}$. Since $\hat{\theta}_j$ is non-negative, there exists a non-negative constant c satisfying $\max\{-\hat{\theta}_j, -\nabla F_k(\hat{\boldsymbol{\theta}})_j/\beta\} = -c\nabla F_k(\hat{\boldsymbol{\theta}})_j/\beta$. Thus, $\Delta^k_{\hat{\theta}_j} = -c\eta(\nabla F_k(\hat{\boldsymbol{\theta}}))_j/\beta$. Inserting $\Delta^k_{\hat{\theta}_j}$ to the upper bound of $E[F(\hat{\boldsymbol{\theta}} + \Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]$ in Lemma 3, we have the following inequality:

$$
\begin{aligned}
&E[F(\hat{\boldsymbol{\theta}} + \Delta\hat{\boldsymbol{\theta}}) - F(\hat{\boldsymbol{\theta}})]\\
&\le PE_j[\sum_{k=1}^{M} \Delta^k_{\hat{\theta}_j} \nabla F(\hat{\boldsymbol{\theta}})_j + \frac{\beta(1+\epsilon)}{2}\sum_{k=1}^{M}(\Delta^k_{\hat{\theta}_j})^2]\\
&\le \frac{Pc\eta}{\beta}E_j\left[-\left(\sum_{k=1}^{M}\nabla F_k(\hat{\boldsymbol{\theta}})_j\right)\nabla F(\hat{\boldsymbol{\theta}})_j + \frac{c\eta(1+\epsilon)}{2}\sum_{k=1}^{M}(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2\right]\\
&= \frac{Pc\eta}{\beta}E_j\left[-(\nabla F(\hat{\boldsymbol{\theta}})_j)^2 + \frac{c\eta(1+\epsilon)}{2}\sum_{k=1}^{M}(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2\right]
\end{aligned}
$$

Let η be a step size satisfying

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

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

$$
\begin{aligned}
&-(\nabla F(\hat{\boldsymbol{\theta}})_j)^2 + \frac{c\eta(1+\epsilon)}{2}\sum_{k=1}^{M}(\nabla F_k(\hat{\boldsymbol{\theta}})_j)^2\\
&\qquad\le -(\nabla F(\hat{\boldsymbol{\theta}})_j)^2 + c\eta(1+\epsilon)\left(\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \lambda\right)
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{(\nabla F(\hat{\boldsymbol{\theta}})_j)^2}{c(1+\epsilon)\left(\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \lambda\right)} &> \eta\\
\iff -(\nabla F(\hat{\boldsymbol{\theta}})_j)^2 + c\eta(1+\epsilon)\left(\|(\mathbf{P}(\hat{\boldsymbol{\theta}}) - \mathbf{I})\mathbf{y}\|_2^2 + \lambda\right) &< 0
\end{aligned}
$$

which finishes the proof. □

2. REFERENCES

[1] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin. Parallel coordinate descent for l1-regularized loss minimization. In *ICML*, 2011.

[2] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng. Efficient l˜ 1 regularized logistic regression. In *AAAI*, 2006.