

Data/Feature Distributed Stochastic Coordinate Descent for Logistic Regression - 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 DF-DSCD. Specifically, we prove that the output from DF-DSCD converges to the solution of the logistic regression problem. Since the loss function is convex [2], it suffices to show that each iteration of DF-DSCD decreases the loss function. Since DF-DSCD 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 DF-DSCD 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 $\text{diag}(\hat{\mathbf{X}}^T \hat{\mathbf{X}}) = \mathbf{1}$, following [1]. The Hessian of $F(\hat{\theta})$ is given by

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

where $\hat{p}_i = 1/(1 + \text{ext}(-y_i \hat{\mathbf{x}}_i^T \hat{\theta}))$. Let $\Delta \hat{\theta}$ be the change of $\hat{\theta}$ at each iteration, and $\Delta_{\hat{\theta}_j}^k$ 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 $\hat{\mathbf{X}}$, $F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta}) \leq (\Delta \hat{\theta})^T \nabla F(\hat{\theta}) + \frac{\beta}{2} (\Delta \hat{\theta})^T \hat{\mathbf{X}}^T \hat{\mathbf{X}} \Delta \hat{\theta}$, where $\beta = \frac{1}{4}$ is a constant.

PROOF. By Taylor's theorem, there exists θ' 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

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

which proves the lemma. \square

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

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

PROOF.

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

\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

$$E[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})] \leq P E_j \left[\sum_{k=1}^M \Delta_{\hat{\theta}_j}^k \nabla F(\hat{\theta})_j + \frac{\beta(1 + \epsilon)}{2} \sum_{k=1}^M (\Delta_{\hat{\theta}_j}^k)^2 \right],$$

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 $E_{\mathbf{J}_t}[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})]$ be the expected difference between $F(\hat{\theta} + \Delta \hat{\theta})$ and $F(\hat{\theta})$. Then by Lemma 1, the upper bound of $E_{\mathbf{J}_t}[F(\hat{\theta} + \Delta \hat{\theta}) - F(\hat{\theta})]$ is given as follows:

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

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:

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

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

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

Here Δ_θ^k is computed by F_k . Next we give $E_{\mathbf{J}_j}[\frac{\beta}{2} \sum_{j \in \mathbf{J}_j} \Delta_\theta^2]$.

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

PROOF. We begin the proof with the matrix notation.

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

Then, we have the following inequality:

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

Consequently, we get the following inequality:

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

□

Now we provide the main theorem which proves that DF-DSCD converges.

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

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

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

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

Let η be a step size satisfying

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

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

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

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

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

which finishes the proof. \square

2. REFERENCES

- [1] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin. Parallel coordinate descent for ℓ_1 -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.