Anonymous Author(s)
Affiliation
Address
email

1 Introduction

In the federated learning and decentralized learning, n participants collaborate to train a global model $\boldsymbol{x}$ over their joint objectives $\min_{\boldsymbol{x}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$. Compared to models trained on individual data silos, this model achieves overall better performance on dataset. There is no guarantee that it performs better than standalone training on some workers. Personalized federated learning is one way to address this problem.

2 Shared setting

(A1) L-smoothness. *For $i \in [n]$, f_i is L-smooth.*

(A2) Lower bound. *For $i \in [n]$, f_i is lower bounded by $f_i^\star$.*

3 Dynamic graph and full gradient

3.1 Problem formulation

In this section, we assume not all workers share same stationary points or minimizers.

(A3) Strong growth condition. *Let $c \subset [n]$ be the a subset of workers that share same stationary point. Then for $\boldsymbol{x} \in \mathbb{R}^d$ and $i \in c$, we have*

$$\left\|\nabla f_i(\boldsymbol{x}) - \nabla \bar{f}_c(\boldsymbol{x})\right\|_2 \leq M\left\|\nabla \bar{f}_c(\boldsymbol{x})\right\|_2.$$

The (A3) indicates that when an iterate $\boldsymbol{x}$ reaches the stationary point of $\bar{f}_c$, then it also reaches the stationary point of f for all $i \in c$. The optimization objective is that

$$\min_{X\in\mathbb{R}^{d\times n}} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i) + \frac{\rho}{2}\sum_{i<j} w_{ij}\|x_i - x_j\|_2^2.$$

We optimize the objective with gradient descent with initialization $\boldsymbol{x}_i^0 = \bar{\boldsymbol{x}}^0$

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t - \eta\left(\nabla f_i(\boldsymbol{x}_i^t) + \rho\sum_{k=1}^{n} w_{ik}^t(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right). \tag{1}$$

We update w_{ij}^t with the following term

$$w_{ij}^{t+1} = \text{sign}\left(\alpha - \|\boldsymbol{x}_i^t - \boldsymbol{x}_j^t\|_2^2\right)$$

3.2 Proof Sketch

Notations. We define the following notations

- Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ be the compact form of iterates.

- Let $\nabla F(\boldsymbol{X}) := [\nabla f_1(\boldsymbol{x}_1), \nabla f_2(\boldsymbol{x}_2), \ldots, \nabla f_n(\boldsymbol{x}_n)]^\top \in \mathbb{R}^{n\times d}$.
- Let $\boldsymbol{W}^t$ be the mixing matrix at time t and $\boldsymbol{D}^t := \text{Diag}(\boldsymbol{W}^t\boldsymbol{1})$.
- Let $\boldsymbol{W}^\star$ be the groundtruth mixing matrix.
- Let $d_{\text{max}} = \max\{D^\star\}$ be the size of largest cluster.

Then we update the iterates as follows

$$\boldsymbol{X}^{t+1} = (\boldsymbol{I} - \eta\rho\left(\boldsymbol{D}^t - \boldsymbol{W}^t\right))\boldsymbol{X}^t - \eta\nabla F(\boldsymbol{X}^t). \tag{2}$$

Lemma 1 (Basic properties)**.** *The following equality and inequalities hold true*

- $(\boldsymbol{D}^\star - \boldsymbol{W}^\star)(\boldsymbol{D}^\star - \boldsymbol{W}^\star) = \boldsymbol{D}^\star(\boldsymbol{D}^\star - \boldsymbol{W}^\star) = (\boldsymbol{D}^\star - \boldsymbol{W}^\star)\boldsymbol{D}^\star$
- $\|AB\|_F \leq \|A\|_2 \|B\|_F$

Note that (1) can be re-written as

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t - \eta\left(\nabla f_i(\boldsymbol{x}_i^t) + \rho\sum_{k=1}^n w_{ik}^\star(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t) + \rho\sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right).$$

After averaging i over the same cluster, we have

$$\bar{\boldsymbol{x}}_c^{t+1} = \bar{\boldsymbol{x}}_c^t - \eta\left(\frac{1}{c}\sum_{i\in c}\nabla f_i(\boldsymbol{x}_i^t) + \frac{\rho}{c}\sum_{i\in c}\sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right). \tag{3}$$

Lemma 2 (Sufficient decrease)**.** *Suppose f_i are L-smooth, then by taking $\eta \leq \frac{1}{L}$, we have*

$$\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq \bar{f}_c\left(\bar{\boldsymbol{x}}_c^t\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^t\right)\right\|_2^2 + \frac{\eta L^2}{c}\sum_{i\in c}\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}_c^t\right\|_2^2 + \eta\left\|\frac{\rho}{c}\sum_{i\in c}\sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2.$$

Here are related equality

- $\sum_c\sum_{i\in c}\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}_c^t\|_2^2 = \left\|(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^t\right\|_F^2$
- $\sum_c c\left\|\frac{1}{c}\sum_{i\in c}\sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 = \left\|\boldsymbol{D}^{-1}\boldsymbol{W}^\star(\boldsymbol{D}^\star - \boldsymbol{W}^\star - \boldsymbol{D}^t + \boldsymbol{W}^t)\boldsymbol{X}^t\right\|_F^2$
- $\sum_c\sum_{i\in c}\|\sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\|_2^2 = \|(\boldsymbol{D}^\star - \boldsymbol{W}^\star - \boldsymbol{D}^t + \boldsymbol{W}^t)\boldsymbol{X}^t\|_F^2$

Let $\mathcal{E}_i := \sum_{k=1}^n (w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)$ be the Misclassification term of worker i.

Lemma 3 (Misclassification error)**.** *Let $\mathcal{E} := \|(\boldsymbol{D}^\star - \boldsymbol{W}^\star - \boldsymbol{D}^t + \boldsymbol{W}^t)\boldsymbol{X}^t\|_F^2$ be the error incurred by misclassification*

$$\mathcal{E} \leq 2\mathcal{E}_{ex} + 2\mathcal{E}_{in}.$$

where $\mathcal{E}_{ex} = \sum_c\sum_{i\in c} c\sum_{k\in c \,\&\, w_{ik}^t=0}\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\|_2^2$ and $\mathcal{E}_{in} = \alpha^t n(n-c)^2$.

Note that as we initialize all models to be the same, the $\mathcal{E}_{\text{ex}}$ is small in the beginning; as we choose $\alpha^t = \mathcal{O}(\frac{1}{t})$, the $\mathcal{E}_{\text{in}}$ is gradually decreasing. The inclusion error eventually vanishes due to heterogeneity across clusters, in which case we can stop decreasing α^t. The exclusion error is bounded by generalized strong-growth condition.

Lemma 4.

Proof.

$$\begin{aligned}\left\|\boldsymbol{x}_i^{t+1} - \boldsymbol{x}_k^{t+1}\right\|_2^2 =& \left\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t - \eta(\nabla f_i(\boldsymbol{x}_i^t) - \nabla f_k(\boldsymbol{x}_k^t)) - \eta\rho c(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t) - \eta\rho(\mathcal{E}_i - \mathcal{E}_k)\right\|_2^2 \\ \leq& (1-\eta\rho c)\left\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\right\|_2^2 + \eta\rho c\left\|\frac{1}{\rho c}\left(\nabla f_i(\boldsymbol{x}_i^t) - \nabla f_k(\boldsymbol{x}_k^t)\right) + \frac{1}{c}(\mathcal{E}_i - \mathcal{E}_k)\right\|_2^2.\end{aligned}$$

Note that

$$\begin{aligned}\left\|\nabla f_i(\boldsymbol{x}_i^t)-\nabla f_k(\boldsymbol{x}_k^t)\right\|_2^2 \leq&\left\|\nabla f_i(\boldsymbol{x}_i^t)\pm\nabla f_i(\bar{\boldsymbol{x}}_c^t)\pm\nabla \bar{f}_c(\bar{\boldsymbol{x}}_c^t)\pm\nabla f_k(\bar{\boldsymbol{x}}_c^t)-\nabla f_k(\boldsymbol{x}_k^t)\right\|_2^2\\ \leq&4L^2\left\|\boldsymbol{x}_i^t-\bar{\boldsymbol{x}}_c^t\right\|_2^2+4L^2\left\|\boldsymbol{x}_k^t-\bar{\boldsymbol{x}}_c^t\right\|_2^2+8M^2\|\nabla\bar{f}_c(\bar{\boldsymbol{x}}_c^t)\|_2^2.\end{aligned}$$

Then

$$\frac{1}{c^2}\sum_{i\in c}\sum_{k\in c}\left\|\nabla f_i(\boldsymbol{x}_i^t)-\nabla f_k(\boldsymbol{x}_k^t)\right\|_2^2\leq 8L^2\frac{1}{c}\sum_{k\in c}\left\|\boldsymbol{x}_k^t-\bar{\boldsymbol{x}}_c^t\right\|_2^2+8M^2\|\nabla\bar{f}_c(\bar{\boldsymbol{x}}_c^t)\|_2^2.$$

Then

$$\begin{aligned}\frac{1}{c^2}\sum_{i\in c}\sum_{k\in c}\left\|\boldsymbol{x}_i^{t+1}-\boldsymbol{x}_k^{t+1}\right\|_2^2\leq&(1-\eta\rho c)\frac{1}{c^2}\sum_{i\in c}\sum_{k\in c}\left\|\boldsymbol{x}_i^t-\boldsymbol{x}_k^t\right\|_2^2\\ &+2\frac{\eta}{\rho c}\left(8L^2\frac{1}{c}\sum_{k\in c}\left\|\boldsymbol{x}_k^t-\bar{\boldsymbol{x}}_c^t\right\|_2^2+8M^2\|\nabla\bar{f}_c(\bar{\boldsymbol{x}}_c^t)\|_2^2\right)\\ &+2\frac{\eta\rho}{c}\frac{1}{c^2}\sum_{i\in c}\sum_{k\in c}\|\mathcal{E}_i-\mathcal{E}_k\|_2^2.\end{aligned}$$

□

Lemma 5 (Consensus distance)**.** *The consensus distance measures the worker model's distance to the center of its own cluster, i.e.* $\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}\right\|_F^2$ *in the compact form.*

Proof. The distance to their own center is

$$\begin{aligned}(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\boldsymbol{X}^{t+1}=&(\boldsymbol{D}^\star-\boldsymbol{W}^\star)(\boldsymbol{I}-\eta\rho\left(\boldsymbol{D}^t-\boldsymbol{W}^t\right))\boldsymbol{X}^t-\eta(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\\ =&(\boldsymbol{I}-\eta\rho\left(\boldsymbol{D}^\star-\boldsymbol{W}^\star\right))(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\boldsymbol{X}^t\\ &+\eta\rho(\boldsymbol{D}^\star-\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\\ &-\eta(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\\ =&(\boldsymbol{I}-\eta\rho\boldsymbol{D}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\boldsymbol{X}^t\\ &+\eta\rho(\boldsymbol{D}^\star-\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\\ &-\eta(\boldsymbol{D}^\star-\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t).\end{aligned}$$

Multiply with $\boldsymbol{D}^{-1}$ to both sides yield

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^{t+1}=&(\boldsymbol{I}-\eta\rho\boldsymbol{D}^\star)(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^t\\ &+\eta\rho(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\\ &-\eta(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t).\end{aligned}$$

Applying Frobenius norm to the above equation, we have

$$\begin{aligned}&\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^{t+1}\right\|_F^2\\ =&3\left\|(\boldsymbol{I}-\eta\rho\boldsymbol{D}^\star)(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^t\right\|_F^2+3\left\|\eta\rho(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\right\|_F^2\\ &+3\left\|\eta(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\right\|_F^2\\ =&3\|I-\eta\rho\boldsymbol{D}^\star\|_2^2\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^t\right\|_F^2+3\left\|\eta\rho(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\right\|_F^2\\ &+3\left\|\eta(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\right\|_F^2\\ \leq&(1-\eta\rho d_{\max})\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\boldsymbol{X}^t\right\|_F^2+3\left\|\eta\rho(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\right\|_F^2\\ &+3\left\|\eta(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\right\|_F^2\end{aligned}$$

where we use $\rho\geq\frac{2}{3\eta d_{\max}}$. The second term can be bounded as follows

$$\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)(\boldsymbol{D}^\star-\boldsymbol{W}^\star-\boldsymbol{D}^t+\boldsymbol{W}^t)\boldsymbol{X}^t\right\|_F^2\leq\|\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star\|_2^2\mathcal{E}.$$

The last term can be bounded as follows

$$\left\|(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}^\star)\nabla F(\boldsymbol{X}^t)\right\|_F^2$$

□

A Proofs

Proof. Apply L-smoothness to each function f_i, we have

$$f_i\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq f_i\left(\bar{\boldsymbol{x}}_c^{t}\right) + \left\langle \nabla f_i\left(\bar{\boldsymbol{x}}_c^{t}\right), \bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t} \right\rangle + \frac{L}{2}\left\|\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}\right\|_2^2.$$

Average the above inequality over $i \in c$, we have

$$\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) + \left\langle \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right), \bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t} \right\rangle + \frac{L}{2}\left\|\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}\right\|_2^2.$$

Expand the intermediate term, we have

$$\begin{aligned}\left\langle \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right), \bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t} \right\rangle = &-\frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 - \frac{1}{2\eta}\left\|\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}\right\|_2^2 \\ &+ \frac{\eta}{2}\left\|\frac{\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}}{\eta} - \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2.\end{aligned}$$

Then by having $\eta \leq \frac{1}{L}$

$$\begin{aligned}\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq &\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 + \frac{\eta}{2}\left\|\frac{\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}}{\eta} - \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 \\ &- \frac{1 - L\eta}{2\eta}\left\|\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}\right\|_2^2 \\ \leq &\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 + \frac{\eta}{2}\left\|\frac{\bar{\boldsymbol{x}}_c^{t+1} - \bar{\boldsymbol{x}}_c^{t}}{\eta} - \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2\end{aligned}$$

Using (3) we have that

$$\begin{aligned}\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq &\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 \\ &+ \frac{\eta}{2}\left\|\frac{1}{c}\sum_{i\in c}\nabla f_i(\boldsymbol{x}_i^t) - \nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) + \frac{\rho}{c}\sum_{i\in c}\sum_{k=1}^{n}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 \\ \leq &\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 + \eta\left\|\frac{1}{c}\sum_{i\in c}\left(\nabla f_i(\boldsymbol{x}_i^t) - \nabla f_i\left(\bar{\boldsymbol{x}}_c^{t}\right)\right)\right\|_2^2 \\ &+ \eta\left\|\frac{\rho}{c}\sum_{i\in c}\sum_{k=1}^{n}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2.\end{aligned}$$

Using the L-smoothness of f_i

$$\begin{aligned}\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t+1}\right) \leq &\bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right) - \frac{\eta}{2}\left\|\nabla \bar{f}_c\left(\bar{\boldsymbol{x}}_c^{t}\right)\right\|_2^2 + \frac{\eta L^2}{c}\sum_{i\in c}\left\|\boldsymbol{x}_i^t - \bar{\boldsymbol{x}}_c^{t}\right\|_2^2 \\ &+ \eta\left\|\frac{\rho}{c}\sum_{i\in c}\sum_{k=1}^{n}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2.\end{aligned}$$

$\square$

A.1 Misclassification Error

Proof.

$$
\begin{aligned}
\mathcal{E} &= \sum_{c}\sum_{i\in c}\left\|\sum_{k=1}^{n}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 \\
&= \sum_{c}\sum_{i\in c}\left\|\sum_{k\in c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t) + \sum_{k\notin c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 \\
&\le 2\underbrace{\sum_{c}\sum_{i\in c}\left\|\sum_{k\in c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2}_{\text{Exclusion Error } \mathcal{E}_{\text{ex}}} + 2\underbrace{\sum_{c}\sum_{i\in c}\left\|\sum_{k\notin c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2}_{\text{Inclusion Error } \mathcal{E}_{\text{in}}}
\end{aligned}
$$

The exclusion error

$$
\begin{aligned}
\mathcal{E}_{\text{ex}} &= \sum_{c}\sum_{i\in c}\left\|\sum_{k\in c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 \\
&\le \sum_{c}\sum_{i\in c}\left(\sum_{k\in c}(w_{ik}^\star - w_{ik}^t)\right)\left(\sum_{k\in c}(w_{ik}^\star - w_{ik}^t)\left\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\right\|_2^2\right) \\
&\le \sum_{c}\sum_{i\in c} c \sum_{k\in c \& w_{ik}^t=0}\left\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\right\|_2^2.
\end{aligned}
$$

This may be further bounded as a function of α_t (but we may not use it)

$$
\mathcal{E}_{\text{ex}} = \sum_{c}\sum_{i\in c} c\frac{\left(\sum_{k\in c\&w_{ik}^t=0}\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\|_2^2\right)^2}{\alpha^t} \text{ Or } \frac{\left(\sum_c\sum_{i\in c} c\sum_{k\in c\&w_{ik}^t=0}\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\|_2^2\right)^2}{\alpha^t}.
$$

On the other hand, the inclusion error can be bounded as follows

$$
\begin{aligned}
\mathcal{E}_{\text{in}} &= \sum_{c}\sum_{i\in c}\left\|\sum_{k\notin c}(w_{ik}^t - w_{ik}^\star)(\boldsymbol{x}_i^t - \boldsymbol{x}_k^t)\right\|_2^2 \\
&\le \sum_{c}\sum_{i\in c}(n-c)\sum_{k\notin c\& w_{ik}^t=1}\left\|\boldsymbol{x}_i^t - \boldsymbol{x}_k^t\right\|_2^2 \\
&\le \alpha^t n(n-c)^2.
\end{aligned}
$$

$\square$