# Improving the Privacy and Accuracy of ADMM-Based Distributed Algorithms (Supplementary materials)

## A. Proof of Simplifying ADMM (Forero et al., 2010)

By KKT condition of (5), there is:

$$0 = \lambda_{ij}^b(t) - \lambda_{ij}^a(t) + \eta(2w_{ij}(t+1) - f_i(t+1) - f_j(t+1))$$

Implies:

$$w_{ij}(t+1) = \frac{1}{2\eta}(\lambda_{ij}^a(t) - \lambda_{ij}^b(t)) + \frac{1}{2}(f_i(t+1) + f_j(t+1)) \tag{27}$$

Plug (27) into (6)(7):

$$\lambda_{ij}^a(t+1) = \frac{1}{2}(\lambda_{ij}^a(t) + \lambda_{ij}^b(t)) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{28}$$

$$\lambda_{ij}^b(t+1) = \frac{1}{2}(\lambda_{ij}^b(t) + \lambda_{ij}^a(t)) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{29}$$

If initialize $\lambda_{ij}^a(0) = \lambda_{ij}^b(0)$ to be zero vectors for all node pairs $(i, j)$, (28)(29) imply that $\lambda_{ij}^a(t) = \lambda_{ij}^b(t)$ and $\lambda_{ji}^k(t) = -\lambda_{ij}^k(t), k \in \{a, b\}$ will hold for all $t$. (27) becomes:

$$w_{ij}(t+1) = \frac{1}{2}(f_i(t+1) + f_j(t+1)) \tag{30}$$

Let $\lambda_{ij}(t) = \lambda_{ij}^a(t) = \lambda_{ij}^b(t)$, (6)(7) can be simplified as:

$$\lambda_{ij}(t+1) = \lambda_{ij}(t) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{31}$$

Plug (30) into the augmented Lagrangian (3) to simplify it:

$$\begin{aligned} L_\eta(\{f_i\}, \{w_{ij}, \lambda_{ij}^k\}) &= \sum_{i=1}^N O(f_i, D_i) + \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} (\lambda_{ij}(t))^T (f_i - f_j) \\ &+ \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \frac{\eta}{2}(||f_i - \frac{1}{2}(f_i(t) + f_j(t))||_2^2) + \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \frac{\eta}{2}(||\frac{1}{2}(f_i(t) + f_j(t)) - f_j||_2^2) \end{aligned} \tag{32}$$

Since $\sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \lambda_{ij}(t) f_j = \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \lambda_{ji}(t) f_i$ and $\lambda_{ij}(t) = -\lambda_{ji}(t)$, the second term in (32) can be simplified:

$$\sum_{i=1}^N \sum_{j \in \mathscr{V}_i} (\lambda_{ij}(t))^T (f_i - f_j) = 2 \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} (\lambda_{ij}(t))^T f_i$$

The last term can be expressed as:

$$\sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \frac{\eta}{2}(||\frac{1}{2}(f_i(t) + f_j(t)) - f_j||_2^2) = \sum_{i=1}^N \sum_{j \in \mathscr{V}_i} \frac{\eta}{2}(||\frac{1}{2}(f_i(t) + f_j(t)) - f_i||_2^2)$$

Therefore, (32) is simplified as:

$$L_\eta(\{f_i\},\{w_{ij},\lambda_{ij}^k\}) = \sum_{i=1}^N O(f_i,D_i) + 2\sum_{i=1}^N\sum_{j\in\mathscr{V}_i}\lambda_{ij}(t)^T f_i + \sum_{i=1}^N\sum_{j\in\mathscr{V}_i}\eta(||f_i - \frac{1}{2}(f_i(t)+f_j(t))||_2^2) \tag{33}$$

Define $\lambda_i(t) = \sum_{j\in\mathscr{V}_i}\lambda_{ij}(t)$. Based on (31)(33), the original ADMM updates (4)-(7) are simplified as:

$$f_i(t+1) = \underset{f_i}{\text{argmin}}\; O(f_i,D_i) + 2\lambda_i(t)^T f_i + \eta\sum_{j\in\mathscr{V}_i}||f_i - \frac{1}{2}(f_i(t)+f_j(t))||_2^2$$

$$\lambda_i(t+1) = \lambda_i(t) + \frac{\eta}{2}\sum_{j\in\mathscr{V}_i}(f_i(t+1)-f_j(t+1))$$

## B. Proof of Theorem 3.1

Subtract (17) from (15) and (18) from (16):

$$\begin{gathered}\nabla\hat{O}(\hat{f}(t+1),D_{all}) - \nabla\hat{O}(\hat{f}^*,D_{all}) + \sqrt{D-A}(Y(t+1)-Y^*) + (W(t+1)-\theta I)(D-A)\hat{f}(t+1)\\ +W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t)) = \mathbf{0}_{N\times d}\end{gathered} \tag{34}$$

$$Y(t+1) = Y(t) + \theta\sqrt{D-A}(\hat{f}(t+1)-\hat{f}^*) \tag{35}$$

By convexity of $O(f_i,D_i)$, for any $f_i^1$ and $f_i^2$, there is:

$$(f_i^1 - f_i^2)^T(\nabla O(f_i^1,D_i) - \nabla O(f_i^2,D_i)) \geq 0$$

Let $\langle\cdot,\cdot\rangle_F$ be frobenius inner product of two matrices, there is:

$$\langle\hat{f}(t+1)-\hat{f}^*, \nabla\hat{O}(\hat{f}(t+1),D_{all}) - \nabla\hat{O}(\hat{f}^*,D_{all})\rangle_F \geq 0$$

Substitute $\nabla\hat{O}(\hat{f}(t+1),D_{all}) - \nabla\hat{O}(\hat{f}^*,D_{all})$ from (34):

$$\begin{gathered}0 \leq \langle\hat{f}(t+1)-\hat{f}^*, -\sqrt{D-A}(Y(t+1)-Y^*)\rangle_F + \langle\hat{f}(t+1)-\hat{f}^*, -(W(t+1)-\theta I)(D-A)\hat{f}(t+1)\rangle_F\\ +\langle\hat{f}(t+1)-\hat{f}^*, -W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))\rangle_F\end{gathered} \tag{36}$$

Consider the right hand side of (36). Since $D-A$ is symmetric and PSD, $\sqrt{D-A}$ is also a symmetric matrix and by (35),

$$\begin{aligned}\langle\hat{f}(t+1)-\hat{f}^*, -\sqrt{D-A}(Y(t+1)-Y^*)\rangle_F &= \langle-\sqrt{D-A}(\hat{f}(t+1)-\hat{f}^*), (Y(t+1)-Y^*)\rangle_F\\ &= -\langle\frac{1}{\theta}(Y(t+1)-Y(t)), Y(t+1)-Y^*\rangle_F\end{aligned} \tag{37}$$

Rearrange (36) and use $(D-A)\hat{f}^* = \mathbf{0}_{N\times d}$

$$0 \geq \langle Z(t+1)-Z^*, J(t+1)(Z(t+1)-Z(t))\rangle_F + \langle\hat{f}(t+1)-\hat{f}^*, (W(t+1)-\theta I)(D-A)(\hat{f}(t+1)-\hat{f}^*)\rangle_F \tag{38}$$

Suppose $\eta_i(t) \geq \theta$ for all $t, i$, i.e., the diagonal matrix $W(t)-\theta I \succeq 0$ for all $t$. Since $D-A \succeq 0$, whose eigenvalues are all non-negative, the eigenvalues of $(W(t+1)-\theta I)(D-A)$ are thus also non-negative, i.e., $(W(t+1)-\theta I)(D-A) \succeq 0$. Then for the second term of the RHS of (38), there is:

$$\langle\hat{f}(t+1)-\hat{f}^*, (W(t+1)-\theta I)(D-A)(\hat{f}(t+1)-\hat{f}^*)\rangle_F \geq 0$$

Therefore,

$$\langle Z(t+1) - Z^*, J(t+1)(Z(t+1) - Z(t))\rangle_F \leq 0 \tag{39}$$

To simplify the notation, for a matrix $X$, let $||X||_J^2 = \langle X, JX\rangle_F$, then (39) can be represented as:

$$\frac{1}{2}||Z(t+1) - Z^*||^2_{J(t+1)} + \frac{1}{2}||Z(t+1) - Z(t)||^2_{J(t+1)} - \frac{1}{2}||Z(t) - Z^*||^2_{J(t+1)} \leq 0$$

implies

$$||Z(t+1) - Z(t)||^2_{J(t+1)} \leq -||Z(t+1) - Z^*||^2_{J(t+1)} + ||Z(t) - Z^*||^2_{J(t)} + ||Z(t) - Z^*||^2_{J(t+1)} - ||Z(t) - Z^*||^2_{J(t)} \tag{40}$$

Suppose $\eta_i(t+1) \geq \eta_i(t)$ for all $t$ and $i$, i.e., the diagonal matrix $W(t+1) - W(t) \succeq 0$ for all $t$. Since $D + A \succeq 0$, implies $(W(t+1) - W(t))(D+A) \succeq 0$. Let $U = \sup\limits_{i,t,k} |(f_i(t) - f_c^*)_k| \in \mathbb{R}$ be the finite upper bound of all nodes $i$, all iterations $t$ and all components $k$, then

$$\begin{aligned} &||Z(t) - Z^*||^2_{J(t+1)} - ||Z(t) - Z^*||^2_{J(t)} = \text{Tr}((Z(t) - Z^*)^T(J(t+1) - J(t))(Z(t) - Z^*)) \\ &= \text{Tr}((\hat{f}(t) - \hat{f}^*)^T(W(t+1) - W(t))(D+A)(\hat{f}(t) - \hat{f}^*)) \leq U^2(||\mathbf{ones}(N,d)||^2_{W(t+1)(D+A)} - \mathbf{ones}(N,d)||^2_{W(t)(D+A)}) \end{aligned} \tag{41}$$

where $\mathbf{ones}(N,d)$ is all one's matrix of size $N \times d$. By (40)(41):

$$\begin{aligned} ||Z(t+1) - Z(t)||^2_{J(t+1)} &\leq ||Z(t) - Z^*||^2_{J(t)} - ||Z(t+1) - Z^*||^2_{J(t+1)} \\ &+ U^2(||\mathbf{ones}(N,d)||^2_{W(t+1)(D+A)} - ||\mathbf{ones}(N,d)||^2_{W(t)(D+A)}) \end{aligned} \tag{42}$$

Sum up (42) over $t$ from 0 to $+\infty$ leads to:

$$\begin{aligned} \sum_{t=0}^{+\infty} ||Z(t+1) - Z(t)||^2_{J(t+1)} &\leq ||Z(0) - Z^*||^2_{J(0)} - ||Z(+\infty) - Z^*||^2_{J(+\infty)} \\ &+ U^2(||\mathbf{ones}(N,d)||^2_{W(+\infty)(D+A)} - ||\mathbf{ones}(N,d)||^2_{W(0)(D+A)}) \end{aligned} \tag{43}$$

Since $\eta_i(t) < +\infty$, the RHS of (43) is finite, implies that $\lim_{t\to+\infty} ||Z(t+1) - Z(t)||^2_{J(t+1)} = 0$ must hold.

By the definition of $Z(t)$, $J(t)$ and $||X||_J^2 = \langle X, JX\rangle_F$, the following must hold

$$\lim_{t\to+\infty} ||\hat{f}(t+1) - \hat{f}(t)||^2_{W(t+1)(D+A)} = 0 \tag{44}$$

$$\lim_{t\to+\infty} ||Y(t+1) - Y(t)||^2_F = 0 \tag{45}$$

(45) shows that $Y(t)$ converges to a stationary point $Y^s$, along with (16) imply $\lim_{t\to+\infty} \sqrt{D-A}\hat{f}(t+1) = 0$. Since $\text{Null}(\sqrt{D-A}) = c\mathbf{1}$, $\hat{f}(t+1)$ must lie in the subspace spanned by $\mathbf{1}$ as $t \to \infty$. To satisfy (44), either of the following two statements must hold:

- $\lim_{t\to+\infty}(\hat{f}(t+1) - \hat{f}(t)) = \mathbf{0}_{N\times d}$
- $\lim_{t\to+\infty} W(t+1)(D+A)\mathbf{1} = \lim_{t\to+\infty} W(t+1)A\mathbf{1} + \lim_{t\to+\infty} \sum_{i=1}^N \eta_i(t+1)V_i = \mathbf{0}_{N\times 1}$

Since $\eta_i(t) \geq \theta > 0$ for all $t$, implies $\lim_{t\to+\infty} \sum_{i=1}^N \eta_i(t+1)V_i > 0$. The second statement can never be true because all elements of $A$ and $W(t+1)$ are non-negative. Hence, $\hat{f}(t)$ should also converge to a stationary point $\hat{f}^s$.

Now show that the stationary point $(Y^s, \hat{f}^s)$ is $(Y^*, \hat{f}^*)$.

Take limit of both sides of (15) (16), substitute $\hat{f}^s, Y^s$ yields

$$\nabla \hat{O}(\hat{f}^s, D_{all}) + \sqrt{D-A}Y^s + (W(t+1) - \theta I)(D-A)\hat{f}^s = \mathbf{0}_{N\times d} \tag{46}$$

$$\sqrt{D-A}\hat{f}^s = \mathbf{0}_{N\times d} \tag{47}$$

By (47), (46) turns into:

$$\nabla \hat{O}(\hat{f}^s, D_{all}) + \sqrt{D-A}Y^s = \mathbf{0}_{N\times d} \tag{48}$$

Compare (47)(48) with (17)(18) in Lemma 3.1 and observe that $(Y^s, \hat{f}^s)$ satisfies the optimality condition (17)(18) and is thus the optimal point. Therefore, $f(t)$ converges to $\hat{f}^*$ and $Y(t)$ converges to $Y^*$.

## C. Proof of Theorem 3.2

According to the Assumption 3 that $O(f_i, D_i)$ is strongly convex and has Lipschitz continues gradients for all $i \in \mathscr{N}$, define diagonal matrices $D_m = \textbf{diag}([m_1; m_2; \cdots; m_N]) \in \mathbb{R}^{N\times N}$ and $D_M = \textbf{diag}([M_1^2; M_2^2; \cdots; M_N^2]) \in \mathbb{R}^{N\times N}$, (20) yield:

$$\langle \hat{f}^1 - \hat{f}^2, \nabla \hat{O}(\hat{f}^1, D_{all}) - \nabla \hat{O}(\hat{f}^2, D_{all})\rangle_F \geq \langle \hat{f}^1 - \hat{f}^2, D_m(\hat{f}^1 - \hat{f}^2)\rangle_F \tag{49}$$

$$||\nabla \hat{O}(\hat{f}^1, D_{all}) - \nabla \hat{O}(\hat{f}^2, D_{all})||_F^2 \leq \langle \hat{f}^1 - \hat{f}^2, D_M(\hat{f}^1 - \hat{f}^2)\rangle_F \tag{50}$$

Since for any $\mu > 1$ and any matrices $C_1, C_2$ with the same dimensions, there is:

$$||C_1 + C_2||_F^2 \leq \mu ||C_1||_F^2 + \frac{\mu}{\mu-1}||C_2||_F^2$$

From (34), there is:

$$\begin{aligned} ||\sqrt{D-A}(Y(t+1) - Y^*)||_F^2 &\leq \mu||\nabla \hat{O}(\hat{f}(t+1), D_{all}) - \nabla \hat{O}(\hat{f}^*, D_{all}) + W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t))||_F^2 \\ &+\frac{\mu}{\mu-1}||(W(t+1) - \theta I)(D-A)\hat{f}(t+1)||_F^2 \leq \frac{\mu^2}{\mu-1}||\nabla \hat{O}(\hat{f}(t+1), D_{all}) - \nabla \hat{O}(\hat{f}^*, D_{all})||_F^2 \\ &+\mu^2||W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t))||_F^2 + \frac{\mu}{\mu-1}||(W(t+1) - \theta I)(D-A)\hat{f}(t+1)||_F^2 \end{aligned} \tag{51}$$

Let $\sigma_{\text{min}}(\cdot)$, $\sigma_{\text{max}}(\cdot)$ denote the smallest nonzero singular value and the largest singular value of a matrix respectively.

For any matrices $C_1$, $C_2$, let $C_1 = U\Sigma V^T$ be SVD of $C_1$, there is:

$$||C_1C_2||_F^2 \leq \sigma_{\text{max}}(C_1)||C_2||_{C_1^T}^2$$

$$\sigma_{\text{min}}(C_1)^2||C_2||_F^2 \leq ||C_1C_2||_F^2 \leq \sigma_{\text{max}}(C_1)^2||C_2||_F^2$$

Denote

$$\bar{\sigma}_{\text{max}}(t+1) = \sigma_{\text{max}}((W(t+1) - \theta I)(D-A))$$

$$\bar{\sigma}_{\text{min}}(t+1) = \sigma_{\text{min}}((W(t+1) - \theta I)(D-A))$$

$$\tilde{\sigma}_{\text{max}}(t+1) = \sigma_{\text{max}}(W(t+1)(D+A))$$

Using (50) and $(D-A)\hat{f}^* = 0$, (51) is turned into:

$$\begin{aligned} \frac{1}{\theta}||Y(t+1) - Y^*||_F^2 &\leq \frac{\mu^2}{\theta\sigma_{\text{min}}(D-A)(\mu-1)}||\hat{f}(t+1) - \hat{f}^*||_{D_M}^2 \\ &+\frac{\mu^2\tilde{\sigma}_{\text{max}}(t+1)}{\theta\sigma_{\text{min}}(D-A)}||\hat{f}(t+1) - \hat{f}(t)||_{W(t+1)(D+A)}^2 + \frac{\mu\bar{\sigma}_{\text{max}}(t+1)^2}{\theta\sigma_{\text{min}}(D-A)(\mu-1)}||(\hat{f}(t+1) - \hat{f}^*)||_F^2 \end{aligned}$$

Adding $||\hat{f}(t+1)-\hat{f}^*||^2_{W(t+1)(D+A)}$ at both sides leads to:

$$\begin{aligned}||Z(t+1)-Z^*||^2_{J(t+1)} \leq \frac{\mu^2\tilde{\sigma}_{\max}(t+1)}{\theta\sigma_{\min}(D-A)}||\hat{f}(t+1)-\hat{f}(t)||^2_{W(t+1)(D+A)} \\ +||\hat{f}(t+1)-\hat{f}^*||^2_{\frac{\mu^2 D_M+\mu\bar{\sigma}_{\max}(t+1)^2\mathbf{I}_N}{\theta\sigma_{\min}(D-A)(\mu-1)}+W(t+1)(D+A)}\end{aligned} \tag{52}$$

Since

$$\frac{\delta(t+1)\mu^2\tilde{\sigma}_{\max}(t+1)}{\theta\sigma_{\min}(D-A)} \leq 1 \tag{53}$$

and

$$\delta(t+1)(\frac{\mu\bar{\sigma}_{\max}(t+1)^2\mathbf{I}_N+\mu^2 D_M}{\theta\sigma_{\min}(D-A)(\mu-1)}+W(t+1)(D+A)) \preceq 2(W(t+1)-\theta I)(D-A)+2D_m \tag{54}$$

It implies from (52) that:

$$\begin{aligned}\delta(t+1)||Z(t+1)-Z^*||^2_{J(t+1)} &\leq ||\hat{f}(t+1)-\hat{f}(t)||^2_{W(t+1)(D+A)}+||\hat{f}(t+1)-\hat{f}^*||^2_{2(W(t+1)-\theta I)(D-A)+2D_m} \\ &\leq ||Z(t+1)-Z(t)||^2_{J(t+1)}+||\hat{f}(t+1)-\hat{f}^*||^2_{2(W(t+1)-\theta I)(D-A)+2D_m}\end{aligned} \tag{55}$$

Substituting $\hat{f}^1$ with $\hat{f}(t+1)$ and $\hat{f}^2$ with $\hat{f}^*$ and the gradient difference from (34) in (49) leads to:

$$\begin{aligned}\langle\hat{f}(t+1)-\hat{f}^*,\sqrt{D-A}(Y(t+1)-Y^*)\rangle_F+\langle\hat{f}(t+1)-\hat{f}^*,W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))\rangle_F \\ +\langle\hat{f}(t+1)-\hat{f}^*,(W(t+1)-\theta I)(D-A)\hat{f}(t+1)\rangle_F \leq -\langle\hat{f}(t+1)-\hat{f}^*,D_m(\hat{f}(t+1)-\hat{f}^*)\rangle_F\end{aligned}$$

Similar to the proof of Theorem 3.1, using the definition of $Z(t+1)$, $Z^*$, $J(t+1)$ and $(D-A)\hat{f}^*=0$, there is:

$$||Z(t+1)-Z^*||^2_{J(t+1)} \leq -||Z(t+1)-Z(t)||^2_{J(t+1)}+||Z(t)-Z^*||^2_{J(t+1)}-||\hat{f}(t+1)-\hat{f}^*||^2_{2D_m+2(W(t+1)-\theta I)(D-A)} \tag{56}$$

Sum up (55) and (56) gives:

$$(1+\delta(t+1))||Z(t+1)-Z^*||^2_{J(t+1)} \leq ||Z(t)-Z^*||^2_{J(t+1)}$$

Let $m_o=\min_{i\in\mathscr{N}}\{m_i\}$, $M_O=\max_{i\in\mathscr{N}}\{M_i\}$. One $\delta(t+1)$ that satisfies (53) and (54) could be:

$$\min\{\frac{\theta\sigma_{\min}(D-A)}{\mu^2\tilde{\sigma}_{\max}(t+1)},\frac{2m_o+2\bar{\sigma}_{\min}(t+1)}{\frac{\mu^2M_O^2+\mu\bar{\sigma}_{\max}(t+1)^2}{\theta\sigma_{\min}(D-A)(\mu-1)}+\tilde{\sigma}_{\max}(t+1)}\}$$

## D. Proof of Theorem 4.1

In the following proof, use the uppercase letters and lowercase letters to denote random variables and the corresponding realizations.

Since the modified ADMM is randomized, denote $F_i(t)$ as the random variable of the result that node $i$ broadcasts in $t$-th iteration, of which the realization is $f_i(t)$. Define $F(t)=\{F_i(t)\}_{i=1}^N$ whose realization is $\{f_i(t)\}_{i=1}^N$.

Let $\mathscr{F}_{F(0:t)}(\cdot)$ be the joint probability distribution of $F(0:t)=\{F(r)\}_{r=0}^t$, and $\mathscr{F}_{F(t)}(\cdot)$ be the distribution of $F(t)$, by chain rule:

$$\begin{aligned}\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^T) &= \mathscr{F}_{F(0:T-1)}(\{f(r)\}_{r=0}^{T-1})\cdot\mathscr{F}_{F(T)}(f(T)|\{f(r)\}_{r=0}^{T-1})=\cdots \\ &= \mathscr{F}_{F(0)}(f(0))\cdot\prod_{t=1}^T\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1})\end{aligned}$$

For two neighboring datasets $D_{all}$ and $\hat{D}_{all}$ of the network, the ratio of joint probabilities is given by:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|\hat{D}_{all})} = \frac{\mathscr{F}_{F(0)}(f(0)|D_{all})}{\mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})} \cdot \prod_{t=1}^{T} \frac{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, D_{all})}{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, \hat{D}_{all})} \tag{57}$$

Since $f_i(0)$ is randomly selected for all $i$, which is independent of dataset, there is $\mathscr{F}_{F(0)}(f(0)|D_{all}) = \mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})$.

First only consider $t$-th iteration, since the primal variable is updated according to (25), by KKT optimality condition, $\nabla_{f_i} L_i^{priv}(t)|_{f_i=f_i(t)} = 0$, implies:

$$\begin{aligned}\epsilon_i(t) = -\frac{1}{2\eta_i(t)V_i}\frac{C}{B_i}\sum_{n=1}^{B_i} y_i^n \mathscr{L}'(y_i^n f_i(t)^T x_i^n)x_i^n - \frac{1}{2\eta_i(t)V_i}(\frac{\rho}{N}\nabla R(f_i(t)) + 2\lambda_i(t-1)) \\ -\frac{1}{2V_i}\sum_{j\in\mathscr{V}_i}(2f_i(t) - f_i(t-1) - f_j(t-1))\end{aligned} \tag{58}$$

Given $\{f_i(r)\}_{r=0}^{t-1}$, $F_i(t)$ and $E_i(t)$ will be bijective:

- For any $F_i(t)$ with the realization $f_i(t)$, $\exists$ an unique $E_i(t) = \epsilon_i(t)$ having the form of (58) such that the KKT condition holds.
- Since the Lagrangian $L_i^{priv}(t)$ is strictly convex (by Assumption 4,5), its minimizer is unique, implies that for any $E_i(t)$ with the realization $\epsilon_i(t)$, $\exists$ an unique $F_i(t) = f_i(t)$ such that the KKT condition holds.

Since each node $i$ generates $\epsilon_i(t)$ independently, $f_i(t)$ is also independent from each other. Let $\mathscr{F}_{F_i(t)}(\cdot)$ be the distribution of $F_i(t)$, there is:

$$\frac{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, D_{all})}{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, \hat{D}_{all})} = \prod_{v=1}^{N} \frac{\mathscr{F}_{F_v(t)}(f_v(t)|\{f_v(r)\}_{r=0}^{t-1}, D_v)}{\mathscr{F}_{F_v(t)}(f_v(t)|\{f_v(r)\}_{r=0}^{t-1}, \hat{D}_v)} = \frac{\mathscr{F}_{F_i(t)}(f_i(t)|\{f_i(r)\}_{r=0}^{t-1}, D_i)}{\mathscr{F}_{F_i(t)}(f_i(t)|\{f_i(r)\}_{r=0}^{t-1}, \hat{D}_i)} \tag{59}$$

Since two neighboring datasets $D_{all}$ and $\hat{D}_{all}$ only have at most one data point that is different, the second equality holds is because of the fact that this different data point could only be possessed by one node, say node $i$. Then there is $D_j = \hat{D}_j$ for $j \neq i$.

Given $\{f_i(r)\}_{r=0}^{t-1}$, let $g_t(\cdot, D_i) : \mathbb{R}^d \to \mathbb{R}^d$ denote the one-to-one mapping from $E_i(t)$ to $F_i(t)$ using dataset $D_i$. Let $\mathscr{F}_{E_i(t)}(\cdot)$ be the probability density of $E_i(t)$, by Jacobian transformation, there is[4]:

$$\mathscr{F}_{F_i(t)}(f_i(t)|D_i) = \mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i)) \cdot |\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))| \tag{60}$$

where $g_t^{-1}(f_i(t), D_i)$ is the mapping from $F_i(t)$ to $E_i(t)$ using data $D_i$ as shown in (58) and $\mathbf{J}(g_t^{-1}(f_i(t), D_i))$ is the Jacobian matrix of it.

Without loss of generality, let $D_i$ and $\hat{D}_i$ be only different in the first data point, say $(x_i^1, y_i^1)$ and $(\hat{x}_i^1, \hat{y}_i^1)$ respectively. Then by (59)(60), (57) yields:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|\hat{D}_{all})} = \prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} \cdot \prod_{t=1}^{T} \frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t), \hat{D}_i)))|} \tag{61}$$

[4] We believe that there is a critical mistake in (Zhang & Zhu, 2017) and the original paper (Chaudhuri et al., 2011) where the objective perturbation method was proposed. A wrong mapping is used in both work:

$$\mathscr{F}_{F_i(t)}(f_i(t)|D_i) = \mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i)) \cdot |\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|^{-1}$$

Consider the first part, $E_i(t) \sim \exp\{-\alpha_i(t)||\epsilon||\}$, let $\hat{\epsilon}_i(t) = g_t^{-1}(f_i(t), \hat{D}_i)$ and $\epsilon_i(t) = g_t^{-1}(f_i(t), D_i)$

$$\prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} = \prod_{t=1}^{T} \exp(\alpha_i(t)(||\hat{\epsilon}_i(t)|| - ||\epsilon_i(t)||)) \leq \exp(\sum_{t=1}^{T} \alpha_i(t)||\hat{\epsilon}_i(t) - \epsilon_i(t)||) \tag{62}$$

By (58), Assumptions 4 and the facts that $||x_i^n||_2 \leq 1$ (pre-normalization), $y_i^n \in \{+1, -1\}$.

$$||\hat{\epsilon}_i(t) - \epsilon_i(t)|| = \frac{1}{2\eta_i(t)V_i} \frac{C}{B_i} \cdot ||y_i^1 \mathscr{L}'(y_i^1 f_i(t)^T x_i^1) x_i^1 - \hat{y}_i^1 \mathscr{L}'(\hat{y}_i^1 f_i(t)^T \hat{x}_i^1) \hat{x}_i^1|| \leq \frac{C}{\eta_i(t)V_i B_i}$$

(62) can be bounded:

$$\prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} \leq \exp(\sum_{t=1}^{T} \frac{C\alpha_i(t)}{\eta_i(t)V_i B_i}) \tag{63}$$

Consider the second part, the Jacobian matrix $\mathbf{J}(g_t^{-1}(f_i(t), D_i))$ is:

$$\mathbf{J}(g_t^{-1}(f_i(t), D_i)) = -\frac{1}{2\eta_i(t)V_i} \frac{C}{B_i} \sum_{n=1}^{B_i} \mathscr{L}''(y_i^n f_i(t)^T x_i^n) x_i^n (x_i^n)^T - \frac{1}{2\eta_i(t)V_i} \frac{\rho}{N} \nabla^2 R(f_i(t)) - \mathbf{I}_d$$

Let $G(t) = \frac{C}{2\eta_i(t)V_i B_i}(\mathscr{L}''(\hat{y}_i^1 f_i(t)^T \hat{x}_i^1) \hat{x}_i^1 (\hat{x}_i^1)^T - \mathscr{L}''(y_i^1 f_i(t)^T x_i^1) x_i^1 (x_i^1)^T)$ and $H(t) = -\mathbf{J}(g_t^{-1}(f_i(t), D_i))$, there is:

$$\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t), \hat{D}_i)))|} = \frac{|\det(H(t))|}{|\det(H(t) + G(t))|} = \frac{1}{|\det(I + H(t)^{-1}G(t))|} = \frac{1}{|\prod_{j=1}^{r}(1 + \lambda_j(H(t)^{-1}G(t)))|}$$

where $\lambda_j(H(t)^{-1}G(t))$ denotes the $j$-th largest eigenvalue of $H(t)^{-1}G(t)$. Since $G(t)$ has rank at most 2, implies $H(t)^{-1}G(t)$ also has rank at most 2.

Because $\theta$ is determined such that $2c_1 < \frac{B_i}{C}(\frac{\rho}{N} + 2\theta V_i)$, and $\theta \leq \eta_i(t)$ holds for all node $i$ and iteration $t$, which implies:

$$\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)} < \frac{1}{2} \tag{64}$$

By Assumptions 4 and 5, the eigenvalue of $H(t)$ and $G(t)$ satisfy:

$$\lambda_j(H(t)) \geq \frac{\rho}{2\eta_i(t)V_i N} + 1 > 0$$

$$-\frac{Cc_1}{2\eta_i(t)V_i B_i} \leq \lambda_j(G(t)) \leq \frac{Cc_1}{2\eta_i(t)V_i B_i}$$

Implies:

$$-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)} \leq \lambda_j(H(t)^{-1}G(t)) \leq \frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)}$$

By (64):

$$-\frac{1}{2} \leq \lambda_j(H(t)^{-1}G(t)) \leq \frac{1}{2}$$

Since $\lambda_{\min}(H(t)^{-1}G(t)) > -1$, there is:

$$\frac{1}{|1 + \lambda_{\max}(H(t)^{-1}G(t))|^2} \leq \frac{1}{|\det(I + H(t)^{-1}G(t))|} \leq \frac{1}{|1 + \lambda_{\min}(H(t)^{-1}G(t))|^2}$$

Therefore,

$$\prod_{t=1}^{T}\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t),D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t),\hat{D}_i)))|}\leq\prod_{t=1}^{T}\frac{1}{|1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)}|^2}=\exp(-\sum_{t=1}^{T}2\ln(1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)})) \tag{65}$$

Since for any real number $x\in[0,0.5]$, $-\ln(1-x)<1.4x$. By condition (64), (65) can be bounded with a simper expression:

$$\prod_{t=1}^{T}\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t),D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t),\hat{D}_i)))|}\leq\exp(\sum_{t=1}^{T}\frac{2.8c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)})\leq\exp(\sum_{t=1}^{T}\frac{1.4Cc_1}{\eta_i(t)V_iB_i}) \tag{66}$$

Combine (63)(66), (61) can be bounded:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|\hat{D}_{all})}\leq\exp(\sum_{t=1}^{T}(\frac{1.4Cc_1}{\eta_i(t)V_iB_i}+\frac{C\alpha_i(t)}{\eta_i(t)V_iB_i}))=\exp(\sum_{t=1}^{T}\frac{C}{\eta_i(t)V_iB_i}(1.4c_1+\alpha_i(t)))$$

Therefore, the total privacy loss during $T$ iterations can be bounded by any $\beta$:

$$\beta\geq\max_{i\in\mathscr{N}}\{\sum_{t=1}^{T}\frac{C}{\eta_i(t)V_iB_i}(1.4c_1+\alpha_i(t))\}$$

## E. Inference of Attackers when $\eta_i(t)$ is Non-private

By KKT optimality condition in each iteration, we have:

$$\begin{aligned}\epsilon_i(t)+\frac{1}{2\eta_i(t)V_i}\frac{C}{B_i}y_i^1\mathscr{L}'(y_i^1f_i(t)^Tx_i^1)x_i^1 &= -\frac{1}{2\eta_i(t)V_i}\frac{C}{B_i}\sum_{n=2}^{B_i}y_i^n\mathscr{L}'(y_i^nf_i(t)^Tx_i^n)x_i^n\\ -\frac{1}{2\eta_i(t)V_i}(\frac{\rho}{N}\nabla R(f_i(t))+2\lambda_i(t-1)) &- \frac{1}{2V_i}\sum_{j\in\mathscr{V}_i}(2f_i(t)-f_i(t-1)-f_j(t-1))\,.\end{aligned}$$

In this case the attacker can compute the RHS of (67) completely. Furthermore, since $E_i(t)$ is zero-mean, over a large number of iterations we will have $\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\approx 0$ with high probability, which then allows the attacker to determine the features of the unknown individual up to a scaling factor, i.e., it can determine the second term on the LHS as a scalar multiplied with $x_i^1$.