

A Bounds needed in the proof in Section 3

In the following sub-sections we provide the upperbounds for Term 1, Term 21 and Term 22 from the main text.

A.1 Upperbound for Term 1

For Term 1 in equation (4) we proceed by observing that conditioned on S_{t-1} , $\mathbf{w}^{(t)}$ is determined while $\mathbf{w}^{(t+1)}$ and $\mathbf{g}^{(t)}$ are random. Thus we compute the following conditional expectation (suppressing the subscripts of $\mathbf{x}_{t_i}, \alpha_{t_i}$),

$$\begin{aligned}
\text{Term 1} &= \mathbb{E} \left[2\eta \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{g}^{(t)} \rangle \middle| S_{t-1} \right] \\
&= 2\frac{\eta}{b} \sum_{i=1}^b \mathbb{E} \left[\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_i}) + \alpha_{t_i} \xi_{t_i} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}) \right) (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i} \middle| S_{t-1} \right] \\
&= 2\frac{\eta}{b} \sum_{i=1}^b \mathbb{E} \left[\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_i}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_i}) \right) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i} \middle| S_{t-1} \right] \\
&\quad + 2\frac{\eta}{b} \sum_{i=1}^b \mathbb{E} \left[\alpha_{t_i} \xi_{t_i} (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i} \middle| S_{t-1} \right] \tag{11}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{-2\eta}{bk} \sum_{i=1}^b \sum_{j=1}^k \mathbb{E} \left[\left(\sigma(\mathbf{w}^{(t)\top} \mathbf{A}_j \mathbf{x}_{t_i}) - \sigma(\mathbf{w}^{*\top} \mathbf{A}_j \mathbf{x}_{t_i}) \right) (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i} \middle| S_{t-1} \right] \\
&\quad + 2\frac{\eta\theta}{b} \sum_{i=1}^b \mathbb{E} \left[\beta(\mathbf{x}_{t_i}) \cdot |(\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i}| \middle| S_{t-1} \right] \tag{12}
\end{aligned}$$

We simplify the first term above by recalling an identity proven in [12], which we have reproduced here as Lemma 3. Thus we get,

$$\begin{aligned}
& \mathbb{E} \left[2\eta \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{g}^{(t)} \rangle \middle| S_{t-1} \right] \\
& \leq \frac{-\eta(1+\alpha)}{bk} \sum_{i=1}^b \sum_{j=1}^k \mathbb{E} \left[(\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{A}_j \mathbf{x}_{t_i} (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \mathbf{M} \mathbf{x}_{t_i} \middle| S_{t-1} \right] \\
& \quad + 2 \frac{\eta \theta_*}{b} \sum_{i=1}^b \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \cdot \mathbb{E} \left[\beta(\mathbf{x}_{t_i}) \|\mathbf{M} \mathbf{x}_{t_i}\| \middle| S_{t-1} \right] \\
& \leq -\eta(1+\alpha) (\mathbf{w}^{(t)} - \mathbf{w}^*)^\top \bar{\mathbf{A}} \mathbb{E} \left[\mathbf{x}_{t_1} \mathbf{x}_{t_1}^\top \middle| S_{t-1} \right] \mathbf{M}^\top (\mathbf{w}^{(t)} - \mathbf{w}^*) \\
& \quad + 2\eta \theta_* \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \sqrt{\lambda_{\max}(\mathbf{M}^\top \mathbf{M})} \cdot \mathbb{E} \left[\beta(\mathbf{x}_{t_1}) \|\mathbf{x}_{t_1}\| \middle| S_{t-1} \right] \\
& \leq -\eta(1+\alpha) \cdot \lambda_{\min} \left(\bar{\mathbf{A}} \mathbb{E} \left[\mathbf{x}_{t_1} \mathbf{x}_{t_1}^\top \middle| S_{t-1} \right] \mathbf{M}^\top \right) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \\
& \quad + 2\eta \theta_* \cdot \mathbb{E} \left[\beta(\mathbf{x}_{t_1}) \|\mathbf{x}_{t_1}\| \middle| S_{t-1} \right] \cdot \sqrt{\lambda_{\max}(\mathbf{M}^\top \mathbf{M})} \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \\
& \leq -\eta(1+\alpha) \cdot \lambda_1 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \\
& \quad + 2\eta \theta \lambda_2 \cdot \mathbb{E} \left[\beta(\mathbf{x}_{t_1}) \|\mathbf{x}_{t_1}\| \middle| S_{t-1} \right] \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|
\end{aligned}$$

We have invoked the i.i.d nature of the data samples to invoke the definition of the λ_1 in above.

A.2 Upperbound for Term 21

For Term 21 in equation (6) we get,

$$\begin{aligned}
\text{Term 21} & \leq \frac{\eta^2 \lambda_2^2}{b} \cdot \mathbb{E} \left[\left(f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) + \alpha_{t_1} \xi_{t_1} - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1}) \right)^2 \cdot \|\mathbf{x}_{t_1}\|^2 \middle| S_{t-1} \right] \\
& \leq \frac{\eta^2 \lambda_2^2}{b} \cdot \mathbb{E} \left[\left((f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1}))^2 + 2\alpha_{t_1} \xi_{t_1} (f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})) + \alpha_{t_1}^2 \xi_{t_1}^2 \right) \cdot \|\mathbf{x}_{t_1}\|^2 \middle| S_{t-1} \right] \\
& \leq \frac{\eta^2 \lambda_2^2 c^2}{b} \mathbb{E} \left[\|\mathbf{x}_{t_1}\|^4 \middle| S_{t-1} \right] \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + \frac{2\eta^2 \lambda_2^2 c \theta}{b} \mathbb{E} \left[\beta(\mathbf{x}_{t_1}) \|\mathbf{x}_{t_1}\|^3 \middle| S_{t-1} \right] \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \\
& \quad + \frac{\eta^2 \lambda_2^2 \theta^2}{b} \mathbb{E} \left[\beta(\mathbf{x}_{t_1}) \|\mathbf{x}_{t_1}\|^2 \middle| S_{t-1} \right] \\
& = \frac{\eta^2 \lambda_2^2}{b} \left(c^2 m_4 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + 2c\theta\beta_3 \|\mathbf{w}^{(t)} - \mathbf{w}^*\| + \theta^2 \beta_2 \right)
\end{aligned}$$

In the above lines we have invoked lemma 4 twice to upperbound the term, $|(f_{\mathbf{w}^*}(\mathbf{x}^{(t)}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}^{(t)}))|$ and we have defined,

$$c^2 := (1 + \alpha)^2 \lambda_3 = \frac{(1 + \alpha)^2}{k} \left(\sum_{i=1}^k \lambda_{\max}(\mathbf{A}_i \mathbf{A}_i^\top) \right).$$

Next we proceed with Term 22 keeping in mind the independence of x_{t_i} and x_{t_j} for $i \neq j$,

A.3 Upperbound for Term 22

For Term 22 in equation (6) we get,

Term 22

$$\begin{aligned} &= \frac{\eta^2(b^2 - b)}{b^2} \mathbb{E} \left[(\alpha_{t_1} \xi_{t_1} + f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})) (\alpha_{t_2} \xi_{t_2} + f_{\mathbf{w}^*}(\mathbf{x}_{t_2}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_2})) \cdot \mathbf{x}_{t_2}^\top \mathbf{M}^\top \mathbf{M} \mathbf{x}_{t_1} \middle| S_{t-1} \right] \\ &\leq \frac{\eta^2(b^2 - b)}{b^2} \left[\theta^2 \left(\mathbb{E}_{x_{t_1}} \left[\beta(x_{t_1}) \|\mathbf{M} x_{t_1}\| \middle| S_{t-1} \right] \right)^2 \right. \\ &\quad + 2\theta \mathbb{E}_{x_{t_1}} \left[(f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})) \|\mathbf{M} x_{t_1}\| \middle| S_{t-1} \right] \mathbb{E}_{x_{t_1}} \left[\beta(x_{t_1}) \|\mathbf{M} x_{t_1}\| \middle| S_{t-1} \right] \\ &\quad \left. + \mathbb{E}_{x_{t_1}} \left[(f_{\mathbf{w}^*}(\mathbf{x}_{t_1}) - f_{\mathbf{w}^{(t)}}(\mathbf{x}_{t_1})) \|\mathbf{M} x_{t_1}\| \middle| S_{t-1} \right]^2 \right] \\ &\leq \frac{\eta^2(b^2 - b)}{b^2} \left[\theta^2 \lambda_2^2 \beta_1^2 + 2\theta_* \lambda_2^2 \beta_1 c m_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\| + \lambda_2^2 c^2 m_2^2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] \end{aligned}$$

B Estimating the necessary recursion

Lemma 2. Suppose we have a sequence of real numbers $\Delta_1, \Delta_2, \dots$ s.t

$$\Delta_{t+1} \leq (1 - \eta' b + \eta'^2 c_1) \Delta_t + \eta'^2 c_2 + \eta' c_3$$

for some fixed parameters $b, c_1, c_2, c_3 > 0$ s.t $\Delta_1 > \frac{c_3}{b}$ and free parameter $\eta' > 0$. Then for,

$$\epsilon'^2 \in \left(\frac{c_3}{b}, \Delta_1 \right), \quad \eta' = \frac{b}{\gamma c_1}, \quad \gamma > \max \left\{ \frac{b^2}{c_1}, \left(\frac{\epsilon'^2 + \frac{c_2}{c_1}}{\epsilon'^2 - \frac{c_3}{b}} \right) \right\} > 1$$

it follows that $\Delta_T \leq \epsilon'^2$ for,

$$T = \mathcal{O} \left(\log \left[\frac{\Delta_1}{\epsilon'^2 - \left(\frac{\frac{c_2}{c_1} + \gamma \cdot \frac{c_3}{b}}{\gamma - 1} \right)} \right] \right)$$

Proof. Let us define $\alpha = 1 - \eta'b + \eta'^2 c_1$ and $\beta = \eta'^2 c_2 + \eta' c_3$. Then by unrolling the recursion we get,

$$\Delta_t \leq \alpha \Delta_{t-1} + \beta \leq \alpha(\alpha \Delta_{t-2} + \beta) + \beta \leq \dots \leq \alpha^{t-1} \Delta_1 + \beta(1 + \alpha + \dots + \alpha^{t-2}).$$

Now suppose that the following are true for ϵ' as given and for α & β (evaluated for the range of η' s as specified in the theorem),

Claim 1 : $\alpha \in (0, 1)$

Claim 2 : $0 < \epsilon'^2(1 - \alpha) - \beta$

We will soon show that the above claims are true. Now if T is s.t we have,

$$\alpha^{T-1} \Delta_1 + \beta(1 + \alpha + \dots + \alpha^{T-2}) = \alpha^{T-1} \Delta_1 + \beta \cdot \frac{1 - \alpha^{T-1}}{1 - \alpha} = \epsilon'^2$$

then $\alpha^{T-1} = \frac{\epsilon'^2(1-\alpha)-\beta}{\Delta_1(1-\alpha)-\beta}$. Note that **Claim 2** along with the assumption that $\epsilon'^2 < \Delta_1$ ensures that the numerator and the denominator of the fraction in the RHS are both positive. Thus we can solve for T as follows,

$$(T-1) \log\left(\frac{1}{\alpha}\right) = \log\left[\frac{\Delta_1(1-\alpha)-\beta}{\epsilon'^2(1-\alpha)-\beta}\right] \implies T = \mathcal{O}\left(\log\left[\frac{\Delta_1}{\epsilon'^2 - \left(\frac{\epsilon'_2 + \gamma \cdot \frac{\epsilon'_3}{b}}{\gamma-1}\right)}\right]\right)$$

In the second equality above we have estimated the expression for T after substituting $\eta' = \frac{b}{\gamma c_1}$ in the expressions for α and β .

Proof of claim 1 : $\alpha \in (0, 1)$

We recall that we have set $\eta' = \frac{b}{\gamma c_1}$. This implies that, $\alpha = 1 - \frac{b^2}{c_1} \cdot \left(\frac{1}{\gamma} - \frac{1}{\gamma^2}\right)$. Hence $\alpha > 0$ is ensured by the assumption that $\gamma > \frac{b^2}{c_1}$. And $\alpha < 1$ is ensured by the assumption that $\gamma > 1$

Proof of claim 2 : $0 < \epsilon'^2(1 - \alpha) - \beta$

We note the following,

$$\begin{aligned}
-\frac{1}{\epsilon'^2} \cdot (\epsilon'^2(1-\alpha) - \beta) &= \alpha - \left(1 - \frac{\beta}{\epsilon'^2}\right) \\
&= 1 - \frac{b^2}{4c_1} + \left(\eta' \sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 - \left(1 - \frac{\beta}{\epsilon'^2}\right) \\
&= \frac{\eta'^2 c_2 + \eta' c_3}{\epsilon'^2} + \left(\eta' \sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 - \frac{b^2}{4c_1} \\
&= \frac{\left(\eta' \sqrt{c_2} + \frac{c_3}{2\sqrt{c_2}}\right)^2 - \frac{c_3^2}{4c_2}}{\epsilon'^2} + \left(\eta' \sqrt{c_1} - \frac{b}{2\sqrt{c_1}}\right)^2 - \frac{b^2}{4c_1} \\
&= \eta'^2 \left(\frac{1}{\epsilon'^2} \cdot \left(\sqrt{c_2} + \frac{c_3}{2\eta' \sqrt{c_2}} \right)^2 + \left(\sqrt{c_1} - \frac{b}{2\eta' \sqrt{c_1}} \right)^2 \right. \\
&\quad \left. - \frac{1}{\eta'^2} \left[\frac{b^2}{4c_1} + \frac{1}{\epsilon'^2} \left(\frac{c_3^2}{4c_2} \right) \right] \right)
\end{aligned}$$

Now we substitute $\eta' = \frac{b}{\gamma c_1}$ for the quantities in the expressions inside the parantheses to get,

$$\begin{aligned}
-\frac{1}{\epsilon'^2} \cdot (\epsilon'^2(1-\alpha) - \beta) &= \alpha - \left(1 - \frac{\beta}{\epsilon'^2}\right) = \eta'^2 \left(\frac{1}{\epsilon'^2} \cdot \left(\sqrt{c_2} + \frac{\gamma c_1 c_3}{2b\sqrt{c_2}} \right)^2 \right. \\
&\quad \left. + c_1 \cdot \left(\frac{\gamma}{2} - 1 \right)^2 - c_1 \frac{\gamma^2}{4} - \frac{1}{\epsilon'^2} \cdot \frac{\gamma^2 c_1^2 c_3^2}{4b^2 c_2} \right) \\
&= \eta'^2 \left(\frac{1}{\epsilon'^2} \cdot \left(\sqrt{c_2} + \frac{\gamma c_1 c_3}{2b\sqrt{c_2}} \right)^2 + c_1(1-\gamma) - \frac{1}{\epsilon'^2} \cdot \frac{\gamma^2 c_1^2 c_3^2}{4b^2 c_2} \right) \\
&= \frac{\eta'^2}{\epsilon'^2} \left(c_2 + \frac{\gamma c_1 c_3}{b} - \epsilon'^2 c_1(\gamma - 1) \right) \\
&= \frac{\eta'^2 c_1}{\epsilon'^2} \left((\epsilon'^2 + \frac{c_2}{c_1}) - \gamma \cdot \left(\epsilon'^2 - \frac{c_3}{b} \right) \right)
\end{aligned}$$

Therefore, $-\frac{1}{\epsilon'^2} (\epsilon'^2(1-\alpha) - \beta) < 0$ since by assumption $\epsilon'^2 > \frac{c_3}{b}$, and $\gamma > \frac{(\epsilon'^2 + \frac{c_2}{c_1})}{\epsilon'^2 - \frac{c_3}{b}}$

C Lemmas For Theorem 1

Lemma 3 (Lemma 1, [12]). *If \mathcal{D} is parity symmetric distribution on \mathbb{R}^n and σ is an α -Leaky ReLU then $\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\sigma(\mathbf{a}^\top \mathbf{x}) \mathbf{b}^\top \mathbf{x} \right] = \frac{1+\alpha}{2} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[(\mathbf{a}^\top \mathbf{x})(\mathbf{b}^\top \mathbf{x}) \right]$$

Lemma 4.

$$(f_{\mathbf{w}_*}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x}))^2 \leq (1+\alpha)^2 \left(\frac{1}{k} \sum_{i=1}^k \lambda_{\max}(A_i A_i^\top) \right) \|\mathbf{w}_* - \mathbf{w}\|^2 \|\mathbf{x}\|^2$$

Proof.

$$\begin{aligned} \left(f_{\mathbf{w}_*}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x}) \right)^2 &\leq \left(\frac{1}{k} \sum_{i=1}^k \sigma(\langle A_i^\top \mathbf{w}_*, \mathbf{x} \rangle) - \frac{1}{k} \sum_{i=1}^k \sigma(\langle A_i^\top \mathbf{w}, \mathbf{x} \rangle) \right)^2 \\ &\leq \frac{1}{k} \sum_{i=1}^k \left(\sigma(\langle A_i^\top \mathbf{w}_*, \mathbf{x} \rangle) - \sigma(\langle A_i^\top \mathbf{w}, \mathbf{x} \rangle) \right)^2 \\ &\leq \frac{(1+\alpha)^2}{k} \sum_{i=1}^k \langle A_i^\top \mathbf{w}_* - A_i^\top \mathbf{w}, \mathbf{x} \rangle^2 = \frac{(1+\alpha)^2}{k} \sum_{i=1}^k \left((\mathbf{w}_* - \mathbf{w})^\top A_i \mathbf{x} \right)^2 \\ &= \frac{(1+\alpha)^2}{k} \sum_{i=1}^k \|\mathbf{w}_* - \mathbf{w}\|^2 \|A_i \mathbf{x}\|^2 \leq \frac{(1+\alpha)^2}{k} \sum_{i=1}^k \|\mathbf{w}_* - \mathbf{w}\|^2 \lambda_{\max}(A_i A_i^\top) \|\mathbf{x}\|^2 \\ &\leq \frac{(1+\alpha)^2}{k} \left(\sum_{i=1}^k \lambda_{\max}(A_i A_i^\top) \right) \|\mathbf{w}_* - \mathbf{w}\|^2 \|\mathbf{x}\|^2 \end{aligned}$$