Proofs of Theorems

Proof of Theorem 1

Pre-multiplying (7) with $(M \otimes I_{d_w})$ where $M = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ the combined dynamics of $\Psi_k \triangleq [\tilde{\mathbf{w}}_k^\top, \alpha\mathbf{d}_k^\top]^\top \in \mathbb{R}^{2nd_w}$ from (7) and (8) can be written as

$$\Psi_{k+1} = \mathcal{W}\Psi_k + \mathbf{e}_k \tag{24}$$

where $\mathcal{W} \triangleq \begin{bmatrix} (\mathcal{W}_\beta \otimes I_{d_w}) & -n(M \otimes I_{d_w}) \\ \mathbf{0}_{nd_w} & (\mathcal{W}_\gamma \otimes I_{d_w}) \end{bmatrix}$ and $\mathbf{e}_k \triangleq \begin{bmatrix} \sqrt{2\alpha n}(M \otimes I_{d_w})\mathbf{v}_k \\ \alpha(\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k) \end{bmatrix}$. Taking the norm of (24) yields

$$\|\Psi_{k+1}\| \leq (1 - \beta\lambda_2)\|\Psi_k\| + \|\mathbf{e}_k\|, \tag{25}$$

where we use the results $\|\mathcal{W}\| \leq 1 - \beta\lambda_2$. Also, we have

$$\|\mathbf{e}_k\|^2 = 2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \tag{26}$$

where we used $\|M \otimes I_{d_w}\| \leq 1$. Squaring (25) and applying the identity $(x + y)^2 \leq (\theta + 1)x^2 + \left(\frac{x+1}{x}\right) y^2$ for any $\theta > 0$ (with $\theta = (1 - \beta\lambda_2)^{-1} - 1 > 0$) yields

$$\begin{aligned} \|\Psi_{k+1}\|^2 &\leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\|\mathbf{e}_k\|^2, & (27) \\ &\leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\Big(2\alpha n\|\mathbf{v}_k\|^2 + \alpha^2\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2\Big), & (28) \end{aligned}$$

where we substituted (26) in (28). Next, we need to establish a bound for $\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2$.

$$\|\mathbf{g}'_{k+1} - \mathbf{g}'_k\|^2 = \sum_{i\in\mathcal{V}} \left[\left(\frac{M_i}{m_i} \sum_{j=1}^{m_i} L_{i,j} + \frac{L'_i}{n} \right)^2 \|\boldsymbol{w}_{i,k+1} - \boldsymbol{w}_{i,k}\|^2 \right] \tag{29}$$

Let $\mathcal{B}_k \in \mathbf{B}$ be the randomness generated by the stochastic gradients and define $Y_i \triangleq \left(\frac{M_i}{m_i}\sum_{j=1}^{m_1} L_{ij} + \frac{L'_i}{n}\right)$. Then, taking the expectation of (29) w.r.t. $\mathbf{B}$ gives

$$\mathbb{E}_{\mathbf{B}}\|\mathbf{g}'_{k+1} - \mathbf{g}'_k\|^2 = \sum_{i\in\mathcal{V}} \left[\mathbb{E}_{\mathbf{B}}[Y_i^2] \times \|\boldsymbol{w}_{i,k+1} - \boldsymbol{w}_{i,k}\|^2 \right], \tag{30}$$

Applying the result from (82) in Lemma 6 to (30) yields

$$\mathbb{E}_{\mathbf{B}}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \leq L_1^2\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \tag{31}$$

Again, from (7) we can write

$$\begin{aligned} \mathbf{w}_{k+1} - \mathbf{w}_k &= \beta(\mathcal{L} \otimes I_{d_w})\mathbf{w}_k - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k, \\ &= \beta(\mathcal{L} \otimes I_{d_w})(\mathbf{w}_k - \mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k, & (32) \\ &= \beta(\mathcal{L} \otimes I_{d_w})\tilde{\mathbf{w}}_k - \alpha n\mathbf{d}_k + \sqrt{2\alpha n}\mathbf{v}_k. & (33) \end{aligned}$$

In (32) we use the fact that $(\mathcal{L} \otimes I_{d_w})(\mathbf{1}_n \otimes \bar{\boldsymbol{w}}_k) = \mathbf{0}_{nd_w}$. Then, taking the norm of (33) yields

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \leq \beta\lambda_n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\| \leq n\|\tilde{\mathbf{w}}_k\| + n\|\alpha\mathbf{d}_k\| + \sqrt{2\alpha n}\|\mathbf{v}_k\|, \quad (34)$$

where we use first inequality in Condition 1. Squaring (34) and noting that $\|\Psi_k\|^2 = \|\tilde{\mathbf{w}}_k\|^2 + \|\alpha\mathbf{d}_k\|^2$ gives

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \leq 4n^2\|\Psi_k\|^2 + 4\alpha n\|\mathbf{v}_k\|^2. \quad (35)$$

Thereby, substituting (33) in(31) results in

$$\mathbb{E}_{\mathbf{B}}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \leq 4n^2L_1^2\|\Psi_k\|^2 + 4\alpha nL_1^2\|\mathbf{v}_k\|^2. \quad (36)$$

Taking $\mathbb{E}_{\mathbf{B}}[\cdot]$ of (28) and substituting (36) in it gives

$$\mathbb{E}_{\mathbf{B}}[\|\Psi_{k+1}\|^2] \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right)\|\Psi_k\|^2 + \frac{2\alpha n(1+2\alpha^2L_1^2)}{\beta\lambda_2}\|\mathbf{v}_k\|^2, \quad (37)$$

Finally, taking the total expectation of (37) yields

$$\mathbb{E}[\|\Psi_{k+1}\|^2] \leq \left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right)\mathbb{E}[\|\Psi_k\|^2] + \frac{2\alpha n^2d_w(1+2\alpha^2L_1^2)}{\beta\lambda_2}, \quad (38)$$

where $\mathbb{E}[\|\mathbf{v}_k\|^2] \leq nd_w$. Note, that $\left(1 - \beta\lambda_2 + \frac{4\alpha^2n^2L_1^2}{\beta\lambda_2}\right) \in (0,1)$ from condition 1, which assures the convergence of $\mathbb{E}[\|\Psi_{k+1}\|^2]$. Further, iteratively using (38) we establish the rate of consensus for the proposed GT-DULA which is presented in Theorem 1.

Proof of Theorem 2

We start off with the average dynamics generated by the GT-DULA. From (5) and noting that $\boldsymbol{d}_{i,0} = \hat{\boldsymbol{g}}_{i,0}$, it is trivial that $\sum_{i\in\mathcal{V}}\boldsymbol{d}_{i,k} = \sum_{i\in\mathcal{V}}\hat{\boldsymbol{g}}_{i,k}$. Thereafter, from (4), the following average dynamics can be established.

$$\bar{\boldsymbol{w}}_{k+1} = \bar{\boldsymbol{w}}_k - \alpha\boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k, \quad (39)$$

where $\boldsymbol{G}_k \triangleq \sum_{i\in\mathcal{V}}\hat{\boldsymbol{g}}_{i,k}$ and $\bar{\boldsymbol{v}} \sim \mathcal{N}(\mathbf{0}_{d_w}, I_{d_w})$. Next, we split the gradient term $\boldsymbol{G}_k$ as

$$\boldsymbol{G}_k = \overline{\nabla E}_k + \xi_k + \zeta_k, \quad (40)$$

where $\overline{\nabla E}_k \triangleq \sum_{i\in\mathcal{V}}\nabla E_i(\bar{\boldsymbol{w}}_k)$, $\xi_k \triangleq \sum_{i\in\mathcal{V}}\left(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k)\right)$ and $\zeta_k \triangleq \sum_{i\in\mathcal{V}}\left(\widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \nabla E_i(\bar{\boldsymbol{w}}_k)\right)$. Hence, in essence, $\overline{\nabla E}_k$ represents the gradient computed at the average sample, ξ_k encompasses the deviation of due to local gradients and is a consequence of the distributed learning setup, and ζ_k is the error incurred by mini-batch gradients. Also, note that since the stochastic gradient $\widehat{\nabla E}_i(\cdot)$ is an unbiased estimator of the true gradient $\nabla E_i(\cdot)$, we have $\mathbb{E}_{\mathbf{B}}[\zeta_k] = \mathbf{0}_{d_w}$. From our succeeding analysis we can conclude that as long as the sources of additional deviation of the net gradient in (39) from $\overline{\nabla E}$ are bounded, the resulting algorithm asymptotically converges to some neighborhood of the target distribution.

With (40) in mind, (39) can be written as a stochastic differential equation in continuous-time as below.

$$d\bar{\boldsymbol{w}}(t) = -\boldsymbol{G}_k dt + \sqrt{2}d\boldsymbol{B}(t) = -\left(\overline{\nabla E}_k + \xi_k + \zeta_k\right)dt + \sqrt{2}d\boldsymbol{B}(t), \tag{41}$$

where $t \in [t_k, t_{k+1})$ such that continuous time $t_k = \alpha k$ corresponds to discrete-time instant k for any $k \geq 0$ and $\boldsymbol{B}(t)$ is a d_w-dimensional Brownian motion. Next, defining $\boldsymbol{y}_{1,k} \triangleq \bar{\boldsymbol{w}}_k$, $\boldsymbol{y}_{2,k} \triangleq \tilde{\mathbf{w}}_k$, $\boldsymbol{y}_{3,k} \triangleq \mathcal{B}_k$ and $\boldsymbol{y}_k \triangleq [\boldsymbol{y}_{1,k}^\top, \boldsymbol{y}_{2,k}^\top, \boldsymbol{y}_{3,k}^\top]^\top$ and following a similar approach as in (33) of Bhar et al. (2022) we can write down the Fokker-Planck (FP) equation for (41) which gives the continuous-time evolution of the distribution of $\bar{\boldsymbol{w}}(t)$ as

$$\frac{\partial p(\bar{\boldsymbol{w}}(t))}{\partial t} = -\nabla \cdot \left[\int \sum_{\mathbf{B}} p(\bar{\boldsymbol{w}}(t)|\boldsymbol{y}_k)\Big(- \overline{\nabla E}_k - \xi_k\Big) p(\boldsymbol{y}_k) d\boldsymbol{y}_k \right] + \nabla^2 p(\bar{\boldsymbol{w}}(t)), \tag{42}$$

where we used the fact that $\sum_{\mathbb{B}} \zeta_k p(\boldsymbol{y}_{k,3}) = \mathbb{E}_{\mathbb{B}}[\zeta_k] = \mathbf{0}_{d_w}$. Thereafter, proceeding with (42) in the same way as in (S101)-(S125) from Parayil et al. (2020) yields

$$\begin{aligned}\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) = -\frac{1}{2}\mathbb{E}\left\|\nabla \log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + \iint \left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 p(\bar{\boldsymbol{w}}(t)) d\boldsymbol{y}_k \\ + \iint \|\xi_k\|^2 p(\bar{\boldsymbol{w}}(t)) d\boldsymbol{y}_k,\end{aligned} \tag{43}$$

where $\overline{\nabla E}_t \triangleq \sum_{i \in \mathcal{V}} \nabla E_i(\bar{\boldsymbol{w}}(t))$. Next, we derive the bounds for $\mathbb{E}\|\overline{\nabla E}_t - \overline{\nabla E}_k\|^2$, $\mathbb{E}\|\xi_k\|^2$ and $\mathbb{E}\|\zeta_k\|^2$ individually. First, from Assumption 1, we have

$$\|\xi_k\|^2 = \left\|\sum_{i \in \mathcal{V}} \left(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k)\right)\right\|^2 \leq n \sum_{i \in \mathcal{V}} \Big(Y_i^2 \|\boldsymbol{w}_{i,k} - \bar{\boldsymbol{w}}_k\|^2\Big), \tag{44}$$

where Y_i is defined in Lemma 6. Taking the expectation of (44) w.r.t. $\mathbf{B}$ and thereby applying (82) from Lemma 6 we get

$$\mathbb{E}_{\mathbf{B}}[\|\xi_k\|^2] \leq n L_1^2 \|\tilde{\mathbf{w}}_k\|^2, \tag{45}$$

which after marginalizing w.r.t. $\boldsymbol{y}_k$ yields

$$\mathbb{E}[\|\xi_k\|^2] \leq n L_1^2 \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2], \tag{46}$$

Next, we analyze $\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2$.

$$\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \leq n\bar{L}^2 \|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2, \tag{47}$$

where $\bar{L}^2 \triangleq \sum_{i \in \mathcal{V}} \left(\sum_{j=1}^{M_i} L_{ij} + \frac{L_i'}{n}\right)^2$. Integrating (41) from t_k to $t \in [t_k, t_{k+1})$ gives

$$\begin{aligned}\|\bar{\boldsymbol{w}}(t) - \bar{\boldsymbol{w}}(t_k)\|^2 &\leq \left\| - \boldsymbol{G}_k(t - t_k) + \sqrt{2}\Big(\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\Big)\right\|^2, \\ &\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \|\boldsymbol{G}_k(t - t_k)\|^2 - 2\sqrt{2}\boldsymbol{S}_k, && (48) \\ &\leq 2\|\boldsymbol{B}(t) - \boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k, && (49)\end{aligned}$$

where $\boldsymbol{S}_k \triangleq \Big(\boldsymbol{B}(t)-\boldsymbol{B}(t_k)\Big)^\top\Big(\boldsymbol{G}_k(t-t_k)\Big)$. In (49) we use $t-t_k < t_{k+1}-t_k = \alpha$ for any $t \in [t_k, t_{k+1})$. Substituting (49) in (47) results in

$$\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le n\bar{L}^2\Big[2\|\boldsymbol{B}(t)-\boldsymbol{B}(t_k)\|^2 + \alpha^2\|\boldsymbol{G}_k\|^2 - 2\sqrt{2}\boldsymbol{S}_k\Big], \tag{50}$$

Now, the $\|\boldsymbol{G}_k\|^2$ can be bound as

$$\begin{aligned}
\|\boldsymbol{G}_k\|^2 &= \left\|\sum_{i\in\mathcal{V}} \widehat{\nabla E}_i(\boldsymbol{w}_{i,k})\right\|^2 = \left\|\sum_{i\in\mathcal{V}}\Big(\widehat{\nabla E}_i(\boldsymbol{w}_{i,k}) - \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) + \widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*)\Big)\right\|^2, \\
&\le 2\,\|\xi_k\|^2 + 2\left\|\sum_{i\in\mathcal{V}}\Big(\widehat{\nabla E}_i(\bar{\boldsymbol{w}}_k) - \widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*)\Big)\right\|^2 \le 2\|\xi_k\|^2 + 2n\left(\sum_{i\in\mathcal{V}} Y_i^2\right)\|\bar{\boldsymbol{w}}_k - \hat{\boldsymbol{w}}^*\|^2, \\
&\le 2n\sum_{i\in\mathcal{V}}\Big(Y_i^2\|\boldsymbol{w}_{i,k}-\bar{\boldsymbol{w}}_k\|^2\Big) + 4n\left(\sum_{i\in\mathcal{V}} Y_i^2\right)\Big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\hat{\boldsymbol{w}}^*\|^2\Big),
\end{aligned} \tag{51}$$

where we used the bound from (44) and $\hat{\boldsymbol{w}}^*$ is some local extremum of $\sum_{i\in\mathcal{V}}\widehat{\nabla E}_i(\cdot)$, i.e., $\sum_{i\in\mathcal{V}}\widehat{\nabla E}_i(\hat{\boldsymbol{w}}^*) = \mathbf{0}_{d_w}$. Thereafter, taking the expectation w.r.t. $\mathbf{B}$ of (51) and using (82) and (83) in Lemma 6 yields

$$\mathbb{E}_{\mathbf{B}}[\|\boldsymbol{G}_k\|^2] \le 2nL_1^2\|\tilde{\mathbf{w}}_k\|^2 + 4nL_2^2\big(\|\bar{\boldsymbol{w}}_k\|^2 + \|\hat{\boldsymbol{w}}^*\|^2\big), \tag{52}$$

Next, taking $\mathbb{E}_{\mathbf{B}}[\cdot]$ of (50) and substituting (52) yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{B}}\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le 2n\bar{L}^2\|\boldsymbol{B}(t)-\boldsymbol{B}(t_k)\|^2 + 2n^2\alpha^2L_1^2\bar{L}^2\|\tilde{\mathbf{w}}_k\|^2 + 4n^2\alpha^2L_2^2\bar{L}^2\|\bar{\boldsymbol{w}}_k\|^2 \\
+ 4n^2\alpha^2L_2^2\bar{L}^2\|\hat{\boldsymbol{w}}^*\|^2 - 2\sqrt{2}n\bar{L}^2\boldsymbol{S}_k',
\end{aligned} \tag{53}$$

where $\boldsymbol{S}_k' \triangleq \Big(\boldsymbol{B}(t)-\boldsymbol{B}(t_k)\Big)^\top\Big(\mathbb{E}_{\mathbf{B}}[\boldsymbol{G}_k](t-t_k)\Big)$. Again, marginalizing (53) w.r.t. $\boldsymbol{y}_k$ gives

$$\mathbb{E}\left\|\overline{\nabla E}_t - \overline{\nabla E}_k\right\|^2 \le 2n\alpha\bar{L}^2d_w + 2n^2\alpha^2L_1^2\bar{L}^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}, \tag{54}$$

where $\mathbb{E}[\|\hat{\boldsymbol{w}}^*\|^2] \le C_{\hat{\boldsymbol{w}}^*}$ for any choice of stochastic gradient. For details on the derivation of (54), refer to (S135)-(S141) in Parayil et al. (2020). Finally, incorporating (46) and (54) in (43) yields

$$\begin{aligned}
\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) &\le -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + 2n\alpha\bar{L}^2d_w + (2n^2\alpha^2L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \\
&\quad + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}, \\
&\le -\frac{1}{2}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2 + (2n^2\alpha^2L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + f,
\end{aligned} \tag{55}$$

where $f \triangleq 2n\alpha\bar{L}^2d_w + 4n^2\alpha^2L_2^2\bar{L}^2C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}$. Here we utilize the LSI assumption in Assumption 3 and putting $g(\bar{\boldsymbol{w}}) = \frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}$ results in the following results

$$F\Big(p(\bar{\boldsymbol{w}}(t))\Big) \le \frac{1}{2\rho_U}\mathbb{E}\left\|\nabla\log\left(\frac{p(\bar{\boldsymbol{w}}(t))}{p^*(\bar{\boldsymbol{w}}(t))}\right)\right\|^2. \tag{56}$$

Using (56) in (55) yields

$$\dot{F}\Big(p(\bar{\boldsymbol{w}}(t))\Big) \leq -\rho_U F\Big(p(\bar{\boldsymbol{w}}(t))\Big) + f + (2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]. \tag{57}$$

Next, integrating (57) w.r.t t within $t \in [t_k, t_{k+1}]$ and utilizing $t_{k+1} - t_k < \alpha$ gives us the evolution of the KL divergence of the posterior generated by GT-DULA samples as follows.

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \exp(-\alpha\rho_U) F\Big(p(\bar{\boldsymbol{w}}_k)\Big) + \frac{1-\exp(-\alpha\rho_U)}{\rho_U}\Big[f_k + (2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big]. \tag{58}$$

Using (58) iteratively yields

$$\begin{aligned} F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) &\leq \exp(-\alpha\rho_U(k+1)) F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1-\exp(-\alpha\rho_U)}{\rho_U}\sum_{\ell=0}^{k}\Big[\exp(-\alpha\rho_U(k-\ell))\times \\ &\qquad \Big(f + (2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_\ell\|^2]\Big)\Big], \\ &\leq \exp(-\alpha\rho_U(k+1)) F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{f}{\rho_U} + \frac{2n^2\alpha^2 L_1^2 \bar{L}^2 + nL_1^2}{\rho_U}\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2], \end{aligned} \tag{59}$$

where we used $\mathbb{E}[\|\tilde{\mathbf{w}}_\ell\|^2] \leq \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$ for any $\ell \in [0,k]$ and $\sum_{\ell=0}^{k}\exp(-\alpha\rho_U(k-\ell)) \leq \sum_{\ell=0}^{\infty}\exp(-\alpha\rho_U\ell) = \frac{1}{1-\exp(-\alpha\rho_U)}$. Finally, substituting (13) in (59), we get the rate of convergence of the KL divergence of the generated posteriors which is presented in Theorem 2

Proof of Corollary 3

From (17), $F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq \epsilon$ can be satisfied if (i) $\exp(-\alpha\rho_U(k+1))F\Big(p(\bar{\boldsymbol{w}}_0)\Big) \leq \frac{\epsilon}{5}$, (ii) $C_d\sigma^k \leq \frac{\epsilon}{5}$ and (iii) $O_{GT} \leq \frac{3\epsilon}{5}$. (i) and (ii) respectively give the minimum k values in (21). Finally, (iii) can be satisfied if the we simultaneously satisfy the following conditions

$$\frac{8d_w L_1^4 \bar{L}^2 n^4 \alpha^5}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{4d_w L_1^2 \bar{L}^2 n^4 \alpha^4}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{4d_w L_1^4 n^3 \alpha^3}{(1-\sigma)\rho_U\beta\lambda_2} \leq \frac{\epsilon}{5}, \tag{60}$$

$$\frac{4L_2^2 \bar{L}^2 (C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}) n^2 \alpha^2}{\rho_U} \leq \frac{\epsilon}{5}, \tag{61}$$

$$\frac{2d_w L_1^2 n^3 \alpha}{(1-\sigma)\rho_U\beta\lambda_2} + \frac{2d_w \bar{L}^2 n\alpha}{\rho_U} \leq \frac{\epsilon}{5}, \tag{62}$$

each of which result in bounds in (20) respectively.

Proof of Theorem 4

From (8) we have

$$\tilde{\mathbf{d}}_{k+1} = (\mathcal{W}_\gamma \otimes I_{d_w})\tilde{\mathbf{d}}_k + (\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k) - \mathbf{1}_n \otimes (\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k). \tag{63}$$

Taking the square of the norm of (63) yields

$$\|\tilde{\mathbf{d}}_{k+1}\|^2 \leq (1-\gamma\lambda_2)\|\tilde{\mathbf{d}}_k\|^2 + \frac{2}{\gamma\lambda_2}\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 + \frac{2n}{\gamma\lambda_2}\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|, \tag{64}$$

and thereafter taking the total expectation of (64) gives

$$\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] + \frac{2}{\gamma\lambda_2}\mathbb{E}[\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2] + \frac{2n}{\gamma\lambda_2}\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|], \tag{65}$$

Next, we derive bounds for the individual terms on the right hand side of (65). First, from (36) we have

$$\mathbb{E}[\|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2] \leq 4n^2 L_1^2 \mathbb{E}[\|\Psi_k\|^2] + 4\alpha n^2 L_1^2 d_w. \tag{66}$$

Then,

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|^2] = \frac{L_2^2}{n}\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2], \tag{67}$$

and from (39) we can write

$$\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k = -\alpha \boldsymbol{G}_k + \sqrt{2\alpha}\bar{\boldsymbol{v}}_k. \tag{68}$$

Thus, from (68)

$$\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2 \leq 2\alpha^2 \|\boldsymbol{G}_k\|^2 + 4\alpha\|\bar{\boldsymbol{v}}_k\|^2. \tag{69}$$

Taking the total expectation of (69) and substituting (52) yields

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1} - \bar{\boldsymbol{w}}_k\|^2] \leq 4n\alpha^2 L_1^2 \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8n\alpha^2 L_2^2 \big(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}\big) + 4n\alpha d_w, \tag{70}$$

where we also use $\mathbb{E}[\|\bar{\boldsymbol{v}}_k\|^2] \leq n d_w$. Substituting (70) in (67) gives

$$\mathbb{E}[\|\bar{\boldsymbol{u}}_{k+1} - \bar{\boldsymbol{u}}_k\|^2] \leq 4\alpha^2 L_1^2 L_2^2 \mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8\alpha^2 L_2^4 \big(C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*}\big) + 4\alpha d_w L_2^2, \tag{71}$$

Finally, substituting (66) and (71) in (65) leads to

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)\mathbb{E}[\|\tilde{\mathbf{d}}_k\|^2] &+ \frac{8nL_1^2(n+\alpha^2 L_2^2)}{\gamma\lambda_2}\mathbb{E}[\|\Psi_k\|^2] + \frac{16n\alpha^2 L_2^4 (C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}{\gamma\lambda_2} \\ &+ \frac{8n\alpha d_w (nL_1^2 + L_2^2)}{\gamma\lambda_2},\end{aligned} \tag{72}$$

where we used $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$. Next, iteratively using (72) yields

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{d}}_{k+1}\|^2] \leq (1-\gamma\lambda_2)^{k+1}\mathbb{E}[\|\tilde{\mathbf{d}}_0\|^2] &+ \frac{8nL_1^2(n+\alpha^2 L_2^2)}{\gamma^2\lambda_2^2}\mathbb{E}[\|\Psi_k\|^2] + \frac{16n\alpha^2 L_2^4 (C_{\bar{\boldsymbol{w}}} + C_{\hat{\boldsymbol{w}}^*})}{\gamma^2\lambda_2^2} \\ &+ \frac{8n\alpha d_w (nL_1^2 + L_2^2)}{\gamma^2\lambda_2^2},\end{aligned} \tag{73}$$

where we use $\mathbb{E}[\|\Psi_\ell\|^2] \leq \mathbb{E}[\|\Psi_k\|^2]$ for all $\ell \in [0,k]$ and $\sum_{\ell=0}^{k}(1-\gamma\lambda_2)^\ell < \sum_{\ell=0}^{\infty}(1-\gamma\lambda_2)^\ell = \frac{1}{\gamma\lambda_2}$ has been used. Next, substituting (13) in (73) yields our final result for the gradient error presented in Theorem 4 below.

Useful Lemmas

Lemma 5 *For any $k \geq 0$ the following bound holds.*

$$\mathbb{E}[\|\bar{\boldsymbol{w}}(t_k)\|^2] \leq C_{\bar{\boldsymbol{w}}}, \tag{74}$$

where

$$C_{\bar{\boldsymbol{w}}} = \max\Bigg\{\mathbb{E}[\|\bar{\boldsymbol{w}}_0\|^2], \frac{1}{\rho_U^2 - 16n^2\alpha^2 L_2^2 \bar{L}^2}\Bigg(2\rho_U^2 c_1 + 4\rho_U F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + 8\big(n\alpha\bar{L}^2 d_w + 2n^2\alpha^2 L_2^2 \bar{L}^2 C_{\hat{\boldsymbol{w}}^*}\big) + 4(2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Psi_0\|^2] + B_{GT}\big)\Bigg)\Bigg\}. \tag{75}$$

Proof: We follow a similar approach as used in Lemma S6 of Parayil et al. (2020) which uses induction to derive this bound. Assuming that $\mathbb{E}[\|\bar{\boldsymbol{w}}_\ell\|^2] \leq C_{\bar{\boldsymbol{w}}}$ for all $\ell \leq k$, we need to prove $\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq C_{\bar{\boldsymbol{w}}}$. From (S252) in Parayil et al. (2020), we have

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big), \tag{76}$$

where $\mathbb{E}_{p^*}[\|\bar{\boldsymbol{w}}\|^2] \leq c_1$. From (59), we can write

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2]\Big). \tag{77}$$

Note, in the right hand side of (77), we have used the bound $C_{\bar{\boldsymbol{w}}}$ since it is assumed to hold up to $\ell \leq k$. Next, from Theorem 1, we can write $\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] \leq \mathbb{E}[\|\Phi_0\|^2] + B_{GT}$ which when substituted in (77) yields

$$F\Big(p(\bar{\boldsymbol{w}}_{k+1})\Big) \leq F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{1}{\rho_U}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big). \tag{78}$$

Next, substituting (78) in (76) results in

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq 2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big). \tag{79}$$

For induction, we need to enforce $\mathbb{E}[\|\bar{\boldsymbol{w}}_{k+1}\|^2] \leq C_{\bar{\boldsymbol{w}}}$, thus, from (79) we have

$$2c_1 + \frac{4}{\rho_U} F\Big(p(\bar{\boldsymbol{w}}_0)\Big) + \frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2 d_w + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\bar{\boldsymbol{w}}} + 4n^2\alpha^2 L_2^2\bar{L}^2 C_{\hat{\boldsymbol{w}}^*} + (2n^2\alpha^2 L_1^2\bar{L}^2 + nL_1^2)\big(\mathbb{E}[\|\Phi_0\|^2] + B_{GT}\big)\Big) \leq C_{\bar{\boldsymbol{w}}}. \tag{80}$$

i.e.,

$$\left(1-\frac{16n^2\alpha^2L_2^2\bar{L}^2}{\rho_U^2}\right)C_{\bar{\boldsymbol{w}}}\geq 2c_1+\frac{4}{\rho_U}F\Big(p(\bar{\boldsymbol{w}}_0)\Big)+\frac{4}{\rho_U^2}\Big(2n\alpha\bar{L}^2d_w+4n^2\alpha^2L_2^2\bar{L}^2C_{\hat{\boldsymbol{w}}^*}$$
$$+\left(2n^2\alpha^2L_1^2\bar{L}^2+nL_1^2\right)\left(\mathbb{E}[\|\Phi_0\|^2]+B_{GT}\right)\Big). \tag{81}$$

For $C_{\bar{\boldsymbol{w}}}$ to exits we need $\alpha<\frac{\rho_U}{4nL_2\bar{L}}$, and the value in (75) follows from (81). ■

Lemma 6 *For any $i\in\mathcal{V}$, let $Y_i\triangleq\left(\frac{M_i}{m_i}\sum_{j=1}^{m_1}L_{ij}+\frac{L_i'}{n}\right)$, then we have the following bounds.*

$$\mathbb{E}_{\mathbf{B}}[Y_i^2]\leq L_1^2,\qquad\forall\, i\in\mathcal{V}, \tag{82}$$

and

$$\mathbb{E}_{\mathbf{B}}\left[\sum_{i\in\mathcal{V}}Y_i^2\right]\leq L_2^2, \tag{83}$$

where L_1^2 and L_2^2 are defined in (89) and (90) respectively.

Proof : We have

$$\mathbb{E}_{\mathbf{B}}[Y_i]=\mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i}L_{i,j}\right]+\frac{L_i'}{n}, \tag{84}$$

Now, it is obvious that for full gradient $\mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i}L_{i,j}\right]=\sum_{j=1}^{M_i}L_{ij}$. In the case of stochastic gradient, let the set of data in the mini-batch of the i-th agent at k-th time step be represented by $\mathcal{B}_{i,k}\in\mathbf{B}_i\subset\mathbf{B}$ where $\mathbf{B}_i$ is the set of all possible mini-batches for the i-th agent. Then the total number of mini-batches possible for any $i\in\mathcal{V}$ is $|\mathbf{B}_i|\triangleq b_i=\binom{M_i}{m_i}$. Now, the probability of choosing any one of these is equal, i.e., $p(\mathcal{B}_{i,k})=b_i^{-1}$. Next, the number of mini-batches that would contain $x_i^j\in\boldsymbol{X}_i$ is $\binom{M_i-1}{m_i-1}$. Thus, the term $L_{i,j}$ for any $j\in\{1,2,\ldots,M_i\}$ will show up in $\binom{M_i-1}{m_i-1}$ of the mini-batches in $\mathbf{B}_i$. Noting that $\mathbf{B}=\bigcup_{i\in\mathcal{V}}\mathbf{B}_i$ and $\mathbf{B}_i\cap\mathbf{B}_i=\emptyset$ for $i\neq j$, we can write

$$\mathbb{E}_{\mathbf{B}}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i}L_{i,j}\right]=\mathbb{E}_{\mathbf{B}_i}\left[\frac{M_i}{m_i}\sum_{j=1}^{m_i}L_{i,j}\right],$$
$$=\frac{M_i}{m_i}b_i^{-1}\binom{M_i-1}{m_i-1}\sum_{j=1}^{M_i}L_{i,j}=\sum_{j=1}^{M_i}L_{i,j}. \tag{85}$$

Thus, substituting (85) in (84) gives

$$\mathbb{E}_{\mathbf{B}}[Y_i]=L_i+\frac{L_i'}{n}, \tag{86}$$

where $L_i \triangleq \sum_{j=1}^{M_i} L_{i,j}$. Next, we have

$$\mathrm{Var}(Y_i) = \mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} + \frac{L_i'}{n} - \sum_{j=1}^{M_i} L_{i,j} - \frac{L_i'}{n}\right)^2\right], \tag{87}$$

$$= \mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} - \sum_{j=1}^{M_i} L_{i,j}\right)^2\right] = \sigma_{L,i}^2. \tag{88}$$

Note $\sigma_{L,i}^2$ sure to exist for all $i \in \mathcal{V}$ since the number of data points are fixed and finite, hence, $\mathbb{E}_{\mathbf{B}}\left[\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i} L_{i,j} - \sum_{j=1}^{M_i} L_{i,j}\right)^2\right]$ is finite. Thus, $\mathbb{E}_{\mathbf{B}}[Y_i^2] = \mathbb{E}_{\mathbf{B}}[Y_i]^2 + \mathrm{Var}[Y_i] \leq \left(L_i + \frac{L_i'}{n}\right)^2 + \sigma_{L,i}^2$. Defining

$$L_1^2 \triangleq L^2 + \sigma_L^2, \tag{89}$$

where $L \triangleq \max_{i\in\mathcal{V}}\left\{L_i + \frac{L_i'}{n}\right\}$ and $\sigma_L^2 \triangleq \max_{i\in\mathcal{V}}\{\sigma_{L,i}^2\}$, we derive (82). Finally, using $\mathbb{E}_{\mathbf{B}}\left[\sum_{i\in\mathcal{V}} Y_i^2\right] = \sum_{i\in\mathcal{V}} \mathbb{E}_{\mathbf{B}}[Y_i^2]$ and denoting

$$L_2^2 \triangleq \sum_{i\in\mathcal{V}}\left(L_i + \frac{L_i'}{n}\right)^2 + \sum_{i\in\mathcal{V}} \sigma_{L,i}^2. \tag{90}$$

the result in (83) can be derived. ■