

## Gradient Tracking based Decentralized Bayesian Learning: Scalability Enhancement and Data-Heterogeneity Awareness in Gradient Tracking based Decentralized Bayesian Learning (Appendix)

### 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

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

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

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

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

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

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

$$\|\Psi_{k+1}\|^2 \leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2}\|\mathbf{e}_k\|^2, \quad (27)$$

$$\leq (1 - \beta\lambda_2)\|\Psi_k\|^2 + \frac{1}{\beta\lambda_2} \left( 2\alpha n \|\mathbf{v}_k\|^2 + \alpha^2 \|\hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k\|^2 \right), \quad (28)$$

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

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

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

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

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

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

Again, from (7) we can write

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

$$= \beta(\mathcal{L} \otimes I_{d_w}) \tilde{\mathbf{w}}_k - \alpha n \mathbf{d}_k + \sqrt{2\alpha n} \mathbf{v}_k. \quad (33)$$

In (32) we use the fact that  $(\mathcal{L} \otimes I_{d_w})(\mathbf{1}_n \otimes \bar{\mathbf{w}}_k) = \mathbf{0}_{nd_w}$ . 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  $\mathbf{d}_{i,0} = \hat{\mathbf{g}}_{i,0}$ , it is trivial that  $\sum_{i \in \mathcal{V}} \mathbf{d}_{i,k} = \sum_{i \in \mathcal{V}} \hat{\mathbf{g}}_{i,k}$ . Thereafter, from (4), the following average dynamics can be established.

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

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

$$\mathbf{G}_k = \bar{\nabla} E_k + \xi_k + \zeta_k, \quad (40)$$

where  $\bar{\nabla} E_k \triangleq \sum_{i \in \mathcal{V}} \nabla E_i(\bar{\mathbf{w}}_k)$ ,  $\xi_k \triangleq \sum_{i \in \mathcal{V}} \left( \widehat{\nabla} E_i(\mathbf{w}_{i,k}) - \widehat{\nabla} E_i(\bar{\mathbf{w}}_k) \right)$  and  $\zeta_k \triangleq \sum_{i \in \mathcal{V}} \left( \widehat{\nabla} E_i(\bar{\mathbf{w}}_k) - \nabla E_i(\bar{\mathbf{w}}_k) \right)$ . Hence, in essence,  $\bar{\nabla} E_k$  represents the gradient computed at the average sample,  $\xi_k$  encompasses the deviation of due to local gradients and is a consequence of the distributed learning setup, and  $\zeta_k$  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  $\bar{\nabla} E$  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.

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

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  $\mathbf{B}(t)$  is a  $d_w$ -dimensional Brownian motion. Next, defining  $\mathbf{y}_{1,k} \triangleq \bar{\mathbf{w}}_k$ ,  $\mathbf{y}_{2,k} \triangleq \tilde{\mathbf{w}}_k$ ,  $\mathbf{y}_{3,k} \triangleq \mathcal{B}_k$  and  $\mathbf{y}_k \triangleq [\mathbf{y}_{1,k}^\top, \mathbf{y}_{2,k}^\top, \mathbf{y}_{3,k}^\top]^\top$  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  $\bar{\mathbf{w}}(t)$  as

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

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

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

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

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

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

$$\mathbb{E}_{\mathbf{B}}[\left\| \xi_k \right\|^2] \leq n L_1^2 \left\| \tilde{\mathbf{w}}_k \right\|^2, \quad (45)$$

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

$$\mathbb{E}[\left\| \xi_k \right\|^2] \leq n L_1^2 \mathbb{E}[\left\| \tilde{\mathbf{w}}_k \right\|^2], \quad (46)$$

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

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

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

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

$$\leq 2 \left\| \mathbf{B}(t) - \mathbf{B}(t_k) \right\|^2 + \alpha^2 \left\| \mathbf{G}_k \right\|^2 - 2\sqrt{2} \mathbf{S}_k, \quad (49)$$

where  $\mathbf{S}_k \triangleq \left( \mathbf{B}(t) - \mathbf{B}(t_k) \right)^\top \left( \mathbf{G}_k(t - t_k) \right)$ . 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

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

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

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

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

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

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

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

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

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

where  $\mathbb{E}[\|\hat{\mathbf{w}}^*\|^2] \leq C_{\hat{\mathbf{w}}^*}$  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

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

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

$$F \left( p(\bar{\mathbf{w}}(t)) \right) \leq \frac{1}{2\rho_U} \mathbb{E} \left\| \nabla \log \left( \frac{p(\bar{\mathbf{w}}(t))}{p^*(\bar{\mathbf{w}}(t))} \right) \right\|^2. \quad (56)$$

Using (56) in (55) yields

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

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.

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

Using (58) iteratively yields

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

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),  $F\left(p(\bar{\mathbf{w}}_{k+1})\right) \leq \epsilon$  can be satisfied if (i)  $\exp(-\alpha\rho_U(k+1)) F\left(p(\bar{\mathbf{w}}_0)\right) \leq \frac{\epsilon}{5}$ , (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

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

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

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

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

### Proof of Theorem 4

From (8) we have

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

Taking the square of the norm of (63) yields

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

and thereafter taking the total expectation of (64) gives

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

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

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

Then,

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

and from (39) we can write

$$\bar{\mathbf{w}}_{k+1} - \bar{\mathbf{w}}_k = -\alpha\mathbf{G}_k + \sqrt{2\alpha}\bar{\mathbf{v}}_k. \quad (68)$$

Thus, from (68)

$$\|\bar{\mathbf{w}}_{k+1} - \bar{\mathbf{w}}_k\|^2 \leq 2\alpha^2\|\mathbf{G}_k\|^2 + 4\alpha\|\bar{\mathbf{v}}_k\|^2. \quad (69)$$

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

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

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

$$\mathbb{E}[\|\bar{\mathbf{u}}_{k+1} - \bar{\mathbf{u}}_k\|^2] \leq 4\alpha^2L_1^2L_2^2\mathbb{E}[\|\tilde{\mathbf{w}}_k\|^2] + 8\alpha^2L_2^4(C_{\bar{\mathbf{w}}} + C_{\bar{\mathbf{w}}^*}) + 4\alpha d_wL_2^2, \quad (71)$$

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

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

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

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

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.*

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

where

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

**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  $\mathbb{E}[\|\bar{\mathbf{w}}_\ell\|^2] \leq C_{\bar{\mathbf{w}}}$  for all  $\ell \leq k$ , we need to prove  $\mathbb{E}[\|\bar{\mathbf{w}}_{k+1}\|^2] \leq C_{\bar{\mathbf{w}}}$ . From (S252) in [Parayil et al. \(2020\)](#), we have

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

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

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

Note, in the right hand side of (77), we have used the bound  $C_{\bar{\mathbf{w}}}$  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

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

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

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

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

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

i.e.,

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

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

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

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

and

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

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

**Proof :** We have

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

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  $x_i^j \in \mathbf{X}_i$  is  $\binom{M_i-1}{m_i-1}$ . Thus, the term  $L_{i,j}$  for any  $j \in \{1, 2, \dots, M_i\}$  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

$$\begin{aligned} \mathbb{E}_{\mathbf{B}} \left[ \frac{M_i}{m_i} \sum_{j=1}^{m_i} L_{i,j} \right] &= \mathbb{E}_{\mathbf{B}_i} \left[ \frac{M_i}{m_i} \sum_{j=1}^{m_i} L_{i,j} \right], \\ &= \frac{M_i}{m_i} b_i^{-1} \binom{M_i-1}{m_i-1} \sum_{j=1}^{M_i} L_{i,j} = \sum_{j=1}^{M_i} L_{i,j}. \end{aligned} \quad (85)$$

Thus, substituting (85) in (84) gives

$$\mathbb{E}_{\mathbf{B}}[Y_i] = L_i + \frac{L'_i}{n}, \quad (86)$$



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

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

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

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,  $\mathbb{E}_{\mathbf{B}}[Y_i^2] = \mathbb{E}_{\mathbf{B}}[Y_i]^2 + \text{Var}[Y_i] \leq \left( L_i + \frac{L'_i}{n} \right)^2 + \sigma_{L,i}^2$ . Defining

$$L_1^2 \triangleq L^2 + \sigma_L^2, \quad (89)$$

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  $\mathbb{E}_{\mathbf{B}} [\sum_{i \in \mathcal{V}} Y_i^2] = \sum_{i \in \mathcal{V}} \mathbb{E}_{\mathbf{B}}[Y_i^2]$  and denoting

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

the result in (83) can be derived. ■