4 Centralized Learning

4.1 In this problem we will demonstrate the linear performance gain promised by relation (4.46) in code for Example 4.2. Formulate a model w^o of your choice, and generate data according to the observation model (4.39), ensuring agents are homogenous by sampling $\boldsymbol{h}_k$ and $\boldsymbol{v}_k$ from identical distributions. Reasonable choices are $\boldsymbol{h}_k \sim \mathcal{N}(0, \sigma_h^2 I_M)$ and $\boldsymbol{v}_k \sim \mathcal{N}(0, \sigma_v^2)$. Implement recursions (4.42)–(4.43) and plot the evolution of the error over time. Compare the performance for $K = 1$, $K = 10$ and $K = 100$ agents and verify whether you observe linear gains in performance.

Solution. The solution is provided as a Jupyter notebook in the separate file `Problem_4_1.ipynb`.

4.2 Consider local empirical risk minimization problems with unbalanced datasets of the form:

$$J_k(w) = \frac{1}{N_k} \sum_{n=1}^{N_k} Q(w; x_{k,n})$$

and the global empirical risk minimization problem:

$$J(w) = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} Q(w; x_{k,n})$$

where we defined $N = \sum_{k=1}^{K} N_k$. Show that $J(w)$ can be written as a *weighted* consensus problem of the form:

$$J(w) = \sum_{k=1}^{K} p_k J_k(w)$$

where p_k are suitably chosen weights satisfying $\sum_{k=1}^{K} p_k = 1$.

Solution. We begin reformulating $J(w)$ as:

$$\begin{aligned} J(w) &= \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} Q(w; x_{k,n}) \\ &= \frac{1}{N} \sum_{k=1}^{K} N_k \cdot \frac{1}{N_k} \sum_{n=1}^{N_k} Q(w; x_{k,n}) \\ &= \sum_{k=1}^{K} \frac{N_k}{N} J_k(w) \\ &= \sum_{k=1}^{K} p_k J_k(w) \end{aligned}$$

where we defined:

$$p_k = \frac{N_k}{N}$$

It follows that the weights p_k are given by the fraction of data available to agent k. Local objectives are weighted higher for agents who have more data, which is reasonable since they better approximate the true data distribution.

5 Federated Learning

5.1 Suppose in the federated averaging recursion we do not employ normalized step-sizes as suggested by (5.26), but instead let $\mu_k = \mu$. Relation (5.25) suggests that in that case we can no longer interpret federated averaging as employing an unbiased gradient approximation to the consensus problem (5.1). Find an expression for which the approximation is unbiased, and use this insight to determine the limiting point of the federated averaging algorithm without step-size normalization.

Solution. From (5.25), we have:

$$\begin{aligned}\mathbb{E}\left\{\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} &= \frac{1}{K}\sum_{k=1}^{K}\frac{\mu_k}{\mu}E_k\nabla J_k(\boldsymbol{w}_{i-1}) \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\mu_k}{\mu}E_k\nabla J_k(\boldsymbol{w}_{i-1}) \\ &\overset{\mu_k=\mu}{=} \frac{1}{K}\sum_{k=1}^{K}E_k\nabla J_k(\boldsymbol{w}_{i-1})\end{aligned}$$

We identify $\mathbb{E}\left\{\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\}$ as the gradient of:

$$J'(w) \triangleq \frac{1}{K}\sum_{k=1}^{K}E_kJ_k(w)$$

We conclude that the gradient federated averaging algorithm without step-size normalization approximates the gradient of a weighted objective $J(w)$, where the weighting is determined by the number E_k of local steps taken.

5.2 Prove Lemma 5.1.

Solution. Using a similar technique as in (5.23)–(5.24), we introduce the selection indicator:

$$\mathbb{1}_k = \begin{cases}1, & \text{if the } k\text{-th sample is picked,} \\ 0, & \text{otherwise.}\end{cases} \tag{5.1}$$

We can then write:

$$\widehat{\boldsymbol{a}} = \frac{1}{L}\sum_{\ell=1}^{L}\boldsymbol{a}_\ell = \frac{1}{L}\sum_{k=1}^{K}\mathbb{1}_k\boldsymbol{a}_k \tag{5.2}$$

It holds that:

$$\mathbb{E}\mathbb{1}_k = \frac{L}{K} \tag{5.3}$$

$$\begin{aligned}\mathbb{E}\left\{\mathbb{1}_k\mathbb{1}_\ell\right\} &= \mathbb{E}\left\{\mathbb{1}_k\mathbb{1}_\ell|\mathbb{1}_\ell=0\right\}\cdot\Pr\left\{\mathbb{1}_\ell=0\right\} + \mathbb{E}\left\{\mathbb{1}_k\mathbb{1}_\ell|\mathbb{1}_\ell=1\right\}\cdot\Pr\left\{\mathbb{1}_\ell=1\right\}\\ &= \mathbb{E}\left\{\mathbb{1}_k\mathbb{1}_\ell|\mathbb{1}_\ell=1\right\}\cdot\Pr\left\{\mathbb{1}_\ell=1\right\}\\ &= \mathbb{E}\left\{\mathbb{1}_k|\mathbb{1}_\ell=1\right\}\cdot\frac{L}{K}\\ &= \frac{L-1}{K-1}\cdot\frac{L}{K}\end{aligned} \tag{5.4}$$

We can then conclude for the mean:

$$\mathbb{E}\widehat{\boldsymbol{a}} = \mathbb{E}\frac{1}{L}\sum_{k=1}^{K}\mathbb{1}_k\boldsymbol{a}_k = \frac{1}{L}\sum_{k=1}^{K}\frac{L}{K}\mathbb{E}\boldsymbol{a}_k = \overline{a} \tag{5.5}$$

For the variance:

$$\begin{aligned}&\mathbb{E}\|\widehat{\boldsymbol{a}}-\overline{a}\|^2\\ &= \mathbb{E}\|\widehat{\boldsymbol{a}}\|^2 - \|\overline{a}\|^2\\ &= \frac{1}{L^2}\sum_{k=1}^{K}\mathbb{E}\left\{\mathbb{1}_k^2\right\}\mathbb{E}\left\{\|\boldsymbol{a}_k\|^2\right\} + \frac{1}{L^2}\sum_{k=1}^{K}\sum_{\ell\neq k}\mathbb{E}\left\{\mathbb{1}_k\mathbb{1}_\ell\right\}\mathbb{E}\left\{\boldsymbol{a}_k^\mathsf{T}\boldsymbol{a}_\ell\right\} - \|\overline{a}\|^2\\ &= \frac{1}{L^2}\sum_{k=1}^{K}\frac{L}{K}\left(\sigma_{a_k}^2 + \|\overline{a}_k\|^2\right) + \frac{1}{L^2}\sum_{k=1}^{K}\sum_{\ell\neq k}\frac{L-1}{K-1}\frac{L}{K}\overline{a}_k^\mathsf{T}\overline{a}_\ell - \|\overline{a}\|^2\\ &= \frac{1}{KL}\sum_{k=1}^{K}\left(\sigma_{a_k}^2 + \|\overline{a}_k\|^2\right) + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\sum_{\ell\neq k}\overline{a}_k^\mathsf{T}\overline{a}_\ell - \|\overline{a}\|^2\end{aligned} \tag{5.6}$$

We are almost done, and now aim to reformulate the expression to be consistent

with Lemma 5.1 by completing the square:

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{a}} - \overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\left(1 - \frac{L-1}{K-1}\right)\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 \\
&\quad + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\sum_{\ell\neq k}\overline{a}_k^{\mathsf{T}}\overline{a}_\ell - \|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\left(\frac{K-1}{K-1} - \frac{L-1}{K-1}\right)\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 \\
&\quad + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\sum_{\ell\neq k}\overline{a}_k^{\mathsf{T}}\overline{a}_\ell - \|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 \\
&\quad + \frac{1}{KL}\frac{L-1}{K-1}\sum_{k=1}^{K}\sum_{\ell\neq k}\overline{a}_k^{\mathsf{T}}\overline{a}_\ell - \|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \frac{1}{KL}\frac{L-1}{K-1}\left\|\sum_{k=1}^{K}\overline{a}_k\right\|^2 - \|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \frac{K}{L}\frac{L-1}{K-1}\|\overline{a}\|^2 - \|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 + \left(\frac{K}{L}\frac{L-1}{K-1} - \frac{L}{L}\frac{K-1}{K-1}\right)\|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a}_k\|^2 - \frac{1}{L}\frac{K-L}{K-1}\|\overline{a}\|^2 \\
&= \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a} - \overline{a}_k\|^2
\end{aligned}
\tag{5.7}
$$

Where the last line can be verified by expanding $\sum_{k=1}^{K}\|\overline{a} - \overline{a}_k\|^2$ and using the fact that $\sum_{k=1}^{K}\overline{a}_k = \overline{a}$.

5.3 Consider the federated averaging algorithm for the least-mean square prob-

lem, which takes the form:

$$\begin{aligned}\boldsymbol{\phi}_{k,e} &= \boldsymbol{\phi}_{k,e-1} + \frac{\mu}{E_k}\frac{1}{B_k}\sum_{b=1}^{B_k}\boldsymbol{h}_{k,i,e,b}\left(\boldsymbol{\gamma}_{k,i,e,b} - \boldsymbol{h}_{k,i,e,b}^{\mathsf{T}}\boldsymbol{\phi}_{k,e-1}\right)\\ \boldsymbol{\psi}_{k,i} &= \boldsymbol{\phi}_{k,E_k}\\ \boldsymbol{w}_i &= \frac{1}{L}\sum_{k\in\mathcal{L}_i}\boldsymbol{\psi}_{k,i}\end{aligned}$$

Determine an expression for the limiting performance $\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$ of the algorithm as a function of data statistics and tuning parameters.

Solution. The local gradient approximation takes the form of a mini-batch of size B_k on the least-mean square problem. We have encountered both mini-batching and least-mean square in Chapter 3, and can hence conlcude from (3.35), (3.36) and (3.46) that:

$$\begin{aligned}\sigma_k^2 &= \frac{\sigma_{\boldsymbol{v}_k}^2\,\mathrm{Tr}\left(R_{\boldsymbol{h}_k}\right)}{B_k}\\ \sigma_k^2 &= \frac{\mathbb{E}\|\boldsymbol{h}_k\boldsymbol{h}_k^{\mathsf{T}} - R_{\boldsymbol{h}_k}\|^2}{B_k}\end{aligned}$$

For the strong-convexity and Lipschitz constants, we can differentiate the least-mean square objective (see, e.g, (5.44)) twice to find:

$$\nabla^2 J_k(w) = R_{\boldsymbol{h}_k}$$

and hence:

$$\begin{aligned}\nu_k &= \lambda_{\min}\left(R_{\boldsymbol{h}_k}\right)\\ \delta_k &= \lambda_{\max}\left(R_{\boldsymbol{h}_k}\right)\end{aligned}$$

We can fuse the local variances using (5.40) to find:

$$\begin{aligned}\sigma_{\mathrm{fed}}^2 = \frac{1}{KL}\sum_{k=1}^{K}\left(2\frac{\mathbb{E}\|\boldsymbol{h}_k\boldsymbol{h}_k^{\mathsf{T}} - R_{\boldsymbol{h}_k}\|^2}{B_kE_k}\|w_k^o - w^o\|^2 + \frac{\sigma_{\boldsymbol{v}_k}^2\,\mathrm{Tr}\left(R_{\boldsymbol{h}_k}\right)}{B_kE_k}\right)\\ + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}4\lambda_{\max}\left(R_{\boldsymbol{h}_k}\right)^2\|w_k^o - w^o\|^2\end{aligned}$$

Putting everything together, we conclude form (5.57):

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \frac{4\mu\sigma_{\mathrm{fed}}^2}{\nu} = \frac{4\mu\sigma_{\mathrm{fed}}^2}{\lambda_{\max}\left(R_{\boldsymbol{h}_k}\right)}$$

where the expression for σ_{fed}^2 is given in the previous line.