5 Federated Learning

In Chapter 4 we developed *centralized* strategies for the optimization of consensus optimization problems:

$$J(w) = \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{5.1}$$

where $J_k(w)$ denotes the local objective function, which can take the form of an expected risk:

$$J_k(w) = \mathbb{E}_{\boldsymbol{x}_k} Q(w; \boldsymbol{x}_k) \tag{5.2}$$

We introduced in Sec. 4.4 the *centralized stochastic gradient* algorithm:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{i-1} - \mu \widehat{\nabla J}_k(\boldsymbol{w}_{i-1}) \tag{5.3}$$

$$\boldsymbol{w}_i = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\psi}_{k,i} \tag{5.4}$$

Here, the intermediate model $\boldsymbol{\psi}_{k,i}$ is generated locally at agent k by adjusting the global model $\boldsymbol{w}_{i-1}$ following the gradient approximation $\widehat{\nabla J}_k(\boldsymbol{w}_{i-1})$. In Sec. 4.5 an asynchronous variant of the centralized stochastic gradient algorithm was studied, where a single agent, indexed by the random variable $\boldsymbol{k}_i$ performs an update:

$$\boldsymbol{\psi}_{\boldsymbol{k}_i,i} = \boldsymbol{w}_{i-1} - \mu \widehat{\nabla J}_{\boldsymbol{k}_i}(\boldsymbol{w}_{i-1}) \tag{5.5}$$

$$\boldsymbol{w}_i = \boldsymbol{\psi}_{\boldsymbol{k}_i,i} \tag{5.6}$$

A common key property of recursions (5.3)–(5.6) is that in both cases:

$$\mathbb{E}\left\{\boldsymbol{w}_i | \boldsymbol{w}_{i-1}\right\} = \boldsymbol{w}_{i-1} - \mu \nabla J(\boldsymbol{w}_{i-1}) = \boldsymbol{w}_{i-1} - \frac{\mu}{K}\sum_{k=1}^{K} \nabla J_k(\boldsymbol{w}_{i-1}) \tag{5.7}$$

It follows that both distributed implementations perform unbiased updates along the aggregate consensus objective (5.1), albeit with differing variances, resulting in different performance properties. We essentially exploited this fact in Chapter 4 to derive convergence guarantees of centralized gradient-based algorithms for distributed optimization.

As communication becomes increasingly costly, intermittent or unreliable, it is no longer reasonable for agents to participate in the model exchange at every iteration, as in (5.3)–(5.4), or to stop and stand by idly if they are not selected as in (5.5)–(5.6). Instead, this calls for flexible schemes for distributed learning, where individual agents may perform more or less local updates depending on their computational capabilities and time between parameter exchanges. Such highly heterogeneous and asynchronous structures have come to be referred to as *federated learning*, and we will introduce in this chapter some popular variants and study their convergence. The key distinction to the centralized algorithms studied in Chapter 4 will be that the presence of multiple local updates introduces a bias to (5.27), requiring an adjustment of the performance analysis.

5.1 FEDERATED AVERAGING

A common structure we observed in the centralized algorithms of Chapter 4 was the presence of a local adaptation step of the form (5.3) or (5.5) followed by an aggregation step (5.4) or (5.6). As we will see, this general structure continuous with federated algorithms, except that we will be adjusting adaptation and aggregation steps to suit the more asynchronous and heterogeneous setting. The parameter server continues to maintain a global model $\boldsymbol{w}_i$. At every iteration i, only a subset of L agents, collected in $\boldsymbol{\mathcal{L}}_i$ will be selected to participate. We will generate the set $\boldsymbol{\mathcal{L}}_i$ by sampling uniformly with equal probablity and *without replacement*, so that:

$$\Pr\left\{k \in \boldsymbol{\mathcal{L}}_i\right\} = \frac{L}{K} \tag{5.8}$$

Now, at time i, each selected agent $k \in \boldsymbol{\mathcal{L}}_i$ receives the model $\boldsymbol{w}_{i-1}$ from the parameter server, initializes $\boldsymbol{\phi}_{k,0} = \boldsymbol{w}_{i-1}$ and then performs E_k local updates of the form:

$$\boldsymbol{\phi}_{k,e} = \boldsymbol{\phi}_{k,e-1} - \mu_k \widehat{\nabla J_k}(\boldsymbol{\phi}_{k,e-1}) \tag{5.9}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,E_k} \tag{5.10}$$

In (5.9) we utilize e to index the inner iteration of local gradient updates, which occurs between every outer time step $i-1$ and i. The final local update $\boldsymbol{\phi}_{k,E_k}$ is then stored in $\boldsymbol{\psi}_{k,i}$, which is then sent back to the parameter server, where aggregation now takes the form:

$$\boldsymbol{w}_i = \frac{1}{L} \sum_{k \in \boldsymbol{\mathcal{L}}_i} \boldsymbol{\psi}_{k,i} \tag{5.11}$$

Before we proceed with studying the learning dynamics of the federated algorithm (5.9)–(5.11), it is important to comment on some important details in the recursions. First, we allow for differing numbers of local updates E_k at different agents. This is important to account for different capabilties across different agents. Some agents may be able to perform more local updates in a given period of time, for example due to increased computational capabilities. Second, we allow for varying local step-size parameters μ_k. This additional degree of freedom will be necessary to control the relative influence of agents with different number of local updates E_k. This is because despite their heterogeneous contributions, we would like the agents to continue to converge to the minimum of the aggregate consensus problem (5.1). If we require all agents to utilize a common step-size μ in a heterogeneous setting, this will allow more active agents to exert increased influence over the network, and bias the limiting point of the algorithm away from the minimizer of (5.1), a phenomenon which we examine more closely in Problem 5.1.

5.1.1 Dynamics of the Federated Averaging Algorithm

While we will present a formal guarantee of convergence of the Federated Averaging algorithm (5.9)–(5.11) in Section 5.1.2 further ahead, we will first reformulate the recursions in an equivalent form that is less amenable to distributed implementation, but provides a high-level intuition behind the learning dynamics of federated averaging. These insights will serve two purposes. First, it will suggest a choice for the local step-sizes μ_k as a function of the number of local updates E_k. Second, it will provide a blueprint for the convergence analysis we conduct in Section 5.1.2.

Iterating (5.9) and plugging the final result into (5.9), we can find for the locally adapted models $\boldsymbol{\psi}_{k,i}$:

$$\begin{aligned}\boldsymbol{\psi}_{k,i} &= \boldsymbol{w}_{i-1} - \mu_k \sum_{e=1}^{E_k} \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) \\ &= \boldsymbol{w}_{i-1} - \mu_k E_k \cdot \frac{1}{E_k} \sum_{e=1}^{E_k} \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) \qquad (5.12)\end{aligned}$$

Inspection of this form reveals that (5.12) resembles a mini-batch update with increased step-size $\mu_k E_k$, where we are averaging E_k mini-batch approximations $\widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1})$. In Chapter 3, we saw that employing mini-batches reduces the variance of the gradient approximation proportionally to the batch size, which then translates into improved steady-state performance. One key detail to note, however, is that the gradient approximations $\widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1})$ are all evaluated at different iterates $\boldsymbol{\phi}_{k,e-1}$, hence

$$\frac{1}{E_k} \sum_{e=1}^{E_k} \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) \neq \frac{1}{E_k} \sum_{e=1}^{E_k} \widehat{\nabla J}_k(\boldsymbol{w}_{i-1}) \qquad (5.13)$$

and we are not performing a true mini-batch gradient update. Instead, we can interpret the federated local update as a perturbed mini-batch update by reformulating:

$$\begin{aligned}\boldsymbol{\psi}_{k,i} &= \boldsymbol{w}_{i-1} - \mu_k E_k \cdot \frac{1}{E_k}\sum_{e=1}^{E_k} \widehat{\nabla J_k}(\boldsymbol{w}_{i-1}) \\ &\quad + \mu_k E_k \cdot \frac{1}{E_k}\sum_{e=1}^{E_k}\Big(\widehat{\nabla J_k}(\boldsymbol{w}_{i-1}) - \widehat{\nabla J_k}(\boldsymbol{\phi}_{k,e-1})\Big) \qquad (5.14)\end{aligned}$$

We can then conjecture that as long as the step-size μ_k and number of local steps E_k are sufficiently small, the iterates $\boldsymbol{\phi}_{k,e-1}$ will not deviate too far from the global model $\boldsymbol{w}_{i-1}$. Under appropriate smoothness conditions on the gradient approximation, we expect the perturbation $\mu_k \sum_{e=1}^{E_k}\Big(\widehat{\nabla J_k}(\boldsymbol{w}_{i-1}) - \widehat{\nabla J_k}(\boldsymbol{\phi}_{k,e-1})\Big)$ to be negligible, and the dynamics to be well-approximated by a mini-batch update with increased effective step-size $\mu_K E_k$ along the mini-batch estimate $\frac{1}{E_k}\sum_{e=1}^{E_k}\widehat{\nabla J_k}(\boldsymbol{w}_{i-1})$, which is an improved estimate of the local gradient $\nabla J_k(\boldsymbol{w}_{i-1})$. We will make this precise in Section 5.1.2.

To develop appropriate expressions for the local step-sizes μ_k, we first introduce the following quantities for ease of notation:

$$\widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) \triangleq \frac{1}{E_k}\sum_{e=1}^{E_k}\widehat{\nabla J_k}(\boldsymbol{w}_{i-1}) \qquad (5.15)$$

$$\boldsymbol{d}_{k,i} = \frac{1}{E_k}\sum_{e=1}^{E_k}\Big(\widehat{\nabla J_k}(\boldsymbol{\phi}_{k,e-1}) - \widehat{\nabla J_k}(\boldsymbol{w}_{i-1})\Big) \qquad (5.16)$$

We can then write (5.12) more compactly:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{i-1} - \mu_k E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) - \mu_k E_k \boldsymbol{d}_{k,i} \qquad (5.17)$$

in terms of the mini-batch gradient approximation $\widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1})$ and the perturbation $\boldsymbol{d}_{k,i}$. After aggregation following (5.11), we then have at the parameter server:

$$\begin{aligned}\boldsymbol{w}_i &= \frac{1}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\Big(\boldsymbol{w}_{i-1} - \mu_k E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) - \mu_k E_k \boldsymbol{d}_{k,i}\Big) \\ &= \boldsymbol{w}_{i-1} - \frac{1}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\mu_k E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) - \frac{1}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\mu_k E_k \boldsymbol{d}_{k,i} \\ &\overset{(a)}{=} \boldsymbol{w}_{i-1} - \frac{\mu}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\frac{\mu_k}{\mu}E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) - \frac{1}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\mu_k E_k \boldsymbol{d}_{k,i} \qquad (5.18)\end{aligned}$$

where in (a) we introduced an arbitrary common step-size $\mu > 0$, and normalized the local step-sizes inside the sum by the same factor μ, leaving the recursion unchanged. We can then introduce the global quantities:

$$\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1}) \triangleq \frac{1}{L} \sum_{k \in \boldsymbol{\mathcal{L}}_i} \frac{\mu_k}{\mu} E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) \tag{5.19}$$

$$\boldsymbol{d}_i \triangleq \frac{1}{L} \sum_{k \in \boldsymbol{\mathcal{L}}_i} \mu_k E_k \boldsymbol{d}_{k,i} \tag{5.20}$$

and write more compactly:

$$\begin{aligned} \boldsymbol{w}_i &= \boldsymbol{w}_{i-1} - \mu \widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1}) - \mu \boldsymbol{d}_i \\ &= \boldsymbol{w}_{i-1} - \mu \nabla J(\boldsymbol{w}_{i-1}) - \mu \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) - \mu \boldsymbol{d}_i \end{aligned} \tag{5.21}$$

with the gradient noise term $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ defined as in 3:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = \widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1}) - \nabla J(\boldsymbol{w}_{i-1}) \tag{5.22}$$

We observe that the federated averaging algorithm (5.9)–(5.11) can be viewed as a perturbed gradient descent recursion on the consensus problem (5.1) with two perturbation terms. The first is the term $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, which is a gradient noise term (as we already encountered in Chapter 3) of the gradient approximation (5.19). The second term is $\boldsymbol{d}_i$, which results from the fact that agents take multiple local updates.

One of the requirements we have so far placed on gradient approximations is that they are unbiased (see (3.22)), which implies that the gradient noise process $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ they induce has mean zero. Let us then see if this also holds for the federated approximation (5.19). To this end, we introduce the participation variable:

$$\mathbb{1}_{k,i} = \begin{cases} 1, & \text{if } k \in \boldsymbol{\mathcal{L}}_i \\ 0, & \text{otherwise.} \end{cases} \tag{5.23}$$

We can then write (5.19) equivalently as:

$$\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1}) = \frac{1}{L} \sum_{k=1}^{K} \mathbb{1}_{k,i} \frac{\mu_k}{\mu} E_k \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) \tag{5.24}$$

After taking expectations, we have:

$$\begin{aligned}
&\mathbb{E}\left\{\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} \\
&= \frac{1}{L}\sum_{k=1}^{K}\mathbb{E}\left\{\mathbb{1}_{k,i}\frac{\mu_k}{\mu}E_k\widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(a)}{=} \frac{1}{L}\sum_{k=1}^{K}\frac{L}{K}\frac{\mu_k}{\mu}E_k\mathbb{E}\left\{\widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(5.15)}{=} \frac{1}{K}\sum_{k=1}^{K}\frac{\mu_k}{\mu}E_k\mathbb{E}\left\{\frac{1}{E_k}\sum_{e=1}^{E_k}\widehat{\nabla J_k}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{\mu_k}{\mu}E_k\frac{1}{E_k}\sum_{e=1}^{E_k}\mathbb{E}\left\{\widehat{\nabla J_k}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(b)}{=} \frac{1}{K}\sum_{k=1}^{K}\frac{\mu_k}{\mu}E_k\nabla J_k(\boldsymbol{w}_{i-1}) \\
&\overset{(c)}{=} \frac{1}{K}\sum_{k=1}^{K}\nabla J_k(\boldsymbol{w}_{i-1}) \\
&= \nabla J(\boldsymbol{w}_{i-1})
\end{aligned} \tag{5.25}$$

where (a) follows since the data used to construct local gradient approximations is independent of whether an agent has been selected or not, and hence $\mathbb{1}_{k,i}$ and $\widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1})$ are independent. Then $\mathbb{E}\left\{\mathbb{1}_{k,i}|\boldsymbol{w}_{i-1}\right\} = \mathbb{E}\mathbb{1}_{k,i} = \frac{K}{L}$. Step (b) holds whenever the local gradient approximations $\widehat{\nabla J_k}(\boldsymbol{w}_{i-1})$ are assumed to be unbiased. Step (c) holds if and only if:

$$\boxed{\frac{\mu_k}{\mu}E_k = 1 \Longleftrightarrow \mu_k = \frac{\mu}{E_k}} \tag{5.26}$$

We conclude that the federated averaging recursions (5.9)–(5.11) can be viewed as employing an unbiased gradient approximation only if the local step-sizes are chosen according to (5.26). For the remainder of this chapter, we will assume this to be the case.

While a simple normalization of local step-sizes μ_k is sufficient to ensure unbiased gradient noise $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, the same is not true for the perturbation term $\boldsymbol{d}_i$. Any $E_k > 1$ will necessarily cause $\mathbb{E}\left\{\boldsymbol{d}_i|\boldsymbol{w}_{i-1}\right\} \neq 0$, irrespecive of the choice of other parameters. However, as the analysis in the sequel will show, the energy of $\boldsymbol{d}_i$ is so small, that even a bias will not result in significant deterioration of performance.

5.1.2 Convergence and Performance

Bound on the gradient noise $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$

While the biased perturbation $\boldsymbol{d}_i$ will cause some complications, our analysis essentially mirrors that of the stochastic gradient algorithm (Chapter 3) and its centralized counterpart (Chapter 4). We will hence need to establish conditions on the gradient approximation (5.19) and the induced gradient noise process $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$. We will assume that each local gradient approximation $\widehat{\nabla J}_k(\boldsymbol{w}_{i-1})$ satisfies the gradient noise conditions (3.22)–(3.23) with $\alpha_k^2 = \gamma_k^2 = 0$ and some $\beta_k^2, \sigma_k^2 \geq 0$. For ease of reference, we repeat:

$$\mathbb{E}\left\{\widehat{\nabla J}_k(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} = \nabla J_k(\boldsymbol{w}_{i-1}) \tag{5.27}$$

$$\mathbb{E}\left\{\left\|\widehat{\nabla J}_k(\boldsymbol{w}_{i-1}) - \nabla J_k(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\} \leq \beta_k^2\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \sigma_k^2 \tag{5.28}$$

We already argued in (5.25) that (5.27) is sufficient to conclude that:

$$\mathbb{E}\left\{\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} = \nabla J(\boldsymbol{w}_{i-1}) \tag{5.29}$$

To establish a bound on the variance of the gradient approximation, work our way through the chain from $\widehat{\nabla J}_k(\boldsymbol{w}_{i-1})$ to $\widehat{\nabla J}_k^{E_k}(\boldsymbol{w}_{i-1})$ to $\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})$. First note that since $\widehat{\nabla J}_k^{E_k}(\boldsymbol{w}_{i-1})$ in (5.15) is identical in structure to a mini-batch approximation of the local gradient $\nabla J_k(\boldsymbol{w}_{i-1})$, we can apply the same argument (3.45) to find:

$$\mathbb{E}\left\{\left\|\widehat{\nabla J}_k^{E_k}(\boldsymbol{w}_{i-1}) - \nabla J_k(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\} \leq \frac{\beta_k^2}{E_k}\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k} \tag{5.30}$$

While the global approximation $\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})$ in (5.24) looks reminiscent of a mini-batch approximation, there are some important details that prevent us from applying the argument of (3.45) in a straightforward manner. First, the local approximations $\widehat{\nabla J}_k^{E_k}(\boldsymbol{w}_{i-1})$ follow different distributions for different k, while mini-batch approximations typically employ identically distributed data. Second, since we are sampling agents without replacement, the participation indicators $\mathbb{1}_{k,i}$ are not independent. To account for this, we present a useful general lemma.

LEMMA 5.1 (**Mean of random variables**). *Let $\{\boldsymbol{a}_k\}_{k=1}^K$ denote a set of independent random variables with means $\overline{a}_k = \mathbb{E}\boldsymbol{a}_k$ and variance $\sigma_{a_k}^2$. We are interested in estimating the global mean:*

$$\overline{a} = \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^K \boldsymbol{a}_k\right\} = \frac{1}{K}\sum_{k=1}^K \overline{a}_k \tag{5.31}$$

To this end, we sample L times without replacement from the set $\{\boldsymbol{a}_k\}_{k=1}^K$ to *obtain* $\{\boldsymbol{a}_\ell\}_{\ell=1}^L$ *and estimate:*

$$\widehat{\boldsymbol{a}} = \frac{1}{L}\sum_{\ell=1}^{L}\boldsymbol{a}_\ell \tag{5.32}$$

Then it holds that $\mathbb{E}\widehat{\boldsymbol{a}} = \overline{a}$ *and:*

$$\mathbb{E}\|\widehat{\boldsymbol{a}} - \overline{a}\|^2 = \frac{1}{KL}\sum_{k=1}^{K}\sigma_{a_k}^2 + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\overline{a} - \overline{a}_k\|^2 \tag{5.33}$$

Proof: The proof is subject of Problem 5.2. □

We can apply Lemma 5.1 to determine the variance of the gradient approximation $\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})$ by setting:

$$\begin{align}
\boldsymbol{a}_k &\longleftarrow \widehat{\nabla J_k}^{E_k}(\boldsymbol{w}_{i-1}) \tag{5.34}\\
\overline{a}_k &\longleftarrow \nabla J_k(\boldsymbol{w}_{i-1}) \tag{5.35}\\
\overline{a} &\longleftarrow \nabla J(\boldsymbol{w}_{i-1}) \tag{5.36}\\
\sigma_{a_k}^2 &\longleftarrow \frac{\beta_k^2}{E_k}\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k} \tag{5.37}
\end{align}$$

Then:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widehat{\nabla J}(\boldsymbol{w}_{i-1}) - \nabla J(\boldsymbol{w}_{i-1})\right\|^2 | \boldsymbol{w}_{i-1}\right\} \\
&\le \frac{1}{KL}\sum_{k=1}^{K}\left(\frac{\beta_k^2}{E_k}\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k}\right) + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\nabla J(\boldsymbol{w}_{i-1}) - \nabla J_k(\boldsymbol{w}_{i-1})\|^2 \\
&\overset{(a)}{=} \frac{1}{KL}\sum_{k=1}^{K}\left(\frac{\beta_k^2}{E_k}\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k}\right) \\
&\quad + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\nabla J(\boldsymbol{w}_{i-1}) - \nabla J(w^o) + \nabla J_k(w_k^o) - \nabla J_k(\boldsymbol{w}_{i-1})\|^2 \\
&\overset{(b)}{\le} \frac{1}{KL}\sum_{k=1}^{K}\left(\frac{\beta_k^2}{E_k}\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k}\right) \\
&\quad + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\left(2\delta^2\|w^o - \boldsymbol{w}_{i-1}\|^2 + 2\delta_k^2\|w_k^o - \boldsymbol{w}_{i-1}\|^2\right) \\
&\overset{(c)}{=} \frac{1}{KL}\sum_{k=1}^{K}\left(\frac{\beta_k^2}{E_k}\|w_k^o - w^o + w^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k}\right) \\
&\quad + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\left(2\delta^2\|w^o - \boldsymbol{w}_{i-1}\|^2 + 2\delta_k^2\|w_k^o - w^o + w^o - \boldsymbol{w}_{i-1}\|^2\right) \\
&\overset{(d)}{=} \frac{1}{KL}\sum_{k=1}^{K}\left(2\frac{\beta_k^2}{E_k}\|w_k^o - w^o\|^2 + 2\frac{\beta_k^2}{E_k}\|w^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\sigma_k^2}{E_k}\right) \\
&\quad + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\left(2\delta^2\|w^o - \boldsymbol{w}_{i-1}\|^2 + 4\delta_k^2\|w_k^o - w^o\|^2 + 4\delta_k^2\|w^o - \boldsymbol{w}_{i-1}\|^2\right) \\
&\overset{(e)}{=} \beta_{\text{fed}}^2\|w^o - \boldsymbol{w}_{i-1}\|^2 + \sigma_{\text{fed}}^2
\end{aligned}
\tag{5.38}
$$

Here, (a) follows since w^o and w_k^o are the minimizing arguments of $J(w)$ and $J_k(w)$ respectively. Step (b) follows from the relation $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ for arbitrary vectors a, b along with the condition that $J(w)$ and $J_k(w)$ have δ and δ_k-Lipschitz gradients respectively (see Sec. 2.1). In (c) we added and subtracted w^o, and applied the relation $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ again to find (d). In (e), we grouped terms and defined:

$$
\beta_{\text{fed}}^2 = \frac{1}{KL}\sum_{k=1}^{K} 2\frac{\beta_k^2}{E_k} + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\left(2\delta^2 + 4\delta_k^2\right) \tag{5.39}
$$

$$
\sigma_{\text{fed}}^2 = \frac{1}{KL}\sum_{k=1}^{K}\left(2\frac{\beta_k^2}{E_k}\|w_k^o - w^o\|^2 + \frac{\sigma_k^2}{E_k}\right) + \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K} 4\delta_k^2\|w_k^o - w^o\|^2 \tag{5.40}
$$

We conclude that the implicit gradient approximation of the federated averaging algorithm satisfies our normal gradient noise conditions with parambers $\beta^2_{\text{fed}}, \sigma^2_{\text{fed}}$ determined a variety of parameters and constants such as the local gradient noise constants β_k^2, σ_k^2, the number of local updates E_k as well as the level of heterogeneity quantified by $\|w_k^o - w^o\|^2$.

Bound on the perturbation $\boldsymbol{d}_i$

We now proceed to establish a bound on the perturbation $\boldsymbol{d}_i$. First, we decompose:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{d}_i\|^2|\boldsymbol{w}_{i-1}\right\} \\
=\, &\mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k\in\mathcal{L}_i}\mu_k E_k \boldsymbol{d}_{k,i}\right\|^2 |\boldsymbol{w}_{i-1}\right\} \\
\overset{(5.26)}{=} &\mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k\in\mathcal{L}_i}\mu \boldsymbol{d}_{k,i}\right\|^2 |\boldsymbol{w}_{i-1}\right\} \\
=\, &\mu^2\mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k\in\mathcal{L}_i}\boldsymbol{d}_{k,i}\right\|^2 |\boldsymbol{w}_{i-1}\right\} \\
=\, &\mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{1}_{k,i}\boldsymbol{d}_{k,i}\right\|^2 |\boldsymbol{w}_{i-1}\right\} \\
\overset{(a)}{\leq} &\mu^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\{\|\mathbb{1}_{k,i}\boldsymbol{d}_{k,i}\|^2 |\boldsymbol{w}_{i-1}\right\} \\
=\, &\mu^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\{\|\boldsymbol{d}_{k,i}\|^2 |\boldsymbol{w}_{i-1}\right\}\cdot\Pr\left\{\mathbb{1}_{k,i}=1\right\} \\
=\, &\mu^2\frac{L}{K^2}\sum_{k=1}^{K}\mathbb{E}\left\{\|\boldsymbol{d}_{k,i}\|^2 |\boldsymbol{w}_{i-1}\right\}
\end{aligned}
\tag{5.41}
$$

where (a) follows from Jensen's inequality for the convex function $\|\cdot\|^2$. For each individual $\boldsymbol{d}_{k,i}$, we have:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{d}_{k,i}\|^2 |\boldsymbol{w}_{i-1}\right\} \\
=\, &\mathbb{E}\left\{\left\|\frac{1}{E_k}\sum_{e=1}^{E_k}\left(\widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{i-1})\right)\right\|^2 |\boldsymbol{w}_{i-1}\right\} \\
\overset{(a)}{\leq} &\frac{1}{E_k}\sum_{e=1}^{E_k}\mathbb{E}\left\{\left\|\widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{i-1})\right\|^2 |\boldsymbol{w}_{i-1}\right\}
\end{aligned}
\tag{5.42}
$$

Here, again (a) is a result of Jensen's inequality. This relation motivates the following mean-square smoothness condition, which is adapted from the regularity conditions of 2.1.

Condition 5.1 (**Mean-square smoothness**). The gradient approximations $\widehat{\nabla J_k}(\cdot)$ are Lipschitz in the mean-square sense, i.e.:

$$\mathbb{E}\|\widehat{\nabla J_k}(w_1) - \widehat{\nabla J_k}(w_2)\|^2 \leq \hat{\delta}_k^2 \|w_1 - w_2\|^2 \tag{5.43}$$

□

Example 5.1 (**Least-Mean Square**) Consider again the least-mean square problems:

$$J_k(w) = \mathbb{E}\,\|\boldsymbol{\gamma}_k - \boldsymbol{h}_k^{\mathsf{T}} w\|^2 \tag{5.44}$$

and ordinary gradient approximations of the form:

$$\widehat{\nabla J_k}(w) = \nabla Q(w; \boldsymbol{h}_{k,i}, \boldsymbol{\gamma}_{k,i}) = -\boldsymbol{h}_{k,i}\left(\boldsymbol{\gamma}_{k,i} - \boldsymbol{h}_{k,i}^{\mathsf{T}} w\right) \tag{5.45}$$

Then, it holds that:

$$\begin{aligned}
&\mathbb{E}\|\widehat{\nabla J_k}(w_1) - \widehat{\nabla J_k}(w_2)\|^2 \\
=\,& \mathbb{E}\| - \boldsymbol{h}_{k,i}\left(\boldsymbol{\gamma}_{k,i} - \boldsymbol{h}_{k,i}^{\mathsf{T}} w_1\right) + \boldsymbol{h}_{k,i}\left(\boldsymbol{\gamma}_{k,i} - \boldsymbol{h}_{k,i}^{\mathsf{T}} w_2\right)\|^2 \\
=\,& \mathbb{E}\|\boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathsf{T}} w_1 - \boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathsf{T}} w_2\|^2 \\
=\,& \mathbb{E}\|\boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathsf{T}}(w_1 - w_2)\|^2 \\
\leq\,& \mathbb{E}\|\boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathsf{T}}\|^2 \|w_1 - w_2\|^2
\end{aligned} \tag{5.46}$$

and hence the approximation (5.45) for (5.44) satisfies Condition 5.1 with $\hat{\delta}_k^2 = \mathbb{E}\|\boldsymbol{h}_{k,i}\boldsymbol{h}_{k,i}^{\mathsf{T}}\|^2$.

Under Condition 5.1, we can then continue from (5.42):

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{d}_{k,i}\|^2 | \boldsymbol{w}_{i-1}\right\} \\
\leq\,& \frac{1}{E_k}\sum_{e=1}^{E_k} \mathbb{E}\left\{\left\|\widehat{\nabla J_k}(\boldsymbol{\phi}_{k,e-1}) - \widehat{\nabla J_k}(\boldsymbol{w}_{i-1})\right\|^2 | \boldsymbol{w}_{i-1}\right\} \\
\overset{(5.43)}{\leq}\,& \frac{\hat{\delta}_k^2}{E_k}\sum_{e=1}^{E_k} \mathbb{E}\left\{\left\|\boldsymbol{\phi}_{k,e-1} - \boldsymbol{w}_{i-1}\right\|^2 | \boldsymbol{w}_{i-1}\right\}
\end{aligned} \tag{5.47}$$

Through this string of inequalities we conclude that, in order to bound $\mathbb{E}\|\boldsymbol{d}_i\|^2$ it is sufficient to bound the distance $\left\|\boldsymbol{\phi}_{k,e-1} - \boldsymbol{w}_{i-1}\right\|^2$ agents travel over the course of their local updates. To this end, we subtract (5.9) from w_k^o and find:

$$\begin{aligned} & w_k^o - \boldsymbol{\phi}_{k,e} \\ &= w_k^o - \boldsymbol{\phi}_{k,e-1} + \mu_k \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) \\ &\overset{(a)}{=} w_k^o - \boldsymbol{\phi}_{k,e-1} + \mu_k \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) - \mu_k \nabla J_k(w_k^o) \\ &\overset{(5.26)}{=} w_k^o - \boldsymbol{\phi}_{k,e-1} + \frac{\mu}{E_k} \widehat{\nabla J}_k(\boldsymbol{\phi}_{k,e-1}) - \frac{\mu}{E_k} \nabla J_k(w_k^o) \end{aligned} \tag{5.48}$$

In step (a) we made use of the fact that w_k^o is the minimizer of $J_k(w)$. Using the same argument that led to (3.61), we can then find:

$$\mathbb{E}\left\{\|w_k^o - \boldsymbol{\phi}_{k,e}\|^2 | \boldsymbol{w}_{i-1}\right\} \leq \lambda_k \mathbb{E}\left\{\|w_k^o - \boldsymbol{\phi}_{k,e-1}\|^2 | \boldsymbol{w}_{i-1}\right\} + \mu^2 \frac{\sigma_k^2}{E_k^2} \tag{5.49}$$

The rate of convergence λ_k is given by:

$$\lambda_k \triangleq 1 - 2\frac{\mu}{E_k}\nu_k + \frac{\mu_k^2}{E_k^2}\left(\delta_k^2 + \beta_k^2\right) \tag{5.50}$$

where ν_k and δ_k correspond to the strong-convexity and smoothness constants of $J_k(w)$ respectively. If μ is chosen small enough so that $\lambda_k \leq 1$, it follows that:

$$\mathbb{E}\left\{\|w_k^o - \boldsymbol{\phi}_{k,e}\|^2 | \boldsymbol{w}_{i-1}\right\} \leq \mathbb{E}\left\{\|w_k^o - \boldsymbol{\phi}_{k,e-1}\|^2 | \boldsymbol{w}_{i-1}\right\} + \mu^2 \frac{\sigma_k^2}{E_k^2} \tag{5.51}$$

and we can iterate starting at $\boldsymbol{\phi}_{k,0} = \boldsymbol{w}_{i-1}$ and find:

$$\begin{aligned} \mathbb{E}\left\{\|w_k^o - \boldsymbol{\phi}_{k,e}\| | \boldsymbol{w}_{i-1}\right\} &\leq \|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \mu^2 \frac{\sigma_k^2 e}{E_k^2} \\ &\leq \|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \mu^2 \frac{\sigma_k^2}{E_k} \end{aligned} \tag{5.52}$$

Returning to (5.47), we bound:

$$\begin{aligned} & \mathbb{E}\left\{\|\boldsymbol{d}_{k,i}\|^2 | \boldsymbol{w}_{i-1}\right\} \\ &\leq \frac{\hat{\delta}_k^2}{E_k} \sum_{e=1}^{E_k} \left\{\|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \mu^2 \frac{\sigma_k^2}{E_k}\right\} \\ &\leq \hat{\delta}_k^2 \|w_k^o - \boldsymbol{w}_{i-1}\|^2 + \mu^2 \hat{\delta}_k^2 \frac{\sigma_k^2}{E_k} \\ &= \hat{\delta}_k^2 \|w_k^o - w^o + w^o - \boldsymbol{w}_{i-1}\|^2 + \mu^2 \hat{\delta}_k^2 \frac{\sigma_k^2}{E_k} \\ &\overset{(a)}{\leq} 2\hat{\delta}_k^2 \|w^o - \boldsymbol{w}_{i-1}\|^2 + 2\hat{\delta}_k^2 \|w^o - w_k^o\|^2 + \mu^2 \hat{\delta}_k^2 \frac{\sigma_k^2}{E_k} \end{aligned} \tag{5.53}$$

where (a) holds due to Jensen's inequality. Returning to (5.41), we find for the global perturbation:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{d}_i\|^2|\boldsymbol{w}_{i-1}\right\} \\
&\leq \mu^2 \frac{L}{K^2}\sum_{k=1}^{K}\mathbb{E}\left\{\left\|\boldsymbol{d}_{k,i}\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&= \mu^2 \frac{L}{K^2}\sum_{k=1}^{K}\left(2\hat{\delta}_k^2\|w^o-\boldsymbol{w}_{i-1}\|^2+2\hat{\delta}_k^2\|w^o-w_k^o\|^2+\mu^2\hat{\delta}_k^2\frac{\sigma_k^2}{E_k}\right) \\
&= \mu^2 \frac{L}{K^2}\left(\sum_{k=1}^{K}2\hat{\delta}_k^2\right)\|w^o-\boldsymbol{w}_{i-1}\|^2+\mu^2\frac{L}{K^2}\sum_{k=1}^{K}\left(2\hat{\delta}_k^2\|w^o-w_k^o\|^2+\mu^2\hat{\delta}_k^2\frac{\sigma_k^2}{E_k}\right) \\
&= O(\mu^2)\cdot\|w^o-\boldsymbol{w}_{i-1}\|^2+O(\mu^2)
\end{aligned}
\tag{5.54}
$$

We have now established bounds on the variance of both perturbation terms in the federated averaging recursion (5.21), and are ready to put things together in a proof of convergence.

Theorem 5.1 (Mean-square-behavior of federated averaging). *Let the individual objectives $J_k(w)$ be ν_k-strongly convex and δ_k-smooth, and the consensus objective $J(w)$ be ν-strongly convex with δ-Lipschitz gradients (see Section 2.1). Suppose further that the local gradient approximations $\widehat{\nabla J}_k(\cdot)$ satisfy the conditions (3.22) and (3.23) with constants $\alpha_k^2=\gamma_k^2=0$ and $\beta_k^2,\sigma_k^2\geq 0$. Suppose further that the gradient approximations are smooth in the mean-square sense, satisfying (5.43). Then, the error $\widetilde{\boldsymbol{w}}_i \triangleq w^o-\boldsymbol{w}_i$ of the iterates generated by the federated averaging algorithm (5.9)–(5.11) satisfy:*

$$
\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \left(\sqrt{\lambda}+O(\mu^3)\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2+2\mu^2\sigma_{\text{fed}}^2+O(\mu^3) \tag{5.55}
$$

where

$$
\sqrt{\lambda}=\sqrt{1-2\mu\nu+\mu^2\left(\delta^2+\beta_{\text{fed}}^2\right)}\leq 1-\mu\nu+\frac{\mu^2}{2}\left(\delta^2+\beta_{\text{fed}}^2\right) \tag{5.56}
$$

Then, for sufficiently small step-sizes it holds that $\sqrt{\lambda}+O(\mu^3)<1$ and we can iterate this relation to find:

$$
\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq (\sqrt{\lambda}^i+O(\mu^3))\|\widetilde{w}_0\|^2+\frac{4\mu\sigma_{\text{fed}}^2}{\nu}+O(\mu^2) \tag{5.57}
$$

Proof: The argument follows a similar structure to that of 3.1, after making the needed adjustments to handle the biased perturbation $\boldsymbol{d}_i$. We subtract recursion (5.21) from w^o and square both sides to obtain for the error $\widetilde{\boldsymbol{w}}_i \triangleq w^o-\boldsymbol{w}_i$:

$$
\begin{aligned}
\|\widetilde{\boldsymbol{w}}_i\|^2 &= \left\|\widetilde{\boldsymbol{w}}_{i-1}+\mu\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})+\mu\boldsymbol{d}_i\right\|^2 \\
&\overset{(a)}{\leq} \frac{1}{\alpha}\left\|\widetilde{\boldsymbol{w}}_{i-1}+\mu\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})\right\|^2+\frac{\mu^2}{1-\alpha}\|\boldsymbol{d}_i\|^2 1-2\mu\nu+\mu^2\left(\delta^2+\beta_{\text{fed}}^2\right)
\end{aligned}
\tag{5.58}
$$

where (a) follows from Jensen's inequality for all $0<\alpha<1$. The first term $\left\|\widetilde{\boldsymbol{w}}_{i-1}+\mu\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})\right\|^2$ is identical to the term we encountered when proving convergence non-cooperative stochastic gradient descent in Theorem 3.1. Applying the same arguement that led to (3.61), we find:

$$\begin{aligned}&\frac{1}{\alpha}\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i-1} + \mu\widehat{\nabla J}^{\text{fed}}(\boldsymbol{w}_{i-1})\right\|^2\right\}\\ &\le \frac{\lambda}{\alpha}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\frac{\sigma_{\text{fed}}^2}{\alpha}\end{aligned} \tag{5.59}$$

where

$$\lambda = 1 - 2\mu\nu + \mu^2\left(\delta^2 + \beta_{\text{fed}}^2\right) \tag{5.60}$$

For the full error recursion (5.58), we then have:

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 &= \frac{\lambda}{\alpha}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\frac{\sigma_{\text{fed}}^2}{\alpha} + \frac{\mu^2}{1-\alpha}\mathbb{E}\|\boldsymbol{d}_i\|^2\\ &\overset{(5.54)}{\le} \frac{\lambda}{\alpha}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\frac{\sigma_{\text{fed}}^2}{\alpha} + \frac{\mu^2}{1-\alpha}\left(O(\mu^2)\cdot\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + O(\mu^2)\right)\end{aligned} \tag{5.61}$$

If we set $\alpha = \sqrt{\lambda} < 1$, we have:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \sqrt{\lambda}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\frac{\sigma_{\text{fed}}^2}{\sqrt{\lambda}} + \frac{\mu^2}{1-\sqrt{\lambda}}\left(O(\mu^2)\cdot\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + O(\mu^2)\right) \tag{5.62}$$

To simplify this relation further, we note that:

$$\sqrt{\lambda} = \sqrt{1 - 2\mu\nu + \mu^2\left(\delta^2 + \beta_{\text{fed}}^2\right)} \le 1 - \mu\nu + \frac{\mu^2}{2}\left(\delta^2 + \beta_{\text{fed}}^2\right) \tag{5.63}$$

and hence:

$$\frac{\mu^2}{1-\sqrt{\lambda}} = \frac{\mu^2}{1 - \left(1 - 2\mu\nu + \mu^2\left(\delta^2 + \beta_{\text{fed}}^2\right)\right)} = \frac{\mu}{2\nu - \mu\left(\delta^2 + \beta_{\text{fed}}^2\right)} = O(\mu) \tag{5.64}$$

As long as μ is small enough so that $\sqrt{\lambda} \ge \frac{1}{2}$, it holds that:

$$\mu^2\frac{\sigma_{\text{fed}}^2}{\sqrt{\lambda}} \le 2\mu^2\sigma_{\text{fed}}^2 \tag{5.65}$$

A sufficient condition for this is $\mu \le \frac{\nu}{2}$. Grouping terms yields the result. □

5.2 PROBLEMS

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.

5.2 Prove Lemma 5.1.

5.3 Consider the federated averaging algorithm for the least-mean square problem, which takes the form:

$$\boldsymbol{\phi}_{k,e} = \boldsymbol{\phi}_{k,e-1} + \frac{\mu}{E_k}\frac{1}{B_k}\sum_{b=1}^{B_k}\boldsymbol{h}_{k,i,e,b}\left(\boldsymbol{\gamma}_{k,i,e,b} - \boldsymbol{h}_{k,i,e,b}^{\mathsf{T}}\boldsymbol{\phi}_{k,e-1}\right) \tag{5.66}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,E_k} \tag{5.67}$$

$$\boldsymbol{w}_i = \frac{1}{L}\sum_{k\in\boldsymbol{\mathcal{L}}_i}\boldsymbol{\psi}_{k,i} \tag{5.68}$$

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.