9 Convergence of Decentralized Algorithms

We developed in Chapter 8 a large number of decentralized algorithms for consensus optimization problems of the form:

$$\min_{w} \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{9.1}$$

We employed different techniques to develop different algorithms, ranging from penalties to primal-dual arguments and gradient-tracking, resulting in a number of different recursions with variations ranging from nuanced to significant. It is reasonable to expect these constructions to result in varying behavior, and the goal of the next two chapters will be to ellucidate and quantify these differences by analyzing the convergence behavior of the various decentralized algorithms. Accounting for these nuanced differences in a manner that allows us to differentiate convergence behavior of different decentralized algorithms will require detailed analysis. Before dwelving into the details, we note that on a high level, all algorithms we have seen thus far employ a combination of *averaging* steps to encourage consensus, and gradient steps to yield optimization. As a result, in analyzing the convergence of algorithms for decentralized optimization, we will be combining techniques from Chapter 7, where we studied consensus techniques over graphs, and Chapters 2 and–3, where we developed single-agent optimization algorithms.

9.1 NETWORK BASIS TRANSFORMATION

When we studied the dynamics of decentralized consensus algorithms in Chapter 7, we found it useful to sperately study the evolution of the network centroid and that of each individual agent's deviation from the centroid. In the context of decentralized optimization algorithm, we will employ the same kind of technique. We illustrate this in the context of the consensus+innovations algorithm, repeated here for reference in network quantities:

$$\mathcal{W}_i = \mathcal{A}^{\mathsf{T}}\,\mathcal{W}_{i-1} - \mu \nabla \mathcal{J}(\mathcal{W}_{i-1}) \tag{9.2}$$

or in terms of node quantities:

$$w_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} w_{\ell,i-1} - \mu \nabla J_k(w_{k,i-1}) \tag{9.3}$$

For generality, we will allow for stochastic gradient approximations, resulting in:

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{w}_{\ell,i-1} - \mu \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{9.4}$$

whre we replaced the true gradient $\nabla J_k(w_{k,i-1})$ by its stochastic approximation $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$, and changed the iterates $\boldsymbol{w}_{k,i}$ to utilize bold font since, as a result of employing stochastic approximations of the gradient, they will now be random themselves. In network notation, we can then write:

$$\boldsymbol{\mathcal{W}}_i = \mathcal{A}^\mathsf{T}\, \boldsymbol{\mathcal{W}}_{i-1} - \mu \widehat{\nabla \mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) \tag{9.5}$$

where we defined the network gradient approximation:

$$\widehat{\nabla \mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) = \begin{pmatrix} \widehat{\nabla J}_1(\boldsymbol{w}_{1,i-1}) \\ \widehat{\nabla J}_2(\boldsymbol{w}_{2,i-1}) \\ \vdots \\ \widehat{\nabla J}_K(\boldsymbol{w}_{K,i-1}) \end{pmatrix} \tag{9.6}$$

Now recall from 6.6 that the weight matrix $\mathcal{A}$, when generated from a strongly connected graph, is primitive, and as a result has a very structured Jordan decomposition decomposition $A = V_\epsilon J V_\epsilon^{-1}$:

$$V_\epsilon = \begin{bmatrix} p & V_R \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & J_\epsilon \end{bmatrix}, \quad V_\epsilon^{-1} = \begin{bmatrix} \mathbb{1}^\mathsf{T} \\ V_L^\mathsf{T} \end{bmatrix} \tag{9.7}$$

In this chapter, we will be employing symmetric combination matrices $A = A^\mathsf{T}$, in which case the Jordan decomposition reduces to the eigendecomposition, and we can more simply write $A = V \Lambda V^\mathsf{T}$ with $V^\mathsf{T} V = V V^\mathsf{T} = I_K$ and:

$$V = \begin{bmatrix} \frac{1}{\sqrt{K}} \mathbb{1} & V_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \tag{9.8}$$

The matrix Λ_2, which corresponds to the Jordan matrix J_ϵ, in (9.8) is now a diagonal matrix, with $\lambda_2(A)$ through $\lambda_K(A)$ on the diagonal and $\rho(\Lambda_2) < 1$. Given the eigendecomposition of A, we can deduce the eigendecomposition of $\mathcal{A} = A \otimes I_M$ through the observation that:

$$A = V \Lambda V^\mathsf{T} \Longleftrightarrow \mathcal{A} = (V \otimes I_M)(\Lambda \otimes I_M)(V \otimes I_M)^\mathsf{T} = \mathcal{V} \mathcal{\Lambda} \mathcal{V}^\mathsf{T} \tag{9.9}$$

where we defined:

$$\mathcal{V} = V \otimes I_M \tag{9.10}$$

$$\mathcal{\Lambda} = \Lambda \otimes I_M \tag{9.11}$$

We use $\mathcal{V}$ to define a basis transformation for the consensus+innovations recursion (9.5):

$$\begin{aligned}
\mathcal{V}^{\mathsf{T}} \mathcal{w}_i &= \mathcal{V}^{\mathsf{T}}\mathcal{A}^{\mathsf{T}} \mathcal{w}_{i-1} - \mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \\
&\overset{(a)}{=} \mathcal{V}^{\mathsf{T}}\mathcal{A}^{\mathsf{T}}\mathcal{V}\mathcal{V}^{\mathsf{T}} \mathcal{w}_{i-1} - \mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \\
&\overset{(b)}{=} \mathcal{V}^{\mathsf{T}}\mathcal{A}\mathcal{V}\mathcal{V}^{\mathsf{T}} \mathcal{w}_{i-1} - \mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \\
&\overset{(c)}{=} \Lambda \mathcal{V}^{\mathsf{T}} \mathcal{w}_{i-1} - \mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \qquad (9.12)
\end{aligned}$$

where (a) follows since $\mathcal{V}$ is orthogonal, (b) follows by symmetry of A and (c) employs the eigendecomposition of A. If we define $\mathcal{w}_i' \triangleq \mathcal{V}^{\mathsf{T}} \mathcal{w}_i$, we can write:

$$\mathcal{w}_i' = \Lambda \mathcal{w}_{i-1}' - \mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \qquad (9.13)$$

We find that the network basis transformation partially diagonalizes the network recursion, since $\Lambda = \Lambda \otimes I_M$ is block-diagonal. We say "partially" here because (9.13) is still driven by the term $\mu \mathcal{V}^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1})$, which causes coupling between individual blocks of $\mathcal{w}_i'$. To more clearly demonstrate this fact, we exploit the block-structure of the weights $\mathcal{w}_i$ and transformation $\mathcal{V}$. We have:

$$\begin{aligned}
&\mathcal{V}^{\mathsf{T}} \mathcal{w}_i \\
&= \left(\begin{bmatrix} \frac{1}{\sqrt{K}} \mathbb{1}^{\mathsf{T}} \\ V_2^{\mathsf{T}} \end{bmatrix} \otimes I_M \right) \mathcal{w}_i \\
&= \begin{bmatrix} \frac{1}{\sqrt{K}} \mathbb{1}^{\mathsf{T}} \otimes I_M \\ V_2^{\mathsf{T}} \otimes I_M \end{bmatrix} \mathcal{w}_i \\
&= \begin{bmatrix} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \boldsymbol{w}_{k,i} \\ \mathcal{V}_2^{\mathsf{T}} \mathcal{w}_i \end{bmatrix} \\
&\overset{(a)}{=} \begin{bmatrix} \sqrt{K} \boldsymbol{w}_{c,i} \\ \mathcal{V}_2^{\mathsf{T}} \mathcal{w}_i \end{bmatrix} \qquad (9.14)
\end{aligned}$$

where in (a) we defined the network centroid:

$$\boldsymbol{w}_{c,i} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{w}_{k,i} \qquad (9.15)$$

We observe that the first block in the transformed network vector $\mathcal{w}_i' = \mathcal{V}^{\mathsf{T}} \mathcal{w}_i$ corresponds to a scaled version of the network centroid $\boldsymbol{w}_{c,i}$. For the right-hand side, we have:

$$\begin{aligned}
&\begin{bmatrix} \sqrt{K}\boldsymbol{w}_{c,i} \\ \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_i \end{bmatrix} \\
&= \begin{bmatrix} I_M & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M \\ \mathcal{V}_2^{\mathsf{T}} \end{bmatrix} \mathscr{w}_{i-1} - \mu \begin{bmatrix} \frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M \\ \mathcal{V}_2^{\mathsf{T}} \end{bmatrix} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \\
&= \begin{bmatrix} \frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M \\ \Lambda_2 \mathcal{V}_2^{\mathsf{T}} \end{bmatrix} \mathscr{w}_{i-1} - \mu \begin{bmatrix} \frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M \\ \mathcal{V}_2^{\mathsf{T}} \end{bmatrix} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \\
&= \begin{bmatrix} \left(\frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathscr{w}_{i-1} \\ \Lambda_2 \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_{i-1} \end{bmatrix} - \mu \begin{bmatrix} \left(\frac{1}{\sqrt{K}}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \\ \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \boldsymbol{w}_{k,i-1} \\ \Lambda_2 \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_{i-1} \end{bmatrix} - \mu \begin{bmatrix} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \\ \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{K}\boldsymbol{w}_{c,i-1} \\ \Lambda_2 \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_{i-1} \end{bmatrix} - \mu \begin{bmatrix} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \\ \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \end{bmatrix} \qquad (9.16)
\end{aligned}$$

We observe that the transformed recursion partially decouples into two recursions:

$$\sqrt{K}\boldsymbol{w}_{c,i} = \sqrt{K}\boldsymbol{w}_{c,i-1} - \frac{\mu}{\sqrt{K}} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \qquad (9.17)$$

$$\mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_i = \Lambda_2 \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_{i-1} - \mu \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \qquad (9.18)$$

We divide (9.17) by $\sqrt{K}$ and find:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \qquad (9.19)$$

$$\mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_i = \Lambda_2 \mathcal{V}_2^{\mathsf{T}}\, \mathscr{w}_{i-1} - \mu \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\mathscr{w}_{i-1}) \qquad (9.20)$$

Examination of (9.19) reveals that the network centroid $\boldsymbol{w}_{c,i}$ evolves *almost* as the centralized stochastic gradient algorithm we studied in Chapter 4, except that stochastic gradient approximations $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$ are evaluated at the local iterates $\boldsymbol{w}_{k,i-1}$ rather than the centroid $\boldsymbol{w}_{c,i-1}$. If it were not for this detail, we could study the evolution of (9.19) independently of (9.20). The deviation between $\boldsymbol{w}_{k,i-1}$ and $\boldsymbol{w}_{c,i-1}$ couples the two recursions and will necessitate more nuanced analysis. If we are able to bound the network disagreement, i.e., the deviation between $\boldsymbol{w}_{c,i-1}$ and $\boldsymbol{w}_{k,i-1}$, we will be able to conclude that (9.19) approximately tracks the dynamics of the centralized stochastic gradient recursion. It turns out that (9.20) allows us to do precisely this. To this end, note that:

$$\mathscr{w}_i' = \mathcal{V}^{\mathsf{T}}\, \mathscr{w}_i \Longrightarrow \mathcal{V}\, \mathscr{w}_i' = \mathcal{V}\mathcal{V}^{\mathsf{T}}\, \mathscr{w}_i = \mathscr{w}_i \qquad (9.21)$$

and hence

$$\begin{aligned}
\boldsymbol{\mathcal{W}}_i &= \left[\begin{array}{cc} \frac{1}{\sqrt{K}}\mathbb{1} \otimes I_M & \mathcal{V}_2 \end{array}\right] \boldsymbol{\mathcal{W}}_i' \\
&= \left[\begin{array}{cc} \frac{1}{\sqrt{K}}\mathbb{1} \otimes I_M & \mathcal{V}_2 \end{array}\right] \left[\begin{array}{c} \sqrt{K}\boldsymbol{w}_{c,i} \\ \mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i \end{array}\right] \\
&= (\mathbb{1} \otimes I_M)\, \boldsymbol{w}_{c,i} + \mathcal{V}_2 \mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i \\
&= \mathbb{1} \otimes \boldsymbol{w}_{c,i} + \mathcal{V}_2 \mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i
\end{aligned} \tag{9.22}$$

After rearranging, we have:

$$\boldsymbol{\mathcal{W}}_i - \mathbb{1} \otimes \boldsymbol{w}_{c,i} = \mathcal{V}_2 \mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i \tag{9.23}$$

and hence $\mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i$, which appears in (9.20) can be seen to measure the network deviation from the centroid. It follows that as long as we can show that $\|\mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i\|^2$ can be shown to be small in some sense, we will be able to argue that (9.19) tracks the centralized stochastic gradient recursion with reasonable accuracy. We will make this precise in Section 9.2 further ahead. First, we comment on the necessary adjustments in the case of alternative decentralized algorithms.

9.1.1 Adapt-then-Combine (ATC) Diffusion

Recal that we motivated in Chapter 8 the Adapt-then-Combine (ATC) diffusion algorithm (8.62)–(8.63) by employing an incremental variation of the argument that led to the consensus+innovations algorithm:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i} - \mu \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{9.24}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell,i} \tag{9.25}$$

We are again allowing for stochastic gradient approximations $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$. In terms of network quantities, we have:

$$\boldsymbol{\mathcal{W}}_i = \mathcal{A}^{\mathsf{T}} \left(\boldsymbol{\mathcal{W}}_{i-1} - \mu \widehat{\nabla \mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) \right) \tag{9.26}$$

Comparing (9.26) and (9.5), the only difference is the presence of a pair of brackets which causes the mixing matrix $\mathcal{A}^{\mathsf{T}}$ to be applied to the gradient update $-\mu \widehat{\nabla \mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1})$ in addition to the weights $\boldsymbol{\mathcal{W}}_{i-1}$ themselves. Applying the same network basis transformation $\boldsymbol{\mathcal{W}}_i' = \mathcal{V}^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i$, and repeating the argument (see Prob. 9.1) that led to the decomposition (9.19)–(9.20), we find for the diffusion algorithm:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{9.27}$$

$$\mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i = \Lambda_2 \mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_{i-1} - \mu \Lambda_2 \mathcal{V}_2^{\mathsf{T}} \widehat{\nabla \mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) \tag{9.28}$$

Comparing now the decomposition (9.27)–(9.28) for the diffusion algorithm with the decomposition (9.19)–(9.20) for the consensus+innovations algorithm, we observe two key insights. First, the recursions for the network centroid (9.27) and (9.19) are identical. Second, the recursions for the network deviation (9.28) and (9.20) are structurally similar, but distinguished by an additional factor Λ_2 multiplying the driving term $-\mu\Lambda_2\mathcal{V}_2^{\mathsf{T}}\widehat{\nabla \mathcal{J}}(\mathsf{w}_{i-1})$ in the case of the diffusion algorithm. This factor results from the fact that the mixing operation in the case of the diffusion algorithm is applied to both the weights themselves as well as the gradient update. As we will see when we derive convergence conditions for both algorithm, this subtle difference is the source of improved stability properties of diffusion type algorithms.

We remark that the fact that the centroid recursions for consensus+innovations and diffusion algorithms are identical is a useful insight, but does imply that the trajectories of both centroids will be identical. This is because the centroid recursions are coupled with the deviation recursions (9.28) and (9.20) through the iterates $\boldsymbol{w}_{k,i-1}$, where the gradient approximations are evaluated. These distinctions will seep into the centroid recursions and cause varying dynamics of the centroid as well. Nevertheless, the structural similarity of the recursion allows us to formulate a general form of penalty-based algorithms, and develop convergence analysis that will apply to both algorithms. In particular, we will study a decomposition of the form:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^{K}\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{9.29}$$

$$\mathcal{V}_2^{\mathsf{T}}\,\mathsf{w}_i = \Lambda_2\mathcal{V}_2^{\mathsf{T}}\,\mathsf{w}_{i-1} - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\widehat{\nabla \mathcal{J}}(\mathsf{w}_{i-1}) \tag{9.30}$$

where we recover the consensus+innovations algorithm by setting $\mathcal{D} = I$, and the diffusion algorithms by setting $\mathcal{D} = \Lambda_2$.

9.1.2 Primal-Dual and Gradient-Tracking Algorithms

For primal-dual and gradient-tracking based algorithms we can apply the same network basis transformation $\mathsf{w}_i' = \mathcal{V}^{\mathsf{T}}\,\mathsf{w}_i$ and again find that the network centroid satisfies the recursion:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^{K}\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{9.31}$$

We explore the detailed derivations in Prob. 9.2. Recursions of the network deviations for all of these algorithms can also be obtained, but require some more effort to account for dual variables and gradient tracking terms. To avoid a digression, we will focus in this chapter on penalty-based algorithms (consensus+innovations and diffusion) and will develop convergence analysis for primal-dual and gradient-tracking based algorithms in a future chapter.

9.2 ERROR RECURSION

We return to the general decomposition for penalty-based algorithms (9.29)–(9.30). To make the coupling explicit, we introduce the error terms:

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{9.32}$$

$$\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) \triangleq \nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1}) \tag{9.33}$$

We can then write:

$$\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) = \nabla J_k(\boldsymbol{w}_{c,i-1}) + \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) \tag{9.34}$$

Here, $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ corresponds to a *gradient noise* term analogous to the ones we have encountered in previous non-cooperative, centralized and federated implementations of stochastic gradient algorithms. The second term $\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})$ represents a second deviation term resulting from lack of consensus across the network. In particular, as long as $\boldsymbol{w}_{k,i-1} \approx \boldsymbol{w}_{c,i-1}$, we would expect $\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})$ to be small. We define the network versions of these quantities as well:

$$\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\} \tag{9.35}$$

$$\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\left\{\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\} \tag{9.36}$$

These definitions allow us to reformulate the recursions (9.29)–(9.30) as:

$$\begin{aligned}\boldsymbol{w}_{c,i} &= \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^{K}\nabla J_k(\boldsymbol{w}_{c,i-1}) - \frac{\mu}{K}\sum_{k=1}^{K}\left(\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right)\\ &= \boldsymbol{w}_{c,i-1} - \mu\nabla J(\boldsymbol{w}_{c,i-1}) - \frac{\mu}{K}\sum_{k=1}^{K}\left(\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right)\end{aligned} \tag{9.37}$$

$$\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{W}}_i = \Lambda_2\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{W}}_{i-1} - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\nabla\mathcal{J}(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\left(\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) + \boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})\right) \tag{9.38}$$

Lemma 9.1 (**Perturbation bounds**). *Suppose each local cost $J_k(\cdot)$ has δ_k-Lipschitz gradients, and the gradient noise terms $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ induced by the local gradient approximations satisfy:*

$$\mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{w}_{k,i-1}\right\} = 0 \tag{9.39}$$

$$\mathbb{E}\left\{\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{w}_{k,i-1}\right\} \le \beta_k^2\|w_k^o - \boldsymbol{w}_{k,i-1}\|^2 + \sigma_k^2 \tag{9.40}$$

Assume additionally that the local gradient noise processes are independent, implying for all $k \neq \ell$:

$$\mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\boldsymbol{s}_{\ell,i}(\boldsymbol{w}_{\ell,i-1})^{\mathsf{T}}|\boldsymbol{w}_{k,i-1},\boldsymbol{w}_{\ell,i-1}\right\} = 0 \tag{9.41}$$

Then, the gradient noise terms satisfy the following bounds:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2|\boldsymbol{w}_{k,i-1}\right\} \\
&\le \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{3}{K^2}\beta_{\max}^2\|\mathcal{V}_2\|^2\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\,\|^2 + \sigma^2 \qquad (9.42)\\
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{\mathcal{w}}_{i-1})\|^2|\,\boldsymbol{\mathcal{w}}_{i-1}\right\} \\
&\le 3\left(\sum_{k=1}^{K}\beta_k^2\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + K^2\beta_{\max}^2\|\mathcal{V}_2\|^2\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\,\|^2 + K^2\sigma^2 \qquad (9.43)
\end{aligned}$$

where we defined $\beta_{\max}^2 = \max_k \beta_k^2$, $\delta_{\max}^2 = \max_k \delta_k^2$ *and:*

$$\sigma^2 = \frac{1}{K^2}\sum_{k=1}^{K}\left(3\beta_k^2\|w_k^o - w^o\|^2 + \sigma_k^2\right) \qquad (9.44)$$

The perturbation terms arising from the network disagreement satisfy the bounds:

$$\begin{aligned}
\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2 &\le \frac{\delta_{\max}^2}{K}\|\mathcal{V}_2\|^2\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\,\|^2 \qquad (9.45)\\
\|\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{w}}_{i-1})\|^2 &\le \delta_{\max}^2\|\mathcal{V}_2\|^2\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\,\|^2 \qquad (9.46)
\end{aligned}$$

Finally, we have the bound:

$$\|\nabla\mathcal{J}(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\|^2 \le 2\left(\sum_{k=1}^{K}\delta_k^2\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 2\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2 \qquad (9.47)$$

Proof: We prove each bound one by one. For the aggregate gradient noise we have:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 |\boldsymbol{w}_{k,i-1}\right\} \\
\overset{(9.41)}{=}\; & \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\left\{\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{w}_{k,i-1}\right\} \\
\overset{(9.40)}{=}\; & \frac{1}{K^2}\sum_{k=1}^{K}\left(\beta_k^2\|w_k^o-\boldsymbol{w}_{k,i-1}\|^2+\sigma_k^2\right) \\
=\; & \frac{1}{K^2}\sum_{k=1}^{K}\left(\beta_k^2\|w_k^o-w^o+w^o-\boldsymbol{w}_{c,i-1}+\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\|^2+\sigma_k^2\right) \\
\overset{(a)}{\le}\; & \frac{1}{K^2}\sum_{k=1}^{K}\left(3\beta_k^2\|w_k^o-w^o\|^2+3\beta_k^2\|w^o-\boldsymbol{w}_{c,i-1}\|^2+3\beta_k^2\|\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\|^2+\sigma_k^2\right) \\
\overset{(b)}{=}\; & \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|w^o-\boldsymbol{w}_{c,i-1}\|^2+\frac{3}{K^2}\sum_{k=1}^{K}\left(\beta_k^2\|\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\|^2\right)+\sigma^2 \\
\overset{(c)}{\le}\; & \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|w^o-\boldsymbol{w}_{c,i-1}\|^2+\frac{3}{K^2}\beta_{\max}^2\sum_{k=1}^{K}\|\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\|^2+\sigma^2 \\
\le\; & \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|w^o-\boldsymbol{w}_{c,i-1}\|^2+\frac{3}{K^2}\beta_{\max}^2\|\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}-\mathcal{w}_{k,i-1}\|^2+\sigma^2 \\
\overset{(9.23)}{=}\; & \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|w^o-\boldsymbol{w}_{c,i-1}\|^2+\frac{3}{K^2}\beta_{\max}^2\left\|\mathcal{V}_2\mathcal{V}_2^{\mathsf{T}}\,\mathcal{w}_{i-1}\right\|^2+\sigma^2 \\
\overset{(d)}{=}\; & \frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2+\frac{3}{K^2}\beta_{\max}^2\|\mathcal{V}_2\|^2\left\|\mathcal{V}_2^{\mathsf{T}}\,\mathcal{w}_{i-1}\right\|^2+\sigma^2 \qquad (9.48)
\end{aligned}
$$

where (a) follows from Jensen's inequality implying $\|a+b+c\|^2 \le 3\|a\|^2+3\|b\|^2+3\|c\|^2$ and in (b) we grouped terms and defined:

$$\sigma^2 = \frac{1}{K^2}\sum_{k=1}^{K}\left(3\beta_k^2\|w_k^o-w^o\|^2+\sigma_k^2\right) \qquad (9.49)$$

Step (d) follows after defining $\widetilde{\boldsymbol{w}}_{c,i-1} \triangleq w^o-\boldsymbol{w}_{c,i-1}$ and sub-multiplicity of norms. The full gradient noise bound (9.43) follows analogously after noting that $\|\boldsymbol{s}_i(\mathcal{w}_{i-1})\|^2 =$

$\sum_{k=1}^{K} \|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2$. For the aggregate deviation perturbation, we have:

$$\begin{aligned}
&\left\| \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) \right\|^2 \\
&\overset{(a)}{\leq} \frac{1}{K} \sum_{k=1}^{K} \|\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\|^2 \\
&\leq \frac{1}{K} \sum_{k=1}^{K} \|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1})\|^2 \\
&\overset{(b)}{\leq} \frac{1}{K} \sum_{k=1}^{K} \delta_k^2 \|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\|^2 \\
&\leq \frac{1}{K} \delta_{\max}^2 \sum_{k=1}^{K} \|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\|^2 \\
&= \frac{\delta_{\max}^2}{K} \|\mathbb{1} \otimes \boldsymbol{w}_{c,i-1} - \boldsymbol{\mathcal{W}}_{k,i-1}\|^2 \\
&\overset{(c)}{\leq} \frac{\delta_{\max}^2}{K} \|\mathcal{V}_2\|^2 \|\mathcal{V}_2^{\mathsf{T}} \boldsymbol{\mathcal{W}}_{i-1}\|^2
\end{aligned} \tag{9.50}$$

where (a) follows from Jensen's inequality, (b) follows from the Lipschitz condition on the gradients, and (c) mirrors the same argument as for the aggregate noise term before. The full deviation bound (9.46) follows analogously after noting that $\|\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})\|^2 = \sum_{k=1}^{K} \|\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\|^2$. For the final perturbation bound, we have:

$$\begin{aligned}
&\|\nabla \mathcal{J}(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1})\|^2 \\
&= \sum_{k=1}^{K} \|\nabla J_k(\boldsymbol{w}_{c,i-1})\|^2 \\
&= \sum_{k=1}^{K} \|\nabla J_k(\boldsymbol{w}_{c,i-1}) - \nabla J_k(w^o) + \nabla J_k(w^o)\|^2 \\
&\overset{(a)}{\leq} 2 \sum_{k=1}^{K} \|\nabla J_k(\boldsymbol{w}_{c,i-1}) - \nabla J_k(w^o)\|^2 + 2 \sum_{k=1}^{K} \|\nabla J_k(w^o)\|^2 \\
&\leq 2 \left(\sum_{k=1}^{K} \delta_k^2 \right) \|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 2 \sum_{k=1}^{K} \|\nabla J_k(w^o)\|^2
\end{aligned} \tag{9.51}$$

where (a) follows from Jensen's inequality. □

Equipped with the perturbation bounds from Lemma 9.1, we are now able develop coupled error recursions. We begin with the network centroid (9.37), and subtract from w^o to obtain:

$$\widetilde{\boldsymbol{w}}_{c,i} = \widetilde{\boldsymbol{w}}_{c,i-1} + \mu \nabla J(\boldsymbol{w}_{c,i-1}) + \frac{\mu}{K} \sum_{k=1}^{K} \left(\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) \right) \tag{9.52}$$

After squaring and taking conditional expectations:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&= \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1}) + \frac{\mu}{K}\sum_{k=1}^{K}\left(\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) + \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right)\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&\overset{(a)}{=} \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1}) + \frac{\mu}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&\quad + \mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&\overset{(a)}{=} \frac{1}{\alpha}\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1})\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\} + \frac{\mu^2}{1-\alpha}\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&\quad + \mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}\\
&\overset{(b)}{=} \frac{1}{\alpha}\|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1})\|^2 + \frac{\mu^2}{1-\alpha}\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2\\
&\quad + \mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 |\, \boldsymbol{\mathsf{w}}_{i-1}\right\}
\end{aligned}
\tag{9.53}
$$

where in (a) we removed cross-terms due to (9.39), and in (b) we removed the expectation around terms involving $\widetilde{\boldsymbol{w}}_{c,i-1}$ or $\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})$ since these terms are fully-determined by, and hence deterministic after conditioning on $\boldsymbol{\mathsf{w}}_{i-1}$. The first term on the right-hand side is identical to the error recursion we encountered in Chapter 2, when proving convergence of gradient descent in Theorem 2.1. We can hence conclude that:

$$
\begin{aligned}
&\|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1})\|^2\\
&= \|\widetilde{\boldsymbol{w}}_{c,i-1} + \mu\nabla J(\boldsymbol{w}_{c,i-1}) - \mu\nabla J(w^o)\|^2\\
&\le \lambda\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2
\end{aligned}
\tag{9.54}
$$

where $\lambda = 1 - 2\mu\nu + \mu^2\delta^2$ and ν, δ are the strong-convexity and smoothness constants of the aggregate objective $J(w)$ respectively. If we set $\alpha = \sqrt{\lambda}$, we find:

$$\begin{aligned}
&\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 | \boldsymbol{\mathcal{w}}_{i-1}\right\} \\
&= \sqrt{\lambda}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{\mu^2}{1-\sqrt{\lambda}}\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 | \boldsymbol{\mathcal{w}}_{i-1}\right\} \\
&\overset{(a)}{\le} \sqrt{\lambda}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{\mu}{\nu-\frac{1}{2}\mu\delta^2}\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mu^2\mathbb{E}\left\{\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2 | \boldsymbol{\mathcal{w}}_{i-1}\right\}
\end{aligned} \tag{9.55}$$

where (a) follows since:

$$\sqrt{\lambda} = \sqrt{1-2\mu\nu+\mu^2\delta^2} \le 1-\mu\nu+\frac{1}{2}\mu^2\delta^2 \tag{9.56}$$

Finally, we apply the perturbation bounds from Lemma 9.1:

$$\begin{aligned}
&\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 | \boldsymbol{\mathcal{w}}_{i-1}\right\} \\
&\le \sqrt{\lambda}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{\mu}{\nu-\frac{1}{2}\mu\delta^2}\frac{\delta_{\max}^2}{K}\|\mathcal{V}_2\|^2\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\right\|^2 \\
&\quad + \mu^2\frac{3}{K^2}\left(\sum_{k=1}^{K}\beta_k^2\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \mu^2\frac{3}{K^2}\beta_{\max}^2\|\mathcal{V}_2\|^2\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\right\|^2 + \mu^2\sigma^2 \\
&= \left(\sqrt{\lambda}+O(\mu^2)\right)\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + O(\mu)\|\mathcal{V}_2\|^2\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\right\|^2 + \mu^2\sigma^2
\end{aligned} \tag{9.57}$$

After taking expectations once more to remove the conditioning, we obtain an error recursion for the network centroid:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \le \left(\sqrt{\lambda}+O(\mu^2)\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + O(\mu)\|\mathcal{V}_2\|^2\mathbb{E}\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\right\|^2 + \mu^2\sigma^2 \tag{9.58}$$

We employ a similar technique to bound the evolution of the network disagreement (9.38). Instead of relying on the descent enabled by the gradient along the aggregate loss, now the contraction will be enabled by the portion of the eigenvalue matrix Λ_2, which has spectral radius strictly less than one. Taking norms and expecations on both sides of (9.38):

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_i\right\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
\overset{(9.39)}{=}\;& \mathbb{E}\left\{\left\|\Lambda_2\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1} - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\nabla\mathcal{J}(\boldsymbol{w}_{i-1}) - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\boldsymbol{d}_{i-1}(\boldsymbol{w}_{i-1})\right\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
&+\mu^2\mathbb{E}\left\{\left\|\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
\overset{(a)}{=}\;& \left\|\Lambda_2\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1} - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\nabla\mathcal{J}(\boldsymbol{w}_{i-1}) - \mu\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\boldsymbol{d}_{i-1}(\boldsymbol{w}_{i-1})\right\|^2\\
&+\mu^2\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
\overset{(b)}{=}\;& \frac{1}{\lambda_2}\left\|\Lambda_2\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1}\right\|^2 + \frac{\mu^2}{1-\lambda_2}\left\|\mathcal{D}\mathcal{V}_2^{\mathsf{T}}\nabla\mathcal{J}(\boldsymbol{w}_{i-1}) + \mathcal{D}\mathcal{V}_2^{\mathsf{T}}\boldsymbol{d}_{i-1}(\boldsymbol{w}_{i-1})\right\|^2\\
&+\mu^2\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
=\;& \lambda_2\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1}\right\|^2 + \frac{\mu^2}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}_2^{\mathsf{T}}\|^2\|\nabla\mathcal{J}(\boldsymbol{w}_{i-1}) + \boldsymbol{d}_{i-1}(\boldsymbol{w}_{i-1})\|^2\\
&+\mu^2\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \mid \boldsymbol{w}_{i-1}\right\}\\
\overset{(c)}{\le}\;& \lambda_2\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1}\right\|^2 + \frac{2\mu^2}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}_2^{\mathsf{T}}\|^2\|\nabla\mathcal{J}(\boldsymbol{w}_{i-1})\|^2 + \frac{2\mu^2}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}_2^{\mathsf{T}}\|^2\|\boldsymbol{d}_{i-1}(\boldsymbol{w}_{i-1})\|^2\\
&+\mu^2\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \mid \boldsymbol{w}_{i-1}\right\} \qquad (9.59)
\end{aligned}
$$

where (a) follows after applying sub-multiplicity of norms and removing redundant expectations. Step (b) again follows from Jensen's inequality with $\lambda_2 = \rho(\Lambda_2)$ and sub-multiplicity of norms. Step (c) follows from $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. Finally, after taking expectations and applying the bounds from Lemma **??**, we find:

$$
\begin{aligned}
\mathbb{E}\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_i\right\|^2 \le \left(\lambda_2 + O(\mu^2)\right)\mathbb{E}\left\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_{i-1}\right\|^2 + O(\mu^2)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2\\
+\frac{4\mu^2}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2 + \mu^2\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2K^2\sigma^2
\end{aligned}
\qquad (9.60)
$$

Having formulated coupled recursive relationships for both the network centroid and the disagreement, we can finally put everything together into one inequality recursion:

$$
\begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_i\,\|^2 \end{bmatrix} \le \Gamma \begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\boldsymbol{w}_i - 1\|^2 \end{bmatrix} + \begin{bmatrix} \mu^2\sigma^2 \\ \mu^2 b \end{bmatrix} \qquad (9.61)
$$

where

$$
\Gamma = \begin{bmatrix} \sqrt{\lambda} + O(\mu^2) & O(\mu) \\ O(\mu^2) & \lambda_2 + O(\mu^2) \end{bmatrix} \qquad (9.62)
$$

$$
b = \frac{4}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2 + \|\mathcal{D}\|^2\|\mathcal{V}^{\mathsf{T}}\|^2K^2\sigma^2 = O(\mu^2) \qquad (9.63)
$$

THEOREM 9.1 (Mean-square-behavior of penalty-based decentralized algorithms). *Suppose all conditions of Lemma 9.1 hold. Then there exists a step-size μ that is small enough, so that:*

$$\rho(\Gamma) \le \|\Gamma\|_1 = \max\left\{\sqrt{\lambda} + O(\mu^2), \lambda_2 + O(\mu)\right\} < 1 \tag{9.64}$$

and the mean-square decentralized primal algorithms converge in the sense that:

$$\limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\mathbf{w}_i\,\|^2 \end{bmatrix} \le (I-\Gamma)^{-1} \begin{bmatrix} \mu^2\sigma^2 \\ \mu^2 b \end{bmatrix} = \begin{bmatrix} \frac{\mu\sigma^2}{\nu} + O(\mu^2) \\ \mu^2\frac{b}{1-\lambda_2} \end{bmatrix} \tag{9.65}$$

Proof: First, note that for any matrix, its spectral radius is bounded by its norm, and hence:

$$\rho(\Gamma) \le \|\Gamma\|_1 = \max\left\{\sqrt{\lambda} + O(\mu^2), \lambda_2 + O(\mu)\right\} \tag{9.66}$$

Now, since $\sqrt{\lambda} = 1 - \mu\nu + O(\mu^2)$, we can always choose μ small enough so that $\sqrt{\lambda} + O(\mu^2) < 1$. Similarly, since $\sqrt{\lambda}$ is bounded away from 1, we can choose μ small enough so that $\lambda_2 + O(\mu) < 1$ and hence $\rho(\Gamma) < 1$. Once we have established that Γ is stable, we can iterate (9.61) to find:

$$\limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\mathbf{w}_i\,\|^2 \end{bmatrix} \le \Gamma \limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\mathbf{w}_i\,\|^2 \end{bmatrix} + \begin{bmatrix} \mu^2\sigma^2 \\ \mu^2 b \end{bmatrix} \tag{9.67}$$

which yields the result after rearranging. To obtain a more accurate description of the limiting behavior, we invert:

$$\begin{aligned}
(I-\Gamma)^{-1} &= \begin{bmatrix} 1-(1-\mu\nu+O(\mu^2)) & O(\mu) \\ O(\mu^2) & 1-\lambda_2+O(\mu^2) \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mu\nu - O(\mu^2) & O(\mu) \\ O(\mu^2) & 1-\lambda_2-O(\mu^2) \end{bmatrix}^{-1} \\
&= \frac{1}{(\mu\nu-O(\mu^2))(1-\lambda_2-O(\mu))-O(\mu^3)} \begin{bmatrix} 1-\lambda_2-O(\mu^2) & -O(\mu) \\ -O(\mu^2) & \mu\nu-O(\mu^2) \end{bmatrix} \\
&\approx \frac{1}{(\mu\nu-O(\mu^2))(1-\lambda_2-O(\mu))} \begin{bmatrix} 1-\lambda_2-O(\mu^2) & -O(\mu) \\ -O(\mu^2) & \mu\nu-O(\mu^2) \end{bmatrix} \\
&\approx \begin{bmatrix} \frac{1}{\mu\nu} & O(1) \\ O(\mu) & \frac{1}{1-\lambda_2} \end{bmatrix}
\end{aligned} \tag{9.68}$$

Then

$$\begin{aligned}
\limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\|\mathcal{V}_2^{\mathsf{T}}\,\mathbf{w}_i\,\|^2 \end{bmatrix} &= \begin{bmatrix} \frac{1}{\mu\nu} & O(1) \\ O(\mu) & \frac{1}{1-\lambda_2} \end{bmatrix} \begin{bmatrix} \mu^2\sigma^2 \\ \mu^2 b \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mu\sigma^2}{\nu} + O(\mu^2) \\ \mu^2\frac{b}{1-\lambda_2} \end{bmatrix}
\end{aligned} \tag{9.69}$$

□

9.3 PROBLEMS

9.1 Verify that the network basis transformation of the diffusion algorithm (9.26) satisfies the decomposition (9.27)–(9.28).

9.2 Show that the network centroid satisfies (9.31) for the EXTRA, Exact diffusion, gradient-tracking and AUG-DGM algorithms.