LECTURE #17: Stability of Multi-Agent Networks

Professor Ali H. Sayed
UCLA Electrical Engineering
Part V: Multi-Agent Network Stability and Performance
Chapter 8 (Evolution of Multi-Agent Networks, pp. 496-506):
Chapter 9 (Stability of Multi-Agent Networks, pp. 507-551):

Distributed Strategies
Table 8.2: Update equations for non-cooperative, diffusion, and consensus strategies.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>update equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-cooperative</td>
<td>[ w_{k,i} = w_{k,i-1} - \mu_k \nabla_{w^*} J_k (w_{k,i-1}) ]</td>
</tr>
</tbody>
</table>
| consensus          | \[
\begin{align*}
\psi_{k,i-1} &= \sum_{\ell \in N_k} a_{\ell k} w_{\ell,i-1} \\
 w_{k,i} &= \psi_{k,i-1} - \mu_k \nabla_{w^*} J_k (w_{k,i-1})
\end{align*}
\] |
| CTA diffusion      | \[
\begin{align*}
\psi_{k,i-1} &= \sum_{\ell \in N_k} a_{\ell k} w_{\ell,i-1} \\
 w_{k,i} &= \psi_{k,i-1} - \mu_k \nabla_{w^*} J_k (\psi_{k,i-1})
\end{align*}
\] |
| ATC diffusion      | \[
\begin{align*}
\psi_{k,i} &= w_{k,i-1} - \mu_k \nabla_{w^*} J_k (w_{k,i-1}) \\
 w_{k,i} &= \sum_{\ell \in N_k} a_{\ell k} \psi_{\ell,i}
\end{align*}
\] |
Unified Description

In a manner similar to (8.5), we can again describe these strategies by means of a single unifying description as follows:

\[
\begin{align*}
\phi_{k,i-1} & = \sum_{\ell \in N_k} a_{1,\ell_k} \ w_{\ell,i-1} \\
\psi_{k,i} & = \sum_{\ell \in N_k} a_{o,\ell_k} \ \phi_{\ell,i-1} - \mu_k \ \nabla^* \ J_k \left( \phi_{k,i-1} \right) \\
w_{k,i} & = \sum_{\ell \in N_k} a_{2,\ell_k} \ \psi_{\ell,i}
\end{align*}
\]  

(8.46)
where \( \{\phi_{k,i-1}, \psi_{k,i}\} \) denote \( M \times 1 \) intermediate variables, while the nonnegative entries of the \( N \times N \) matrices \( A_o = [a_{o,\ell k}] \), \( A_1 = [a_{1,\ell k}] \), and \( A_2 = [a_{2,\ell k}] \) satisfy the same conditions (7.10) and, hence, the matrices \( \{A_o, A_1, A_2\} \) are left-stochastic

\[
A_o^T 1 = 1, \quad A_1^T 1 = 1, \quad A_2^T 1 = 1
\]  

(8.47)

We assume that each of these combination matrices defines an underlying connected network topology so that none of their rows are identically zero.
Again, different choices for \( \{A_o, A_1, A_2\} \) correspond to different distributed strategies, as indicated earlier by (8.7)–(8.10), and where the left-stochastic matrix \( P \) represents the product:

\[
P \triangleq A_1 A_o A_2 \tag{8.48}
\]
Perron Eigenvector

(a) The matrix $P$ has a single eigenvalue at one.

(b) All other eigenvalues of $P$ are strictly inside the unit circle so that $\rho(P) = 1$.

(c) With proper sign scaling, all entries of the right-eigenvector of $P$ corresponding to the single eigenvalue at one are positive. Let $p$ denote this right-eigenvector, with its entries $\{p_k\}$ normalized to add up to one, i.e.,

$$ Pp = p, \quad 1^T p = 1, \quad p_k > 0, \quad k = 1, 2, \ldots, N $$

We refer to $p$ as the Perron eigenvector of $P$. 
Weighted Aggregate Cost

Following [68-70], we next introduce the vector:

\[ q \triangleq \text{diag}\{\mu_1, \mu_2, \ldots, \mu_N\}A_2p \quad (8.50) \]

It is clear that all entries of \( q \) are strictly positive since each \( \mu_k > 0 \) and the entries of \( A_2p \) are all positive. The latter statement follows from the fact that each entry of \( A_2p \) is a linear combination of the positive entries of \( p \). Therefore, if we denote the individual entries of the vector \( q \) by \( \{q_k\} \), then it holds that

\[ q_k > 0, \quad k = 1, 2, \ldots, N \quad (8.51) \]
Weighted Aggregate Cost

We also represent the step-sizes as scaled multiples of the same factor $\mu_{\text{max}}$, namely,

$$
\mu_k \triangleq \tau_k \mu_{\text{max}}, \quad k = 1, 2, \ldots, N
$$

(8.52)

where $0 < \tau_k \leq 1$. In this way, it becomes clear that all step-sizes become smaller as $\mu_{\text{max}}$ is reduced in size.

We further introduce the weighted aggregate cost

$$
J_{\text{glob,*}}(w) \triangleq \sum_{k=1}^{N} q_k J_k(w)
$$

(8.53)
Weighted Aggregate Cost

\[ J_{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w) \]

\[ \nabla_w J_{\text{glob}}(w^o) = 0 \iff \sum_{k=1}^{N} \nabla_w J_k(w^o) = 0 \]

\[ J_{\text{glob},*}(w) \triangleq \sum_{k=1}^{N} q_k J_k(w) \]

\[ \nabla_w J_{\text{glob},*}(w^*) = 0 \iff \sum_{k=1}^{N} q_k \nabla_w J_k(w^*) = 0 \]

(is also strongly convex)
Error Vectors

From this point onwards, we shall therefore measure the performance of the distributed strategy (8.46) by using $w^*$ as the reference vector (instead of $w^0$) and define the error vectors as:

$$\tilde{w}_{k,i} \triangleq w^* - w_{k,i} \quad (8.106)$$

$$\tilde{\psi}_{k,i} \triangleq w^* - \psi_{k,i} \quad (8.107)$$

$$\tilde{\phi}_{k,i-1} \triangleq w^* - \phi_{k,i-1} \quad (8.108)$$
Moreover, with each agent $k$, we associate a gradient noise vector in addition to a mismatch (or bias) vector, namely,

$$s_{k,i}(\phi_{k,i-1}) \triangleq \widehat{\nabla_w J_k(\phi_{k,i-1})} - \nabla_w J_k(\phi_{k,i-1}) \quad (8.109)$$

and

$$b_k \triangleq -\nabla_w J_k(w^*) \quad (8.110)$$
Example #8.11 (Gradient noise over MSE networks). Let us continue with the setting of Example 8.8, which deals with a variation of MSE networks where the data model at each agent is instead assumed to be given by

$$d_k(i) = u_{k,i} w^o_k + v_k(i)$$  \hspace{1cm} (8.126)

with the model vectors, $w^o_k$, being possibly different at the various agents. In a manner similar to (8.15), we can verify that if the distributed strategy (8.5) is employed at the agents, then the resulting gradient noise process at each agent $k$ is now given by:

$$s_{k,i}(\phi_{k,i-1}) = \frac{2}{h} \left( R_{u,k} - u^*_{k,i} u_{k,i} \right) \left( w^o_k - \phi_{k,i-1} \right) - \frac{2}{h} u^*_{k,i} v_k(i)$$  \hspace{1cm} (8.127)
Extended Error Dynamics
Error Dynamics

We explained earlier after (8.45) that because the Hessian matrices, $\nabla^2_w J_k(w)$, are not generally block diagonal, we will need to introduce extended versions of the error quantities $\{\tilde{w}_{k,i}, \tilde{\psi}_{k,i}, \tilde{\phi}_{k,i-1}\}$ in order to fully capture the dynamics of the network in the general case. This is in contrast to the mean-square-error case studied in Example 8.1 where these errors were sufficient to arrive at the state recursions (8.22) or (8.25) for the evolution of the network dynamics.
To motivate the need for extended error vectors, let us first introduce some notation. Thus, note that if we express any column vector $w \in \mathbb{C}^M$ in terms of its real and imaginary parts $x, y \in \mathbb{R}^M$, then

\begin{align*}
  w &= x + jy \quad \text{(a column vector)} \quad (8.128) \\
  w^* &= x^T - jy^T \quad \text{(a row vector)} \quad (8.129) \\
  (w^*)^T &= x - jy \quad \text{(a column vector)} \quad (8.130)
\end{align*}

In other words, the quantity $(w^*)^T$ is again a column vector, just like $w$, except that its complex representation is obtained by replacing $j$ by $-j$. 
Error Dynamics

The reason why we need to introduce the quantity \((w^*)^T\) is because, as the discussion will reveal, we will need to track the evolution of both quantities \(w_{k,i}\) and \((w^*_{k,i})^T\) in the general case in order to examine how the network is performing. Thus, using equations (8.46), we can deduce similar relations for the evolution of the complex conjugate iterates, namely,
Error Dynamics

\[
\begin{align*}
\left( \phi_{k,i-1}^* \right)^T &= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \left( w_{\ell,i-1}^* \right)^T \\
\left( \psi_{k,i}^* \right)^T &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \left( \phi_{\ell,i-1}^* \right)^T - \mu_k \nabla_w J_k \left( \phi_{k,i-1} \right) \\
\left( w_{k,i}^* \right)^T &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \left( \psi_{\ell,i}^* \right)^T
\end{align*}
\]

(8.131)
Error Dynamics

Observe how the gradient vector approximation that appears in the second equation now involves differentiation relative to $w^T$ and not $w^*$. Representations (8.46) and (8.131) can be grouped together into a single set of equations by introducing extended vectors of dimensions $2M \times 1$ as follows:
Error Dynamics

\[
\begin{align*}
\begin{bmatrix}
\phi_{k,i-1} \\
(\phi_{k,i-1}^*)^T \\
\psi_{k,i} \\
(\psi_{k,i}^*)^T \\
\omega_{k,i} \\
(\omega_{k,i}^*)^T
\end{bmatrix}
&= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \\
&= \sum_{\ell \in \mathcal{N}_k} a_{o,\ell k} \\
&= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k}
\begin{bmatrix}
\omega_{\ell,i-1} \\
(\omega_{\ell,i-1}^*)^T \\
\phi_{\ell,i-1} \\
(\phi_{\ell,i-1}^*)^T \\
\psi_{\ell,i} \\
(\psi_{\ell,i}^*)^T
\end{bmatrix}
- \mu_k \\
&= \nabla_w \mathcal{J}_k(\phi_{k,i-1}) \\
&= \nabla_w^T \mathcal{J}_k(\phi_{k,i-1})
\end{align*}
\]
Error Dynamics

We therefore extend the error vectors into size $2M \times 1$ and introduce

$$\tilde{w}_{k,i}^e \triangleq \begin{bmatrix} \tilde{w}_{k,i} \\ (\tilde{w}_{k,i}^*)^T \end{bmatrix}, \quad \tilde{\psi}_{k,i}^e \triangleq \begin{bmatrix} \tilde{\psi}_{k,i} \\ (\tilde{\psi}_{k,i}^*)^T \end{bmatrix}, \quad \tilde{\phi}_{k,i-1}^e \triangleq \begin{bmatrix} \tilde{\phi}_{k,i-1} \\ (\tilde{\phi}_{k,i-1}^*)^T \end{bmatrix}$$

(8.133)

where we are using the superscript “e” to refer to extended quantities of size $2M \times 1$. We also introduce extended versions of the limit vector, the gradient noise vector, and the bias vector:
Error Dynamics

\[(w^*)^e \triangleq \begin{bmatrix} w^* \\ ((w^*)^*)^T \end{bmatrix}, \quad s_{k,i}^e \triangleq \begin{bmatrix} s_{k,i}(\phi_{k,i-1}) \\ (s_{k,i}(\phi_{k,i-1})^* \end{bmatrix}^T, \quad b_k^e \triangleq \begin{bmatrix} b_k \\ (b^*)^T \end{bmatrix}\]

(8.134)

where the vector \( s_{k,i}^e \) in (8.134) should have been written more explicitly as \( s_{k,i}(\phi_{k,i-1}) \); we are dropping the argument for compactness of notation. Now, subtracting \((w^*)^e\) from both sides of the equations in (8.132) and using (8.109) gives
Error Dynamics

\[
\begin{align*}
\tilde{\phi}^e_{k,i-1} &= \sum_{\ell \in N_k} a_{1,\ell k} \tilde{w}^e_{\ell,i-1} \\
\tilde{\psi}^e_{k,i} &= \sum_{\ell \in N_k} a_{o,\ell k} \tilde{\phi}^e_{\ell,i-1} + \mu_k \\
\tilde{w}^e_{k,i} &= \sum_{\ell \in N_k} a_{2,\ell k} \tilde{\psi}^e_{\ell,i}
\end{align*}
\]

\[
\begin{bmatrix}
\nabla_{w^*} J_k(\phi_{k,i-1}) \\
\nabla_{w^T} J_k(\phi_{k,i-1})
\end{bmatrix} + \mu_k s^e_{k,i}
\]

(8.135)
Error Dynamics

We observe that the gradient vectors in (8.135) are being evaluated at the intermediate variable, $\phi_{k,i-1}$, and not at any of the error variables. For this reason, equation (8.135) is still not an actual recursion. To transform it into a recursion that only involves error variables, we call upon the mean-value theorem (D.20) from the appendix, which allows us to write:

\[
\begin{bmatrix}
\nabla_w^* J_k(\phi_{k,i-1}) \\
\nabla_w J_k(\phi_{k,i-1})
\end{bmatrix}
= 
\begin{bmatrix}
\nabla_w^* J_k(w^*) \\
\nabla_w J_k(w^*)
\end{bmatrix}
- \left[ \int_0^1 \nabla_w^2 J_k(w^* - t\tilde{\phi}_{k,i-1}) dt \right] \tilde{\phi}_{k,i-1}
\]

\[
\Delta = -b_k^e \\
\Delta = H_{k,i-1}
\]
Error Dynamics

That is,

\[
\begin{bmatrix}
\nabla_{w^+} J_k(\phi_{k,i-1}) \\
\nabla_{w^\top} J_k(\phi_{k,i-1})
\end{bmatrix}
= -b_k^e - H_{k,i-1} \tilde{\phi}_{k,i-1}^e
\tag{8.137}
\]

in terms of a \(2M \times 2M\) stochastic matrix \(H_{k,i-1}\) defined in terms of the integral of the \(2M \times 2M\) Hessian matrix of agent \(k\):

\[
H_{k,i-1} \triangleq \int_0^1 \nabla^2_{w} J_k(w^* - t\tilde{\phi}_{k,i-1}) dt
\tag{8.138}
\]

Substituting (8.137) into (8.135) leads to
Error Dynamics

\[
\begin{aligned}
\tilde{\phi}^e_{k,i-1} &= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \tilde{w}^e_{\ell,i-1} \\
\tilde{\psi}^e_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{0,\ell k} \tilde{\phi}^e_{\ell,i-1} - \mu_k H_{k,i-1} \tilde{\phi}^e_{k,i-1} - \mu_k b^e_k + \mu_k s^e_{k,i} \\
\tilde{w}^e_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \tilde{\psi}^e_{\ell,i}
\end{aligned}
\]

(8.139)
These equations describe the evolution of the error quantities at the individual agents for \( k = 1, 2, \ldots, N \). Observe that when the matrix \( H_{k,i-1} \) happens to be block diagonal, which occurs when the Hessian matrix function itself is block diagonal (as happened in (8.4) with the quadratic costs in Example 8.1), then the last term in (8.137) decouples into two separate terms in the variables

\[
\left\{ \tilde{\phi}_{k,i-1}, \left( \tilde{\phi}_{k,i-1}^* \right)^T \right\}
\]  
(8.140)
since then
\[ H_{k,i-1} \tilde{e} = \begin{bmatrix} H_{k,i-1}^{11} & 0 \\ 0 & H_{k,i-1}^{22} \end{bmatrix} \begin{bmatrix} \tilde{\phi}_{k,i-1} \\ (\tilde{\phi}_{k,i-1}^*)^T \end{bmatrix} \] (8.141)

In that case, it becomes unnecessary to propagate the extended vectors \( \{\tilde{\omega}_{k,i}, \tilde{\psi}_{k,i}, \tilde{\phi}_{k,i-1}\} \) using (8.139); the dynamics of the network can be studied by examining solely the evolution of the original error vectors \( \{\tilde{\omega}_{k,i}, \tilde{\psi}_{k,i}, \tilde{\phi}_{k,i-1}\} \), namely,
Error Dynamics

\[
\begin{align*}
\tilde{\phi}_{k,i-1} &= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \tilde{w}_{\ell,i-1} \\
\tilde{\psi}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{o,\ell k} \tilde{\phi}_{\ell,i-1} - \mu_k H_{k,i-1}^{11} \tilde{\phi}_{k,i-1} - \mu_k b_k + \mu_k s_{k,i} \\
\tilde{w}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \tilde{\psi}_{\ell,i}
\end{align*}
\]

(8.142)
Error Dynamics

We continue our discussion by treating the general case \((8.139)\). We collect the extended error vectors from all agents into the following \(N \times 1\) block error vectors (whose individual entries are of size \(2M \times 1\) each):

\[
\tilde{w}_i^e \triangleq \begin{bmatrix}
\tilde{w}_{1,i}^e \\
\tilde{w}_{2,i}^e \\
\vdots \\
\tilde{w}_{N,i}^e
\end{bmatrix}, \quad \tilde{\phi}_{i-1}^e \triangleq \begin{bmatrix}
\tilde{\phi}_{1,i-1}^e \\
\tilde{\phi}_{2,i-1}^e \\
\vdots \\
\tilde{\phi}_{N,i-1}^e
\end{bmatrix}, \quad \tilde{\psi}_i^e \triangleq \begin{bmatrix}
\tilde{\psi}_{1,i}^e \\
\tilde{\psi}_{2,i}^e \\
\vdots \\
\tilde{\psi}_{N,i}^e
\end{bmatrix}
\]

\((8.143)\)
Error Dynamics

We also define the following block gradient noise and bias vectors:

\[
\begin{align*}
\mathbf{s}^e_i & \triangleq \begin{bmatrix}
\mathbf{s}^e_{1,i} \\
\mathbf{s}^e_{2,i} \\
\vdots \\
\mathbf{s}^e_{N,i}
\end{bmatrix}, &
\mathbf{b}^e & \triangleq \begin{bmatrix}
b^e_1 \\
b^e_2 \\
\vdots \\
b^e_N
\end{bmatrix}
\end{align*}
\] (8.144)
Error Dynamics

Now recall from the explanation after (8.134) that each entry, $s_{k,i}^c$, in (8.144) is dependent on $\phi_{k,i-1}$. Recall also from the distributed algorithm (8.46) that $\phi_{k,i-1}$ is a combination of various $\{w_{\ell,i-1}\}$. Therefore, the block gradient vector, $s_i^c$, defined in (8.144) is dependent on the network vector, $w_{i-1}^c$, namely,
Error Dynamics

\[ w_{i-1}^e \overset{\triangle}{=} \begin{bmatrix} w_{1,i-1}^e \\ w_{2,i-1}^e \\ \vdots \\ w_{N,i-1}^e \end{bmatrix}, \quad w_{k,i-1}^e \overset{\triangle}{=} \begin{bmatrix} w_{k,i-1} \\ (w_{k,i-1}^*)^T \end{bmatrix} \]  

(8.145)

For this reason, we shall also write \( s_i^e(w_{i-1}^e) \) rather than simply \( s_i^e \) when it is desired to highlight the dependency of \( s_i^e \) on \( w_{i-1}^e \).
Error Dynamics

We further introduce the Kronecker products

\[
\begin{align*}
A_o & \triangleq A_o \otimes I_{2M} \\
A_1 & \triangleq A_1 \otimes I_{2M} \\
A_2 & \triangleq A_2 \otimes I_{2M}
\end{align*}
\]  

(8.146)

The matrix $A_o$ is an $N \times N$ block matrix whose $(\ell, k)$--th block is equal to $a_{o,\ell k}I_{2M}$. Similarly, for $A_1$ and $A_2$. Likewise, we introduce the following $N \times N$ block diagonal matrices, whose individual entries are of size $2M \times 2M$ each:
Error Dynamics

\[ \mathcal{M} \triangleq \text{diag}\{ \mu_1 I_{2M}, \mu_2 I_{2M}, \ldots, \mu_N I_{2M} \} \quad (8.147) \]

\[ \mathcal{H}_{i-1} \triangleq \text{diag}\{ H_{1,i-1}, H_{2,i-1}, \ldots, H_{N,i-1} \} \quad (8.148) \]

We then conclude from (8.139) that the following relations hold for the network variables:

\[
\begin{align*}
\tilde{\phi}^e_{i-1} &= A_1^T \tilde{w}^e_{i-1} \\
\tilde{\psi}^e_i &= [A_o^T - \mathcal{M}\mathcal{H}_{i-1}] \tilde{\phi}^e_{i-1} + \mathcal{M}s_i^e(w^e_{i-1}) - \mathcal{M}b^e \\
\tilde{w}^e_i &= A_2^T \tilde{\psi}^e_i
\end{align*}
\]  

(8.149)
Error Dynamics

so that the network weight error vector, $\tilde{w}_i^e$, ends up evolving according to the following stochastic recursion over $i \geq 0$:

$$\tilde{w}_i^e = A_2^T \left( A_o^T - M H_{i-1} \right) A_1^T \tilde{w}_{i-1}^e + A_2^T Ms_i^e(w_{i-1}^e) - A_2^T Mb^e$$

(8.150)

For comparison purposes, if each agent operates individually and uses the non-cooperative strategy, then the weight error vectors across all $N$ agents would instead evolve according to the following stochastic recursion:
Error Dynamics

\[ \tilde{w}_i^c = (I_{2MN} - \mathcal{H}_{i-1}) \tilde{w}_{i-1}^c + \mathcal{M}s_i^e(w_{i-1}^c) - \mathcal{M}b^e \quad (8.151) \]

where the matrices \( \{A_0, A_1, A_2\} \) do not appear since, in this case, \( A_0 = A_1 = A_2 = I_N \). We summarize the discussion so far in the following statement for complex data (we show how these results simplify for real data in the example after the lemma).
Error Dynamics

**Lemma 8.1** (Network error dynamics). Consider a network of $N$ interacting agents running the distributed strategy (8.46). The evolution of the error dynamics across the network relative to the reference vector $w^*$ defined by (8.55) is described by the following recursion:

$$
\tilde{w}_i^e = B_{i-1} \tilde{w}_{i-1}^e + A_{2}^T M s_i^e(w_{i-1}^e) - A_{2}^T M b^e, \quad i \geq 0
$$

(8.152)

where
\[ \mathcal{B}_{i-1} \triangleq \mathcal{A}_2^T (\mathcal{A}_o^T - \mathcal{M} \mathcal{H}_{i-1}) \mathcal{A}_1^T \]  
(8.153)

\[ \mathcal{A}_o \triangleq \mathcal{A}_o \otimes I_{2M}, \quad \mathcal{A}_1 \triangleq \mathcal{A}_1 \otimes I_{2M}, \quad \mathcal{A}_2 \triangleq \mathcal{A}_2 \otimes I_{2M} \]  
(8.154)

\[ \mathcal{M} \triangleq \text{diag}\{ \mu_1 I_{2M}, \mu_2 I_{2M}, \ldots, \mu_N I_{2M} \} \]  
(8.155)

\[ \mathcal{H}_{i-1} \triangleq \text{diag}\{ \mathbf{H}_{1,i-1}, \mathbf{H}_{2,i-1}, \ldots, \mathbf{H}_{N,i-1} \} \]  
(8.156)

\[ \mathbf{H}_{k,i-1} \triangleq \int_0^1 \nabla_w^2 J_k (w^* - t\bar{\phi}_{k,i-1}) dt \]  
(8.157)

where \( \nabla_w^2 J_k (w) \) denotes the \( 2M \times 2M \) Hessian matrix of \( J_k (w) \) relative to \( w \). Moreover, the extended vectors \( \{ \tilde{w}_i^e, s_i^e, b^e \} \) are defined by (8.143) and (8.144).
Example 8.12 (Mean-square-error costs). Let us re-consider the scenario studied in Example 8.1 and verify that result (8.152) collapses to (8.25). Indeed, in this case, we have $w^* = w^0$ and the bias vector, $b_k^c$, will be zero for all agents $k = 1, 2, \ldots, N$. Moreover since the Hessian matrix is now block diagonal, we can easily verify from the definition (8.137) that

$$H_{k,i-1} = \begin{bmatrix} R_{u,k} & 0 \\ 0 & R_{u,k}^T \end{bmatrix}$$

(8.158)

Substituting these facts into the expressions in Lemma 8.1 we recover (8.25).
Example 8.13 (Simplifications in the real case). The network error model of Lemma 8.1 can be simplified in the case of real data. This is because when $w \in \mathbb{R}^M$ is real-valued, we do not need to introduce the extended vectors (8.133) and (8.134) any longer. The simplifications that occur are described below.

To begin with, the distributed strategy (8.46) will be given by

$$\begin{align*}
\phi_{k,i-1} &= \sum_{\ell \in N_k} a_{1,\ell k} w_{\ell,i-1} \\
\psi_{k,i} &= \sum_{\ell \in N_k} a_{o,\ell k} \phi_{\ell,i-1} - \mu_k \nabla w^T J_k (\phi_{k,i-1}) \\
w_{k,i} &= \sum_{\ell \in N_k} a_{2,\ell k} \psi_{\ell,i}
\end{align*}$$

(8.159)
Example #8.13

where the gradient vector approximation in the second equation is now relative to \( w^T \) and not \( w^* \). Subtracting the limit vector \( w^* \) directly from both sides of the above equations gives

\[
\begin{align*}
\tilde{\phi}_{k,i-1} &= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \tilde{w}_{\ell,i-1} \\
\tilde{\psi}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{o,\ell k} \tilde{\phi}_{\ell,i-1} + \mu_k \nabla_{w^T} J_k(\phi_{k,i-1}) + \mu_k s_{k,i} \\
\tilde{w}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \tilde{\psi}_{\ell,i}
\end{align*}
\] (8.160)
Example #8.13

where now

\[ s_{k,i} \overset{\Delta}{=} \nabla_{w^T} J_k(\phi_{k,i-1}) - \nabla_{w^T} J_k(\phi_{k,i-1}) \quad (8.161) \]

and the error vectors are measured relative to the same limit vector \( w^* \):

\[ \tilde{w}_{k,i} = w^* - w_{k,i}, \quad \tilde{\psi}_{k,i} = w^* - \psi_{k,i}, \quad \tilde{\phi}_{k,i-1} = w^* - \phi_{k,i-1} \quad (8.162) \]

We then call upon the real-version of the mean-value theorem, namely, expression (D.9) in the appendix, to write
Example #8.13

\[
\nabla_{w^T} J_k(\phi_{k,i-1}) = \nabla_{w^T} J_k(w^*) - \left[ \int_0^1 \nabla_{w^2} J_k(w^* - t\phi_{k,i-1}) dt \right] \phi_{k,i-1} \\
\Delta = -b_k \\
\Delta = H_{k,i-1}
\]

\[
= -b_k - H_{k,i-1} \tilde{\phi}_{k,i-1}
\]

(8.163)

where we introduced the \( M \times 1 \) constant vector \( b_k \) and the (now) \( M \times M \) stochastic matrix \( H_{k,i} \). Substituting (8.163) into (8.160) leads to
Example #8.13

\[
\begin{align*}
\tilde{\phi}_{k,i-1} &= \sum_{\ell \in \mathcal{N}_k} a_{1,\ell k} \tilde{w}_{\ell,i-1} \\
\tilde{\psi}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{o,\ell k} \tilde{\phi}_{\ell,i-1} - \mu_k H_{k,i-1} \tilde{\phi}_{k,i-1} - \mu_k b_k + \mu_k s_{k,i} \\
\tilde{w}_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{2,\ell k} \tilde{\psi}_{\ell,i}
\end{align*}
\]

so that the network error vector

\[
\tilde{w}_i = \text{col}\{\tilde{w}_{1,i}, \tilde{w}_{2,i}, \ldots, \tilde{w}_{N,i}\}
\]

evolves according to the recursion

\[
\tilde{w}_i = \mathcal{B}_{i-1} \tilde{w}_{i-1} + \mathcal{A}_2^T \mathcal{M} s_i(\tilde{w}_{i-1}) - \mathcal{A}_2^T \mathcal{M} b, \quad i \geq 0
\]

where now
Example #8.13

\[ \mathcal{B}_{i-1} \triangleq A_2^T \left( A_o^T - \mathcal{M} \mathcal{H}_{i-1} \right) A_1^T \quad (8.167) \]

\[ \mathcal{A}_o \triangleq A_o \otimes I_M, \quad \mathcal{A}_1 \triangleq A_1 \otimes I_M, \quad \mathcal{A}_2 \triangleq A_2 \otimes I_M \quad (8.168) \]

\[ \mathcal{M} \triangleq \text{diag}\{ \mu_1 I_M, \mu_2 I_M, \ldots, \mu_N I_M \} \quad (8.169) \]

\[ \mathcal{H}_{i-1} \triangleq \text{diag}\{ H_{1,i-1}, H_{2,i-1}, \ldots, H_{N,i-1} \} \quad (8.170) \]

\[ H_{k,i-1} \triangleq \int_0^1 \nabla_w^2 J_k(w^* - t\tilde{\phi}_{k,i-1}) dt \quad (8.171) \]

\[ w_{i-1} \triangleq \text{col}\{ w_{1,i-1}, w_{2,i-1}, \ldots, w_{N,i-1} \} \quad (8.172) \]

and \( \nabla_w^2 J_k(w) \) denotes the \( M \times M \) Hessian matrix of \( J_k(w) \) relative to \( w \).
Gradient Noise Model
Moreover, with each agent $k$, we associate a gradient noise vector in addition to a mismatch (or bias) vector, namely,

$$s_{k,i}(\phi_{k,i-1}) \triangleq \nabla_{w^*} J_k(\phi_{k,i-1}) - \nabla_{w^*} J_k(\phi_{k,i-1}) \quad (8.109)$$

and

$$b_k \triangleq -\nabla_{w^*} J_k(w^*) \quad (8.110)$$
Gradient Noise Model

In the special case when all individual costs, $J_k(w)$, have the same minimizer at $w_k^o \equiv w^o$ (which is the situation considered in Example 8.1 over MSE networks), then $w^* = w^o$ and the vector $b_k$ will be identically zero. In general, though, the vector $b_k$ is nonzero. Let $\mathcal{F}_{i-1}$ represent the collection of all random events generated by the processes $\{w_{k,j}\}$ at all agents $k = 1, 2, \ldots, N$ up to time $i - 1$:

$$\mathcal{F}_{i-1} \triangleq \text{filtration}\{w_{k,-1}, w_{k,0}, w_{k,1}, \ldots, w_{k,i-1}, \text{ all } k\} \quad (8.111)$$
Assumption 8.1 (Conditions on gradient noise). It is assumed that the first and second-order conditional moments of the individual gradient noise processes, \( s_{k,i}(\phi) \), satisfy the following conditions for any iterates \( \phi \in \mathcal{F}_{i-1} \) and for all \( k, \ell = 1, 2, \ldots, N \):

\[
\mathbb{E} \left[ s_{k,i}(\phi) \bigg| \mathcal{F}_{i-1} \right] = 0 \tag{8.112}
\]

\[
\mathbb{E} \left[ s_{k,i}(\phi)s_{\ell,i}(\phi) \bigg| \mathcal{F}_{i-1} \right] = 0, \quad k \neq \ell \tag{8.113}
\]

\[
\mathbb{E} \left[ s_{k,i}(\phi)s_{\ell,i}(\phi)^\top \bigg| \mathcal{F}_{i-1} \right] = 0, \quad k \neq \ell \tag{8.114}
\]

\[
\mathbb{E} \left[ \|s_{k,i}(\phi)\|^2 \bigg| \mathcal{F}_{i-1} \right] \leq \left( \frac{\bar{\beta}_k}{h} \right)^2 \|\phi\|^2 + \bar{\sigma}_{s,k}^2 \tag{8.115}
\]

almost surely, for some nonnegative scalars \( \bar{\beta}_k^2 \) and \( \bar{\sigma}_{s,k}^2 \) and where \( h = 1 \) for real data and \( h = 2 \) for complex data.
Using the above conditions, and in a manner similar to the derivation (3.28), it is straightforward to verify that the gradient noise processes satisfy:

\begin{align}
\mathbb{E} \left[ s_{k,i}(\phi_{k,i-1}) | \mathcal{F}_{i-1} \right] &= 0 \\
\mathbb{E} \left[ \|s_{k,i}(\phi_{k,i-1})\|^2 | \mathcal{F}_{i-1} \right] &\leq (\beta_k^2/h^2)\|\tilde{\phi}_{k,i-1}\|^2 + \sigma_{s,k}^2 \\
\mathbb{E} \|s_{k,i}(\phi_{k,i-1})\|^2 &\leq (\beta_k^2/h^2)\mathbb{E} \|\tilde{\phi}_{k,i-1}\|^2 + \sigma_{s,k}^2
\end{align}

in terms of the scalars

\begin{align}
\beta_k^2 &\triangleq 2\overline{\beta}_k^2 \\
\sigma_{s,k}^2 &\triangleq 2(\overline{\beta}_k/h)^2\|w^*\|^2 + \overline{\sigma}_{s,k}^2
\end{align}
Gradient Noise Model

We shall use conditions (8.116)–(8.118) more frequently in lieu of (8.112)–(8.115). We could have required these conditions directly in the statement of Assumption 8.1. We instead opted to state conditions (8.112)–(8.115) in that manner, in terms of a generic $\phi \in \mathcal{F}_{i-1}$ rather than $\tilde{w}_{k,i-1}$, so that the upper bound in (8.115) is independent of the unknown $w^*$. 
Gradient Noise Model

Conditions (8.116)–(8.118) will be useful in establishing the mean-square stability of the second-order moment of the error vector, $\mathbb{E} \|\tilde{w}_{k,i}\|^2$, in the next chapter. Later, in Sec. 9.2, when we examine the stability of the fourth-order moment of the same error vector, $\mathbb{E} \|\tilde{w}_{k,i}\|^4$, we will need to replace the bound (8.115) by a condition similar to (5.36) on the fourth-order moments of the individual gradient noise processes, namely, by the following condition:

$$\mathbb{E} \left[ \|s_{k,i}(\phi)\|^4 \mid \mathcal{F}_{i-1} \right] \leq (\bar{\beta}_k/h)^4 \|\phi\|^4 + \bar{\sigma}^4_{s,k}$$

(8.121)
Gradient Noise Model

almost surely, for nonnegative scalars $\{\beta_k^4, \sigma_{s,k}^4\}$. Using an argument similar to (3.56), we can similarly conclude from these conditions that

$$
\mathbb{E} \left[ \| s_{k,i}(\phi_{k,i-1}) \|^4 \mid \mathcal{F}_{i-1} \right] \leq (\beta_{4,k}^4/h^4) \| \phi_{k,i-1} \|^4 + \sigma_{s4,k}^4 \quad (8.122)
$$

for some non-negative parameters defined by:

$$
\beta_{4,k}^4 \overset{\Delta}{=} 8\beta_k^4 \quad (8.123)
$$

$$
\sigma_{s4,k}^4 \overset{\Delta}{=} 8(\beta_k/h)^4 \| w^* \|^4 + \sigma_{s4,k}^4 \quad (8.124)
$$
Gradient Noise Model

We will not need to introduce condition (8.121) in addition to the second-order moment condition (8.115). This is because, as explained earlier following (3.50), condition (8.121) implies that condition (8.115) also holds, namely, it follows from (8.121) that

\[
\mathbb{E} \left[ \| s_{k,i}(\phi) \|^2 \mid \mathcal{F}_{i-1} \right] \leq \left( \frac{\beta_k}{h} \right)^2 \| \phi \|^2 + \sigma_{s,k}^2
\]  
(8.125)
Stability Results
Key Stability Results

We are now ready to examine the stability of the mean-error process, $\mathbb{E}\tilde{\mathbf{w}}_i$, the mean-square-error, $\mathbb{E}\|\tilde{\mathbf{w}}_i\|^2$, and the fourth-order moment, $\mathbb{E}\|\tilde{\mathbf{w}}_i\|^4$, by using the network error recursion (8.152). The key results proven in the current chapter are that for sufficiently small step-sizes, and for each agent $k$, it will hold that

$$\limsup_{i \to \infty} \|\mathbb{E}\tilde{\mathbf{w}}_{k,i}\| = O(\mu_{\text{max}})$$ (9.1)

$$\limsup_{i \to \infty} \mathbb{E}\|\tilde{\mathbf{w}}_{k,i}\|^2 = O(\mu_{\text{max}})$$ (9.2)

$$\limsup_{i \to \infty} \mathbb{E}\|\tilde{\mathbf{w}}_{k,i}\|^4 = O(\mu_{\text{max}}^2)$$ (9.3)
Network Stability

Theorems 9.1, 9.2, 9.6: For sufficiently small step-sizes:

\[ \limsup_{i \to \infty} \mathbb{E} \| \bar{w}_{k,i} \| \leq O(\mu_{\text{max}}) \]

\[ \limsup_{i \to \infty} \mathbb{E} \| \bar{w}_{k,i} \|^2 \leq O(\mu_{\text{max}}) \]

\[ \limsup_{i \to \infty} \mathbb{E} \| \bar{w}_{k,i} \|^4 \leq O(\mu_{\text{max}}^2) \]
where $\mu_{\text{max}}$ is an upper bound on the largest step-size parameter across the network since, from (8.52), we parameterized all step-sizes as scaled multiples of $\mu_{\text{max}}$, namely,

$$
\mu_k \triangleq \tau_k \mu_{\text{max}}, \quad k = 1, 2, \ldots, N
$$

(9.4)

where $0 < \tau_k \leq 1$. The error vectors, $\{\tilde{w}_{k,i}\}$, in the above expressions are measured relative to the limit vector, $w^*$:

$$
\tilde{w}_{k,i} = w^* - w_{k,i}
$$

(9.5)
where \( w^* \) was defined by (8.55) as the unique minimum of the weighted aggregate cost function, \( J_{\text{glob},*}(w) \), from (8.53), namely,

\[
J_{\text{glob},*}(w) \triangleq \sum_{k=1}^{N} q_k J_k(w) \tag{9.6}
\]

and the \( \{q_k\} \) are positive scalars corresponding to the entries of the vector:

\[
q \triangleq \text{diag}\{\mu_1, \mu_2, \ldots, \mu_N\} A_{2p} \tag{9.7}
\]
Here, the vector $p$ refers to the Perron eigenvector of the matrix product

$$P \triangleq A_1 A_0 A_2$$  \hspace{1cm} (9.8)

and is defined through the relations:

$$Pp = p, \quad 1^T p = 1, \quad p_k > 0$$ \hspace{1cm} (9.9)

For ease of reference, we recall the definition of the original aggregate cost function (8.44), namely,

$$J_{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w)$$  \hspace{1cm} (9.10)
Second-Order Stability
Recall: Error Dynamics

\[ \tilde{w}_i^e = B_{i-1} \tilde{w}_{i-1}^e + A_2^T M s_i^e(w_{i-1}^e) - A_2^T M b^e, \quad i \geq 0 \]  

(8.152)

\[
\begin{align*}
B_{i-1} & \triangleq A_2^T \left( A_o^T - M H_{i-1} \right) A_1^T \\
A_o & \triangleq A_o \otimes I_{2M}, \quad A_1 \triangleq A_1 \otimes I_{2M}, \quad A_2 \triangleq A_2 \otimes I_{2M} \\
M & \triangleq \text{diag}\{ \mu_1 I_{2M}, \mu_2 I_{2M}, \ldots, \mu_N I_{2M} \} \\
H_{i-1} & \triangleq \text{diag}\{ H_{1,i-1}, H_{2,i-1}, \ldots, H_{N,i-1} \} \\
H_{k,i-1} & \triangleq \int_0^1 \nabla^2_w J_k(w^* - t\tilde{\phi}_{k,i-1}) dt
\end{align*}
\]  

Second-Order Error Moment

**Theorem 9.1** (Network mean-square-error stability). Consider a network of $N$ interacting agents running the distributed strategy (8.46) with a primitive matrix $P = A_1 A_o A_2$. Assume the aggregate cost (9.10) and the individual costs, $J_k(w)$, satisfy the conditions in Assumption 6.1. Assume further that the first and second-order moments of the gradient noise process satisfy the conditions in Assumption 8.1. Then, the network is mean-square stable for sufficiently small step-sizes, namely, it holds that

$$
\limsup_{i \to \infty} \mathbb{E} \left\| \tilde{w}_{k,i} \right\|^2 = O(\mu_{\text{max}}), \quad k = 1, 2, \ldots, N
$$

(9.11)

for any $\mu_{\text{max}} < \mu_o$, for some small enough $\mu_o$. 
**Proof.** The derivation is demanding. We follow arguments motivated by the analysis in [70, 277] and they involve, as an initial step, transforming the error recursion (9.12) shown below into a more convenient form shown later in (9.60). We establish the result for the general case of complex data and, therefore, $h = 2$ throughout this derivation.

We start from the network error recursion (8.152):

$$
\tilde{w}_i^e = \mathcal{B}_{i-1} \tilde{w}_{i-1}^e + A_2^T \mathcal{M}s_i^e(w_{i-1}^e) - A_2^T \mathcal{M}b^e, \ i \geq 0
$$

(9.12)

where
Proof

\[ \mathcal{B}_{i-1} = A_2^T (A_o^T - MH_{i-1}) A_1^T \]
\[ = A_2^T A_o^T A_1^T - A_2^T MH_{i-1} A_1^T \]
\[ \triangleq P^T - A_2^T MH_{i-1} A_1^T \]  \hspace{1cm} (9.13)

in terms of the matrix

\[ P^T \triangleq A_2^T A_o^T A_1^T \]
\[ = (A_2^T \otimes I_{2M})(A_o^T \otimes I_{2M})(A_1^T \otimes I_{2M}) \]
\[ = (A_2^T A_o^T A_1^T \otimes I_{2M}) \]
\[ = P^T \otimes I_{2M} \] \hspace{1cm} (9.14)
The matrix $P = A_1 A_0 A_2$ is left-stochastic and assumed primitive. It follows that it has a single eigenvalue at one while all other eigenvalues are strictly inside the unit circle. We let $p$ denote its Perron eigenvector, which is already defined by (9.9). This vector determines the entries of $q$ defined by (9.7). Note, for later reference, that the $k$—entry of $q$ can be extracted by computing the inner product of $q$ with the $k$—th basis vector, $e_k$, which has a unit entry at the $k$—th location and zeros elsewhere, i.e.,

$$q_k = \mu_k (e_k^T A_2 p)$$

(9.4)

$$= \mu_{\text{max}} \tau_k (e_k^T A_2 p)$$

(9.15)
Proof

Obviously, it holds for the extended matrices \( \{ \mathcal{P}, \mathcal{A}_2 \} \) that

\[
\mathcal{P}(p \otimes I_{2M}) = (p \otimes I_{2M}) \quad \text{(9.16)}
\]

\[
\mathcal{M} \mathcal{A}_2(p \otimes I_{2M}) = (q \otimes I_{2M}) \quad \text{(9.17)}
\]

\[
(1^T \otimes I_{2M})(p \otimes I_{2M}) = I_{2M} \quad \text{(9.18)}
\]

Moreover, since \( A_1 \) and \( A_2 \) are left-stochastic, it holds that

\[
\mathcal{A}_1^T(1 \otimes I_{2M}) = (1 \otimes I_{2M}) \quad \text{(9.19)}
\]

\[
\mathcal{A}_2^T(1 \otimes I_{2M}) = (1 \otimes I_{2M}) \quad \text{(9.20)}
\]
Proof

The derivation that follows exploits the eigen-structure of \( P \). We start by noting that the \( N \times N \) matrix \( P \) admits a Jordan canonical decomposition of the form [113, p.128]:

\[
P \triangleq V_\epsilon J V_\epsilon^{-1}
\]  \hspace{1cm} (9.21)

\[
J = \begin{bmatrix}
1 & 0 \\
0 & J_\epsilon
\end{bmatrix}
\]  \hspace{1cm} (9.22)

\[
V_\epsilon = \begin{bmatrix}
p \\
V_R
\end{bmatrix}
\]  \hspace{1cm} (9.23)

\[
V_\epsilon^{-1} = \begin{bmatrix}
1^T \\
V_L^T
\end{bmatrix}
\]  \hspace{1cm} (9.24)
Proof

where the matrix $J_\epsilon$ consists of Jordan blocks, with each one of them having the generic form (say, for a Jordan block of size $4 \times 4$):

$$
\begin{bmatrix}
\lambda & \epsilon & \lambda & \\
\epsilon & \lambda & \epsilon & \\
\epsilon & \lambda & \epsilon & \\
& & \epsilon & \\
\end{bmatrix}
$$

(9.25)

with $\epsilon > 0$ appearing on the lower diagonal, and where the eigenvalue $\lambda$ may be complex but has magnitude strictly less than one. The scalar $\epsilon$ is any small positive number that is independent of $\mu_{\max}$. Obviously, since $V_\epsilon^{-1}V_\epsilon = I_N$, it holds that
Proof

For any $N \times N$ matrix $A$, the traditional Jordan decomposition $A = T J' T^{-1}$ involves Jordan blocks in $J'$ that have ones on the lower diagonal instead of $\epsilon$. However, if we introduce the diagonal matrix $E = \text{diag}\{1, \epsilon, \epsilon^2, \ldots, \epsilon^{N-1}\}$, then $A = T E^{-1} E J' E^{-1} E T^{-1}$, which we rewrite as $A = V_\epsilon J V_\epsilon^{-1}$ with $V_\epsilon = T E^{-1}$ and $J = E J' E^{-1}$. The matrix $J$ now has $\epsilon$ values instead of ones on the lower diagonal.
Proof

\[ 1^T V_R = 0 \]  \hspace{2cm} (9.26)
\[ V_L^T P = 0 \]  \hspace{2cm} (9.27)
\[ V_L^T V_R = I_{N-1} \]  \hspace{2cm} (9.28)

The matrices \( \{V_\epsilon, J, V_\epsilon^{-1}\} \) have dimensions \( N \times N \) while the matrices \( \{V_L, J_\epsilon, V_R\} \) have dimensions \( (N - 1) \times (N - 1) \). The Jordan decomposition of the extended matrix \( \mathcal{P} = P \otimes I_{2M} \) is given by

\[ \mathcal{P} = (V_\epsilon \otimes I_{2M})(J \otimes I_{2M})(V_\epsilon^{-1} \otimes I_{2M}) \]  \hspace{2cm} (9.29)

so that substituting into (9.13) we obtain

\[ \mathcal{B}_{i-1} = ((V_\epsilon^{-1})^T \otimes I_{2M}) \{(J^T \otimes I_{2M}) - \mathcal{D}_{i-1}^T\} (V_\epsilon^T \otimes I_{2M}) \]  \hspace{2cm} (9.30)
Proof

where

\[ \mathcal{D}_{i-1}^T \triangleq (V_\varepsilon^T \otimes I_{2M}) \mathcal{A}_2^T \mathcal{M} \mathcal{H}_{i-1} \mathcal{A}_1^T ((V_{\varepsilon^{-1}})^T \otimes I_{2M}) \]

\[ \equiv \begin{bmatrix} D_{11,i-1}^T & D_{21,i-1}^T \\ D_{12,i-1}^T & D_{22,i-1}^T \end{bmatrix} \] (9.31)

Using the partitioning (9.23)–(9.24) and the fact that

\[ \mathcal{A}_1 = A_1 \otimes I_{2M}, \quad \mathcal{A}_2 = A_2 \otimes I_{2M} \] (9.32)

we find that the block entries \( \{D_{mn,i-1}\} \) in (9.31) are given by
Proof

\[
D_{11,i-1} = \sum_{k=1}^{N} q_k H_{k,i-1}^T
\]

(9.33)

\[
D_{12,i-1} = (1^T \otimes I_{2M}) \mathcal{H}_{i-1}^T \mathcal{M}(A_2 V_R \otimes I_{2M})
\]

(9.34)

\[
D_{21,i-1} = (V_L^T A_1 \otimes I_{2M}) \mathcal{H}_{i-1}^T (q \otimes I_{2M})
\]

(9.35)

\[
D_{22,i-1} = (V_L^T A_1 \otimes I_{2M}) \mathcal{H}_{i-1}^T \mathcal{M}(A_2 V_R \otimes I_{2M})
\]

(9.36)

Let us now show that the entries in each of these matrices is in the order of \(O(\mu_{\text{max}})\), as well as verify that the matrix norm sequences of these matrices are uniformly bounded from above for all \(i\). To begin with, recall from (8.157) that

\[
H_{k,i-1} \triangleq \int_0^1 \nabla_w^2 J_k(w^* - t\tilde{\phi}_{k,i-1}) dt
\]

(9.37)
and, moreover, by assumption, all individual costs $J_k(w)$ are convex functions with at least one of them, say, the cost function of index $k_o$, being $\nu_d$—strongly-convex. This fact implies that, for any $w$,

$$\nabla^2_w J_{k_o}(w) \geq \frac{\nu_d}{h} I_{hM} > 0, \quad \nabla^2_w J_k(w) \geq 0, \quad k \neq k_o$$  \hspace{1cm} (9.38)

Consequently,

$$H_{k_o,i-1} \geq \frac{\nu_d}{h} I_{hM} > 0, \quad H_{k,i-1} \geq 0, \quad k \neq k_o$$  \hspace{1cm} (9.39)

and, therefore, $D_{11,i-1} > 0$. More specifically, the matrix sequence $D_{11,i-1}$ is uniformly bounded from below as follows:
Proof

\[
D_{11,i-1} \geq q_{k_o} \frac{\nu_d}{h} I_{hM}
\]

\[
= \mu_{\text{max}} \tau_{k_o} (e_{k_o}^T A_{2p}) \frac{\nu_d}{h} I_{hM}
\]

\[
= O(\mu_{\text{max}})
\]

On the other hand, from the upper bound on the sum of the Hessian matrices in (6.13), and since each individual Hessian matrix is at least non-negative definite, we get

\[
H_{k,i-1} \leq \frac{\delta_d}{h} I_{hM}
\]

so that the matrix sequence \(D_{11,i-1}\) is uniformly bounded from above as well:
Proof

\[
    D_{11,i-1} \leq q_{\text{max}} N \frac{\delta_d}{h} I_{\text{hM}}
\]

\[
    (9.15) \quad \Rightarrow \quad \mu_{\text{max}} \tau_{k_{\text{max}}} (e_{k_{\text{max}}}^\top A_{2p}) N \frac{\delta_d}{h} I_{\text{hM}}
\]

\[
    = O(\mu_{\text{max}})
\]

(9.42)

where \( k_{\text{max}} \) denotes the \( k \)-index of the largest \( q_k \) entry. Combining results (9.40)–(9.42) we conclude that

\[
    D_{11,i-1} = O(\mu_{\text{max}})
\]

(9.43)

Actually, since \( D_{11,i-1} \) is Hermitian positive-definite, we also conclude that its eigenvalues (which are positive and real) are \( O(\mu_{\text{max}}) \). This is because from the relation
Proof

\[ \mu_{\max} \tau_{k_0} (e_{k_0}^T A_2 p)^{\nu_d h} I_{hM} \leq D_{11,i-1} \leq \mu_{\max} \tau_{k_{\max}} (e_{k_{\max}}^T A_{2p}) N^{\frac{\delta_d h}{h}} I_{hM} \]

we can write, more compactly,

\[ c_1 \mu_{\max} I_{hM} \leq D_{11,i-1} \leq c_2 \mu_{\max} I_{hM} \]

for some positive constants \( c_1 \) and \( c_2 \) that are independent of \( \mu_{\max} \) and \( i \). Accordingly, for the eigenvalues of \( D_{11,i-1} \), we can write

\[ c_1 \mu_{\max} \leq \lambda(D_{11,i-1}) \leq c_2 \mu_{\max} \]

It follows that the eigenvalues of \( I_{2M} - D_{11,i-1}^T \) are \( 1 - O(\mu_{\max}) \) so that, in terms of the 2–induced norm and for sufficiently small \( \mu_{\max} \):
Proof

\[ \| I_{2M} - D_{11,i-1}^T \| = \rho(I_{2M} - D_{11,i-1}^T) \leq 1 - \sigma_{11} \mu_{\text{max}} = 1 - O(\mu_{\text{max}}) \] (9.47)

for some positive constant \( \sigma_{11} \) that is independent of \( \mu_{\text{max}} \) and \( i \).

Similarly, from (9.39) and (9.41), and since each \( H_{k,i-1} \) is bounded from below and from above, we can conclude that

\[ D_{12,i-1} = O(\mu_{\text{max}}), \quad D_{21,i-1} = O(\mu_{\text{max}}), \quad D_{22,i-1} = O(\mu_{\text{max}}) \] (9.48)

and that the norms of these matrix sequences are also uniformly bounded from above. For example, using the 2–induced norm (i.e., maximum singular value):
Proof

\[
\| D_{21,i-1} \| \leq \| V_L^T A_1 \otimes I_{2M} \| \| q \otimes I_{2M} \| \| H_{i-1}^T \|
\]

\[
\leq \| V_L^T A_1 \otimes I_{2M} \| \| q \otimes I_{2M} \| \left( \max_{1 \leq k \leq N} \| H_{k,i-1} \| \right)
\]

(9.41)

\[
\leq \| V_L^T A_1 \otimes I_{2M} \| \| q \otimes I_{2M} \| \left( \frac{\delta_d}{h} \right)
\]

\[
= \| V_L^T A_1 \otimes I_{2M} \| \| q \| \left( \frac{\delta_d}{h} \right)
\]

\[
\leq \| V_L^T A_1 \otimes I_{2M} \| \sqrt{N} q_{\max}^2 \left( \frac{\delta_d}{h} \right)
\]

\[
= \| V_L^T A_1 \otimes I_{2M} \| \sqrt{N} \mu_{\max} \tau_{k_{\max}} (e_{k_{\max}}^T A_2 p) \left( \frac{\delta_d}{h} \right)
\]

(9.49)
Proof

so that

\[ \|D_{21,i-1}\| \leq \sigma_{21}\mu_{\text{max}} = O(\mu_{\text{max}}) \quad (9.50) \]

for some positive constant \(\sigma_{21}\). In the above derivation we used the fact that \(\|q \otimes I_{2M}\| = \|q\|\) since, from Table F.1 in the appendix, the singular values of a Kronecker product are given by all possible products of the singular values of the individual matrices. A similar argument applies to \(D_{12,i-1}\) and \(D_{22,i-1}\) for which we can verify that

\[ \|D_{12,i-1}\| \leq \sigma_{12}\mu_{\text{max}} = O(\mu_{\text{max}}), \quad \|D_{22,i-1}\| \leq \sigma_{22}\mu_{\text{max}} = O(\mu_{\text{max}}) \quad (9.51) \]

for some positive constants \(\sigma_{21}\) and \(\sigma_{22}\). Let

\[ \mathcal{V}_\epsilon \triangleq V_\epsilon \otimes I_{2M}, \quad \mathcal{J}_\epsilon \triangleq J_\epsilon \otimes I_{2M} \quad (9.52) \]
Proof: Recall

\[ \mathcal{B}_{i-1} = \left( (V_{\epsilon}^{-1})^T \otimes I_{2M} \right) \left\{ (J^T \otimes I_{2M}) - D_{i-1}^T \right\} \left( V_{\epsilon}^T \otimes I_{2M} \right) \]  

(9.30)

where

\[ D_{i-1}^T \triangleq \left( V_{\epsilon}^T \otimes I_{2M} \right) A_2^T M H_{i-1} A_1^T \left( (V_{\epsilon}^{-1})^T \otimes I_{2M} \right) \]

\[ \equiv \begin{bmatrix} D_{11,i-1}^T & D_{21,i-1}^T \\ D_{12,i-1}^T & D_{22,i-1}^T \end{bmatrix} \]  

(9.31)
Proof

Then, using (9.30), we can write

\[
\mathcal{B}_{i-1} = (\mathcal{V}_e^{-1})^T \begin{bmatrix}
I_{2M} & -D_{11,i-1}^T & -D_{21,i-1}^T \\
-D_{12,i-1}^T & \mathcal{J}_e^T & -D_{22,i-1}^T
\end{bmatrix} \mathcal{V}_e^T
\]  

(9.53)

To simplify the notation, we drop the argument \( \mathbf{w}_{i-1}^e \) in (9.12) and write \( \mathbf{s}_i^e \) instead of \( \mathbf{s}_i^e(\mathbf{w}_{i-1}^e) \) from this point onwards. We now multiply both sides of the error recursion (9.12) from the left by \( \mathcal{V}_e^T \):

\[
\mathcal{V}_e^T \tilde{\mathbf{w}}_i^e = \mathcal{V}_e^T \mathcal{B}_{i-1} (\mathcal{V}_e^{-1})^T \mathcal{V}_e^T \tilde{\mathbf{w}}_{i-1}^e + \mathcal{V}_e^T \mathcal{A}_2^T \mathcal{M} \mathbf{s}_i^e - \mathcal{V}_e^T \mathcal{A}_2^T \mathcal{M} \mathbf{b}^e, \quad i \geq 0
\]

(9.54)
and let

\[ \forall \epsilon \tilde{w}_i^e = \begin{bmatrix} (p^T \otimes I_{2M}) \tilde{w}_i^e \\ (V_R^T \otimes I_{2M}) \tilde{w}_i^e \end{bmatrix} \triangleq \begin{bmatrix} \tilde{w}_i^e \\ \tilde{w}_i^e \end{bmatrix} \quad (9.55) \]

\[ \forall \epsilon A_2^T M s_i^e = \begin{bmatrix} (p^T \otimes I_{2M}) A_2^T M s_i^e \\ (V_R^T \otimes I_{2M}) A_2^T M s_i^e \end{bmatrix} \triangleq \begin{bmatrix} \tilde{s}_i^e \\ \tilde{s}_i^e \end{bmatrix} \quad (9.56) \]

\[ \forall \epsilon A_2^T M b^e = \begin{bmatrix} (p^T \otimes I_{2M}) A_2^T M b^e \\ (V_R^T \otimes I_{2M}) A_2^T M b^e \end{bmatrix} \triangleq \begin{bmatrix} 0 \\ \tilde{b}^e \end{bmatrix} \quad (9.57) \]

where the zero entry in the last equality is due to the fact that
Proof

\[(p^T \otimes I_{2M}) A_2^T M b^e = (q^T \otimes I_{2M}) b^e\]

\[= \sum_{k=1}^{N} q_k b_k^e\]

\[= -\sum_{k=1}^{N} q_k \begin{bmatrix}
\nabla_{w^*} J_k(w^*) \\
\nabla_{w^*}^T J_k(w^*)
\end{bmatrix}\]

\[= -\sum_{k=1}^{N} q_k \begin{bmatrix}
\nabla_{w} J_k(w^*)^* \\
\nabla_{w}^T J_k(w^*)^T
\end{bmatrix}\]

\[(8.55) \quad \Rightarrow \quad 0 \quad (9.58)\]
Moreover, from the expression for $\dot{b}^e$ in (9.57), we note that it depends on $\mathcal{M}$ and $b^e$. Recall from (8.110) and (8.144) that the entries of $b^e$ are defined in terms of the gradient vectors $\nabla_{w^*} J_k(w^*)$.

It follows that $b^e$ has bounded norm and we conclude that

$$\dot{b}^e = O(\mu_{\text{max}}) \quad (9.59)$$

Using the just introduced transformed variables, we can rewrite (9.54) in the form
Proof

\[
\begin{bmatrix}
\tilde{w}_i^e \\
\tilde{\tilde{w}}_i^e
\end{bmatrix} =
\begin{bmatrix}
I_{2M} - D_{11,i-1}^T & -D_{21,i-1}^T \\
-D_{12,i-1}^T & \mathcal{J}_\epsilon^T - D_{22,i-1}^T
\end{bmatrix}
\begin{bmatrix}
\tilde{w}_{i-1}^e \\
\tilde{\tilde{w}}_{i-1}^e
\end{bmatrix} +
\begin{bmatrix}
\bar{s}_i^e \\
\tilde{s}_i^e
\end{bmatrix} -
\begin{bmatrix}
0 \\
\bar{b}^e
\end{bmatrix}
\]

or, in expanded form,

\[
\begin{align*}
\tilde{w}_i^e &= (I_{2M} - D_{11,i-1}^T)\tilde{w}_{i-1}^e - D_{21,i-1}^T\tilde{\tilde{w}}_{i-1}^e + \bar{s}_i^e \quad (9.61) \\
\tilde{\tilde{w}}_i^e &= (\mathcal{J}_\epsilon^T - D_{22,i-1}^T)\tilde{\tilde{w}}_{i-1}^e - D_{12,i-1}^T\tilde{w}_{i-1}^e + \bar{s}_i^e - \bar{b}^e \quad (9.62)
\end{align*}
\]

Conditioning both sides on \( \mathcal{F}_{i-1} \), computing the conditional second-order moments, and using the conditions from Assumption 8.1 on the gradient noise process we get
Proof

\[
\mathbb{E} \left[ \| \mathbf{\bar{w}}^e_i \|^2 \mid \mathcal{F}_{i-1} \right] = \| (I_{2M} - D^{T}_{11,i-1}) \mathbf{\bar{w}}^e_{i-1} - D^{T}_{21,i-1} \mathbf{\bar{w}}^e_{i-1} \|^2 + \mathbb{E} \left[ \| \mathbf{s}^e_i \|^2 \mid \mathcal{F}_{i-1} \right]
\]

(9.63)

and

\[
\mathbb{E} \left[ \| \mathbf{\bar{w}}^e_i \|^2 \mid \mathcal{F}_{i-1} \right] = \| (\mathcal{J}_e^T - D^{T}_{22,i-1}) \mathbf{\bar{w}}^e_{i-1} - D^{T}_{12,i-1} \mathbf{\bar{w}}^e_{i-1} - \mathbf{\bar{b}}^e \|^2 + \\
\mathbb{E} \left[ \| \mathbf{s}^e_i \|^2 \mid \mathcal{F}_{i-1} \right]
\]

(9.64)
Proof

Computing the expectations again we conclude that

$$\mathbb{E} \left\| \hat{w}^e_i \right\|^2 = \mathbb{E} \left\| (I_{2M} - D_{11,i-1}^T) \bar{w}^e_{i-1} - D_{21,i-1}^T \hat{w}^e_{i-1} \right\|^2 + \mathbb{E} \left\| \bar{s}_i^e \right\|^2$$  \hspace{1cm} (9.65)$$

and

$$\mathbb{E} \left\| \hat{\dot{w}}^e_i \right\|^2 = \mathbb{E} \left\| (J_{\epsilon}^T - D_{22,i-1}^T) \bar{w}^e_{i-1} - D_{12,i-1}^T \bar{w}^e_{i-1} - \hat{b}^e \right\|^2 + \mathbb{E} \left\| \bar{s}_i^e \right\|^2$$  \hspace{1cm} (9.66)$$

Continuing with the first variance (9.65), we can appeal to Jensen’s inequality (F.26) from the appendix and apply it to the function $f(x) = \|x\|^2$ to bound the variance as follows:
Proof

\[ \mathbb{E} \| \bar{w}_i^e \|^2 \]

\[ = \mathbb{E} \left\| (1 - t) \frac{1}{1 - t} (I_{2M} - D_{11,i-1}^T) \bar{w}_{i-1}^e - t \frac{1}{t} D_{21,i-1}^T \bar{w}_{i-1}^e \right\|^2 + \mathbb{E} \| \bar{s}_i^e \|^2 \]

\[ \leq (1 - t) \mathbb{E} \left\| \frac{1}{1 - t} (I_{2M} - D_{11,i-1}^T) \bar{w}_{i-1}^e \right\|^2 + t \mathbb{E} \left\| \frac{1}{t} D_{21,i-1}^T \bar{w}_{i-1}^e \right\|^2 + \mathbb{E} \| \bar{s}_i^e \|^2 \]

\[ \leq \frac{1}{1 - t} \mathbb{E} \left[ \| I_{2M} - D_{11,i-1}^T \| \| \bar{w}_{i-1}^e \|^2 \right] + \frac{1}{t} \mathbb{E} \left[ \| D_{21,i-1}^T \| \| \bar{w}_{i-1}^e \|^2 \right] + \mathbb{E} \| \bar{s}_i^e \|^2 \]

\[ \leq \frac{(1 - \sigma_{11} \mu_{\text{max}})^2}{1 - t} \mathbb{E} \| \bar{w}_{i-1}^e \|^2 + \frac{\sigma_{21}^2 \mu_{\text{max}}^2}{t} \mathbb{E} \| \bar{w}_{i-1}^e \|^2 + \mathbb{E} \| \bar{s}_i^e \|^2 \]  \quad (9.67)

for any arbitrary positive number \( t \in (0, 1) \). We select

\[ t = \sigma_{11} \mu_{\text{max}} \]  \quad (9.68)
Proof

Then, the last inequality can be written as

$$
\mathbb{E} \| \bar{w}_i^c \|^2 \leq (1 - \sigma_1 \mu_{\text{max}}) \mathbb{E} \| \bar{w}_{i-1}^c \|^2 + \left( \frac{\sigma_2^2 \mu_{\text{max}}}{\sigma_1} \right) \mathbb{E} \| \tilde{w}_{i-1} \|^2 + \mathbb{E} \| \bar{s}_i^c \|^2
$$

(9.69)

We now repeat a similar argument for the second variance relation (9.66). Thus, using Jensen’s inequality again we have
Proof

\[ \mathbb{E} \left\| \dot{\hat{w}}_i \right\|^2 = \]
\[ = \mathbb{E} \left\| J_e^T \hat{w}_i^{e-1} - \left[ D_{22,i-1}^T \hat{w}_{i-1} + D_{12,i-1}^T \bar{w}_{i-1} + \bar{b}^e \right] \right\|^2 + \mathbb{E} \left\| \bar{s}_i^e \right\|^2 \]
\[ = \mathbb{E} \left\| t \frac{1}{t} J_e^T \hat{w}_i^{e-1} - (1-t) \frac{1}{1-t} \left[ D_{22,i-1}^T \hat{w}_{i-1} + D_{12,i-1}^T \bar{w}_{i-1} + \bar{b}^e \right] \right\|^2 + \mathbb{E} \left\| \bar{s}_i^e \right\|^2 \]
\[ \leq \frac{1}{t} \mathbb{E} \left\| J_e^T \hat{w}_i^{e-1} \right\|^2 + \frac{1}{1-t} \mathbb{E} \left\| D_{22,i-1}^T \hat{w}_{i-1} + D_{12,i-1}^T \bar{w}_{i-1} + \bar{b}^e \right\|^2 + \mathbb{E} \left\| \bar{s}_i^e \right\|^2 \]

for any arbitrary positive number \( t \in (0, 1) \). Now note that
Proof

\[ \| \mathcal{J}_\epsilon^T \hat{w}_i \|^2 = (\hat{w}_{i-1})^* (\mathcal{J}_\epsilon^T)^* \mathcal{J}_\epsilon^T \hat{w}_{i-1} \]

\[ = (\hat{w}_{i-1})^* (\mathcal{J}_\epsilon \mathcal{J}_\epsilon^*)^T \hat{w}_{i-1} \]

\[ \leq \rho (\mathcal{J}_\epsilon \mathcal{J}_\epsilon^*) \| \hat{w}_{i-1} \|^2 \]

(9.71)

where we called upon the Rayleigh-Ritz characterization of the eigenvalues of Hermitian matrices [104, 113], namely,

\[ \lambda_{\text{min}}(C) \| x \|^2 \leq x^* C x \leq \lambda_{\text{max}}(C) \| x \|^2 \]

(9.72)
Proof

for any Hermitian matrix $C$. Applying this result to the Hermitian and non-negative definite matrix $C = (J_\epsilon J_\epsilon^*)^T$, and noting that $\rho(C) = \rho(C^T)$, we obtain (9.71). From definition (9.52) for $J_\epsilon$ we further get

\[
\rho \left( J_\epsilon J_\epsilon^* \right) = \rho \left[ (J_\epsilon \otimes I_{2M})(J_\epsilon^* \otimes I_{2M}) \right] \\
= \rho \left[ (J_\epsilon J_\epsilon^* \otimes I_{2M}) \right] \\
= \rho(J_\epsilon J_\epsilon^*) \\
(9.73)
\]

The matrix $J_\epsilon$ is block diagonal and consists of Jordan blocks. Assume initially that it consists of a single Jordan block, say, of size $4 \times 4$, for illustration purposes. Then, we can write:
Proof

$$J_\epsilon J^*_\epsilon = \begin{bmatrix}
\lambda & \epsilon \\
\epsilon & \lambda
\end{bmatrix}
\begin{bmatrix}
\lambda^* & \epsilon \\
\epsilon & \lambda^*
\end{bmatrix}
= \begin{bmatrix}
|\lambda|^2 & \epsilon\lambda \\
\epsilon\lambda^* & |\lambda|^2 + \epsilon^2
\end{bmatrix}
\begin{bmatrix}
|\lambda|^2 + \epsilon^2 & \epsilon\lambda \\
\epsilon\lambda^* & |\lambda|^2 + \epsilon^2
\end{bmatrix}
= (9.74)$$
Proof

Using the property that the spectral radius of a matrix is bounded by any of its norms, and using the 1–norm (maximum absolute column sum), we get for the above example

\[
\rho(J_\epsilon J_\epsilon^*) \leq \|J_\epsilon J_\epsilon^*\|_1 \\
= |\lambda|^2 + \epsilon^2 + \epsilon |\lambda^*| + \epsilon |\lambda| \\
= (|\lambda| + \epsilon)^2
\]  \hspace{1cm} (9.75)

If \( J_\epsilon \) consists of multiple Jordan blocks, say, \( L \) of them with eigenvalue \( \lambda_\ell \) each, then

\[
\rho(J_\epsilon J_\epsilon^*) \leq \max_{1 \leq \ell \leq L} (|\lambda_\ell| + \epsilon)^2 = (\rho(J_\epsilon) + \epsilon)^2
\]  \hspace{1cm} (9.76)
Proof

where $\rho(J_\epsilon)$ does not depend on $\epsilon$ and is equal to the second largest eigenvalue in magnitude in $J$, which we know is strictly less than one in magnitude. Substituting this conclusion into (9.70) gives

$$
\mathbb{E} \left\| \tilde{w}_i^e \right\|^2 \leq \frac{1}{t} (\rho(J_\epsilon) + \epsilon)^2 \mathbb{E} \left\| \tilde{w}_{i-1}^e \right\|^2 + \\
\frac{1}{1-t} \mathbb{E} \left\| D_{22,i-1}^T \tilde{w}_{i-1}^e + D_{12,i-1}^T \tilde{w}_{i-1}^e + \tilde{b}_e^e \right\|^2 + \mathbb{E} \left\| \tilde{s}_i^e \right\|^2
$$

(9.77)

Since we know that $\rho(J_\epsilon) \in (0, 1)$, then we can select $\epsilon$ small enough to ensure $\rho(J_\epsilon) + \epsilon \in (0, 1)$. We then select

$$
t = \rho(J_\epsilon) + \epsilon
$$

(9.78)
and rewrite (9.77) as

\[ \mathbb{E} \| \tilde{w}_i^e \|^2 \leq (\rho(J_\varepsilon) + \varepsilon) \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + \mathbb{E} \| \tilde{s}_i^e \|^2 + \left( \frac{1}{1 - \rho(J_\varepsilon) - \varepsilon} \right) \mathbb{E} \left\| D_{22,i-1}^\top \tilde{w}_{i-1}^e + D_{12,i-1}^\top \tilde{w}_{i-1}^e + \tilde{b}^e \right\|^2 \]

(9.79)

We can bound the last term on the right-hand side of the above expression as follows:
Proof

\[ \mathbb{E} \left\| D_{22,i-1}^T \ddot{w}_{i-1}^e + D_{12,i-1}^T \ddot{w}_{i-1}^e + \dot{b}^e \right\|^2 = \]

\[ = \mathbb{E} \left\| \frac{1}{3} 3D_{22,i-1}^T \ddot{w}_{i-1}^e + \frac{1}{3} 3D_{12,i-1}^T \ddot{w}_{i-1}^e + \frac{1}{3} 3\dot{b}^e \right\|^2 \]

\[ \leq \frac{1}{3} \mathbb{E} \left\| 3D_{22,i-1}^T \ddot{w}_{i-1}^e \right\|^2 + \frac{1}{3} \mathbb{E} \left\| 3D_{12,i-1}^T \ddot{w}_{i-1}^e \right\|^2 + \frac{1}{3} \| 3\dot{b}^e \|^2 \]

\[ \leq 3\mathbb{E} \left\| D_{22,i-1}^T \ddot{w}_{i-1}^e \right\|^2 + 3\mathbb{E} \left\| D_{12,i-1}^T \ddot{w}_{i-1}^e \right\|^2 + 3\| \dot{b}^e \|^2 \]

\[ \leq 3\sigma_{22}^2 \mu_{\text{max}}^2 \mathbb{E} \| \ddot{w}_{i-1}^e \|^2 + 3\sigma_{12}^2 \mu_{\text{max}}^2 \mathbb{E} \| \ddot{w}_{i-1}^e \|^2 + 3\| \dot{b}^e \|^2 \]
Proof

Substituting into (9.79) we obtain

\[
\mathbb{E} \left\| \tilde{w}_i^e \right\|^2 \leq \left( \rho(J_\epsilon) + \epsilon + \frac{3\sigma_{22}^2 \mu_{\text{max}}^2}{1 - \rho(J_\epsilon) - \epsilon} \right) \mathbb{E} \left\| \tilde{w}_{i-1}^e \right\|^2 + \\
\left( \frac{3\sigma_{12}^2 \mu_{\text{max}}^2}{1 - \rho(J_\epsilon) - \epsilon} \right) \mathbb{E} \left\| \tilde{w}_{i-1}^e \right\|^2 + \\
\left( \frac{3}{1 - \rho(J_\epsilon) - \epsilon} \right) \left\| \bar{b}^e \right\|^2 + \mathbb{E} \left\| \tilde{s}_i^e \right\|^2
\]

(9.81)
Proof

We now bound the noise terms, $\mathbb{E}\|\bar{s}^e_i\|^2$ in (9.69) and $\mathbb{E}\|\bar{s}^e_i\|^2$ in (9.81). For that purpose, we first note that

\[
\mathbb{E}\|\bar{s}^e_i\|^2 + \mathbb{E}\|\bar{s}^e_i\|^2 = \mathbb{E}\left\|\begin{bmatrix} \bar{s}^e_i \\ \bar{s}^e_i \end{bmatrix}\right\|^2 \\
= \mathbb{E}\left\|\begin{bmatrix} \bar{v}_e^T \bar{A}_2^T \bar{M} & \bar{s}^e_i \end{bmatrix}\right\|^2 \\
\leq \left\|\begin{bmatrix} \bar{v}_e^T \bar{A}_2^T \end{bmatrix}\right\|^2 \left\|\bar{M}\right\|^2 \mathbb{E}\|\bar{s}^e_i\|^2 \\
\leq v_1^2 \mu_{\text{max}}^2 \mathbb{E}\|\bar{s}^e_i\|^2
\]

(9.82)
where the positive constant $v_1$ is independent of $\mu_{\text{max}}$ and is equal to the following norm

$$v_1 \triangleq \left\| V_\epsilon^T A_2^T \right\|$$  \hspace{1cm} (9.83)

On the other hand, using (8.113)–(8.114), we have

$$\mathbb{E} \left\| s_i \right\|^2 = \sum_{k=1}^{N} \mathbb{E} \left\| s_{k,i} \right\|^2 = 2 \left( \sum_{k=1}^{N} \mathbb{E} \left\| s_{k,i} \right\|^2 \right)$$  \hspace{1cm} (9.84)
in terms of the variances of the individual gradient noise processes, $\mathbb{E}\|s_{k,i}\|^2$, and where we used the fact that

$$\|s_{k,i}^e\|^2 = 2\|s_{k,i}\|^2$$  \hfill (9.85)

Now, for each term $s_{k,i}$ we have
Proof

\[ \mathbb{E} \| s_{k,i} \|^2 \leq (\beta_k^2 / h^2) \mathbb{E} \| \phi_{k,i-1} \|^2 + \sigma_{s,k}^2 \]

\[ = (\beta_k^2 / h^2) \mathbb{E} \left\| \sum_{\ell \in \mathcal{N}_k} a_{1,\ell} \tilde{w}_{\ell,i-1} \right\|^2 + \sigma_{s,k}^2 \]

\[ \leq (\beta_k^2 / h^2) \sum_{\ell \in \mathcal{N}_k} a_{1,\ell} \mathbb{E} \| \tilde{w}_{\ell,i-1} \|^2 + \sigma_{s,k}^2 \]

\[ \leq (\beta_k^2 / h^2) \sum_{\ell=1}^{N} \mathbb{E} \| \tilde{w}_{\ell,i-1} \|^2 + \sigma_{s,k}^2 \]
Proof

\[
= \left( \beta_k^2 / 2h^2 \right) \mathbb{E} \left\| \tilde{w}_{i-1} \right\|^2 + \sigma_{s,k}^2
\]

\[
= \left( \beta_k^2 / 2h^2 \right) \mathbb{E} \left\| \left( \mathcal{V}_\epsilon^{-1} \right)^T \mathcal{V}_\epsilon^T \tilde{w}_{i-1} \right\|^2 + \sigma_{s,k}^2
\]

\[
\leq \left( \beta_k^2 / 2h^2 \right) \left\| \left( \mathcal{V}_\epsilon^{-1} \right)^T \right\|^2 \mathbb{E} \left\| \mathcal{V}_\epsilon^T \tilde{w}_{i-1} \right\|^2 + \sigma_{s,k}^2
\]

\[
\overset{(9.55)}{=} \left( \beta_k^2 / 2h^2 \right) v_2^2 \left[ \mathbb{E} \left\| \tilde{w}_{i-1} \right\|^2 + \mathbb{E} \left\| \tilde{w}_{i-1} \right\|^2 \right] + \sigma_{s,k}^2
\]

\[
(9.86)
\]
Proof

where \( h = 2 \) for complex data, while the positive constant \( v_2 \) is independent of \( \mu_{\text{max}} \) and denotes the norm

\[
v_2 \triangleq \left\| (\mathcal{N}_{\epsilon}^{-1})^T \right\|
\]

(9.87)

In this way, we can bound the term \( \mathbb{E} \| s^c_i \|^2 \) as follows:

\[
\mathbb{E} \| s^c_i \|^2 = 2 \left( \sum_{k=1}^{N} \mathbb{E} \| s_{k,i} \|^2 \right) \\
\leq v_2 \beta_d^2 \left[ \mathbb{E} \| \bar{w}_{i-1}^c \|^2 + \mathbb{E} \| \bar{w}_{i-1}^e \|^2 \right] + \sigma_s^2
\]

(9.88)

where we introduced the scalars:
Proof

\[ \beta_d^2 \triangleq \sum_{k=1}^{N} \beta_k^2 / h^2 \quad (9.89) \]

\[ \sigma_s^2 \triangleq \sum_{k=1}^{N} 2 \sigma_{s,k}^2 \quad (9.90) \]

Substituting into (9.82) we get

\[ \mathbb{E} \| \tilde{s}_i^e \|^2 + \mathbb{E} \| \tilde{s}_i^e \|^2 \leq v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + v_1^2 \mu_{\text{max}}^2 \sigma_s^2 \quad (9.91) \]
Proof

Using this bound in (9.69) and (9.81) we find that

\[ \mathbb{E} \| \tilde{w}_i^e \|^2 \leq \left( 1 - \sigma_{11} \mu_{\text{max}} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \right) \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + \left( \frac{\sigma_{21}^2 \mu_{\text{max}}}{\sigma_{11}} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \right) \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + v_1^2 \mu_{\text{max}}^2 \sigma_s^2 \]

(9.92)

and
Proof

$$\mathbb{E} \| \tilde{w}_i \|^2 \leq \left( \rho(J_\epsilon) + \epsilon + \frac{3\sigma_{12}^2 \mu_{\max}^2}{1 - \rho(J_\epsilon) - \epsilon} + v_1^2 v_2^2 \beta_d^2 \mu_{\max}^2 \right) \mathbb{E} \| \tilde{w}_{i-1} \|^2 +$$

$$\left( \frac{3\sigma_{12}^2 \mu_{\max}^2}{1 - \rho(J_\epsilon) - \epsilon} + v_1^2 v_2^2 \beta_d^2 \mu_{\max}^2 \right) \mathbb{E} \| \tilde{w}_{i-1} \|^2 +$$

$$\left( \frac{3}{1 - \rho(J_\epsilon) - \epsilon} \right) \| \tilde{b} \|^2 + v_1^2 \mu_{\max}^2 \sigma_s^2$$

(9.93)

We introduce the scalar coefficients
Proof

\[ a = 1 - \sigma_{11} \mu_{\text{max}} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 = 1 - O(\mu_{\text{max}}) \quad (9.94) \]

\[ b = \frac{\sigma_{21}^2 \mu_{\text{max}}}{\sigma_{11}} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 = O(\mu_{\text{max}}) \quad (9.95) \]

\[ c = \frac{3\sigma_{12}^2 \mu_{\text{max}}^2}{1 - \rho(J_\epsilon) - \epsilon} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 = O(\mu_{\text{max}}^2) \quad (9.96) \]

\[ d = \rho(J_\epsilon) + \epsilon + \frac{3\sigma_{22}^2 \mu_{\text{max}}^2}{1 - \rho(J_\epsilon) - \epsilon} + v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \]
\[ = \rho(J_\epsilon) + \epsilon + O(\mu_{\text{max}}^2) \quad (9.97) \]

\[ e = v_1^2 \mu_{\text{max}}^2 \sigma_s^2 = O(\mu_{\text{max}}^2) \quad (9.98) \]

\[ f = \left( \frac{3}{1 - \rho(J_\epsilon) - \epsilon} \right) \| \hat{b}^e \|^2 = O(\mu_{\text{max}}^2) \quad (9.99) \]
Proof

since $\|\ddot{\mathbf{b}}^e\| = O(\mu_{\text{max}})$. Using these parameters, we can combine (9.92) and (9.93) into a single compact inequality recursion as follows:

$$
\begin{bmatrix}
\mathbb{E} \|\mathbf{w}^e_i\|^2 \\
\mathbb{E} \|\dot{\mathbf{w}}^e_i\|^2
\end{bmatrix} \prec
\begin{bmatrix}
a & b \\
c & d
\end{bmatrix}
\begin{bmatrix}
\mathbb{E} \|\mathbf{w}^e_{i-1}\|^2 \\
\mathbb{E} \|\dot{\mathbf{w}}^e_{i-1}\|^2
\end{bmatrix} +
\begin{bmatrix}
e \\
e + f
\end{bmatrix}
$$

(9.100)

in terms of the $2 \times 2$ coefficient matrix $\Gamma$ indicated above and whose entries are of the form

$$
\Gamma =
\begin{bmatrix}
1 - O(\mu_{\text{max}}) & O(\mu_{\text{max}}) \\
O(\mu^2_{\text{max}}) & \rho(J_\epsilon) + \epsilon + O(\mu^2_{\text{max}})
\end{bmatrix}
$$

(9.101)
Now, we invoke again the property that the spectral radius of a matrix is upper bounded by any of its norms, and use the $1$–norm (maximum absolute column sum), to conclude that

$$\rho(\Gamma) \leq \max \left\{ 1 - O(\mu_{\text{max}}) + O(\mu_{\text{max}}^2), \; \rho(J_\epsilon) + \epsilon + O(\mu_{\text{max}}) + O(\mu_{\text{max}}^2) \right\}$$

Since $\rho(J_\epsilon) < 1$ is independent of $\mu_{\text{max}}$, and since $\epsilon$ and $\mu_{\text{max}}$ are small positive numbers that can be chosen arbitrarily small and independently of each other, it is clear that the right-hand side of the above expression can be made strictly smaller than one for sufficiently small $\epsilon$ and $\mu_{\text{max}}$. In that case, $\rho(\Gamma) < 1$ so that $\Gamma$ is stable. Moreover, it holds that
Proof

\[
(I_2 - \Gamma)^{-1} = \left[ \begin{array}{cc}
1 - a & -b \\
-c & 1 - d
\end{array} \right]^{-1}
\]

\[
= \frac{1}{(1 - a)(1 - d) - bc} \left[ \begin{array}{cc}
1 - d & b \\
c & 1 - a
\end{array} \right]
\]

\[
= \left[ \begin{array}{cc}
O(1/\mu_{\max}) & O(1) \\
O(\mu_{\max}) & O(1)
\end{array} \right]
\]

(9.103)

If we now iterate (9.100), and since \( \Gamma \) is stable, we conclude that
Proof

\[
\limsup_{i \to \infty} \left[ \frac{\mathbb{E} \left\| \tilde{w}_i^e \right\|^2}{\mathbb{E} \left\| \tilde{w}_i \right\|^2} \right] \preceq (I_2 - \Gamma)^{-1} \begin{bmatrix} e \\ e + f \end{bmatrix} = \begin{bmatrix} O(1/\mu_{\text{max}}) & O(1) \\ O(\mu_{\text{max}}) & O(1) \end{bmatrix} \begin{bmatrix} O(\mu_{\text{max}}^2) \\ O(\mu_{\text{max}}^2) \end{bmatrix} = \begin{bmatrix} O(\mu_{\text{max}}) \\ O(\mu_{\text{max}}^2) \end{bmatrix}
\]

(9.104)

from which we conclude that

\[
\limsup_{i \to \infty} \mathbb{E} \left\| \tilde{w}_i^e \right\|^2 = O(\mu_{\text{max}}), \quad \limsup_{i \to \infty} \mathbb{E} \left\| \tilde{w}_i \right\|^2 = O(\mu_{\text{max}}^2)
\]

(9.105)
and, therefore,

\[
\limsup_{i \to \infty} E \left\| \tilde{w}_i \right\|^2 = \limsup_{i \to \infty} E \left\| (V_\epsilon^{-1})^T \begin{bmatrix} \tilde{w}_i^c \\ \tilde{w}_i^e \end{bmatrix} \right\|^2 \\
\leq \limsup_{i \to \infty} v_2^2 \left[ E \left\| \tilde{w}_i^c \right\|^2 + E \left\| \tilde{w}_i^e \right\|^2 \right] \\
= O(\mu_{\text{max}}) \quad (9.106)
\]

which leads to the desired result (9.11).
Remark

We remark that the type of derivation used in the above proof, which starts from a stochastic recursion of the form (9.60) and transforms it into a deterministic recursion of the form (9.100), with the sizes of the parameters specified in terms of $\mu_{\text{max}}$ and with a $\Gamma$ matrix of the form (9.101), will be a recurring technique in our presentation. For example, we will encounter a similar derivation in two more locations in the current chapter while establishing Theorems 9.2 and 9.6 further ahead — see expressions (9.153) and (9.301); these theorems deal with the stability of the fourth and first-order moments of the error vector. We will also encounter a similar derivation in the next chapter — see expressions (10.48), (10.77), and (10.89).
Fourth-Order Stability
In the next chapter we will derive a long-term model to approximate the behavior of the network in the long term, as $i \to \infty$, and for sufficiently small step-sizes. The long-term model will be more tractable for performance analysis in the steady-state regime. At that point, we will argue that performance results that are derived from analyzing the long-term model provide accurate expressions for the performance results of the original network model to first-order in the step-size parameters. This is
Fourth-Order Error Moment

A reassuring conclusion that will lead to useful closed-form performance expressions. These results will be established under the condition that the fourth-order moment of the error vector, $\mathbb{E} \| \tilde{w}_{k,i} \|^4$, is asymptotically stable. We therefore establish this fact here and call upon it later in the analysis. To do so, we will rely on condition (8.121) on the fourth-order moments of the individual gradient noise processes.
**Fourth-Order Error Moment**

**Theorem 9.2** (Fourth-order moment stability). Consider a network of $N$ interacting agents running the distributed strategy (8.46) with a primitive matrix $P = A_1 A_o A_2$. Assume the aggregate cost (9.10) and the individual costs, $J_k(w)$, satisfy the conditions in Assumption 6.1. Assume further that the first and fourth-order moments of the gradient noise process satisfy the conditions of Assumption 8.1 with the second-order moment condition (8.115) replaced by the fourth-order moment condition (8.121). Then, the fourth-order moments of the network error vectors are stable for sufficiently small step-sizes, namely, it holds that

$$
\limsup_{i \to \infty} \mathbb{E} \left\| \tilde{w}_{k,i} \right\|^4 = O(\mu_{\text{max}}^2), \quad k = 1, 2, \ldots, N
$$

(9.107)
Sketch of Proof

Following similar arguments to the second-order case, it can be verified that:

\[
\begin{align*}
\begin{bmatrix}
\mathbb{E} \| \bar{w}_i^e \|_4^4 \\
\mathbb{E} \| \bar{\bar{w}}_i^e \|_4^4
\end{bmatrix} & \preceq \begin{bmatrix}
a & b \\
c & d
\end{bmatrix} \begin{bmatrix}
\mathbb{E} \| \bar{w}_{i-1}^e \|_4^4 \\
\mathbb{E} \| \bar{\bar{w}}_{i-1}^e \|_4^4
\end{bmatrix} + \begin{bmatrix}
a' & b' \\
c' & d'
\end{bmatrix} \begin{bmatrix}
\mathbb{E} \| \bar{w}_{i-1}^e \|_2^2 \\
\mathbb{E} \| \bar{\bar{w}}_{i-1}^e \|_2^2
\end{bmatrix} + \begin{bmatrix}
e \\
f
\end{bmatrix}
\end{align*}
\]

where

\[
\Gamma = \begin{bmatrix}
1 - O(\mu_{\max}) & O(\mu_{\max}) \\
O(\mu_{\max}^2) & \rho(J_\varepsilon) + \varepsilon + O(\mu_{\max}^2)
\end{bmatrix}, \quad (I - \Gamma)^{-1} = \begin{bmatrix}
O(1/\mu_{\max}) & O(1) \\
O(\mu_{\max}) & O(1)
\end{bmatrix}
\]
Sketch of Proof

\[
\limsup_{i \to \infty} \begin{bmatrix}
\mathbb{E} \| \tilde{\mathbf{w}}_i^e \|^4 \\
\mathbb{E} \| \tilde{\mathbf{w}}_i^e \|^4
\end{bmatrix} \leq \begin{bmatrix}
O(1/\mu_{\text{max}}) & O(1) \\
O(\mu_{\text{max}}) & O(1)
\end{bmatrix} \left( \begin{bmatrix}
O(\mu_{\text{max}}^3) \\
O(\mu_{\text{max}}^4)
\end{bmatrix} + \begin{bmatrix}
O(\mu_{\text{max}}^4) \\
O(\mu_{\text{max}}^4)
\end{bmatrix} \right)
\]

\[
= \begin{bmatrix}
O(\mu_{\text{max}}^2) \\
O(\mu_{\text{max}}^4)
\end{bmatrix}
\]
Sketch of Proof

\[
\limsup_{i \to \infty} \mathbb{E} \| \tilde{w}_i^e \|^4 = \limsup_{i \to \infty} \mathbb{E} \left( \left\| \left( \chi^{-1}_\varepsilon \right)^T \begin{bmatrix} \tilde{w}_i^e \\ \tilde{w}_i^e \end{bmatrix} \right\|^2 \right)^2 \\
\leq \left\| \left( \chi^{-1}_\varepsilon \right)^T \right\|^4 \left( \limsup_{i \to \infty} \mathbb{E} \left( \| \tilde{w}_i^e \|^2 + \| \tilde{w}_i^e \|^2 \right)^2 \right) \\
\leq \limsup_{i \to \infty} 2 \nu_2^4 \left( \mathbb{E} \| \tilde{w}_i^e \|^4 + \mathbb{E} \| \tilde{w}_i^e \|^4 \right) \\
= O(\mu_{\text{max}}^2) \quad (9.157)
\]

which leads to the desired result \((9.107)\).
Detailed Proof

*Proof.* We again establish the result for the general case of complex data and, therefore, $h = 2$ throughout this derivation. We recall relations (9.61)–(9.62), namely,

\[
\tilde{w}_i^e = (I_{2M} - D^T_{11,i-1})\tilde{w}_{i-1}^e - D^T_{21,i-1}\tilde{v}_{i-1}^e + \tilde{s}_i^e \quad (9.108)
\]

\[
\tilde{v}_i^e = (J_{\epsilon}^T - D^T_{22,i-1})\tilde{w}_{i-1}^e - D^T_{12,i-1}\tilde{w}_{i-1}^e + \tilde{s}_i^e - \tilde{v}_i^c \quad (9.109)
\]

Now note that, for any (deterministic or random) column vectors $a$ and $b$, it holds that

\[
\|a + b\|^4 = \|a\|^4 + \|b\|^4 + 2\|a\|^2 \|b\|^2 + 4\text{Re}(a^*b) \left[\|a\|^2 + \|b\|^2 + \text{Re}(a^*b)\] \quad (9.110)
\]
Proof

so that using the vector inequalities

\[
[\text{Re}(a^*b)]^2 \leq |a^*b|^2 \leq \|a\|^2 \|b\|^2 \tag{9.111}
\]

and

\[
2\text{Re}(a^*b) \leq \|a\|^2 + \|b\|^2 \tag{9.112}
\]

we get

\[
\|a + b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2 \|b\|^2 + 4\|a\|^2 \text{Re}(a^*b) \tag{9.113}
\]

Applying this inequality to (9.108) with the identifications
Proof

\begin{align*}
    a & \leftarrow (I_{2M} - D_{11,i-1}^T)\bar{w}^e_{i-1} - D_{21,i-1}^T \tilde{w}^e_{i-1} \\
    b & \leftarrow \bar{s}^e_i
\end{align*}  \quad (9.114)

we obtain

\begin{align*}
    \|\bar{w}^e_i\|^4 & \leq \|(I_{2M} - D_{11,i-1}^T)\bar{w}^e_{i-1} - D_{21,i-1}^T \tilde{w}^e_{i-1}\|^4 + 3\|\bar{s}^e_i\|^4 + \\
    & \quad 8\|(I_{2M} - D_{11,i-1}^T)\bar{w}^e_{i-1} - D_{21,i-1}^T \tilde{w}^e_{i-1}\|^2 \|\bar{s}^e_i\|^2 + \\
    & \quad 4\|(I_{2M} - D_{11,i-1}^T)\bar{w}^e_{i-1} - D_{21,i-1}^T \tilde{w}^e_{i-1}\|^2 \text{Re}(a^* \bar{s}^e_i)
\end{align*}  \quad (9.116)
Proof

Conditioning on $\mathcal{F}_{i-1}$ and computing the expectations of both sides, we will find that the expectation of the last term on the right-hand side of the above expression is zero due to the assumed properties on the gradient noise process. Taking expectations again we then conclude that

$$\mathbb{E}\|\bar{w}_i^e\|^4 \leq \mathbb{E}\|(I_{2M} - D_{11,i-1}^T)\bar{w}_{i-1}^e - D_{21,i-1}^T\tilde{w}_{i-1}^e\|^4 + 3(\mathbb{E}\|\bar{s}_i^e\|^4) + 8\left(\mathbb{E}\|(I_{2M} - D_{11,i-1}^T)\bar{w}_{i-1}^e - D_{21,i-1}^T\tilde{w}_{i-1}^e\|^2\right)(\mathbb{E}\|\bar{s}_i^e\|^2)$$
Proof

\[
\begin{align*}
= \mathbb{E} \left\| (1-t) \frac{1}{1-t} (I_{2M} - D_{11,i-1}^T) \tilde{w}_{i-1}^e - t \frac{1}{t} D_{21,i-1}^T \tilde{w}_{i-1}^e \right\|^4 + \\
3 \left( \mathbb{E} \| \bar{s}_i^e \|^4 \right) + 8 \left( \mathbb{E} \| \bar{s}_i^e \|^2 \right) \times \\
\left( \mathbb{E} \left\| (1-t) \frac{1}{1-t} (I_{2M} - D_{11,i-1}^T) \tilde{w}_{i-1}^e - t \frac{1}{t} D_{21,i-1}^T \tilde{w}_{i-1}^e \right\|^2 \right) \\
\leq \frac{(1 - \sigma_{11} \mu_{\max})^4}{(1-t)^3} \mathbb{E} \| \tilde{w}_{i-1}^e \|^4 + \frac{\sigma_{21}^4 \mu_{\max}^4}{t^3} \mathbb{E} \| \tilde{w}_{i-1}^e \|^4 + 3 \mathbb{E} \| \bar{s}_i^e \|^4 + \\
8 \left( \mathbb{E} \| \bar{s}_i^e \|^2 \right) \left( \frac{(1 - \sigma_{11} \mu_{\max})^2}{1-t} \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 + \frac{\sigma_{21}^2 \mu_{\max}^2}{t} \mathbb{E} \| \tilde{w}_{i-1}^e \|^2 \right) \\
(9.117)
\end{align*}
\]
for any arbitrary positive number $t \in (0, 1)$. Similarly, using the identifications

$$a \leftarrow (\mathcal{J}_e^T - D_{22,i-1}^T) \tilde{w}_{i-1}^e - D_{12,i-1}^T \tilde{w}_{i-1}^e - \tilde{b}^e \quad (9.118)$$

$$b \leftarrow \tilde{s}_i^e \quad (9.119)$$

for relation (9.109), we can establish the inequality
Proof

\[ \| \tilde{w}_i^e \|^4 \leq \left\| (J_{\epsilon}^T - D_{22,i-1}^T) \tilde{w}_{i-1}^e - D_{12,i-1}^T \bar{w}_{i-1}^e - \tilde{b}^e \right\|^4 + 3 \| \tilde{s}_i^e \|^4 + \\
8 \left\| (J_{\epsilon}^T - D_{22,i-1}^T) \tilde{w}_{i-1}^e - D_{12,i-1}^T \bar{w}_{i-1}^e - \tilde{b}^e \right\|^2 \| \tilde{s}_i^e \|^2 + \\
4 \| a \|^2 \text{Re}(a^*b) \]  

(9.120)

from which we conclude that, again for any positive scalar \( t \in (0, 1) \):
Proof

\[ \mathbb{E} \left\| \tilde{\mathbf{w}}_i \right\|^4 \]

\[ \leq \mathbb{E} \left\| (\mathcal{J}_e^T - \mathbf{D}_{22,i-1}^T) \tilde{\mathbf{w}}_{i-1}^e - \mathbf{D}_{12,i-1}^T \bar{\mathbf{w}}_{i-1}^e - \tilde{b}^e \right\|^4 + 3\mathbb{E} \left\| \tilde{s}_i^e \right\|^4 + 

8 \left( \mathbb{E} \left\| \tilde{s}_i^e \right\|^2 \right) \left( \mathbb{E} \left\| (\mathcal{J}_e^T - \mathbf{D}_{22,i-1}^T) \tilde{\mathbf{w}}_{i-1}^e - \mathbf{D}_{12,i-1}^T \bar{\mathbf{w}}_{i-1}^e - \tilde{b}^e \right\|^2 \right) \]

\[ \leq \mathbb{E} \left\| t \frac{1}{t} \mathcal{J}_e^T \tilde{\mathbf{w}}_{i-1}^e - (1 - t) \frac{1}{1 - t} \left[ \mathbf{D}_{22,i-1}^T \tilde{\mathbf{w}}_{i-1}^e + \mathbf{D}_{12,i-1}^T \bar{\mathbf{w}}_{i-1}^e + \tilde{b}^e \right] \right\|^4 + 

3 \left( \mathbb{E} \left\| \tilde{s}_i^e \right\|^4 \right) + 8 \left( \mathbb{E} \left\| \tilde{s}_i^e \right\|^2 \right) \times 

\left( \mathbb{E} \left\| t \frac{1}{t} \mathcal{J}_e^T \tilde{\mathbf{w}}_{i-1}^e - (1 - t) \frac{1}{1 - t} \left[ \mathbf{D}_{22,i-1}^T \tilde{\mathbf{w}}_{i-1}^e + \mathbf{D}_{12,i-1}^T \bar{\mathbf{w}}_{i-1}^e + \tilde{b}^e \right] \right\|^2 \right) \]
Proof

\[
\leq \frac{1}{t^3} \|J_\epsilon\|^4 \mathbb{E} \left\| \tilde{w}_{i-1}^e \right\|^4 + \\
\frac{1}{(1-t)^3} \mathbb{E} \left\| D_{22,i-1}^T \tilde{w}_{i-1}^e + D_{12,i-1}^T \tilde{w}_{i-1}^e + \tilde{b}^e \right\|^4 + \\
3 \left( \mathbb{E} \left\| s_i^e \right\|^4 \right) + 8 \left( \mathbb{E} \left\| s_i^e \right\|^2 \right) \times \\
\left( \frac{1}{t} \|J_\epsilon\|^2 \mathbb{E} \left\| \tilde{w}_{i-1}^e \right\|^2 + \frac{1}{1-t} \mathbb{E} \left\| D_{22,i-1}^T \tilde{w}_{i-1}^e + D_{12,i-1}^T \tilde{w}_{i-1}^e + \tilde{b}^e \right\|^2 \right)
\]
Proof

\[ \leq \frac{(\rho(J_\epsilon) + \epsilon) \cdot \mathbb{E} \|\tilde{w}_{i-1}^e\|^4}{t^3} + \]

\[ + \frac{27\sigma_{22}^4\mu_{\text{max}}^4}{(1-t)^3} \mathbb{E} \|\tilde{w}_{i-1}^e\|^4 + \frac{27\sigma_{12}^4\mu_{\text{max}}^4}{(1-t)^3} \mathbb{E} \|\bar{w}_{i-1}^e\|^4 + 27\|\bar{b}^e\|_4^4 + \]

\[ 3 \left( \mathbb{E} \|\tilde{s}_i^e\|_4^4 \right) + 8 \left( \frac{\rho(J_\epsilon) + \epsilon}{t} \right) \left( \mathbb{E} \|\tilde{w}_{i-1}^e\|_2^2 \right) \mathbb{E} \|\tilde{s}_i^e\|_2^2 + \]

\[ + 8 \left( \frac{3\sigma_{22}^2\mu_{\text{max}}^2}{1-t} \mathbb{E} \|\tilde{w}_{i-1}^e\|_2^2 + \frac{3\sigma_{12}^2\mu_{\text{max}}^2}{1-t} \mathbb{E} \|\bar{w}_{i-1}^e\|_2^2 + 3\|\bar{b}^e\|_2^2 \right) \mathbb{E} \|\tilde{s}_i^e\|_2^2 \]

(9.121)
Proof

where in the last inequality we used the result

\[
\|a + b + c\|^4 = \left\| \frac{1}{3}3a + \frac{1}{3}3b + \frac{1}{3}3c \right\|^4 \\
\leq \frac{1}{3}\|3a\|^4 + \frac{1}{3}\|3b\|^4 + \frac{1}{3}\|3c\|^4 \\
= 27\|a\|^4 + 27\|b\|^4 + 27\|c\|^4
\] (9.122)
Proof

We now bound the fourth-order noise terms that appear in expressions (9.121) and (9.121) for $\mathbb{E} \| \tilde{w}_i^e \|^4$ and $\mathbb{E} \| \tilde{w}_i^e \|^4$, namely, $\mathbb{E} \| \tilde{s}_i^e \|^4$ and $\mathbb{E} \| \tilde{s}_i^e \|^4$. Thus, note that

$$
\mathbb{E} \| \tilde{s}_i^e \|^4 + \mathbb{E} \| \tilde{s}_i^e \|^4 \leq \mathbb{E} (\| \tilde{s}_i^e \|^2 + \| \tilde{s}_i^e \|^2)^2
$$

$$
= \mathbb{E} \left( \left\| \begin{bmatrix} \tilde{s}_i^e \\ \tilde{s}_i^e \end{bmatrix} \right\|^2 \right)^2
$$

$$
= \mathbb{E} \| V_T^T A_T^T M s_i^e \|^4
$$

$$
\leq \| V_T^T A_T \|^4 \| M \|^4 \mathbb{E} \| s_i^e \|^4
$$

$$
\leq \nu_1^4 \mu_\max \mathbb{E} \| s_i^e \|^4
$$

(9.123)
Proof

On the other hand, using Jensen’s inequality (F.26) and applying it to the convex function \( f(x) = x^2 \),

\[
\mathbb{E} \left\| s_i^e \right\|^4 = \left( \mathbb{E} \left\| s_i^e \right\|^2 \right)^2
\]

\[
= 4 \mathbb{E} \left( \sum_{k=1}^{N} \left\| s_{k,i} \right\|^2 \right)^2
\]

\[
= 4 \mathbb{E} \left( \frac{1}{N} N \left\| s_{1,i} \right\|^2 + \frac{1}{N} N \left\| s_{2,i} \right\|^2 + \ldots + \frac{1}{N} N \left\| s_{N,i} \right\|^2 \right)^2
\]

\[
(\text{F.26}) \leq \frac{4}{N} \mathbb{E} \left( \left( N \left\| s_{1,i} \right\|^2 \right)^2 + \left( N \left\| s_{2,i} \right\|^2 \right)^2 + \ldots + \left( N \left\| s_{N,i} \right\|^2 \right)^2 \right)
\]

\[
= 4N \sum_{k=1}^{N} \mathbb{E} \left\| s_{k,i} \right\|^4
\]

(9.124)
Proof

in terms of the fourth-order moments of the individual gradient noise processes, \( \mathbb{E} \left\| s_{k,i} \right\|^4 \). Likewise, we have

\[
\sum_{\ell=1}^{N} \left\| \tilde{w}_{\ell,i-1} \right\|^4 \leq \left( \left\| \tilde{w}_{1,i-1} \right\|^2 + \left\| \tilde{w}_{2,i-1} \right\|^2 + \ldots + \left\| \tilde{w}_{N,i-1} \right\|^2 \right)^2
\]

\[
= \left( \left\| \tilde{w}_{i-1} \right\|^2 \right)^2
\]

\[
= \left( \frac{1}{2} \left\| \tilde{w}^e_{i-1} \right\|^2 \right)^2
\]

\[
= \frac{1}{4} \left\| \tilde{w}^e_{i-1} \right\|^4
\]

(9.125)
Therefore, for each term $s_{k,i}$ in (9.124) we can write

$$E \| s_{k,i} \|^4 \overset{(8.122)}{\leq} (\beta_{4,k}^4 / h^4) E \| \phi_{k,i-1} \|^4 + \sigma_{s_{4,k}}^4$$

$$= (\beta_{4,k}^4 / h^4) E \left\| \sum_{\ell \in N_k} a_{1,\ell k} \tilde{w}_{\ell,i-1} \right\|^4 + \sigma_{s_{4,k}}^4$$

$$\overset{(F.26)}{\leq} (\beta_{4,k}^4 / h^4) \sum_{\ell \in N_k} a_{1,\ell k} E \| \tilde{w}_{\ell,i-1} \|^4 + \sigma_{s_{4,k}}^4$$
Proof

\[
\begin{align*}
\leq & \quad (\beta_{4,k}^4 / h^4) \sum_{\ell=1}^{N} \mathbb{E} \left\| \tilde{w}_{i-1} \right\|^4 + \sigma_{s4,k}^4 \\
(9.125) & \leq (\beta_{4,k}^4 / 4h^4) \mathbb{E} \left\| \tilde{w}^e_{i-1} \right\|^4 + \sigma_{s4,k}^4 \\
= & \quad (\beta_{4,k}^4 / 4h^4) \mathbb{E} \left\| (\mathcal{V}_e^{-1})^T \mathcal{V}_e^T \tilde{w}^e_{i-1} \right\|^4 + \sigma_{s4,k}^4 \\
\leq & \quad (\beta_{4,k}^4 / 4h^4) \mathbb{E} \left\| (\mathcal{V}_e^{-1})^T \right\|^4 \mathbb{E} \left\| \mathcal{V}_e^T \tilde{w}^e_{i-1} \right\|^4 + \sigma_{s4,k}^4 \\
(a) & \quad 2(\beta_{4,k}^4 / 4h^4) v_2^4 \left[ \mathbb{E} \left\| \bar{w}^e_{i-1} \right\|^4 + \mathbb{E} \left\| \bar{w}^e_{i-1} \right\|^4 \right] + \sigma_{s4,k}^4 \\
(9.126)
\end{align*}
\]
Proof

where in step (a) we used (9.55) to conclude that

$$
\| \nu_e^T \tilde{w}_{i-1}^e \|^4 = \left( \left\| \begin{bmatrix} \tilde{w}_{i-1}^e \\ \tilde{w}_{i-1}^e \end{bmatrix} \right\|^2 \right)^2
$$

$$
= \left( \| \tilde{w}_{i-1}^e \|^2 + \| \tilde{w}_{i-1}^e \|^2 \right)^2

\leq 2 \| \tilde{w}_{i-1}^e \|^4 + 2 \| \tilde{w}_{i-1}^e \|^4
$$

(9.127)

In this way, we can bound the term $\mathbb{E} \| s_i^e \|^4$ as follows:

$$
\mathbb{E} \| s_i^e \|^4 \leq v_2^4 \beta_d^4 \left[ \mathbb{E} \| \tilde{w}_{i-1}^e \|^4 + \mathbb{E} \| \tilde{w}_{i-1}^e \|^4 \right] + \sigma_{s_4}^4
$$

(9.128)
Proof

where we introduced the scalars:

\[
\beta_{d4}^4 \triangleq 2N \left( \sum_{k=1}^{N} \frac{\beta_{4,k}^4}{h^4} \right) \\
\sigma_{s4}^4 \triangleq 4N \left( \sum_{k=1}^{N} \sigma_{s4,k}^4 \right)
\]
Proof

Substituting into (9.123) we get

$$
\mathbb{E} \left\| \bar{s}_i^e \right\|^4 + \mathbb{E} \left\| \bar{s}_i^e \right\|^4 \leq v_1^4 v_2^4 \beta d_4 \mu_{\text{max}}^4 \left[ \mathbb{E} \left\| \bar{w}_{i-1}^e \right\|^4 + \mathbb{E} \left\| \bar{w}_{i-1}^e \right\|^4 \right] + v_1^4 \mu_{\text{max}}^4 \sigma_{s4}^4
$$

(9.131)

Returning to (9.117), selecting $t = \sigma_{11} \mu_{\text{max}}$, and using the bounds (9.91) and (9.131), we then find that
Proof

\[
\mathbb{E} \| \mathbf{w}_i^e \|^4 \leq (1 - \sigma_{11} \mu_{\text{max}}) \mathbb{E} \| \mathbf{w}_{i-1}^e \|^4 + \frac{\sigma_{21}^4 \mu_{\text{max}}^4}{\sigma_{11}^3} \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^4 + \\
3v_1^4 v_2^4 \beta_d^2 \mu_{\text{max}}^4 \left[ \mathbb{E} \| \mathbf{w}_{i-1}^e \|^4 + \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^4 \right] + 3v_1^4 \mu_{\text{max}}^4 \sigma_{s4}^4 + \\
8v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 (1 - \sigma_{11} \mu_{\text{max}}) \left( \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 \right)^2 + \\
8v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 (1 - \sigma_{11} \mu_{\text{max}}) \left( \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 \right) \left( \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^2 \right) + \\
8v_1^2 v_2^2 \mu_{\text{max}}^2 \sigma_{s}^2 (1 - \sigma_{11} \mu_{\text{max}}) \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 + \\
8 \frac{\sigma_{21}^4 \mu_{\text{max}}^4}{\sigma_{11}} v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \left( \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^2 \right)^2 + \\
8 \frac{\sigma_{21}^4 \mu_{\text{max}}^4}{\sigma_{11}} v_1^2 v_2^2 \beta_d^2 \mu_{\text{max}}^2 \left( \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^2 \right) \left( \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 \right) + \\
8 \frac{\sigma_{21}^4 \mu_{\text{max}}^4}{\sigma_{11}} v_1^2 v_2^2 \mu_{\text{max}}^2 \sigma_{s}^2 \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^e \|^2 \]

(9.132)
Proof

Now, for any real random variables $a$ and $b$ it holds that

$$(E a)^2 \leq E a^2$$  \hspace{1cm} (9.133)

and

$$2 (E a^2) \cdot (E b^2) \leq E a^4 + E b^4$$  \hspace{1cm} (9.134)

This latter property can be established as follows. Using $(E a^2 - E b^2)^2 \geq 0$, we get

$$2 (E a^2) \cdot (E b^2) \leq (E a^2)^2 + (E b^2)^2$$

$$\leq E a^4 + E b^4$$  \hspace{1cm} (9.135)
Proof

These properties enable us to write

\[ 2 \left( \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^2 \right) \left( \mathbb{E} \left| \bar{w}_{i-1}^e \right|^2 \right) \leq \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^4 + \mathbb{E} \left| \bar{w}_{i-1}^e \right|^4 \quad (9.136) \]

\[ \left( \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^2 \right)^2 \leq \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^4 \quad (9.137) \]

\[ \left( \mathbb{E} \left| \bar{w}_{i-1}^e \right|^2 \right)^2 \leq \mathbb{E} \left| \bar{w}_{i-1}^e \right|^4 \quad (9.138) \]

so that

\[ \mathbb{E} \left| \tilde{w}_{i}^e \right|^4 \leq a \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^4 + b \mathbb{E} \left| \bar{w}_{i-1}^e \right|^4 + a' \mathbb{E} \left| \bar{w}_{i-1}^e \right|^2 + b' \mathbb{E} \left| \tilde{w}_{i-1}^e \right|^2 + e \quad (9.139) \]
Proof

where the constant parameters \( \{a, b, a', b', e\} \) have the following form

\[
a = 1 - \sigma_{11} \mu_{\max} + O(\mu_{\max}^2) \\
b = O(\mu_{\max}) \\
a' = O(\mu_{\max}^2) \\
b' = O(\mu_{\max}^3) \\
e = O(\mu_{\max}^4)
\]  

(9.140) \hspace{1cm} (9.141) \hspace{1cm} (9.142) \hspace{1cm} (9.143) \hspace{1cm} (9.144)

In a similar manner, using (9.121) and selecting

\[
t = \rho(J_e) + \epsilon < 1
\]

(9.145)
Proof

\[ \mathbb{E} \| \tilde{w}_i^c \|^4 \leq (\rho(J_\epsilon) + \epsilon) \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 + \]

\[ \frac{27 \mu_{\text{max}}^4}{(1 - \rho(J_\epsilon) - \epsilon)^3} \left[ \sigma_{22}^4 \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 + \sigma_{12}^4 \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 \right] + \]

\[ 3v_1^4 v_2^4 \beta_{d4}^4 \mu_{\text{max}}^4 \left[ \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 + \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 \right] + 3v_1^4 \mu_{\text{max}}^4 \sigma_s^4 \]

\[ 8 (\rho(J_\epsilon) + \epsilon) v_1^2 \mu_{\text{max}}^2 \sigma_s^2 \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 + 27 \| \tilde{b}^e \|^4 + \]

\[ 4 (\rho(J_\epsilon) + \epsilon) v_1^2 v_2^2 \beta_{d2}^2 \mu_{\text{max}}^2 \left[ \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 + 3 \mathbb{E} \| \tilde{w}_{i-1}^c \|^4 \right] + \]

\[ \frac{24 \mu_{\text{max}}^4 v_1^2 \sigma_s^2}{1 - \rho(J_\epsilon) - \epsilon} \left[ \sigma_{22}^2 \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 + \sigma_{12}^2 \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 \right] + \]

\[ \frac{12 \mu_{\text{max}}^4 v_1^2 v_2^2 \beta_{d2}^2}{1 - \rho(J_\epsilon) - \epsilon} \times \]

\[ \left[ (\sigma_{22}^2 + 3 \sigma_{12}^2) \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 + (\sigma_{12}^2 + 3 \sigma_{22}^2) \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 \right] + \]

\[ 24 \| \tilde{b}^e \|^2 v_1^2 v_2^2 \beta_{d2}^2 \mu_{\text{max}}^2 \left[ \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 + \mathbb{E} \| \tilde{w}_{i-1}^c \|^2 \right] + \]

\[ 24 \| \tilde{b}^e \|^2 v_1^2 \mu_{\text{max}}^2 \sigma_s^2 \]

(9.146)
Proof

so that

\[ \mathbb{E} \| \dot{\mathbf{w}}_i^c \|^4 \leq c \mathbb{E} \| \mathbf{w}_{i-1}^c \|^4 + d \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^c \|^4 + c' \mathbb{E} \| \mathbf{w}_{i-1}^c \|^2 + d' \mathbb{E} \| \dot{\mathbf{w}}_{i-1}^c \|^2 + f \]  \tag{9.147}

where the constant parameters \( \{c, d, c', d', f\} \) have the following form

\[
\begin{align*}
    c &= O(\mu_{\text{max}}^2) \quad \tag{9.148} \\
    d &= \rho(J_e) + \epsilon + O(\mu_{\text{max}}^2) \quad \tag{9.149} \\
    c' &= O(\mu_{\text{max}}^4) \quad \tag{9.150} \\
    d' &= O(\mu_{\text{max}}^2) \quad \tag{9.151} \\
    f &= O(\mu_{\text{max}}^4) \quad \tag{9.152}
\end{align*}
\]
Proof

In other words, we can write

\[
\left[ \frac{\mathbb{E} \| \bar{w}_i^e \|^4}{\mathbb{E} \| \bar{w}_i^e \|^4} \right] \preceq \left[ \begin{array}{cc} a & b \\ c & d \end{array} \right] \left[ \frac{\mathbb{E} \| \bar{w}_{i-1}^e \|^4}{\mathbb{E} \| \bar{w}_{i-1}^e \|^4} \right] + \left[ \begin{array}{cc} a' & b' \\ c' & d' \end{array} \right] \left[ \frac{\mathbb{E} \| \bar{w}_{i-1}^e \|^2}{\mathbb{E} \| \bar{w}_{i-1}^e \|^2} \right] + \left[ \begin{array}{c} e \\ f \end{array} \right]
\]

\[\Delta \equiv \Gamma \]

in terms of the $2 \times 2$ coefficient matrix $\Gamma$ indicated above and whose entries are of the form

\[\Gamma = \begin{bmatrix} 1 - O(\mu_{\max}) & O(\mu_{\max}) \\ O(\mu_{\max}^2) & \rho(J_\epsilon) + \epsilon + O(\mu_{\max}^2) \end{bmatrix}, \quad (I - \Gamma)^{-1} = \begin{bmatrix} O(1/\mu_{\max}) & O(1) \\ O(\mu_{\max}) & O(1) \end{bmatrix} \quad (9.154)\]
Proof

We again find that $\Gamma$ is a stable matrix for sufficiently small $\mu_{\text{max}}$ and $\epsilon$. Moreover, using (9.105) we have

$$\limsup_{i \to \infty} \left[ \begin{array}{cc} a' & b' \\ c' & d' \end{array} \right] \left[ \begin{array}{c} \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 \\ \mathbb{E} \| \mathbf{w}_{i-1}^e \|^2 \end{array} \right] = \left[ \begin{array}{c} O(\mu_{\text{max}}^3) \\ O(\mu_{\text{max}}^4) \end{array} \right] (9.155)$$

In this case, we can iterate (9.153) and use (9.103) to conclude that

$$\limsup_{i \to \infty} \mathbb{E} \| \mathbf{w}_i^e \|^4 = O(\mu_{\text{max}}^2), \quad \limsup_{i \to \infty} \mathbb{E} \| \mathbf{w}_i^e \|^4 = O(\mu_{\text{max}}^4) (9.156)$$

and, therefore,
Proof

\[
\lim_{i \to \infty} \mathbb{E} \left\| \tilde{w}_i^e \right\|^4 \leq \left( \left\| (\mathcal{V}_\epsilon^{-1})^T \begin{bmatrix} \tilde{w}_i^e \\ \tilde{w}_i^e \end{bmatrix} \right\|^2 \right)^2 \\
\leq \left\| (\mathcal{V}_\epsilon^{-1})^T \right\|^4 \left( \lim_{i \to \infty} \mathbb{E} \left( \left\| \tilde{w}_i^e \right\|^2 + \left\| \tilde{w}_i^e \right\|^2 \right)^2 \right) \\
\leq \lim_{i \to \infty} 2\nu_2^4 \left( \mathbb{E} \left\| \tilde{w}_i^e \right\|^4 + \mathbb{E} \left\| \tilde{w}_i^e \right\|^4 \right) \\
= O(\mu_{\text{max}}^2) \\
\tag{9.157}
\]

which leads to the desired result (9.107).
End of Lecture