8 Algorithms for Decentralized Optimization

8.1 Establish an analogous argument to the one in Section 8.3.2 for the gradient-tracking recursion (8.58)–(8.59).

Solution. In network quantities, the gradient tracking recursion takes the form:

$$\begin{aligned} \mathcal{W}_i &= \mathcal{A}^{\mathsf{T}} \mathcal{W}_{i-1} - \mu g_{i-1} \\ g_i &= \mathcal{A}^{\mathsf{T}} g_{i-1} + \nabla \mathcal{J}(\mathcal{W}_i) - \nabla \mathcal{J}(\mathcal{W}_{i-1}) \end{aligned}$$

Assuming the algorithm converges to some set of fixed-points $\mathcal{W}_\infty, g_\infty$, we have:

$$\begin{aligned} \mathcal{W}_\infty &= \mathcal{A}^{\mathsf{T}} \mathcal{W}_\infty - \mu g_\infty \\ g_\infty &= \mathcal{A}^{\mathsf{T}} g_\infty + \nabla \mathcal{J}(\mathcal{W}_\infty) - \nabla \mathcal{J}(\mathcal{W}_\infty) = \mathcal{A}^{\mathsf{T}} g_\infty \end{aligned}$$

It follows that g_∞ is consensual. We can then examine the evolution of the centroid:

$$\frac{1}{K}\sum_{k=1}^{K} g_{k,i} = \frac{1}{K}\sum_{k=1}^{K} g_{k,i-1} + \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i}) - \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i-1})$$

After iterating and telescoping:

$$\frac{1}{K}\sum_{k=1}^{K} g_{k,i} = \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i})$$

Since g_∞ is consensual, it follows that $g_{k,\infty} = \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,\infty})$. Then, for the centroid of the weights, we have:

$$\frac{1}{K}\sum_{k=1}^{K} w_{k,\infty} = \frac{1}{K}\sum_{k=1}^{K} w_{k,\infty} - \frac{\mu}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,\infty}) \Longrightarrow \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,\infty}) = 0$$

Finally, we have:

$$\mathcal{W}_\infty = \mathcal{A}^{\mathsf{T}} \mathcal{W}_\infty - \mu g_\infty = \mathcal{A}^{\mathsf{T}} \mathcal{W}_\infty$$

It then follows that $\mathcal{W}_\infty$ is consensual, hence $w_{k,\infty} = w_\infty$ for all k and from $\sum_{k=1}^{K} \nabla J_k(w_\infty) = 0$ that it is optimal.

8.2 For a deterministic optimization problem of your choice, implement the consensus+innovation, EXTRA and gradient-tracking algorithms and show that consensus+innovations exhibits a bias, while EXTRA and gradient-tracking converge exactly. How do these findings change with the choice of the step-size μ?

Solution. The solution is provided as a Jupyter notebook in the separate file `Problem_8_2.ipynb`.

8.3 Show the exact incremental adjustments to the derivation of the EXTRA algorithm in Section 8.3 that lead to Exact diffusion (8.65)–(8.67).

Solution. We begn with (8.41), repeated here for reference:

$$\mathcal{L}(\mathcal{w}, \lambda) = \mathcal{J}(\mathcal{w}) + \eta\lambda^{\mathsf{T}}\mathcal{B}\,\mathcal{w} + \frac{\eta}{2}\,\mathcal{w}^{\mathsf{T}}\,\mathcal{L}\,\mathcal{w}$$

Instead of performing straight gradient descent ascent as in (8.42)–(8.43), we descend incrementally, first along $\mathcal{J}(\mathcal{w})$ and subsequently along the remaining terms. This yields:

$$\begin{aligned}
\psi_i &= \mathcal{w}_{i-1} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) \\
\mathcal{w}_i &= \psi_i - \mu\eta\mathcal{L}\psi_i - \mu\eta\mathcal{B}^{\mathsf{T}}\lambda_{i-1} \\
\lambda_i &= \lambda_{i-1} + \mu\eta\mathcal{B}\,\mathcal{w}_i
\end{aligned}$$

With the choice $\eta = \mu^{-1}$ and $\mathcal{A}^{\mathsf{T}} = I - \mu\eta\mathcal{L} = I - \mathcal{L}$, we have:

$$\begin{aligned}
\psi_i &= \mathcal{w}_{i-1} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) \\
\mathcal{w}_i &= \mathcal{A}^{\mathsf{T}}\psi_i - \mathcal{B}^{\mathsf{T}}\lambda_{i-1} \\
\lambda_i &= \lambda_{i-1} + \mathcal{B}\,\mathcal{w}_i
\end{aligned}$$

We can write this compactly as:

$$\begin{aligned}
\mathcal{w}_i &= \mathcal{A}^{\mathsf{T}}\left(\mathcal{w}_{i-1} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1})\right) - \mathcal{B}^{\mathsf{T}}\lambda_{i-1} \\
\lambda_i &= \lambda_{i-1} + \mathcal{B}\,\mathcal{w}_i
\end{aligned}$$

We now follow a similar argument to EXTRA to eliminate the dual variable. The primal update at time $i-1$ is evaluated to:

$$\mathcal{w}_{i-1} = \mathcal{A}^{\mathsf{T}}\left(\mathcal{w}_{i-2} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-2})\right) - \mathcal{B}^{\mathsf{T}}\lambda_{i-2}$$

Subtracting:

$$\begin{aligned}
\mathcal{w}_i - \mathcal{w}_{i-1} &= \mathcal{A}^{\mathsf{T}}\left(\mathcal{w}_{i-1} - \mathcal{w}_{i-2} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) + \mu\nabla\mathcal{J}(\mathcal{w}_{i-2})\right) - \mathcal{B}^{\mathsf{T}}\left(\lambda_{i-1} - \lambda_{i-2}\right) \\
&= \mathcal{A}^{\mathsf{T}}\left(\mathcal{w}_{i-1} - \mathcal{w}_{i-2} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) + \mu\nabla\mathcal{J}(\mathcal{w}_{i-2})\right) - \mathcal{B}^{\mathsf{T}}\mathcal{B}\,\mathcal{w}_{i-1} \\
&= \mathcal{A}^{\mathsf{T}}\left(\mathcal{w}_{i-1} - \mathcal{w}_{i-2} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) + \mu\nabla\mathcal{J}(\mathcal{w}_{i-2})\right) - \mathcal{L}\,\mathcal{w}_{i-1}
\end{aligned}$$

After rearranging:

$$\mathcal{w}_i = \mathcal{A}^{\mathsf{T}}\left(2\,\mathcal{w}_{i-1} - \mathcal{w}_{i-2} - \mu\nabla\mathcal{J}(\mathcal{w}_{i-1}) + \mu\nabla\mathcal{J}(\mathcal{w}_{i-2})\right)$$

We can formulate this relation in multiple steps as:

$$\begin{aligned}\psi_i &= \mathcal{W}_{i-1} - \mu\nabla\mathcal{J}(\mathcal{W}_{i-1})\\ \phi_i &= \mathcal{W}_{i-1} + \psi_i - \psi_{i-1}\\ \mathcal{W}_i &= \mathcal{A}^{\mathsf{T}}\phi_i\end{aligned}$$

which is the Exact diffusion algorithm in network form.

8.4 Show the exact incremental adjustments to the derivation of the NEXT algorithm in Section 8.4 that lead to Aug-DGM (8.68)–(8.70).

Solution. Let us examine the next algorithm, which we repeat here for reference:

$$\begin{aligned}w_{k,i} &= \sum_{\ell\in\mathcal{N}_k} a_{\ell k} w_{\ell,i-1} - \mu g_{k,i-1}\\ g_{k,i} &= \sum_{\ell\in\mathcal{N}_k} a_{\ell k} g_{k,i-1} + \nabla J_k(w_{k,i}) + \nabla J_k(w_{k,i-1})\end{aligned}$$

The first update in $w_{k,i}$ is reminiscent of a consensus+innovations update, where the local gradient in the innovation is replaced by the gradient tracking variable $g_{k,i-1}$. This motivates a diffusion-type update, where the averaging operation is applied to both the weight and the innovation term, of the form:

$$w_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\left(w_{\ell,i-1} - \mu g_{\ell,i-1}\right)$$

Similarly, we may apply the averaging opperations to the driving term of the dynamic consensus algorithm in $g_{k,i}$, resulting in:

$$g_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\left(g_{k,i-1} + \nabla J_k(w_{k,i}) + \nabla J_k(w_{k,i-1})\right)$$

Together, these recursions correspond exactly to Aug-DGM (8.68)–(8.70).

9 Convergence of Decentralized Algorithms

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

Solution. Analogously to (9.12), we have for the diffusion recursion:

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

We note that the only difference to (9.12) for the consensus+innovations algorithm is Λ pre-multiplying the gradient term $\mathcal{V}^\mathsf{T}\widehat{\nabla\mathcal{J}}(\mathcal{w}_{i-1})$. Then, proceeding with (9.16) accordingly, we find:

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

Relations (9.27)–(9.28) then follow directly after normalization and substitutions.

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

Solution. For the EXTRA and Exact diffusion algorithms, this is most immediately seen from the primal-dual recursions, namely (8.42) for EXTRA and the recursions in Problem 8.3. We repeat both recursions here for reference:

$$\begin{aligned}
\mathcal{w}_i &= \mathcal{A}^\mathsf{T}\, \mathcal{w}_{i-1} - \mu \nabla \mathcal{J}(\mathcal{w}_{i-1}) - \mu\eta \mathcal{B}^\mathsf{T} \lambda_{i-1} \\
\mathcal{w}_i &= \mathcal{A}^\mathsf{T} \left(\mathcal{w}_{i-1} - \mu \nabla \mathcal{J}(\mathcal{w}_{i-1})\right) - \mu\eta \mathcal{B}^\mathsf{T} \lambda_{i-1}
\end{aligned}$$

The two recursions correspond to those of the consensus+innovations and diffusion algorithms respectively, with an additional common correction term $-\mu\eta\mathcal{B}^\mathsf{T}\lambda_{i-1}$ for bias correction. Recall that $\mathcal{B} = \mathcal{B}^\mathsf{T}$ is the square root of $\mathcal{L}$, and hence shares eigenvectors with $\mathcal{L}$, where the eigenvalues are the square roots of the eigenvalues of $\mathcal{L}$. Then:

$$\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{B}^\mathsf{T} = 0$$

and hence:

$$\begin{aligned}
&\frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{w}_i \\
&= \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{A}^\mathsf{T}\, \mathcal{w}_{i-1} - \mu \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \nabla \mathcal{J}(\mathcal{w}_{i-1}) - \mu\eta \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{B}^\mathsf{T}\lambda_{i-1} \\
&= \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{A}^\mathsf{T}\, \mathcal{w}_{i-1} - \mu \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \nabla \mathcal{J}(\mathcal{w}_{i-1}) \\
&\frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{w}_i \\
&= \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{A}^\mathsf{T} \left(\mathcal{w}_{i-1} - \mu \nabla \mathcal{J}(\mathcal{w}_{i-1})\right) - \mu\eta \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{B}^\mathsf{T}\lambda_{i-1} \\
&= \frac{1}{K}\left(\mathbb{1}^\mathsf{T} \otimes I_M\right) \mathcal{A}^\mathsf{T} \left(\mathcal{w}_{i-1} - \mu \nabla \mathcal{J}(\mathcal{w}_{i-1})\right)
\end{aligned}$$

It follows that the network centroids of the EXTRA and Exact diffusion algorithms satisfy the same recursions as those of the consensus+innovations and diffusion algorithms respectively, and hence satisfy (9.31).

For the NEXT and Aug-DGM algorithms, we follow a different argument. Recall recursions (8.58) and (8.68)–(8.69) for NEXT and Aug-DGM respectively:

$$\begin{aligned}
w_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{\ell k} w_{\ell,i-1} - \mu g_{k,i-1} \\
w_{k,i} &= \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \left(w_{\ell,i-1} - \mu g_{\ell,i-1}\right)
\end{aligned}$$

For both, we find for the network centroid:

$$w_{c,i} = \frac{1}{K}\sum_{k=1}^{K} w_{k,i} = \frac{1}{K}\sum_{k=1}^{K} w_{k,i-1} - \frac{\mu}{K}\sum_{k=1}^{K} g_{k,i-1} = w_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^{K} g_{k,i-1}$$

We hence need to evaluate the centroid of the gradient tracking term $g_{k,i-1}$, for which we have recursions (8.59) and (8.70):

$$g_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} g_{\ell,i-1} + \nabla J_k(w_{k,i}) - \nabla J_k(w_{k,i-1})$$
$$g_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \left(g_{\ell,i-1} + \nabla J_k(w_{\ell,i}) - \nabla J_k(w_{\ell,i-1})\right)$$

In both cases, we find for the centroid:

$$\frac{1}{K}\sum_{k=1} g_{k,i} = \frac{1}{K}\sum_{k=1} g_{k,i-1} + \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i}) - \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i-1})$$

Iterating starting at $i = 0$ and telescoping, we have:

$$\frac{1}{K}\sum_{k=1} g_{k,i} = \frac{1}{K}\sum_{k=1}^{K} \nabla J_k(w_{k,i})$$

We then have the result after substitutions.