10 Convergence of Decentralized Primal-Dual Algorithms

In Chapter 9 we noted that penalty-based algorithms exhibit a bias in steady-state, which is a result of the fact that penalty-based algorithms solve only an approximation to the consensus optimization problem:

$$w^o \triangleq \arg\min_{w} \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{10.1}$$

This bias made its way into the performance expression of Theorem 9.1, from which we can infer that:

$$\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_{k,i}\|^2 \le O\left(\frac{\mu\sigma^2}{\nu} + \frac{\mu^2 b}{1-\lambda_2}\right) \tag{10.2}$$

Here, μ denotes the step-size of the algorithm, λ_2 denotes the second-largest eigenvalue of the weight matrix A and ν is the strong-convexity constant of the aggregate objective (10.1). The remaining constants are:

$$\sigma^2 = \frac{1}{K^2}\sum_{k=1}^{K}\left(3\beta_k^2\|w_k^o - w^o\|^2 + \sigma_k^2\right) \tag{10.3}$$

$$\begin{aligned} b &= \frac{4}{1-\lambda_2}\|\mathcal{D}\|^2\|\mathcal{V}^\mathsf{T}\|^2\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2 + \|\mathcal{D}\|^2\|\mathcal{V}^\mathsf{T}\|^2 K^2\sigma^2 \\ &= O\left(\frac{\|\mathcal{D}\|^2\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2}{1-\lambda_2} + \|\mathcal{D}\|^2 K^2\sigma^2\right) \end{aligned} \tag{10.4}$$

A number of different trade-offs are captured in these expressions. The first term $\frac{\mu\sigma^2}{\nu}$ is proportional to the step-size μ and inversely proportional to the "signal-to-noise" ratio $\frac{\nu}{\sigma^2}$. We encountered this exact term as the steady-state-error expression for stochastic gradient descent in Chapter 3, and centralized stochastic gradient descent in Chapter 4. Indeed, after specializing to the homogenous scenario where all local gradient noise profiles and minimizers are the same, we can recover linear performance gain via:

$$\frac{\mu\sigma^2}{\nu} = \frac{\mu\sigma_k^2}{K\nu} \tag{10.5}$$

The second term $\frac{\mu^2 b}{1-\lambda_2}$ is new and a result of our decentralized implementation. It essentially corresponds to the loss in performance we endure since we are implementing out decentralized algorithm over a graph and rely on the local diffusion of estimate rather than central aggregation. For the diffusion algorithm, we have $\mathcal{D} = \Lambda_2$, and hence

$$\frac{\mu^2 b}{1-\lambda_2} = O\left(\frac{\mu^2\lambda_2^2\sum_{k=1}^K \|\nabla J_k(w^o)\|^2}{(1-\lambda_2)^2} + \frac{\mu^2\lambda_2^2 K^2\sigma^2}{1-\lambda_2}\right) \tag{10.6}$$

This entire term is multiplied by μ^2. Assuming all other constants are fixed and finite, this means that as $\mu \to 0$, the bias term $\frac{\mu^2 b}{1-\lambda_2}$ will eventually be dominated by the noise term $\frac{\mu\sigma^2}{\nu}$. Similarly, if the network is very densely connected, this will imply that $\lambda_2 \to 0$, and again the bias term will be dominated by the noise term, since both expression on the right handside of (10.6) are scaled by λ_2^2.

There are, however, important settings where the bias $\frac{\mu^2 b}{1-\lambda_2}$ is non-trivial. One of these is when the network is very sparsely connected, resulting in λ_2 close to one and hence $1-\lambda_2 \to 0$. The fact that the two terms on the right handside of (10.6) are divided by $(1-\lambda_2)^2$ and $1-\lambda_2$ respectively has the potential to significantly amplify the bias term for sparse networks. A second important setting is one where exact gradients are used, or the gradient approximation is of very high quality, resulting in $\sigma^2 \to 0$. In that case the term $\frac{\mu^2\lambda_2^2\sum_{k=1}^K \|\nabla J_k(w^o)\|^2}{(1-\lambda_2)^2}$ will dominate all terms involving the gradient noise variance σ^2 and cause a bottleneck. This insight is consistent with the discussion in Section 8.2.3, which concluded that the consensus+innovations algorithm is unbiased if, and only if, all objectives $J_k(w)$ are minimized at a common minimizer w^o, which implies $\sum_{k=1}^K \|\nabla J_k(w^o)\|^2 = 0$.

This same observation of a bias in penalty-based algorithms was the motivation for introducing primal-dual and gradient-tracking based algorithms. We will now proceed to develop convergence guarantees for this new class of algorithms with the aim of demonstrating removal of the bias term resulting from $\frac{\mu^2\lambda_2^2\sum_{k=1}^K \|\nabla J_k(w^o)\|^2}{(1-\lambda_2)^2}$.

10.1 Unified Formulation

In this section, we describe a unifying and generalized framework that includes *all* the decentralized primal-dual and gradient tracking-based methods derived so far as special cases.

Let $\mathcal{B} \in \mathbb{R}^{KM\times KM}$ and $\mathcal{C} \in \mathbb{R}^{KM\times KM}$ denote two general *symmetric* matrices that satisfy the following conditions:

$$\begin{cases} \mathcal{B}\,\mathcal{w} = 0 \Longleftrightarrow w_1 = w_2 \dots, w_K \\ \mathcal{C}\,\mathcal{w} = 0 \Longleftrightarrow \mathcal{B}\,\mathcal{w} = 0 \text{ or } \mathcal{C} = 0 \\ \mathcal{C} \text{ is positive semi-definite} \end{cases} \tag{10.7}$$

For example, $\mathcal{C} = \mathcal{L} = I_{KM} - \mathcal{A}$ and $\mathcal{B} = \mathcal{L}^{1/2}$ is one choice (as in Section 8.3.1), but many other choices are possible including beyond what we have encountered so far. We will provide more examples in the sequel. Let also

$$\bar{\mathcal{A}} = \bar{A} \times I_M \tag{10.8}$$

where $\bar{A}$ is some symmetric doubly-stochastic matrix. For example, $\bar{A} = \frac{1}{2}(I_K + A)$ is one possibility. Assuming the matrices $\{\bar{\mathcal{A}}, \mathcal{B}, \mathcal{C}\}$ have been chosen, we can then reformulate problem (10.1) in the equivalent form

$$w^\star \triangleq \underset{\mathcal{W}\in\mathbb{R}^{KM}}{\text{argmin}} \left\{ \mathcal{J}(\mathcal{W}) + \frac{1}{2\mu}\| \mathcal{W} \|^2_{\mathcal{C}} \right\}, \quad \text{subject to } \mathcal{B}\,\mathcal{W} = 0 \tag{10.9}$$

and introduce the corresponding saddle-point formulation

$$\min_{\mathcal{W}} \max_{\lambda} \; \mathcal{J}(\mathcal{W}) \; + \; \frac{1}{2\mu}\| \mathcal{W} \|^2_{\mathcal{C}} \; + \; \frac{1}{\mu}\lambda^{\mathsf{T}}\mathcal{B}\,\mathcal{W} \tag{10.10}$$

where $\lambda \in \mathbb{R}^{KM}$ is a Lagrangian factor and $\mu > 0$. To solve the above problem, we introduce the following *unified decentralized algorithm* (UDA), which consists of three successive steps (primal-descent, dual-ascent, and combination):

$$\begin{aligned}
\boldsymbol{z}_i &= (I_{KM} - \mathcal{C})\,\boldsymbol{\mathcal{W}}_{i-1} - \mu\,\widehat{\nabla\,\mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) - \mathcal{B}\boldsymbol{\lambda}_{i-1} && (10.11a)\\
\boldsymbol{\lambda}_i &= \boldsymbol{\lambda}_{i-1} + \mathcal{B}\,\boldsymbol{z}_i && (10.11b)\\
\boldsymbol{\mathcal{W}}_i &= \bar{\mathcal{A}}\,\boldsymbol{z}_i && (10.11c)
\end{aligned}$$

The first step is a primal-descent step over $\mathcal{W}$ applied to the Lagrangian function. The result is denoted by the intermediate variable $\boldsymbol{z}_i$. The second equation is a dual-ascent step over λ; it uses the updated iterate $\boldsymbol{z}_i$ instead of $\boldsymbol{\mathcal{W}}_{i-1}$ as benefits from an incremental implementation. The last equation represents a combination step.

If desired, we can eliminate the dual variable from the above equations. Indeed, note that over two successive time instants we get

$$\begin{aligned}
\boldsymbol{z}_i - \boldsymbol{z}_{i-1} = (I_{KM} - \mathcal{C})(\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{i-2}) - \\
\mu\left(\widehat{\nabla\mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) - \widehat{\nabla\mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-2})\right) - \mathcal{B}^2\,\boldsymbol{z}_{i-1}
\end{aligned} \tag{10.12}$$

or, after rearrangements,

$$\begin{aligned}
\boldsymbol{z}_i &= (I_{KM} - \mathcal{B}^2)\,\boldsymbol{z}_{i-1} + (I_{KM} - \mathcal{C})(\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{i-2}) - \\
&\quad \mu\left(\widehat{\nabla\mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-1}) - \widehat{\nabla\mathcal{J}}(\boldsymbol{\mathcal{W}}_{i-2})\right) && (10.13a)\\
\boldsymbol{\mathcal{W}}_i &= \bar{\mathcal{A}}\,\boldsymbol{z}_i && (10.13b)
\end{aligned}$$

with initial condition

$$\boldsymbol{z}_0 = (I_{KM} - \mathcal{C})\,\boldsymbol{\mathcal{W}}_{-1} - \mu\,\widehat{\nabla\,\mathcal{J}}(\boldsymbol{\mathcal{W}}_{-1}), \quad \boldsymbol{\mathcal{W}}_0 = \bar{\mathcal{A}}\,\boldsymbol{z}_0 \tag{10.13c}$$

for any $\boldsymbol{\mathcal{W}}_{-1}$. Now, it is straightforward to see that recursions (10.13a)–(10.13b) reduce to the various decentralized algorithms presented in the earlier chapters for different choices of the triplet $\{\bar{\mathcal{A}}, \mathcal{B}, \mathcal{C}\}$. This is illustrated in Table 10.1. Obviously, many other possibilities can be considered. Observe that EXTRA and

NEXT employ $\mathcal{A} = I_{KM}$.

Table 10.1 Obtaining several decentralized methods as special cases of the unified decentralized algorithm (UDA) described by (10.13a)–(10.13b). Following the convention of Chapter 8, we define $A = I - L$ in terms of the Laplacian matrix L.

Algorithm	$\bar{\mathcal{A}}$	$\mathcal{B}$	$\mathcal{C}$
EXTRA (8.46)	I_{KM}	$\mathcal{L}^{1/2}$	$\mathcal{L} = I_{KM} - \mathcal{A}$
EXACT diffusion (8.65)–(8.67)	$\mathcal{A} = I_{KM} - \mathcal{L}$	$\mathcal{L}^{1/2}$	0
Gradient-tracking (NEXT) (8.58)–(8.59)	I_{KM}	$\mathcal{L} = I_{KM} - \mathcal{A}$	$I_{KM} - \mathcal{A}^2$
Aug-DGM (8.68)–(8.70)	$\mathcal{A}^2$	$\mathcal{L} = I_{KM} - \mathcal{A}$	0

Remark 10.1 (**Consensus+innovations and diffusion strategies**). The consensus+innovations and diffusion strategies were shown earlier to correspond to penalty-based methods. They do not fit into the unified decentralized formulation of this section, which is specific to primal-dual methods. Nevertheless, it can still be seen from recursions (10.11a)–(10.11c) for the unified decentralized algorithm, that we can recover the consensus+innovations recursion (8.18) by setting $\mathcal{B} = 0$, $\mathcal{C} = I_{KM} - \mathcal{A}^{\mathsf{T}}$ and $\bar{\mathcal{A}} = I_{KM}$ and the ATC diffusion strategy (8.60)–(8.61) by setting $\mathcal{B} = 0$, $\mathcal{C} = 0$, and $\bar{\mathcal{A}} = \mathcal{A}^{\mathsf{T}}$. These choices do not satisfy conditions (10.7).

□

10.2 Convergence Analysis

The argument essentially mirrors the one for penalty-based methods in Chapter 9. We will again decompose the network recursion into a recursion for the network centroid, and a second coupled recursion for the network deviation. The additional technical challenge now will be to account for the presence and impact of the dual variable $\boldsymbol{\lambda}_i$, which will complicate the recursions but ultimately be instrumental in allowing for bias correction. We can combine (10.11a) and (10.11c)

$$\mathbf{w}_i = \bar{\mathcal{A}}(I_{KM} - \mathcal{C})\,\mathbf{w}_{i-1} - \mu\bar{\mathcal{A}}\widehat{\nabla \mathcal{J}}(\mathbf{w}_{i-1}) - \bar{\mathcal{A}}\mathcal{B}\boldsymbol{\lambda}_{i-1} \tag{10.14}$$

Note that for all choices of $\bar{\mathcal{A}}, \mathcal{B}, \mathcal{C}$ in Table 10.1, we have:

$$\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\bar{\mathcal{A}} = \mathbb{1}^{\mathsf{T}} \otimes I_M \tag{10.15}$$

$$\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\mathcal{B} = 0 \tag{10.16}$$

$$\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\left(I_{KM} - \mathcal{C}\right) = \mathbb{1}^{\mathsf{T}} \otimes I_M \tag{10.17}$$

We can then conclude:

$$\begin{aligned}
&\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathcal{w}_i \\
=&\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}}(I_{KM} - \mathcal{C})\, \mathcal{w}_{i-1} - \mu \left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) - \left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}} \mathcal{B} \boldsymbol{\lambda}_{i-1} \\
=&\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathcal{w}_{i-1} - \mu \left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1})
\end{aligned} \tag{10.18}$$

Hence:

$$\boldsymbol{w}_{c,i} \triangleq \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{w}_{k,i} = \boldsymbol{w}_{c,i-1} - \frac{\mu}{K} \sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{10.19}$$

We conclude that the network centroid for all primal-dual and gradient tracking-based algorithms evolve according to an approximate centralized stochastic gradient recursion, provided that the matrices $\bar{\mathcal{A}}, \mathcal{B}, \mathcal{C}$ satisfy conditions (10.15)–(10.17). We established this fact already in the problems of Chapter 8 for individual instances of the uniform decentralized formulation. We will now need to bound the deviation of $\boldsymbol{w}_{k,i-1}$ from $\boldsymbol{w}_{c,i-1}$.

To this end, note that the network deviation can be written as:

$$\mathcal{w}_i - \mathbb{1} \otimes \boldsymbol{w}_{c,i} = \mathcal{w}_i - \left(\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathcal{w}_i = \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathcal{w}_i \tag{10.20}$$

Applying this linear transformation to (10.14), we have:

$$\begin{aligned}
&\mathcal{w}_i - \mathbb{1} \otimes \boldsymbol{w}_{c,i} \\
=& \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}}(I_{KM} - \mathcal{C})\, \mathcal{w}_{i-1} \\
&- \mu \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}} \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) \\
&- \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}} \mathcal{B} \boldsymbol{\lambda}_{i-1} \\
\overset{(a)}{=}& \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}}(I_{KM} - \mathcal{C}) \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \mathcal{w}_{i-1} \\
&- \mu \left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) - \bar{\mathcal{A}} \mathcal{B} \boldsymbol{\lambda}_{i-1} \\
=& \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}}(I_{KM} - \mathcal{C}) \left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) \\
&- \mu \left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) - \bar{\mathcal{A}} \mathcal{B} \boldsymbol{\lambda}_{i-1} \\
\overset{(b)}{=}& \mathcal{D} \left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) - \mu \left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \widehat{\nabla \mathcal{J}}(\mathcal{w}_{i-1}) - \bar{\mathcal{A}} \mathcal{B} \boldsymbol{\lambda}_{i-1}
\end{aligned} \tag{10.21}$$

where in (a) we again made use of the spectral properties (10.15)–(10.17), and in (b) we defined:

$$\mathcal{D} = \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right) \bar{\mathcal{A}}(I_{KM} - \mathcal{C}) \tag{10.22}$$

We can verify for the choices in Table 10.1 that $\rho(\mathcal{D}) < 1$ and hence the recursion for the disagreement exhibits contractive behavior, except for the driving terms that appear on the right hand-side. When we developed error recursions (9.38) for penalty-based decentralized algorithms in Chapter 9 we encountered similar driving terms.

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{10.23}$$

$$\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) \triangleq \nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1}) \tag{10.24}$$

$$\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\} \tag{10.25}$$

$$\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\left\{\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})\right\} \tag{10.26}$$

With this definition, we can write:

$$\begin{aligned} &\boldsymbol{\mathcal{W}}_i - \mathbb{1} \otimes \boldsymbol{w}_{c,i} \\ &= \mathcal{D}\left(\boldsymbol{\mathcal{W}}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) - \mu\left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\nabla\,\mathcal{J}(\mathcal{W}^o) - \bar{\mathcal{A}}\mathcal{B}\boldsymbol{\lambda}_{i-1} \\ &\quad - \mu\left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1}) - \mu\left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) \end{aligned} \tag{10.27}$$

Previously, in bounding (9.38), we accepted $\nabla\,\mathcal{J}(\mathcal{W}^o)$ and as a non-vanishing term which is the source of the bias of penalty-based algorithms. In (10.27), however, we have an additional dual term $\bar{\mathcal{A}}\mathcal{B}\boldsymbol{\lambda}_{i-1}$. It turns out that we are able to show that this dual term asymptotically leads to cancellation of the bias arising from $\nabla\,\mathcal{J}(\mathcal{W}^o)$, resulting in a vanishing driving term. To make this argument precise, we first recall that the Karush-Kuhn-Tucker (KKT) conditions conditions for the saddle-point problem (10.10) ensure that the optimal solution satisfies:

$$\mu\nabla\mathcal{J}(\mathcal{W}^o) + \mathcal{B}\lambda^o = 0 \tag{10.28}$$

$$\mathcal{B}\,\mathcal{W}^o = 0 \tag{10.29}$$

Condition (10.29) implies in light of (10.7) that $\mathcal{W}^o = \mathbb{1} \otimes w^o$. Relation (10.28) on the other hand implies that:

$$\mu\left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\nabla\mathcal{J}(\mathcal{W}^o) + \left(\mathbb{1}^{\mathsf{T}} \otimes I_M\right)\mathcal{B}\lambda^o = \mu\sum_{k=1}^{K}\nabla J_k(w^o) = 0 \tag{10.30}$$

and hence w^o is indeed an optimal solution to the consensus problem (10.1).

Again appealing to (10.7), relation (10.28) also implies that:

$$\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\nabla\mathcal{J}(w^o)+\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\mathcal{B}\lambda^o$$
$$=\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\nabla\mathcal{J}(w^o)+\bar{\mathcal{A}}\mathcal{B}\lambda^o=0 \tag{10.31}$$

Adding (10.31) to (10.27), we have:

$$\begin{aligned}
&\boldsymbol{\mathcal{W}}_i-\mathbb{1}\otimes\boldsymbol{w}_{c,i}\\
&=\mathcal{D}\left(\boldsymbol{\mathcal{W}}_{i-1}-\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}\right)+\bar{\mathcal{A}}\mathcal{B}\left(\lambda^o-\boldsymbol{\lambda}_{i-1}\right)\\
&\quad-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})\\
&=\mathcal{D}\left(\boldsymbol{\mathcal{W}}_{i-1}-\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}\right)+\bar{\mathcal{A}}\mathcal{B}\widetilde{\boldsymbol{\lambda}}_{i-1}\\
&\quad-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})
\end{aligned} \tag{10.32}$$

where we defined:

$$\widetilde{\boldsymbol{\lambda}}_{i-1}=\lambda^o-\boldsymbol{\lambda}_{i-1} \tag{10.33}$$

We note that the constant driving term arising from $\nabla\mathcal{J}(w^o)$ has cancelled and we are only left with the perturbations terms $\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})$ and $\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})$ resulting from the network disagreement and gradient noise respectively. Putting everything back together, we find the following coupled recursions:

$$\begin{aligned}
\boldsymbol{w}_{c,i}=\boldsymbol{w}_{c,i-1}-\frac{\mu}{K}\sum_{k=1}^{K}\widehat{\nabla J}_k(\boldsymbol{w}_{c,i-1})&\\
-\frac{\mu}{K}\sum_{k=1}^{K}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})-\frac{\mu}{K}\sum_{k=1}^{K}\boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1})&
\end{aligned} \tag{10.34}$$

$$\begin{aligned}
\boldsymbol{\mathcal{W}}_i-\mathbb{1}\otimes\boldsymbol{w}_{c,i}=\mathcal{D}\left(\boldsymbol{\mathcal{W}}_{i-1}-\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}\right)+\bar{\mathcal{A}}\mathcal{B}\widetilde{\boldsymbol{\lambda}}_{i-1}-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{d}_{i-1}(\boldsymbol{\mathcal{W}}_{i-1})&\\
-\mu\left(\bar{\mathcal{A}}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})&
\end{aligned} \tag{10.35}$$

$$\widetilde{\boldsymbol{\lambda}}_i=\widetilde{\boldsymbol{\lambda}}_{i-1}-\mathcal{B}\boldsymbol{z}_i \tag{10.36}$$

The coupling between the first two recursions is very similar to what we observed in Chapter 9. The challenging coupling now is between (10.35) and (10.36). We can reformulate (10.36):

$$\widetilde{\boldsymbol{\lambda}}_i=\widetilde{\boldsymbol{\lambda}}_{i-1}-\mathcal{B}\left(I_{KM}-\frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I_M\right)\boldsymbol{z}_i \tag{10.37}$$

Following the same argument that led to (10.32), we find:

$$\begin{aligned}
&\left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right) \boldsymbol{z}_i \\
=& \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right) (I_{KM} - \mathcal{C}) \left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) + \mathcal{B}\widetilde{\boldsymbol{\lambda}}_{i-1} \\
&- \mu \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right) \boldsymbol{d}_{i-1}(\mathcal{w}_{i-1}) - \mu \left(I_{KM} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right) \boldsymbol{s}_i(\mathcal{w}_{i-1})
\end{aligned} \tag{10.38}$$

We can then expand:

$$\begin{aligned}
&\widetilde{\boldsymbol{\lambda}}_i \\
=& \widetilde{\boldsymbol{\lambda}}_{i-1} - \mathcal{B}(I_{KM} - \mathcal{C}) \left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) - \mathcal{B}^2\widetilde{\boldsymbol{\lambda}}_{i-1} + \mu\mathcal{B}\boldsymbol{d}_{i-1}(\mathcal{w}_{i-1}) + \mu\mathcal{B}\boldsymbol{s}_i(\mathcal{w}_{i-1}) \\
=& \left(I - \mathcal{B}^2\right) \widetilde{\boldsymbol{\lambda}}_{i-1} - \mathcal{B}(I_{KM} - \mathcal{C}) \left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) - \mu\mathcal{B}\boldsymbol{d}_{i-1}(\mathcal{w}_{i-1}) - \mu\mathcal{B}\boldsymbol{s}_i(\mathcal{w}_{i-1})
\end{aligned} \tag{10.39}$$

Ultimately, we arrive at the coupled set of recursions:

$$\begin{aligned}
\boldsymbol{w}_{c,i} &= \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{k=1}^{K} \widehat{\nabla J}_k(\boldsymbol{w}_{c,i-1}) \\
&\quad - \frac{\mu}{K}\sum_{k=1}^{K} \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) - \frac{\mu}{K}\sum_{k=1}^{K} \boldsymbol{d}_{k,i-1}(\boldsymbol{w}_{k,i-1}) && (10.40)\\
\mathcal{w}_i - \mathbb{1} \otimes \boldsymbol{w}_{c,i} &= \mathcal{D}\left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) + \bar{\mathcal{A}}\mathcal{B}\widetilde{\boldsymbol{\lambda}}_{i-1} - \mu\left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right)\boldsymbol{d}_{i-1}(\mathcal{w}_{i-1}) \\
&\quad - \mu\left(\bar{\mathcal{A}} - \frac{1}{K}\mathbb{1}\mathbb{1}^\top \otimes I_M\right)\boldsymbol{s}_i(\mathcal{w}_{i-1}) && (10.41)\\
\widetilde{\boldsymbol{\lambda}}_i &= \left(I - \mathcal{B}^2\right)\widetilde{\boldsymbol{\lambda}}_{i-1} - \mathcal{B}(I_{KM} - \mathcal{C})\left(\mathcal{w}_{i-1} - \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}\right) \\
&\quad - \mu\mathcal{B}\boldsymbol{d}_{i-1}(\mathcal{w}_{i-1}) - \mu\mathcal{B}\boldsymbol{s}_i(\mathcal{w}_{i-1}) && (10.42)
\end{aligned}$$

This system of equalities can be shown to be convergent for all primal-dual and gradient-tracking based algorithms we have encountered so far, using techniques similar to those in Chapter 9. A detailed derivation of the convergence guarantee is beyond the scope of the module. Instead, we will list in the sequel one performance guarantee from the literature and disucss its implications.

10.3 Convergence of the Exact Diffusion Algorithm

THEOREM 10.1 (Mean-square-behavior of the Exact diffusion algorithm [Yuan et al., 2020). [1]] *Suppose all conditions of Lemma 9.1 hold. Then there exists a*

[1] K. Yuan, S. A. Alghunaim, B. Ying and A. H. Sayed, "On the Influence of Bias-Correction on Distributed Stochastic Optimization," in IEEE Transactions on Signal Processing, vol. 68, pp. 4352-4367, 2020, doi: 10.1109/TSP.2020.3008605.

step-size μ that is small enough, so that iterates generated by a stochastic implementation of the Exact diffusion algorithm 8.65–(8.67) converge in the sense mean-square sense and:

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu\sigma^2}{\nu} + \frac{\mu^2\lambda_2^2 K^2\sigma^2}{1-\lambda_2}\right) \tag{10.43}$$

Comparing this expression with (10.6), notably the dependence on gradient noise is generally preserved, while the bias term proportional to the variability $\sum_{k=1}^{K}\|\nabla J_k(w^o)\|^2$ has disappeared. This results in significantly improved performance when $\sigma^2 \to 0$ and the step-sizes are chosen moderately large.