A LEMMAS

Proof of Lemma 3.3:

Proof. This proof can be easily adapted from Moulines and Bach (2011). From the recursive definition of θ_t one has

$$\mathbb{E}\left[||\theta_t - \theta^*||^2\right] \leq \left(1 - 2\eta(\mu - L^2\eta)\right) \cdot \mathbb{E}\left[||\theta_{t-1} - \theta^*||^2\right] + 2G^2\eta^2.$$

This inequality can be recursively applied to obtain the desired result

$$\begin{aligned}\mathbb{E}\left[||\theta_t - \theta^*||^2\right] &\leq \left(1 - 2\eta(\mu - L^2\eta)\right)^t \cdot \mathbb{E}\left[||\theta_0 - \theta^*||^2\right] + 2G^2\eta^2 \sum_{j=0}^{t-1} \left(1 - 2\eta(\mu - L^2\eta)\right)^j \\ &\leq \left(1 - 2\eta(\mu - L^2\eta)\right)^t \cdot \mathbb{E}\left[||\theta_0 - \theta^*||^2\right] + \frac{G^2\eta}{\mu - L^2\eta}\end{aligned}$$

□

This lemma represents the dynamic of SGD with constant learning rate, where the dependence from the starting point vanishes exponentially fast, but there is a term dependent on η that is not vanishing even for large t.

Lemma A.1. *If Assumption 3.4 with $m = 4$ holds, then for any $t, i \in \mathbb{N}$ one has*

$$\mathbb{E}\left[||\theta_{t+i} - \theta_t||^4 \mid \mathcal{F}_t\right] \leq \eta^4 i^4 G^4$$

Proof. For any $j = 1, ..., l$, let x_j be a vector of length n. Applying Cauchy-Schwarz inequality twice, we get

$$\begin{aligned}||\sum_{j=1}^{l} x_j||^4 = ||\sum_{j=1}^{l} x_j||^2 \cdot ||\sum_{j=1}^{l} x_j||^2 &\leq \left(l \cdot \sum_{j=1}^{l} ||x_j||^2\right)^2 \\ &= l^2 \left(\sum_{j=1}^{l} ||x_j||^2\right)^2 \leq l^3 \cdot \sum_{j=1}^{l} ||x_j||^4 \end{aligned} \tag{10}$$

Since

$$\theta_{t+i} = \theta_t - \eta \sum_{j=0}^{i-1} g(\theta_{t+j}, Z_{t+j+1}),$$

then we can use the fact that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for any k, together with Assumption 3.4 and (10), to get that

$$\begin{aligned}\mathbb{E}\left[||\theta_{t+i} - \theta_t||^4 \mid \mathcal{F}_t\right] &= \eta^4 \cdot \mathbb{E}\left[||\sum_{j=0}^{i-1} g(\theta_{t+j}, Z_{t+j+1})||^4 \mid \mathcal{F}_t\right] \\ &\leq \eta^4 i^3 \sum_{j=0}^{i-1} \mathbb{E}\left[||g(\theta_{t+j}, Z_{t+j+1})||^4 \mid \mathcal{F}_t\right] \\ &= \eta^4 i^3 \sum_{j=0}^{i-1} \mathbb{E}\Big[\underbrace{\mathbb{E}\left[||g(\theta_{t+j}, Z_{t+j+1})||^4 \mid \mathcal{F}_{t+j}\right]}_{\leq G^4} \mid \mathcal{F}_t\Big] \\ &\leq \eta^4 i^4 G^4\end{aligned}$$

Note that this is a bound that considers the worst case in which all the noisy gradient updates point in the same direction and are of norm G. □

Remark A.2. *We can obviously use the same bound for the unconditional squared norm, since*

$$\mathbb{E}\left[||\theta_{t+i} - \theta_t||^4\right] = \mathbb{E}\left[\mathbb{E}\left[||\theta_{t+i} - \theta_t||^4 \mid \mathcal{F}_t\right]\right] \leq \eta^4 i^4 G^4.$$

Lemma A.3. *If Assumption 3.2 and 3.4 with $m = 2$ hold, then for any $i = 1, ..., l$ and $k = 1, 2$ we have that*

$$\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] \leq (L||\theta_t - \theta_0|| + L\eta G i)^2$$

Proof. By adding and subtracting $\nabla F(\theta_t)$, and by Lemma A.1, we get.

$$\begin{aligned}
\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] &\leq \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t) + \nabla F(\theta_t) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] \\
&\leq ||\nabla F(\theta_t) - \nabla F(\theta_0)||^2 + \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t)||^2 \mid \mathcal{F}_t\right] \\
&\quad + 2||\nabla F(\theta_t) - \nabla F(\theta_0)|| \cdot \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t)|| \mid \mathcal{F}_t\right] \\
&\leq L^2||\theta_t - \theta_0||^2 + L^2\mathbb{E}\left[||\theta_{t+i}^{(k)} - \theta_t||^2 \mid \mathcal{F}_t\right] + 2L^2||\theta_t - \theta_0|| \cdot \mathbb{E}\left[||\theta_{t+i}^{(k)} - \theta_t|| \mid \mathcal{F}_t\right] \\
&\leq L^2||\theta_t - \theta_0||^2 + L^2\eta^2G^2i^2 + 2L^2||\theta_t - \theta_0||\eta G i \\
&= (L||\theta_t - \theta_0|| + L\eta G i)^2
\end{aligned}$$

□

Remark A.4. *When we consider the unconditional distance of the gradients, we can simply use smoothness and Remark A.2 to get*

$$\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0)||^2\right] \leq L^2\mathbb{E}\left[||\theta_{t+i}^{(k)} - \theta_0||^2\right] \leq L^2\eta^2G^2(t+i)^2$$

which is the same result that we obtain from Lemma A.3 if at the end we bound $||\theta_t - \theta_0||$ with its expectation, and use the fact that $\mathbb{E}\left[||\theta_t - \theta_0||\right] \leq \eta G t$.

Lemma A.5. *If Assumption 3.2 and 3.4 with $m = 2$ hold, then for any $i = 1, ..., l$ and $k = 1, 2$ we have that*

$$i)\ \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)})||^2 \mid \mathcal{F}_t\right] \leq (||\nabla F(\theta_0)|| + L||\theta_t - \theta_0|| + L\eta G i)^2$$

$$ii)\ \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)})||^2 \mid \mathcal{F}_t\right] \leq (L||\theta_t - \theta^*|| + L\eta G i)^2$$

Proof. We add and subtract $\nabla F(\theta_t)$ to the gradient on the left hand side, and apply Lemma A.1.

$$\begin{aligned}
\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)})||^2 \mid \mathcal{F}_t\right] &= \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t) + \nabla F(\theta_t)||^2 \mid \mathcal{F}_t\right] \\
&\leq ||\nabla F(\theta_t)||^2 + \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t)||^2 \mid \mathcal{F}_t\right] \\
&\quad + 2||\nabla F(\theta_t)|| \cdot \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_t)|| \mid \mathcal{F}_t\right] \\
&\leq ||\nabla F(\theta_t)||^2 + L^2\mathbb{E}\left[||\theta_{t+i}^{(k)} - \theta_t||^2 \mid \mathcal{F}_t\right] + 2L||\nabla F(\theta_t)|| \cdot \mathbb{E}\left[||\theta_{t+i}^{(k)} - \theta_t|| \mid \mathcal{F}_t\right] \\
&\leq ||\nabla F(\theta_t)||^2 + L^2\eta^2G^2i^2 + 2||\nabla F(\theta_t)|| \cdot L\eta G i \qquad (11)
\end{aligned}$$

To get part $i)$ we repeat the same trick, this time adding and subtracting $\nabla F(\theta_0)$ to the terms that contain $\nabla F(\theta_t)$.

$$\begin{aligned}
(11) \leq\ & ||\nabla F(\theta_0)||^2 + ||\nabla F(\theta_t) - \nabla F(\theta_0)||^2 + 2||\nabla F(\theta_0)|| \cdot ||\nabla F(\theta_t) - \nabla F(\theta_0)|| \\
&+ L^2\eta^2G^2i^2 + 2||\nabla F(\theta_0)|| \cdot L\eta G i + 2||\nabla F(\theta_t) - \nabla F(\theta_0)|| \cdot L\eta G i \\
\leq\ & ||\nabla F(\theta_0)||^2 + L^2||\theta_t - \theta_0||^2 + 2L||\nabla F(\theta_0)|| \cdot ||\theta_t - \theta_0|| \\
&+ L^2\eta^2G^2i^2 + 2||\nabla F(\theta_0)|| \cdot L\eta G i + 2||\theta_t - \theta_0|| \cdot L^2\eta G i \\
=\ & (||\nabla F(\theta_0)|| + L||\theta_t - \theta_0|| + L\eta G i)^2
\end{aligned}$$

To get part $ii)$, instead, we can add $\nabla f(\theta^*)$ and get

$$\begin{aligned}
(11) &\leq L^2||\theta_t - \theta^*||^2 + L^2\eta^2G^2i^2 + 2||\theta_t - \theta^*|| \cdot L^2\eta G i \\
&= (L||\theta_t - \theta^*|| + L\eta G i)^2
\end{aligned}$$

□

Remark A.6. *For the unconditional squared norm of the gradient we again obtain the same bound as if in Lemma A.5 we were considering $\mathbb{E}[||\theta_t - \theta_0||] \leq \eta G t$ instead of just the argument of the expectation.*

$$\begin{aligned}
\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)})||^2\right] &= \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0) + \nabla F(\theta_0)||^2\right] \\
&\leq ||\nabla F(\theta_0)||^2 + \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0)||^2\right] \\
&\quad + 2||\nabla F(\theta_0)|| \cdot \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(k)}) - \nabla F(\theta_0)||\right] \\
&\leq ||\nabla F(\theta_0)||^2 + L^2\eta^2G^2(t+i)^2 + 2||\nabla F(\theta_0)||L\eta G(t+i) \\
&= (||\nabla F(\theta_0)|| + L\eta G(t+i))^2
\end{aligned}$$

B Proof of Theorem 3.1

Proof. To slightly simplify the notation, we consider only Q_1. For the following windows, the calculations are equal and just involve some more terms, that are negligible if η is small enough. We assume that the Splitting Diagnostic starts after t iterations have already been made. We use the idea that, for a fixed t, if the learning rate is sufficiently small, the SGD iterate θ_t and θ_0 will not be very far apart. In particular we will use η small enough such that $\eta \cdot (t+l)$ is small, making every term of order $O(\eta^k(t+l)^k)$ negligible for $k > 1$. Thanks to the conditional independence of the errors, the expectation of Q_1 can be written only in terms of the true gradients.

$$\begin{aligned}
\mathbb{E}[Q_1] &= \frac{1}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\mathbb{E}\left[\langle g(\theta_{t+i}^{(1)}), g(\theta_{t+j}^{(2)})\rangle\right] \\
&= \frac{1}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\mathbb{E}\left[\langle \nabla F(\theta_{t+i}^{(1)}) + \epsilon(\theta_{t+i}^{(1)}), \nabla F(\theta_{t+j}^{(2)}) + \epsilon(\theta_{t+j}^{(2)})\rangle\right] \\
&= \frac{1}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\mathbb{E}\left[\langle \nabla F(\theta_{t+i}^{(1)}), \nabla F(\theta_{t+j}^{(2)})\rangle\right] \qquad (12)
\end{aligned}$$

We now add and subtract $\nabla F(\theta_0)$, and use L-smoothness and Remark A.2 to provide a lower bound for $\mathbb{E}[Q_1]$. From (12) we get

$$\begin{aligned}
\mathbb{E}[Q_1] = \frac{1}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}&\Big\{\langle \nabla F(\theta_0), \nabla F(\theta_0)\rangle + \mathbb{E}\left[\langle \nabla F(\theta_{t+i}^{(1)}) - \nabla F(\theta_0), \nabla F(\theta_{t+j}^{(2)}) - \nabla F(\theta_0)\rangle\right] \\
&+\mathbb{E}\left[\langle \nabla F(\theta_0), \nabla F(\theta_{t+j}^{(2)}) - \nabla F(\theta_0)\rangle\right] + \mathbb{E}\left[\langle \nabla F(\theta_{t+i}^{(1)}) - \nabla F(\theta_0), \nabla F(\theta_0)\rangle\right]\Big\} \\
\geq ||\nabla F(\theta_0)||^2 &- \frac{1}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\mathbb{E}\left[||\nabla F(\theta_{t+i}^{(1)}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta_{t+j}^{(2)}) - \nabla F(\theta_0)||\right] \\
&- \frac{1}{l}\sum_{j=0}^{l-1}\mathbb{E}\left[||\nabla F(\theta_0)|| \cdot ||\nabla F(\theta_{t+j}^{(2)}) - \nabla F(\theta_0)||\right] - \frac{1}{l}\sum_{i=0}^{l-1}\mathbb{E}\left[||\nabla F(\theta_0)|| \cdot ||\nabla F(\theta_{t+i}^{(1)}) - \nabla F(\theta_0)||\right] \\
\geq ||\nabla F(\theta_0)||^2 &- \frac{L^2}{l^2}\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\sqrt{\mathbb{E}\left[||\theta_{t+i}^{(1)} - \theta_0||^2\right] \cdot \mathbb{E}\left[||\theta_{t+j}^{(2)} - \theta_0||^2\right]} \\
&- \frac{2L}{l}\sum_{i=0}^{l-1}||\nabla F(\theta_0)|| \cdot \mathbb{E}\left[||\theta_{t+i}^{(1)} - \theta_0||\right] \\
\geq ||\nabla F(\theta_0)||^2 &- L^2\eta^2G^2(t+l)^2 - 2L||\nabla F(\theta_0)||\eta G(t+l) \\
= ||\nabla F(\theta_0)||^2 &- 2L||\nabla F(\theta_0)||\eta G(t+l) + O(\eta^2(t+l)^2) \qquad (13)
\end{aligned}$$

Notice that, in the extreme case where $\eta = 0$, we simply have $\mathbb{E}[Q_1] \geq ||\nabla F(\theta_0)||^2$ which is actually an equality, since we would have $\theta_t = \theta_0$ and the noisy gradient at step t would be $g(\theta_0, Z_t)$, whose expectation is just $\nabla F(\theta_0)$. We now expand the second moment, and there are a lot of terms to be considered separately.

$$
\begin{aligned}
l^4 \cdot \mathbb{E}\left[Q_1^2\right] &= \mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} g(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} g(\theta_{t+j}^{(2)}) \right\rangle^2\right] \\
&= \mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \left(\nabla F(\theta_{t+i}^{(1)}) + \epsilon(\theta_{t+i}^{(1)})\right), \sum_{j=0}^{l-1} \left(\nabla F(\theta_{t+j}^{(2)}) + \epsilon(\theta_{t+j}^{(2)})\right) \right\rangle^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle^2\right]}_{I} + \underbrace{\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \epsilon(\theta_{t+j}^{(2)}) \right\rangle^2\right]}_{II} \\
&\quad + \underbrace{\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \epsilon(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle^2\right]}_{III} + \underbrace{\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \epsilon(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \epsilon(\theta_{t+j}^{(2)}) \right\rangle^2\right]}_{IV} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \nabla F(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \epsilon(\theta_{t+k}^{(2)}) \right\rangle\right]}_{V} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \nabla F(\theta_{t+k}^{(2)}) \right\rangle\right]}_{VI} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \epsilon(\theta_{t+k}^{(2)}) \right\rangle\right]}_{VII} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \epsilon(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \nabla F(\theta_{t+k}^{(2)}) \right\rangle\right]}_{VIII} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \epsilon(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \epsilon(\theta_{t+k}^{(2)}) \right\rangle\right]}_{IX} \\
&\quad + \underbrace{2\,\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \epsilon(\theta_{t+i}^{(1)}), \sum_{j=0}^{l-1} \nabla F(\theta_{t+j}^{(2)}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta_{t+h}^{(1)}), \sum_{k=0}^{l-1} \epsilon(\theta_{t+k}^{(2)}) \right\rangle\right]}_{X}
\end{aligned}
$$

In the squared terms I to IV, the errors are independent from the other argument of the dot product, conditional on $\mathcal{F}_t$, since they are evaluated on different threads. However, in the double products (V to X), some errors are used to generate the subsequent values of the SGD iterates on the same thread. This means that we cannot just ignore them, but we instead have to carefully find an upper bound for each one.

- In I we use the Cauchy-Schwarz inequality and Lemma A.5, after exploiting the independence of the two threads conditional on $\mathcal{F}_t$.

$$\begin{aligned}
\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \nabla F(\theta^{(2)}_{t+j}) \right\rangle^2\right] &\leq l^4 \cdot \max_{i,j}\, \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}), \nabla F(\theta^{(2)}_{t+j}) \right\rangle^2\right] \\
&\leq l^4 \cdot \max_{i,j}\, \mathbb{E}\left[\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i})||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j})||^2 \mid \mathcal{F}_t\right]\right] \\
&\leq l^4 \cdot \mathbb{E}\left[(||\nabla F(\theta_0)|| + L||\theta_t - \theta_0|| + L\eta G l)^4\right] \\
&\lesssim l^4 \cdot \mathbb{E}\left[||\nabla F(\theta_0)||^4 + 4L||\nabla F(\theta_0)||^3 \cdot ||\theta_t - \theta_0|| + 4||\nabla F(\theta_0)||^3 \cdot L\eta G l + O(\eta^2 (t+l)^2)\right] \\
&\lesssim l^4 \cdot \left(||\nabla F(\theta_0)||^4 + 4L\eta G||\nabla F(\theta_0)||^3 (t+l) + O(\eta^2(t+l)^2)\right)
\end{aligned}$$

In the first approximate inequality denoted by $\lesssim$, we have included most of the terms of the expansion in the $O(\eta^2(t+l)^2)$, even if technically we could have done it only after taking the expected value. Notice that here it was important to have a bound in Remark A.2 up to the fourth order.

- Terms II and III are equal, since the two threads are identically distributed, and the errors in one thread are a martingale difference sequence independent from the updates in the other thread. We will use the bound for the error norm

$$\mathbb{E}\left[||\epsilon_t||^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\epsilon_t^T \epsilon_t \mid \mathcal{F}_t\right] = \mathbb{E}\left[\text{tr}(\epsilon_t \epsilon_t^T) \mid \mathcal{F}_t\right] \leq d \cdot \sigma_{max} \tag{14}$$

which is a consequence of Assumption 3.3, and condition on $\mathcal{F}_t$ to use independence of the errors. In the last line we use Remark A.6.

$$\begin{aligned}
\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \epsilon^{(2)}_{t+j} \right\rangle^2\right] &= \sum_{j=0}^{l-1} \mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \epsilon^{(2)}_{t+j} \right\rangle^2\right] \\
&\leq l^2 \max_i \sum_{j=0}^{l-1} \mathbb{E}\left[\left|\left|\nabla F(\theta^{(1)}_{t+i})\right|\right|^2 \cdot ||\epsilon^{(2)}_{t+j}||^2\right] \\
&= l^3 \cdot \max_i\, \mathbb{E}\left[\mathbb{E}\left[||\epsilon^{(2)}_t||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[\left|\left|\nabla F(\theta^{(1)}_{t+i})\right|\right|^2 \mid \mathcal{F}_t\right]\right] \\
&\leq l^3 \cdot d\sigma_{max} \cdot \max_i\, \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i})||^2\right] \\
&\lesssim l^3 \cdot d\sigma_{max} \cdot \left(||\nabla f(\theta_0)||^2 + 2||\nabla f(\theta_0)|| L G \eta (t+l) + O(\eta^2(t+l)^2)\right)
\end{aligned}$$

- In IV, we use the conditional independence of the two threads, and the fact that the errors are a martingale difference sequence, to cancel out all the cross products. An upper bound is then

$$\begin{aligned}
\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \epsilon^{(1)}_{t+i}, \sum_{j=0}^{l-1} \epsilon^{(2)}_{t+j} \right\rangle^2\right] &= \sum_{i=0}^{l-1}\sum_{j=0}^{l-1} \mathbb{E}\left[\left\langle \epsilon^{(1)}_{t+i}, \epsilon^{(2)}_{t+j} \right\rangle^2\right] \\
&\leq \sum_{i=0}^{l-1}\sum_{j=0}^{l-1} \mathbb{E}\left[||\epsilon^{(1)}_{t+i}||^2 \cdot ||\epsilon^{(2)}_{t+j}||^2\right] \\
&= \sum_{i=0}^{l-1}\sum_{j=0}^{l-1} \mathbb{E}\left[\mathbb{E}\left[||\epsilon^{(1)}_{t+i}||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon^{(2)}_{t+j}||^2 \mid \mathcal{F}_t\right]\right] \\
&\leq l^2 d^2 \sigma^2_{max}
\end{aligned}$$

Now we start dealing with the double products. The problem here is that these terms are not all null, since the errors are used in the subsequent updates in the same thread, and they are then not independent.

- V and VI are distributed in the same way. We can cancel out some terms using the conditional independence given $\mathcal{F}_t$, and use the conditional version of Cauchy-Schwarz inequality separately on the two threads.

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \nabla F(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \nabla F(\theta^{(1)}_{t+h}), \sum_{k=0}^{l-1} \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad= \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}), \nabla F(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \nabla F(\theta^{(1)}_{t+h}), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad= \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta_0) + \left(\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)\right), \nabla F(\theta_0) + \left(\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)\right) \right\rangle \times \right. \\
&\qquad\qquad\qquad \left. \times \left\langle \nabla F(\theta_0) + \left(\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)\right), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad= \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta_0), \nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0) \right\rangle \cdot \left\langle \nabla F(\theta_0), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\qquad+ \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta_0), \nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0) \right\rangle \cdot \left\langle \nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\qquad+ \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0), \nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0) \right\rangle \cdot \left\langle \nabla F(\theta_0), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\qquad+ \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0), \nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0) \right\rangle \times \right. \\
&\qquad\qquad\qquad \left. \times \left\langle \nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad\le l^2 ||\nabla F(\theta_0)||^2 \sum_{j,k=0}^{l-1} \mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] \\
&\qquad+ l||\nabla F(\theta_0)|| \sum_{j,h,k=0}^{l-1} \mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] \\
&\qquad+ l||\nabla F(\theta_0)|| \sum_{i,j,k=0}^{l-1} \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] \\
&\qquad+ \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \times \right. \\
&\qquad\qquad\qquad \left. \times ||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right]
\end{aligned}$$

We bound the four pieces separately. For the first, we can just apply Cauchy-Schwarz and L-smoothness, together with Remark A.2

$$\begin{aligned}
\mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] &\le L\sqrt{\mathbb{E}\left[||\theta^{(2)}_{t+j} - \theta_0||^2\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(2)}_{t+k})||^2\right]} \\
&\le \sqrt{d\sigma_{max}} \cdot L\eta G(t+l)
\end{aligned}$$

The bound for the second and third term is equal. We use the conditional independence of the two threads and Lemma A.3.

$$\begin{aligned}
&\mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] = \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})|| \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \mid \mathcal{F}_t\right]\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}\left[\sqrt{\mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(2)}_{t+k})||^2 \mid \mathcal{F}_t\right]} \times \right. \\
&\qquad \left. \times\, \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \mid \mathcal{F}_t\right]\right] \\
&\leq \sqrt{d\sigma_{max}} \cdot \mathbb{E}\left[(L||\theta_t - \theta_0|| + L\eta G l)^2\right] \\
&\leq \sqrt{d\sigma_{max}} \cdot L^2 \eta^2 G^2 (t+l)^2
\end{aligned}$$

The last term again makes use of conditional independence and Lemma A.3.

$$\begin{aligned}
&\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] = \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)|| \mid \mathcal{F}_t\right] \times \right. \\
&\qquad \left. \times\, \mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+k})|| \mid \mathcal{F}_t\right]\right] \\
&\quad \leq \mathbb{E}\left[\sqrt{\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+h}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right]} \times \right. \\
&\qquad \left. \times \sqrt{\mathbb{E}\left[||\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(2)}_{t+k})||^2 \mid \mathcal{F}_t\right]}\right] \\
&\quad \leq \sqrt{d\sigma_{max}} \cdot \mathbb{E}\left[(L||\theta_t - \theta_0|| + L\eta G l)^3\right] \\
&\quad \leq \sqrt{d\sigma_{max}} \cdot L^3 \eta^3 G^3 (t+l)^3
\end{aligned}$$

The last inequality follows from the use of Remark A.2 to bound the moments of $||\theta_t - \theta_0||$ up to order three.

- The upper bound for VII and $VIII$ is the same, even if the error terms are in different positions. Again we invoke conditional independence to get rid of the dot products that only contain $\nabla F(\theta_0)$, and subsequently apply Cauchy-Schwarz inequality.

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \nabla F(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \sum_{i=0}^{l-1} \epsilon(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \epsilon(\theta^{(2)}_{t+j}) \right\rangle\right] \\
&\quad = \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta_0) + \left(\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)\right), \nabla F(\theta_0) + \left(\nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0)\right) \right\rangle \times \right. \\
&\qquad\qquad \left. \times \left\langle \epsilon(\theta^{(1)}_{t+h}), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad = \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0), \nabla F(\theta^{(2)}_{t+j}) - \nabla F(\theta_0) \right\rangle \cdot \left\langle \epsilon(\theta^{(1)}_{t+h}), \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad \leq L^2 \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[||\theta^{(1)}_{t+i} - \theta_0|| \cdot ||\theta^{(2)}_{t+j} - \theta_0|| \cdot ||\epsilon(\theta^{(1)}_{t+h})|| \cdot ||\epsilon(\theta^{(2)}_{t+k})||\right] \\
&\quad \leq L^2 \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\mathbb{E}\left[||\theta^{(1)}_{t+i} - \theta_0|| \cdot ||\epsilon(\theta^{(1)}_{t+h})|| \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\theta^{(2)}_{t+j} - \theta_0|| \cdot ||\epsilon(\theta^{(2)}_{t+k})|| \mid \mathcal{F}_t\right]\right] \\
&\quad \leq L^2 \sum_{i,j,h,k=0}^{l-1} \mathbb{E}\left[\sqrt{\mathbb{E}\left[||\theta^{(1)}_{t+i} - \theta_0||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(1)}_{t+h})||^2 \mid \mathcal{F}_t\right]} \times \right. \\
&\qquad \left. \times \sqrt{\mathbb{E}\left[||\theta^{(2)}_{t+j} - \theta_0||^2 \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(2)}_{t+k})||^2 \mid \mathcal{F}_t\right]}\right] \\
&\quad \leq l^4 L^2 \eta^2 G^2 (t+l)^2 d\sigma_{max}
\end{aligned}$$

- Also the upper bounds for IX and X are equal. In the first one, when $k \neq j$ we can condition on $\mathcal{F}^{(1)}_{t+l}$ and $\mathcal{F}^{(2)}_{t+\max\{k,j\}}$ to get that the expectation is null. Then we are only left with a sum on three indexes i, j, h and $k = j$. In the last passage we again condition on the appropriate σ-algebras to bound separately the two threads.

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{i=0}^{l-1} \nabla F(\theta^{(1)}_{t+i}), \sum_{j=0}^{l-1} \epsilon(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \sum_{h=0}^{l-1} \epsilon(\theta^{(1)}_{t+h}), \sum_{k=0}^{l-1} \epsilon(\theta^{(2)}_{t+k}) \right\rangle\right] \\
&\quad = \sum_{i,j,h=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta_0) + \left(\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)\right), \epsilon(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \epsilon(\theta^{(1)}_{t+h}), \epsilon(\theta^{(2)}_{t+j}) \right\rangle\right] \\
&\quad = \sum_{i,j,h=0}^{l-1} \mathbb{E}\left[\left\langle \nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0), \epsilon(\theta^{(2)}_{t+j}) \right\rangle \cdot \left\langle \epsilon(\theta^{(1)}_{t+h}), \epsilon(\theta^{(2)}_{t+j}) \right\rangle\right] \\
&\quad \leq \sum_{i,j,h=0}^{l-1} \mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(2)}_{t+j})||^2 \cdot ||\epsilon(\theta^{(1)}_{t+h})||\right] \\
&\quad \leq \sum_{i,j,h=0}^{l-1} \mathbb{E}\left[\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+i}) - \nabla F(\theta_0)|| \cdot ||\epsilon(\theta^{(1)}_{t+h})|| \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[||\epsilon(\theta^{(2)}_{t+j})||^2 \mid \mathcal{F}_t\right]\right] \\
&\quad \leq l^3 L\eta G(t+l)\left(d\sigma_{max}\right)^{3/2}
\end{aligned}$$

We put together all these upper bounds, leaving in extended form all the terms that are more significant than $O(\eta^2(t+l)^2)$. We get

$$\begin{aligned}
Var\left(Q_1\right) &= \mathbb{E}\left[Q_1^2\right] - \mathbb{E}\left[Q_1\right]^2 \\
&\lesssim \frac{2||\nabla F(\theta_0)||^2 d\sigma_{max}}{l} + \frac{d^2\sigma_{max}^2}{l^2} \\
&\quad + \eta \cdot \left(\frac{4d\sigma_{max}||\nabla F(\theta_0)||LG(t+l)}{l} + \frac{2LG(t+l)(d\sigma_{max})^{3/2}}{l}\right) \\
&\quad + \eta \cdot \left(8LG||\nabla F(\theta_0)||^3(t+l) + 2||\nabla F(\theta_0)||^2 LG(t+l)\sqrt{d\sigma_{max}}\right) + O(\eta^2(t+l)^2)
\end{aligned}$$

which immediately translates to a bound for the standard deviation of the following form

$$\begin{aligned}
\mathrm{sd}\left(Q_1\right) &\lesssim \frac{||\nabla F(\theta_0)||\sqrt{2d\sigma_{max}}}{\sqrt{l}} + \frac{d\sigma_{max}}{l} \\
&\quad + \sqrt{\eta} \cdot \left(8LG||\nabla F(\theta_0)||^3(t+l) + 2||\nabla F(\theta_0)||^2 LG(t+l)\sqrt{d\sigma_{max}}\right)^{1/2} \\
&\quad + \sqrt{\eta} \cdot \left(\frac{4d\sigma_{max}||\nabla F(\theta_0)||LG(t+l)}{l} + \frac{2LG(t+l)(d\sigma_{max})^{3/2}}{l}\right)^{1/2} + O(\eta(t+l))
\end{aligned} \tag{15}$$

We combine (15) with the fact, consequence of (13), that $\mathbb{E}[Q_1]/||\nabla F(\theta_0)||^2 \gtrsim 1 + O(\eta(t+l))$, to get the desired inequality

$$\mathrm{sd}(Q_1) \lesssim C_1(\eta, l) \cdot \mathbb{E}[Q_1]$$

where

$$\begin{aligned}
C_1(\eta, l) = \frac{1}{||\nabla F(\theta_0)||^2} \cdot \Bigg\{ &\frac{||\nabla F(\theta_0)||\sqrt{2d\sigma_{max}}}{\sqrt{l}} + \frac{d\sigma_{max}}{l} \\
&+ \sqrt{\eta} \cdot \left(8LG||\nabla F(\theta_0)||^3(t+l) + 2||\nabla F(\theta_0)||^2 LG(t+l)\sqrt{d\sigma_{max}}\right)^{1/2} \\
&+ \sqrt{\eta} \cdot \left(\frac{4d\sigma_{max}||\nabla F(\theta_0)||LG(t+l)}{l} + \frac{2LG(t+l)(d\sigma_{max})^{3/2}}{l}\right)^{1/2} \Bigg\}
\end{aligned}$$

This confirms that $C_1(\eta, l) = O(1/\sqrt{l}) + O(\sqrt{\eta(t+l)})$. □

C Proof of Theorem 3.2

Proof. As before, we only consider Q_1 for simplicity. To provide an upper bound for $|\mathbb{E}[Q_1]|$, we use the fact that $\nabla F(\theta^*) = 0$ together with Assumption 3.2. Starting from (12) we have

$$\begin{aligned}
|\mathbb{E}[Q_1]| &= \frac{1}{l^2}\left|\sum_{j=0}^{l-1}\sum_{k=0}^{l-1}\mathbb{E}\left[\langle\nabla F(\theta^{(1)}_{t+j}), \nabla F(\theta^{(2)}_{t+k})\rangle\right]\right| \\
&\le \frac{1}{l^2}\sum_{j=0}^{l-1}\sum_{k=0}^{l-1}\mathbb{E}\left[||\nabla F(\theta^{(1)}_{t+j}) - \nabla F(\theta^*)|| \cdot ||\nabla F(\theta^{(2)}_{t+k}) - \nabla F(\theta^*)||\right] \\
&\le \frac{L^2}{l^2}\sum_{j=0}^{l-1}\sum_{k=0}^{l-1}\mathbb{E}\left[||\theta^{(1)}_{t+j} - \theta^*|| \cdot ||\theta^{(2)}_{t+k} - \theta^*||\right] \\
&\le \frac{L^2}{l^2}\sum_{j=0}^{l-1}\sum_{k=0}^{l-1}\sqrt{\mathbb{E}\left[||\theta^{(1)}_{t+j} - \theta^*||^2\right] \cdot \mathbb{E}\left[||\theta^{(2)}_{t+k} - \theta^*||^2\right]}
\end{aligned}$$

Now we can use Lemma 3.3 that states that, for $\eta \le \frac{\mu}{L^2}$,

$$\mathbb{E}\left[||\theta_t - \theta^*||^2\right] \le \left(1 - 2\eta(\mu - L^2\eta)\right)^t \cdot \mathbb{E}\left[||\theta_0 - \theta^*||^2\right] + \frac{G^2\eta}{\mu - L^2\eta}. \tag{16}$$

As $t \to \infty$ we have that $\mathbb{E}\left[||\theta_t - \theta^*||^2\right] \lesssim \frac{G^2\eta}{\mu - L^2\eta}$. L-smoothness combined with (16) also gets

$$\mathbb{E}\left[||\nabla F(\theta_t)||^2\right] \lesssim \frac{L^2G^2\eta}{\mu - L^2\eta} \qquad \text{as } t \to \infty. \tag{17}$$

Since the first term of (16) is decreasing in t, our bound on the expectation of Q_1 is

$$|\mathbb{E}[Q_1]| \le L^2 \cdot \left(\left(1 - 2\eta(\mu - L^2\eta)\right)^t \cdot \mathbb{E}\left[||\theta_0 - \theta^*||^2\right] + \frac{G^2\eta}{\mu - L^2\eta}\right) \tag{18}$$

To deal with the second moment, we introduce the notation

$$S_k := \sum_{i=0}^{l-1} g(\theta^{(k)}_{t+i}, Z^{(k)}_{t+i+1}) = \sum_{i=0}^{l-1}\nabla F(\theta^{(k)}_{t+i}) + \sum_{i=0}^{l-1}\epsilon(\theta^{(k)}_{t+i}) =: G_k + e_k.$$

where G_k is the true signal in the first window of thread k and e_k the related noise. Conditional on $\mathcal{F}_t$, the random variables S_1 and S_2 are independent and identically distributed. Then we can write

$$\begin{aligned}
l^4 \cdot \mathbb{E}[Q_1^2] &= \mathbb{E}\left[\langle S_1, S_2\rangle^2\right] = \mathbb{E}\left[S_2^T S_1 S_1^T S_2\right] \\
&= \mathbb{E}\left[\mathrm{Tr}(S_2^T S_1 S_1^T S_2)\right] = \mathbb{E}\left[\mathrm{Tr}(S_1 S_1^T S_2 S_2^T)\right] \\
&= \mathrm{Tr}\left(\mathbb{E}\left[S_1 S_1^T S_2 S_2^T\right]\right) = \mathrm{Tr}\left(\mathbb{E}\{\mathbb{E}\left[S_1 S_1^T \mid \mathcal{F}_t\right] \cdot \mathbb{E}\left[S_2 S_2^T \mid \mathcal{F}_t\right]\}\right) \\
&= \mathrm{Tr}\left(\mathbb{E}\{\mathbb{E}\left[S_1 S_1^T \mid \mathcal{F}_t\right]^2\}\right)
\end{aligned}$$

The goal is now to show that the matrix $\mathbb{E}\left[S_1 S_1^T | \mathcal{F}_t\right]$ is positive definite, and provide a lower bound for its second moment using the fact that if $A \succeq \lambda I$ for $\lambda \ge 0$, then $A^2 \succeq \lambda^2 I$. We can write

$$\begin{aligned}
\mathbb{E}\left[S_1 S_1^T | \mathcal{F}_t\right] &= \mathbb{E}\left[(G_1 + e_1)(G_1 + e_1)^T | \mathcal{F}_t\right] \\
&= \mathbb{E}\left[G_1 G_1^T | \mathcal{F}_t\right] + \mathbb{E}\left[G_1 e_1^T | \mathcal{F}_t\right] + \mathbb{E}\left[e_1 G_1^T | \mathcal{F}_t\right] + \mathbb{E}\left[e_1 e_1^T | \mathcal{F}_t\right]
\end{aligned}$$

We immediately have that $\mathbb{E}\left[G_1 G_1^T | \mathcal{F}_t\right] \succeq 0$, because, for any $x \in \mathbb{R}^d$,

$$x^T \mathbb{E}\left[G_1 G_1^T | \mathcal{F}_t\right] x = \mathbb{E}\left[x^T G_1 G_1^T x | \mathcal{F}_t\right] = \mathbb{E}\left[||x^T G_1||^2 | \mathcal{F}_t\right] \ge 0.$$

Moreover we can also find an easy lower bound for the error term using Assumption 3.3,

$$\mathbb{E}\left[e_1 e_1^T | \mathcal{F}_t\right] = \mathbb{E}\left[\left(\sum_{i=0}^{l-1}\epsilon(\theta^{(1)}_{t+i})\right)\left(\sum_{j=0}^{l-1}\epsilon(\theta^{(1)}_{t+j})\right)^T \middle| \mathcal{F}_t\right]$$

$$
\begin{aligned}
&= \sum_{i=0}^{l-1} \mathbb{E}\left\{ \mathbb{E}\left[\epsilon(\theta_{t+i}^{(1)})\epsilon(\theta_{t+i}^{(1)})^T \; | \mathcal{F}_{t+i-1} \right] \right\} \\
&\succeq l \cdot \sigma_{min} \cdot I
\end{aligned}
$$

To lower bound the remaining terms we introduce a simple Lemma.

Lemma C.1. *If* $u, v \in \mathbb{R}^d$*, then* $uv^T + vu^T \succeq -2||u|| \cdot ||v|| \cdot I$

Proof. We apply the Cauchy-Schwarz inequality and get, for any $x \in \mathbb{R}^d$,

$$
\begin{aligned}
x^T(uv^T + vu^T + 2||u|| \cdot ||v|| \cdot I)x &= x^T uv^T x + x^T vu^T x + 2||u|| \cdot ||v|| \cdot x^T x \\
&= \langle x, u\rangle\langle v, x\rangle + \langle x, v\rangle\langle u, x\rangle + 2||u|| \cdot ||v|| \cdot ||x||^2 \geq 0
\end{aligned}
$$

□

Using Lemma C.1, and Lemma A.5 *ii*) in the last inequality, we immediately get that

$$
\begin{aligned}
\mathbb{E}\left[G_1 e_1^T \; |\mathcal{F}_t\right] + \mathbb{E}\left[e_1 G_1^T \; |\mathcal{F}_t\right] &\succeq -2\mathbb{E}\left[||G_1|| \cdot ||e_1|| \; |\mathcal{F}_t\right] \cdot I \\
&\succeq -2 \sum_{i=1}^{l} \sum_{j=1}^{l} \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(1)})|| \cdot ||\epsilon(\theta_{t+j}^{(1)})|| \; |\mathcal{F}_t \right] \cdot I \\
&\succeq -2 \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} \sqrt{ \mathbb{E}\left[||\nabla F(\theta_{t+i}^{(1)})||^2 \; |\mathcal{F}_t \right] \cdot \mathbb{E}\left[||\epsilon(\theta_{t+j}^{(1)})||^2 \; |\mathcal{F}_t \right] } \cdot I \\
&\succeq -2l^2 \cdot \sqrt{d\sigma_{max}} \cdot (L||\theta_t - \theta^*|| + L\eta G l) \cdot I
\end{aligned}
$$

Notice that we could improve the bound using the fact that $\epsilon(\theta_{t+j}^{(1)})$ is independent from $\nabla F(\theta_{t+i}^{(1)})$ for any $j \geq i$. Putting the pieces together we get that

$$
\begin{aligned}
&\mathbb{E}\left[S_1 S_1^T \; |\mathcal{F}_t\right] \succeq \left(l\sigma_{min} - 2l^2 \cdot \sqrt{d\sigma_{max}} \cdot (L||\theta_t - \theta^*|| + L\eta G l) \right) \cdot I \\
\Rightarrow \quad &\mathbb{E}\left[S_1 S_1^T \; |\mathcal{F}_t\right]^2 \succeq \left(l\sigma_{min} - 2l^2 \cdot \sqrt{d\sigma_{max}} \cdot (L||\theta_t - \theta^*|| + L\eta G l) \right)^2 \cdot I \\
&\succeq \Big\{ l^2\sigma_{min}^2 + 4l^4 d\sigma_{max} \cdot (L||\theta_t - \theta^*|| + L\eta G l)^2 \\
&\quad - 4l^3 \sigma_{min} \sqrt{d\sigma_{max}} \cdot (L||\theta_t - \theta^*|| + L\eta G l) \Big\} \cdot I
\end{aligned}
$$

and then, using the asymptotic bound in (16),

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[S_1 S_1^T \; |\mathcal{F}_t\right]^2 \right] &\succeq \left\{ l^2\sigma_{min}^2 - 4l^3 \sigma_{min} \sqrt{d\sigma_{max}} \cdot (L \cdot \mathbb{E}\left[||\theta_t - \theta^*||\right] + L\eta G l) \right\} \cdot I \\
&\overset{t\to\infty}{\succeq} \left\{ l^2\sigma_{min}^2 - 4l^3 \sigma_{min} \sqrt{d\sigma_{max}} \cdot \left(\frac{LG\sqrt{\eta}}{\sqrt{\mu - L^2\eta}} + L\eta G l \right) \right\} \cdot I
\end{aligned}
$$

which finally gives the bound on the second moment, which is

$$
\begin{aligned}
l^4 \cdot \mathbb{E}[Q_1^2] &\gtrsim d \cdot \left(l^2\sigma_{min}^2 - 4l^3 \sigma_{min} \sqrt{d\sigma_{max}} LG\sqrt{\eta} \cdot \left(\frac{1}{\sqrt{\mu - L^2\eta}} + l\sqrt{\eta} \right) \right) \\
&\geq dl^2\sigma_{min}^2 - K_1 l^3 \sqrt{\eta} - K_2 l^4 \eta
\end{aligned}
$$

Using the fact shown before, that

$$
\mathbb{E}[Q_1]^2 \lesssim \frac{L^4 G^4 \eta^2}{(\mu - L^2\eta)^2} \qquad \text{as } \; t \to \infty,
$$

we can bound the variance of Q_1 from below with

$$Var(Q_1) = \mathbb{E}[Q_1^2] - \mathbb{E}[Q_1]^2 \geq \frac{d\sigma_{min}^2}{l^2} - \frac{K_1\sqrt{\eta}}{l} - K_2\eta - \frac{L^4G^4\eta^2}{(\mu - L^2\eta)^2}$$

and then

$$Var(Q_1) \gtrsim \left(\frac{d\sigma_{min}^2}{l^2} - \frac{K_1\sqrt{\eta}}{l} + O(\eta)\right) \cdot \frac{\mathbb{E}[Q_1]^2(\mu - L^2\eta)^2}{L^4G^4\eta^2}.$$

The desired inequality is finally

$$|\mathbb{E}[Q_1]| \lesssim C_2(\eta) \cdot \mathrm{sd}(Q_1)$$

with

$$C_2(\eta) = \frac{L^2G^2\eta}{(\mu - L^2\eta)} \cdot \left(\frac{d\sigma_{min}^2}{l^2} - \frac{K_1\sqrt{\eta}}{l} + O(\eta)\right)^{-1/2} = C_2 \cdot \eta + o(\eta).$$

□

D Proof of Proposition 3.4

Proof. We first notice that the averaging at the end of each diagnostic can be ignored, and replaced by simply considering each diagnostic as a single thread made of wl iterates. For the first diagnostic, for example, we have that

$$\begin{aligned}
\mathbb{E}\big[\|\theta_{D_1} - \theta^*\|^2\big] &\leq \mathbb{E}\left[\left\|\frac{\theta^{(1)}_{t_1+wl} + \theta^{(2)}_{t_1+wl} - 2\theta^*}{2}\right\|^2\right] \\
&= \frac{1}{4}\left(\mathbb{E}\big[\|\theta^{(1)}_{t_1+wl} - \theta^*\|^2\big] + \mathbb{E}\big[\|\theta^{(2)}_{t_1+wl} - \theta^*\|^2\big] + 2 \cdot \mathbb{E}\left[\langle\theta^{(1)}_{t_1+wl} - \theta^*, \theta^{(2)}_{t_1+wl} - \theta^*\rangle\right]\right) \\
&\leq \mathbb{E}\big[\|\theta^{(1)}_{t_1+wl} - \theta^*\|^2\big]
\end{aligned}$$

where we have used the fact that each thread is identically distributed, together with the Cauchy-Schwarz inequality. The same inequality, with appropriate indexes, is true for all the diagnostics.

Our proof is now divided in two parts. First we show that, in the extreme case where each diagnostic detects stationarity deterministically, the learning rate does not decay too fast and we still have convergence to θ^*. Then we prove that eventually the learning rate decreases to zero when the number of diagnostics goes to infinity. We initially notice that

$$\left(1 - 2\eta\gamma^b(\mu - L^2\eta\gamma^b)\right)^{t_1/\gamma^b} \leq e^{-2\eta(\mu - L^2\eta)t_1} =: c_1$$

where $c_1 \in (0, 1)$. We also have

$$\frac{G^2\eta\gamma^b}{\mu - L^2\eta\gamma^b} \leq \frac{G^2\eta\gamma^b}{\mu - L^2\eta} =: c_2 \cdot \gamma^b$$

We define L_b to be the expected square distance from the minimizer, $\mathbb{E}\big[\|\theta_{D_b} - \theta^*\|^2\big]$, at the end of the bth diagnostic, and $L_0 = \mathbb{E}\big[\|\theta_0 - \theta^*\|^2\big]$. If the learning rate decreases deterministically, then we have that after the bth diagnostic, the learning rate is $\eta\gamma^b$ and the length of the single thread is $\lfloor t_1/\gamma^b \rfloor$. By recursion, using Lemma 3.3 in the main text, we have that

$$\begin{aligned}
L_{b+1} &\leq \left(1 - 2\eta\gamma^b(\mu - L^2\eta\gamma^b)\right)^{t_1/\gamma^b} \cdot L_b + \frac{G^2\eta\gamma^b}{\mu - L^2\eta\gamma^b} \\
&\leq c_1 \cdot L_b + c_2 \cdot \gamma^b \\
&\leq c_1^{b+1} \cdot L_0 + c_2 \cdot \sum_{i=0}^{b} \gamma^{b-i} c_1^i \\
&\leq c_1^{b+1} \cdot L_0 + c_2 \cdot b \cdot \max\{\gamma, c_1\}^b
\end{aligned}$$

Since $\gamma, c_1 \in (0, 1)$, this proves that $L_b \to 0$ as the number of diagnostics $b \to \infty$.

To prove that it is impossible for the learning rate to remain fixed on a certain value for infinite many iterations, we show that the probability that the learning rate reaches a point where it never decreases is zero. We assume by contradiction that there exists a point in the SplitSGD procedure where the learning rate is η^* and, from that moment on, it is never reduced again. Following Dieuleveut et al. (2017), we know that the Markov chain $\{\theta_t\}$ defined as (2 in the main text) with constant learning rate η^* will converge in distribution to its stationary distribution π_{η^*}. This means that

$$\sup_{s,t\geq T} \|\mathbb{E}[\theta_t] - \mathbb{E}[\theta_s]\| \to 0 \qquad \text{as } T \to \infty \tag{19}$$

and if we let $s = t+1$ we realise that $\|\mathbb{E}[g(\theta_t, Z_{t+1})]\| \to 0$ as $t \to \infty$. Notice that also the Markov chain $\{g(\theta_t, Z_{t+1})\}$ converges to a stationarity distribution when $\{\theta_t\}$ does, so we can use the Central Limit Theorem for Markov chains (Maxwell and Woodroofe, 2000) to get that

$$\frac{1}{\sqrt{l}} \sum_{j=1}^{l} g(\theta_{t+j}, Z_{t+j+1}) \xrightarrow{d} N(0, \sigma^2) \qquad \text{as } l \to \infty \tag{20}$$

where $\sigma^2 > 0$. We are now going to use the fact that $\text{sign}\,(Q_i) = \text{sign}\,(l \cdot Q_i)$. Thanks to (20) we can now write

$$\begin{aligned} l \cdot Q_i &= \left\langle \frac{1}{\sqrt{l}} \sum_{j=1}^{l} g^{(1)}_{t+(i-1)l+j}, \frac{1}{\sqrt{l}} \sum_{j=1}^{l} g^{(2)}_{t+(i-1)l+k} \right\rangle \\ &= \langle X_1 + o_p(1), X_2 + o_p(1) \rangle \\ &= \langle X_1, X_2 \rangle + o_p(1) \end{aligned}$$

where X_1, X_2 are independent $N(0, \sigma^2)$ (the independence being true for $l \to \infty$ and $i = 2, ..., w$) and the $o_p(1)$ are defined as $l \to \infty$. Since $l \cdot Q_i$ is approximately distributed as $\langle X_1, X_2 \rangle$, which has mean zero and positive variance, then for any choice of $q < (w-1)/w$ we know that there is a positive probability $\alpha > 0$ that the proportion of negative gradient coherences observed is greater than q, which means that stationarity is detected. The probability that the learning rate η^* never decays is then bounded above by $\lim_{b\to\infty}(1-\alpha)^b = 0$, so the learning rate gets eventually reduced with probability 1. □

E More Comments on the Experiments Section

In this section we discuss some topics that for reasons of space did not fit in the main paper.

E.1 Description of the convex setting and choice of the tolerance parameter q

For the experiments in the convex setting we use a feature matrix $X \in \mathbb{R}^{n\times d}$ with standard normal entries and $n = 1000$, $d = 20$. We set $\theta^*_j = 5 \cdot e^{-j/2}$ for $j = 1, ..., 20$ to guarantee some difference in the entries. We generate the linear data as $y_i = X_i \cdot \theta^* + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$, and the data for logistic regression from a Bernoulli with probability $(1 + e^{-X_i \cdot \theta^*})^{-1}$. The other parameters that are used through all Section 4.1 are the numbers of windows $w = 20$ of size $l = 50$ (so that each diagnostic consists of one epoch), the length of the first single thread $t_1 = 4$ epochs, and the acceptance proportion $q = 0.4$.

As we say in the main text, in general we would like w, l, t_1 and the number of diagnostics B to be as large as possible, given the computational budget that we have. The tolerance q, instead, is more tricky. In Theorem 3.2 and Figure 3 we shown that, as $t_1 \to \infty$, the distribution of the sign of the gradient coherence is approximately a coin flip, provided that η is small enough. This means that, once stationarity is reached, we want q not to be too big, so that we will not observe a proportion of negative gradient coherences smaller than q just by chance too often (and erroneously think that stationarity has not been reached yet). If we were then to assume independence between the Q_i, we should set q to control the probability of a type I error (returning $T_D = N$ even though stationarity has been reached), which is

$$\frac{1}{2^w} \sum_{i=0}^{\lfloor w \cdot q \rfloor - 1} \binom{w}{i}$$

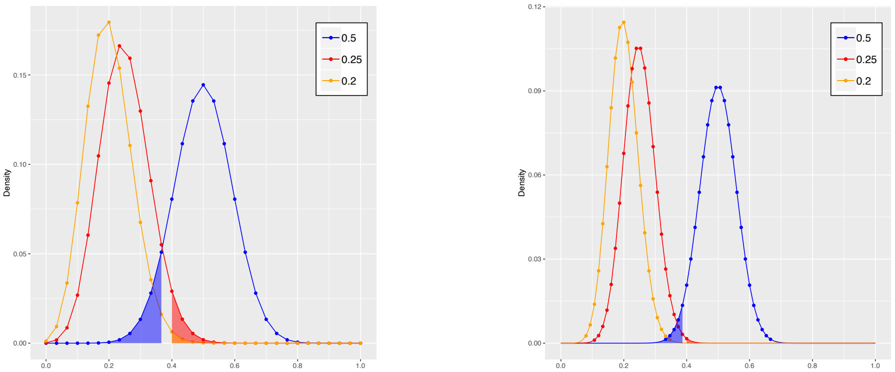

Figure 7: Continuous representation of the probability mass function of Binomial distributions. On the left we set $w = 30$ and $q = 0.4$, on the right $w = 75$ and $q = 0.4$, for both the probability of success (observing a negative gradient coherence) is $p \in \{0.2, 0.25, 0.5\}$. When $p = 0.5$ (stationarity) the type I error happens with probability approximated by the shaded blue region. When $p < 0.5$ (non stationarity) we erroneously declare stationarity with probability approximated by the shaded red and orange region.

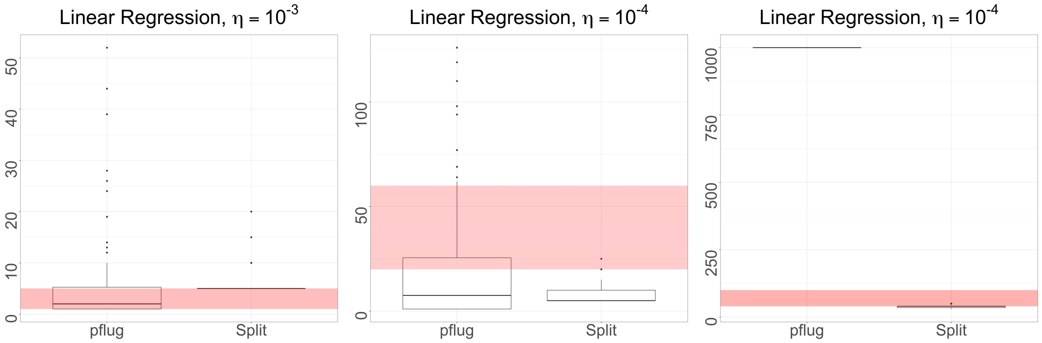

Figure 8: (left) starting around θ^*, large learning rate. (middle) starting around θ^*, small learning rate. (right) starting around θ_s, small learning rate.

However, if we set q to be too small, then in the initial phases of the procedure we might think that we have already reached stationarity only because by chance we observed a proportion of negative dot products larger than q. This trade-off, represented in Figure 7, is particularly relevant if we cannot afford a large number of windows w, but it loses importance as w grows.

E.2 Comparison with `pflug` Diagnostic with different parameters

In Figure 8 and Figure 9 we see other configurations for the experiment reported in the left panels of Figure 4. There, the starting point was set to be around θ_s, where $\theta_{s,j} = 5 \cdot e^{-(d-j)/2}$ for $j = 1, ..., 20$. Here we consider the same starting point for the panels on the right (for both linear and logistic regression) but a smaller learning rate. In both cases it is extremely clear that the `pflug` Diagnostic is detecting stationarity too late, and often (in the case of linear regression) running to the end of the budget. This can be a big problem in practice, because after stationarity has been reached all the iterations that keep using the same learning rate are not going to improve convergence, and are fundamentally wasted. In the left and middle panel of both figures we consider a starting point for the procedures around the minimizer θ^*. In this scenario, for both larger and smaller learning rates, we see that both procedure are either very precise or detect stationarity a bit too early. This is a smaller problem in practice, since at that point the learning rate is reduced but the SGD procedures keep running, even if with a smaller learning rate. The speed of convergence is then slower, but the steps that we make are still important towards convergence.

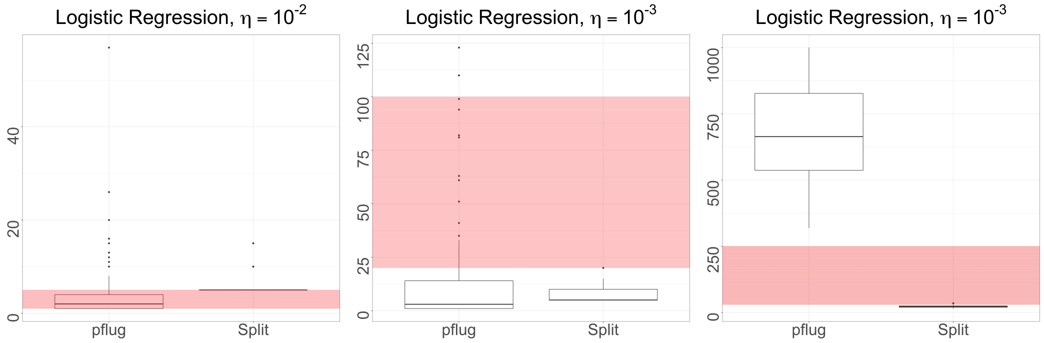

Figure 9: (left) starting around θ^*, large learning rate. (middle) starting around θ^*, small learning rate. (right) starting around θ_s, small learning rate.

E.3 Changes to the SplitSGD procedure in deep learning

The differences between the SplitSGD procedure that we analysed in Section 2 and its adaptation to deep learning are the following:

- **momentum of SGD:** while in the convex setting we study the behavior of vanilla SGD, when training deep neural networks the standard choice is to use SGD with momentum (Sutskever et al., 2013), which updates as

$$\begin{aligned}\Delta\theta_t &= \beta \cdot \Delta\theta_{t-1} - \eta_t \cdot g(\theta_t, Z_{t+1}) \\ \theta_{t+1} &= \theta_t + \Delta\theta_t\end{aligned} \tag{21}$$

  If the learning rate is kept constant, SGD with momentum still goes through a transient phase before reaching stationarity. Even in this case, as we already saw in (3), we have that $\mathbb{E}_{\theta\sim\pi_\eta}[g(\theta, Z)] = 0$ since we see from (21) that $\mathbb{E}_{\theta\sim\pi_\eta}[\Delta\theta] = 0$. This justifies the use of the gradient coherence as defined in (6) also when considering SGD with momentum.

- **gradient coherence on layers:** when considering a parameter space of dimension d, the gradient coherence is a dot product of two d-dimensional vectors. In deep learning, the parameter space is usually extremely large, so we decided to divide these vectors into pieces to try to extract more information about the stationarity of the SGD updates, by computing the dot product of each of the pieces. In practice, let's divide the vectors u and v into p pieces not necessarily of equal length, so that $v = (v_1, v_2, ..., v_p)$ and $u = (u_1, u_2, ..., u_p)$. Instead of computing the single dot product $\langle v, u\rangle$ we store the p dot products $\langle v_i, u_i\rangle$. In this way, we can also relax the trade-off between l and w (remember that we want to allocate a single epoch to the diagnostic, so that $2lw$ is fixed to be the size of the training set). By computing more than a single value of Q for each pair of vectors, we can allow to set w smaller.

  A natural division of the parameter space into smaller pieces comes from the layers of the network, so each time we compute the gradient coherence of the two threads we actually compute a separate value for each layer and then store all of them together. In the final count, as we did in the non-convex setting, we look at the proportion of these values that are negative to decide whether to decay the learning rate.

- **length of the single thread:** since training deep neural networks is usually computationally expensive, we decided not to increase the length of the single thread after stationarity was detected. This is made simply to avoid situations where stationarity is detected early and the length of the single thread increases so fast that we do not have time to decay the learning rate by much before reaching the end of the computational budget that we allocated.

- **hyperparameters for the diagnostic:** we set the relevant hyperparameters w and q to take value $w = 4$ and $q = 0.25$. The value of w is much smaller than the one used in the convex setting for the reason explained above that we compute the gradient coherence separately for each layer. With this choice, we can dedicate $1/8$ of the updates of each epoch to compute for each thread the average of the gradients and be sure that we averaged out a lot of the noise. The choice of setting $q = 0.25$ comes from the empirical results that we observed, and a deeper study of this parameter is probably needed. We performed a sensitivity analysis on these two parameters in Section 4.3 and noticed that a departure from these values is not changing the performance by much.

E.4 Other experiments in deep learning

We add here the description of two more experiments in Deep Learning that did not fit in the main body of the paper, together with the plots that we already included in Figure 5 but this time with the addition of the 90% confidence bands. We see that the Splitting Diagnostic increases the variability of SplitSGD with respect to other methods in some settings, but the interpretation that we gave in Section 4.2 of the better performance of SplitSGD and the lack of overfitting holds.

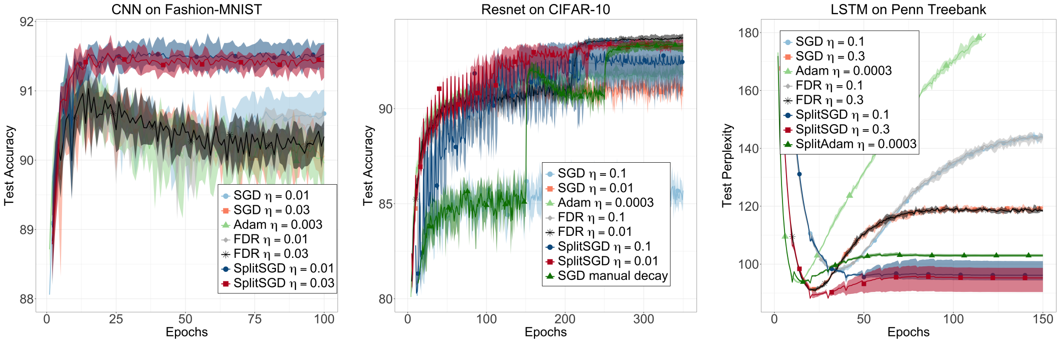

Figure 10: This is the same as Figure 5 but here we also added 90% confidence bands for each method.

Feedforward neural networks (FNNs). We train a FNN with three hidden layers of size $256, 128$ and 64 on the Fashion-MNIST dataset (Xiao et al., 2017). The network is fully connected, with ReLu activation functions. The initial learning rates are $\eta \in \{1\text{e}{-2}, 3\text{e}{-2}, 1\text{e}{-1}\}$ for SGD and SplitSGD and $\eta \in \{3\text{e}{-4}, 1\text{e}{-3}, 3\text{e}{-3}\}$ for Adam. In the first panel of Figure 11 we see that most methods achieve very good accuracy, but SplitSGD reaches the overall best test accuracy when $\eta = 1\text{e}{-1}$ and great accuracy with small oscillations when $\eta = 3\text{e}{-2}$. The peaks in the SplitSGD performance are usually due to the averaging, while the smaller oscillations are due to the learning rate decay.

VGG19. When training the neural network VGG19[2] on CIFAR-10, we observe a similar behavior to what already shown when training ResNet (second panel of Figure 5). SplitSGD, with both learning rates $1\text{e}{-1}$ and $1\text{e}{-2}$ achieves the same test accuracy of the manually tuned SGD, but in less epochs, and beats the performance of SGD and Adam. Also here it is possible to see the spikes given by the averaging, followed by the smoothing caused by the learning rate decay.

[2] More details can be found in `https://pytorch.org/docs/stable/torchvision/models.html`

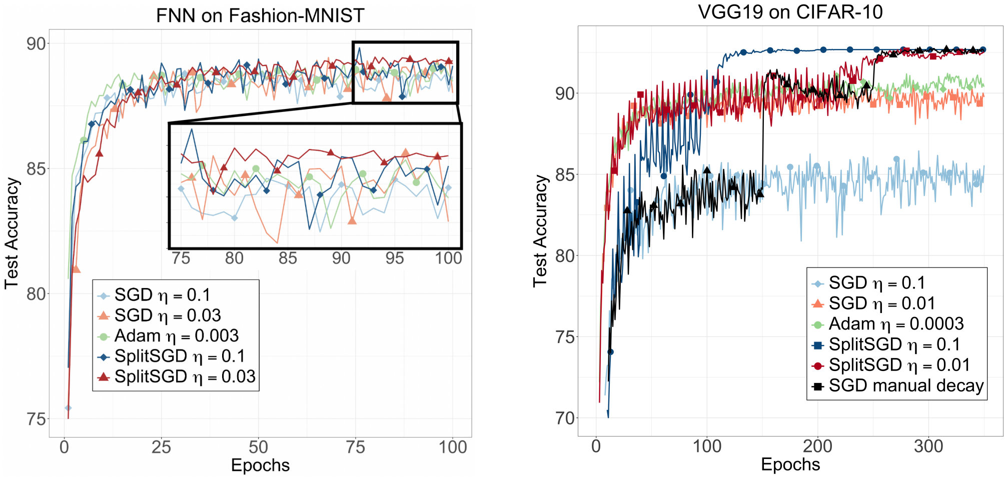

Figure 11: Compare the accuracy of SGD, Adam and SplitSGD in training FNN on Fashion-MNIST (left) and VGG19 on CIFAR-10 (right).