A On the proof of Theorem 1

For the proof of Theorem 1 we resort to techniques and results that have been developed in prior works and that can easily be adapted to the setting considered here, in particular (Stich and Karimireddy, 2020, Theorem 15). This theorem addresses the particular case when the gradients in Algorithm 1 are all delayed by a delay of *exactly* τ (a.k.a. *delayed SGD*). In contrast, we consider here the setting were coordinates of the gradients can be delayed independently, delays do not follow a particular order and reading of the variable $\mathbf{x}$ from the memory can be inconsistent. However, the proof in (Stich and Karimireddy, 2020) can easily be adapted to our more general setting (as also observed in Elalamy et al., 2020) and we do not claim much novelty here—except of explicitly stating this generalization.

A.1 Proof Overview

The proof in (Stich and Karimireddy, 2020) for the convex case follows by refining the perturbed iterated framework, developed in (Mania et al., 2017) and extended in (Leblond et al., 2018). A key ingredient in the proof is to consider a (virtual, ghost) sequence

$$\tilde{\mathbf{x}}_{t+1} := \tilde{\mathbf{x}}_t - \gamma_t \mathbf{g}_t$$

with $\mathbf{g}_t = \mathbf{g}_t(\mathbf{x}_t)$. In the following we resort—for the ease of presentation—to constant step sizes $\gamma_t \equiv \gamma$.

For instance for convex functions, it can be shown (Stich and Karimireddy, 2020, Lemma 7) that the perturbed iterates satisfy

$$\mathbb{E}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu\gamma}{2}\right) \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 - \frac{\gamma}{2}(\mathbb{E}f(\mathbf{x}_t) - f^\star) + \gamma^2\sigma^2 + 3L\gamma \underbrace{\mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2}_{=:R_t}, \tag{10}$$

and for non-convex functions (Stich and Karimireddy, 2020, Lemma 8):

$$\mathbb{E}f(\tilde{\mathbf{x}}_{t+1}) \leq \mathbb{E}f(\mathbf{x}_t) - \frac{\gamma}{4}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L}{2}\sigma^2 + \frac{\gamma L}{2} \underbrace{\mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2}_{=:R_t}. \tag{11}$$

A.2 Bound on R_t in (Stich and Karimireddy, 2020)

Stich and Karimireddy (2020) analyze the convergence of a *delayed gradient method*, as introduced in (Arjevani et al., 2020) and provide an upper bound for the value of R_t.

Lemma 3 (Stich and Karimireddy 2020, Lemma 10)**.** *Let* $\gamma \leq \frac{1}{10L(\tau+M)}$ *and* $\mathbf{x}_t$ *defined as* $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \mathbf{g}_{t-\tau}$ *for* $t \geq \tau$*, and* $\mathbf{x}_t = \mathbf{x}_0$ *for* $t \in \{0, \dots, \tau - 1\}$ *(*delayed SGD*). Then*

$$R_t := \mathbb{E}\left[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2\right] \leq \frac{1}{30L^2\tau} \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla f(\mathbf{x}_k)\|^2 + \frac{2}{3L}\gamma\sigma^2 =: \Theta_{\text{SK}}. \tag{12}$$

A.3 Bound on R_t under τ bounded parallelism

We now switch to our setting and derive a similar bound on R_t that holds for the more general class of algorithms considered in Theorem 1.

Lemma 4. *It holds*

$$R_t = \mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 \leq 2\gamma^2(\tau + M) \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla f(\mathbf{x}_k)\|^2 + 2\gamma^2\tau\sigma^2.$$

and in particular for $\gamma \leq \gamma_{\text{crit}} = \frac{1}{10L(M+\tau)}$

$$R_t \leq \frac{1}{50L^2\tau} \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{5L}\gamma\sigma^2 =: \Theta_{\text{SMJ}}. \tag{13}$$

We observe that our bound provided in (13) is smaller than the bound provided in (12), i.e., $\Theta_{\mathrm{SMJ}} \leq \Theta_{\mathrm{SK}}$. Therefore, the proof of Theorem 1 now follows from (Stich and Karimireddy, 2020, Theorem 16) (that only relies on the weaker bound Θ_{SK}).

Proof of Lemma 4. First, we observe that by definition of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$ and the maximal overlap τ, we can write

$$\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 := \|\gamma \sum_{k<t} (\mathbf{J}_k^t - \mathbf{I}_d)\mathbf{g}_k\|^2 = \|\gamma \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\mathbf{g}_k\|^2 \,, \tag{14}$$

where $\mathbf{g}_k := \nabla f(\mathbf{x}_k) + \boldsymbol{\xi}_k$ for zero-mean noise terms. Therefore

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 &\overset{\textcircled{1}}{\leq} 2\gamma^2 \left(\mathbb{E}\| \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\|^2 + \mathbb{E}\| \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\boldsymbol{\xi}_k\|^2 \right) \\
&\overset{\textcircled{2}}{\leq} 2\gamma^2 \left(\tau \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|(\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\|^2 + \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|(\mathbf{J}_k^t - \mathbf{I}_d)\boldsymbol{\xi}_k\|^2 \right) \\
&\overset{\textcircled{3}}{\leq} 2\gamma^2 \left(\tau \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla f(\mathbf{x}_k)\|^2 + \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\boldsymbol{\xi}_k\|^2 \right) \\
&\overset{\textcircled{4}}{\leq} 2\gamma^2 \left((\tau + M) \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla f(\mathbf{x}_k)\|^2 + \tau\sigma^2 \right) \,,
\end{aligned}$$

where we used ① $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, ② $\|\sum_{i=1}^{\tau} \mathbf{a}_i\|^2 \leq \tau \sum_{i=1}^{\tau} \|\mathbf{a}_i\|^2$, and $\mathbb{E}\|\sum_{i=1}^{\tau} \boldsymbol{\xi}_i\|^2 = \sum_{i=1}^{\tau} \mathbb{E}\|\boldsymbol{\xi}_k\|^2$, ③ $\|(\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\|^2 \leq \|\mathbf{J}_k^t - \mathbf{I}_d\|^2 \|\mathbf{g}_k\|^2 \leq \|\nabla f(\mathbf{x}_k)\|^2$, ④ $\mathbb{E}\|\boldsymbol{\xi}_k\|^2 \leq M\|\nabla f(\mathbf{x}_k)\|^2 + \sigma^2$. □

A.4 Concluding the proof

As mentioned above, the proof now follows directly from (Stich and Karimireddy, 2020, Theorem 16). To make this paper more self-contained, we illustrate the remaining steps for the case of non-convex functions.

For the non-convex case, equation (11) gives us the progress of one step. Using notation $r_t := 4\mathbb{E}(f(\tilde{\mathbf{x}}_t) - f^\star)$, $s_t := \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2$, and $c = 4L\sigma^2$ we have

$$\begin{aligned}
\frac{1}{4T}\sum_{t=0}^{T} s_t &\overset{(11)}{\leq} \frac{1}{T}\sum_{t=0}^{T} \left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma} + \frac{\gamma c}{8} \right) + \frac{L^2}{2T}\sum_{t=0}^{T} \mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 \\
&\overset{\substack{(\Theta_{\mathrm{SMJ}} \leq \Theta_{\mathrm{SK}}) \\ (12)}}{\leq} \frac{1}{T}\sum_{t=0}^{T} \left(\frac{r_t}{4\gamma} - \frac{r_{t+1}}{4\gamma} + \frac{\gamma c}{8} \right) + \frac{L^2}{2T}\sum_{t=0}^{T} \left(\frac{1}{15L^2} s_t + \frac{\gamma c}{4L^2} \right) .
\end{aligned}$$

The above equation can be simplified as:

$$\frac{1}{5T}\sum_{t=0}^{T} s_t \leq \frac{1}{T}\sum_{t=0}^{T} \left(\frac{r_t}{4\gamma} - \frac{r_{t+1}}{4\gamma} + \frac{\gamma c}{4} \right) \leq \frac{r_0}{4\gamma T} + \frac{\gamma c}{4} \,.$$

Now, the claimed bound follows by choosing the optimal stepsize $\gamma \leq \gamma_{\mathrm{crit}}$ that minimizes the right hand side. For this refer e.g. to (Stich and Karimireddy, 2020, Lemma 14) or (Arjevani et al., 2020).

The proof for the convex cases start from the one step progress provided in (10) instead, and proceed similarly.

B Additional numerical experiments

In this section we report additional empirical results for the setting considered in Section 7.1. We consider three algorithms with the same level of parallelism: mini-batch SGD as considered in the main text, and two implementations of SGD with delayed updates.

B.1 On the estimator $\hat{b}(\mathbf{x})$

Note that

$$1 + \frac{\sigma_\star^2}{\max\{\hat{\epsilon}_T, \|\nabla f(\mathbf{x})\|^2\}} + M \geq 1 + \frac{M\|\nabla f(\mathbf{x})\|^2 + \sigma_\star^2}{\|\nabla f(\mathbf{x})\|^2 + \hat{\epsilon}_T} \overset{(5)}{\geq} \hat{b}(\mathbf{x})\,. \tag{15}$$

Moreover, for $\tilde{\epsilon} := \max\{\hat{\epsilon}_T, \|\nabla f(\mathbf{x})\|^2\}$,

$$\begin{aligned} 1 + \frac{\sigma_\star^2}{\tilde{\epsilon}} + M &\leq 1 + \sup_{\|\nabla f(\mathbf{x})\|^2 \leq \tilde{\epsilon}} \frac{\mathbb{E}\|\boldsymbol{\xi}(\mathbf{x})\|^2}{\tilde{\epsilon}} + \sup_{\|\nabla f(\mathbf{x})\|^2 \geq \tilde{\epsilon}} \frac{\mathbb{E}\|\boldsymbol{\xi}(\mathbf{x})\|^2}{\|\nabla f(\mathbf{x})\|^2} && (16)\\ &\leq 1 + 2\sup_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbb{E}\|\boldsymbol{\xi}(\mathbf{x})\|^2}{\tilde{\epsilon} + \|\nabla f(\mathbf{x})\|^2} + 2\sup_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbb{E}\|\boldsymbol{\xi}(\mathbf{x})\|^2}{\tilde{\epsilon} + \|\nabla f(\mathbf{x})\|^2} && (17)\\ &\leq 4\sup_{\mathbf{x}\in\mathbb{R}^d} \hat{b}(\mathbf{x})\,. && (18) \end{aligned}$$

This method of measuring the critical batch size might not be too accurate. We use this estimator only to show that our theoretical findings match our observations in practice and to show how they can be used to explain phenomena such as critical batch size and scaling of learning rate. We leave finding a more accurate and online method for measuring the critical batch size as a possible future work.

B.2 Mini-batch SGD

We consider standard mini-batch SGD, for batch size $b \geq 1$,

$$\mathbf{x}_{t+b} = \mathbf{x}_t - \frac{\gamma_{\text{bm}}}{b} \sum_{i=1}^{b} \mathbf{g}^i(\mathbf{x}_t)\,,$$

where $\mathbf{g}^i(\mathbf{x}_t)$ for $i \in [b]$ denotes independently sampled stochastic gradients.

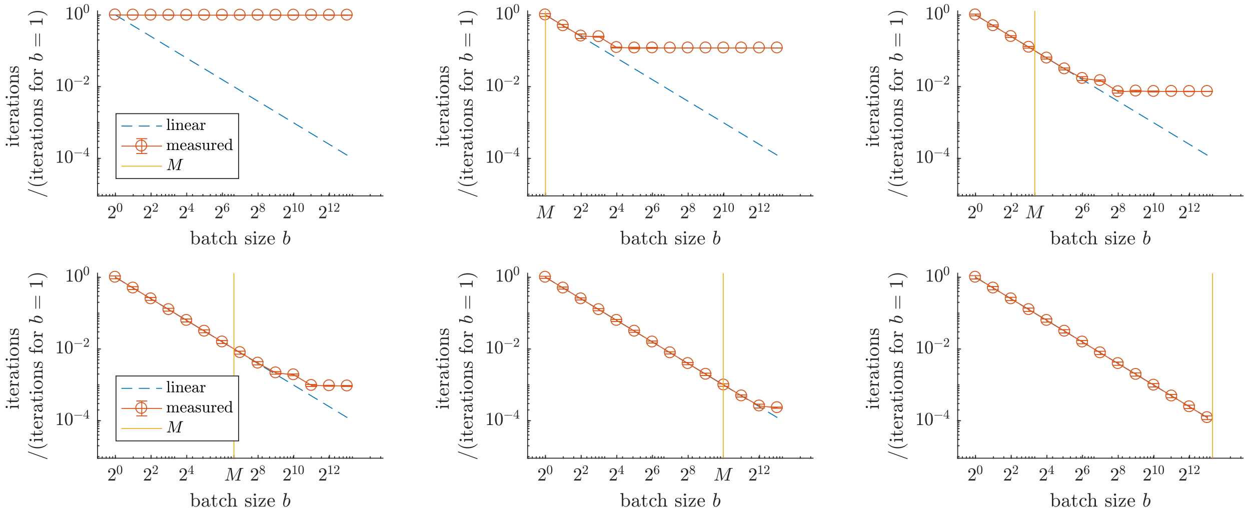

Figure 9: **Scaling (Mini-batch SGD)**. Parallel speedup for various batch sizes $b \in \{2^0, \ldots, 2^{14}\}$ and problem instances with $M \in \{0, 1, 10\}$ (top) and $M \in \{100, 1000, 10000\}$ (bottom), on the synthetic optimization problem described in Section 7.1, averaged over three random seeds (depicting mean and ±SD). Plots depict number of iterations (i.e. parallel running time $\frac{1}{b}T(b,\epsilon)$), normalized by $T(1,\epsilon)$, required to reach the target accuracy with tuned optimal learning rates.

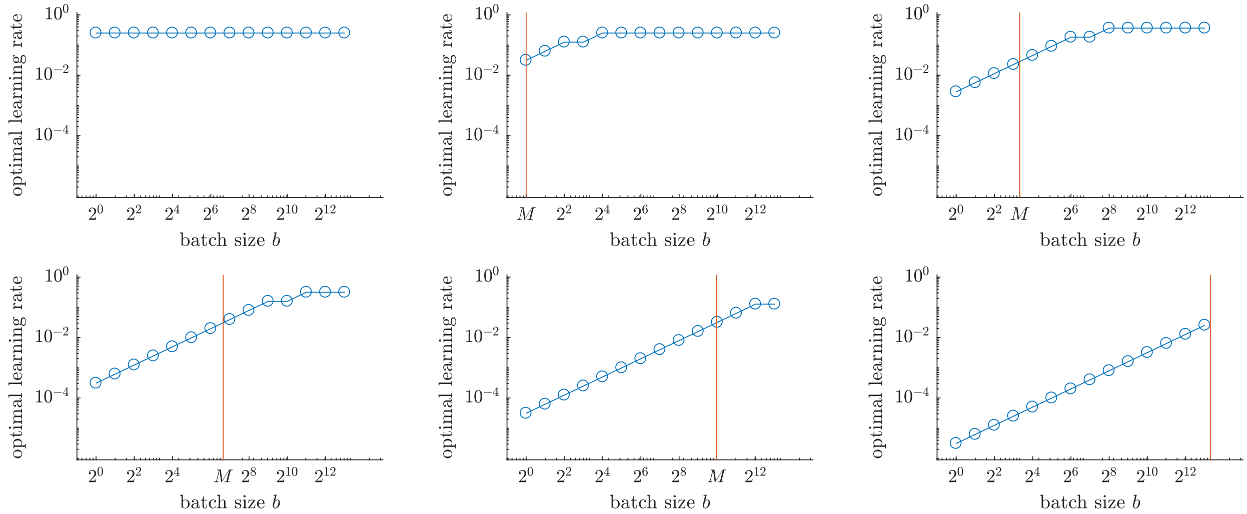

Figure 10: **Optimal learning rate** γ_{mb} **for mini-batch SGD** for the results reported in Figure 9.

B.3 Delayed SGD (coordinate-wise random delays)

In this section we consider SGD with delayed updates. Concretely, we simulate the case where each coordinate $[\mathbf{g}(\mathbf{x}_t)]_v$, $v \in [d]$ is delayed for a delay $\tau_{t,v} \sim_{\text{u.a.r.}} [\tau]$. This can be seen as a simplistic modeling of Hogwild! (Niu et al., 2011), though in practical settings the delays might be correlated. The update can be written as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_{\text{HW}}}{\tau} \sum_{i=t+1-\tau}^{t} \mathbf{P}_i^t \mathbf{g}_i$$

where $\mathbf{g}_t := \mathbf{g}(\mathbf{x}_t)$ stochastic gradients (sampled at iteration t), and $\mathbf{P}_i^t$ are diagonal matrices with $\sum_{k \geq t} \mathbf{P}_i^k = \mathbf{I}_d$, $(\mathbf{P}_i^t)_{vv} = 1$ if $[\mathbf{g}_t]_v$ is written at iteration $i \geq t$ and $(\mathbf{P}_i^t)_{vv} = 0$ otherwise.

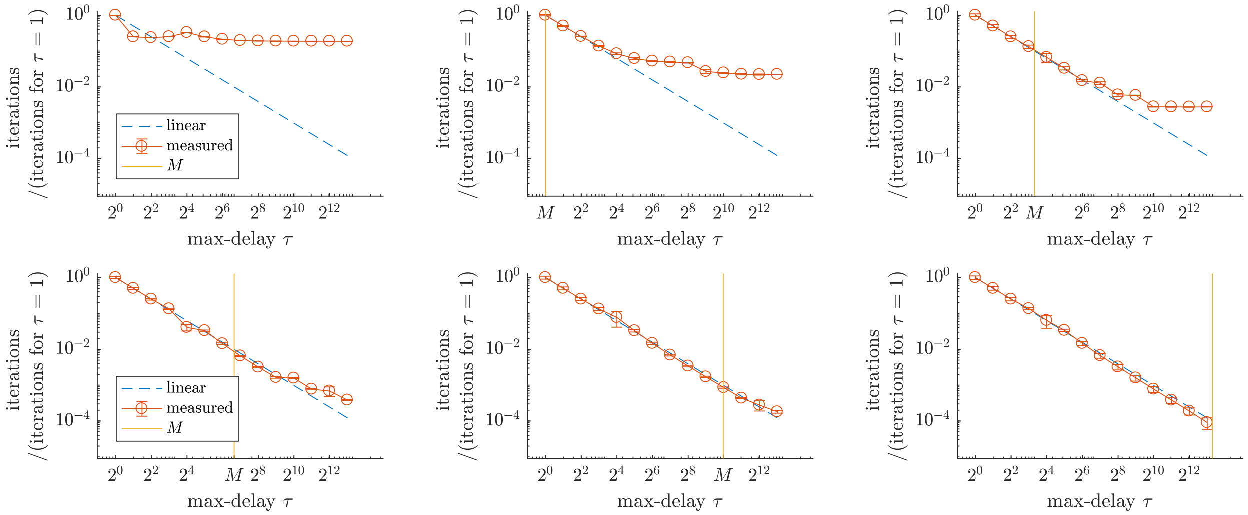

Figure 11: **Scaling of SGD with coordinate-wise delayed updates (Hogwild!).** Parallel speedup for various batch sizes/delay values $b \in \{2^0, \ldots, 2^{14}\}$ and problem instances with $M \in \{0, 1, 10\}$ (top) and $M \in \{100, 1000, 10000\}$ (bottom), on the synthetic optimization problem described in Section 7.1, averaged over three random seeds (depicting mean and ±SD). Plots depict number of iterations (i.e. parallel running time $\frac{1}{b}T(b,\epsilon)$), normalized by $T(1,\epsilon)$, required to reach the target accuracy with tuned optimal learning rates.

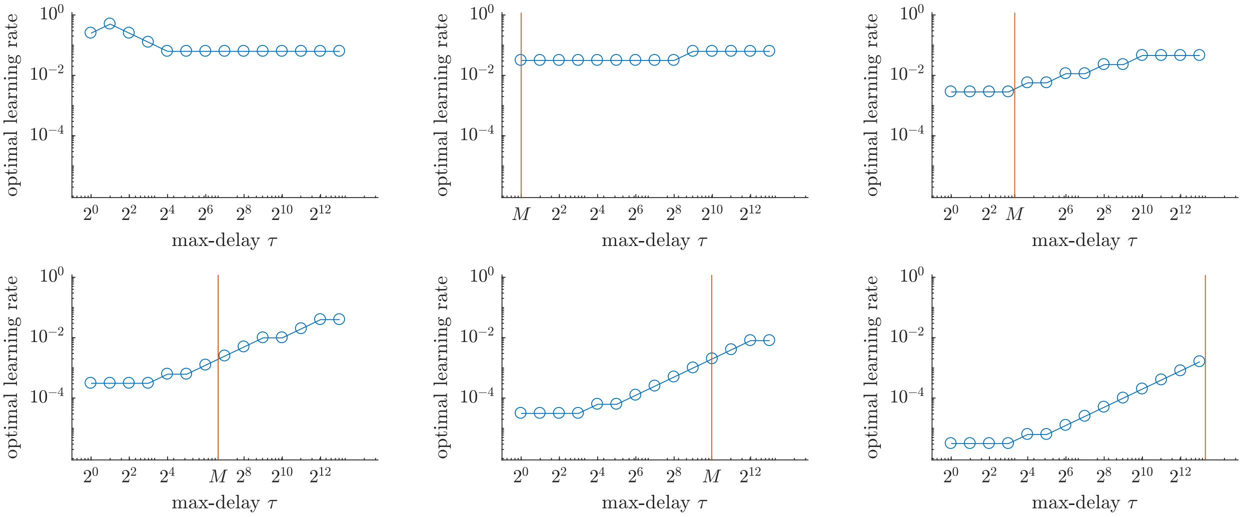

Figure 12: **Optimal learning rate γ_{HW} for Hogwild!** for the results reported in Figure 11.

B.4 Delayed SGD (worst case delays)

In this section we consider SGD with delayed updates (Arjevani et al., 2020). Concretely, we assume each gradient update is delayed by exactly τ iterations. For $t \geq \tau$, the update can be written as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_{\mathrm{d}}}{\tau}\mathbf{g}(\mathbf{x}_{t+1-\tau})\,,$$

with $\mathbf{x}_i = \mathbf{x}_0$ for $i \in [\tau - 1]$.

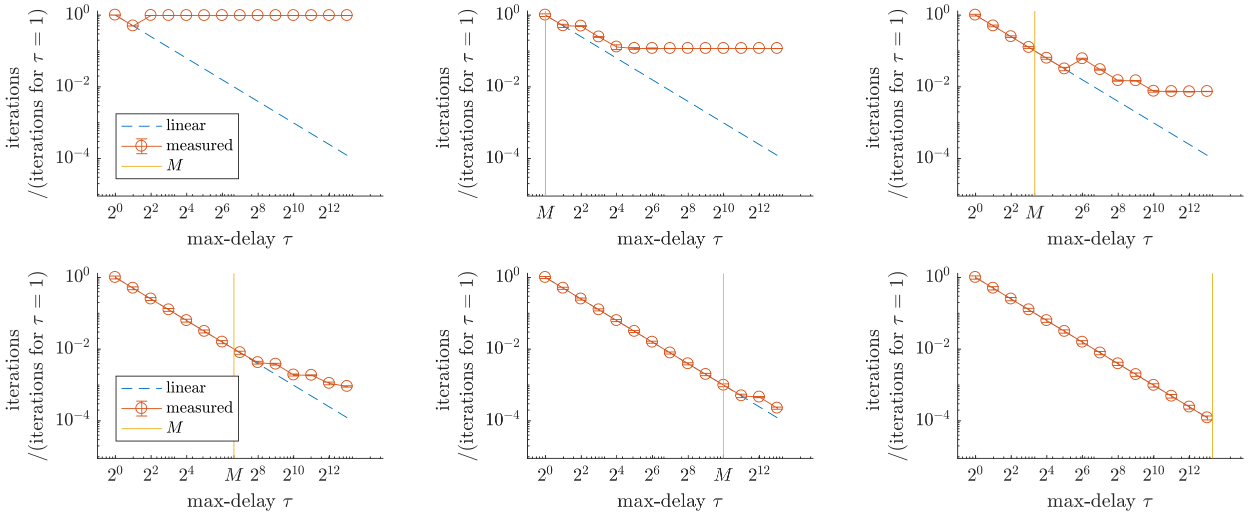

Figure 13: **Scaling of delayed SGD.** Parallel speedup for various delay values $\tau \in \{2^0, \dots, 2^{14}\}$ and problem instances with $M \in \{0, 1, 10\}$ (top) and $M \in \{100, 1000, 10000\}$ (bottom), on the synthetic optimization problem described in Section 7.1, averaged over three random seeds (depicting mean and ±SD). Plots depict number of iterations (i.e. parallel running time $\frac{1}{b}T(b,\epsilon)$), normalized by $T(1,\epsilon)$, required to reach the target accuracy with tuned optimal learning rates.

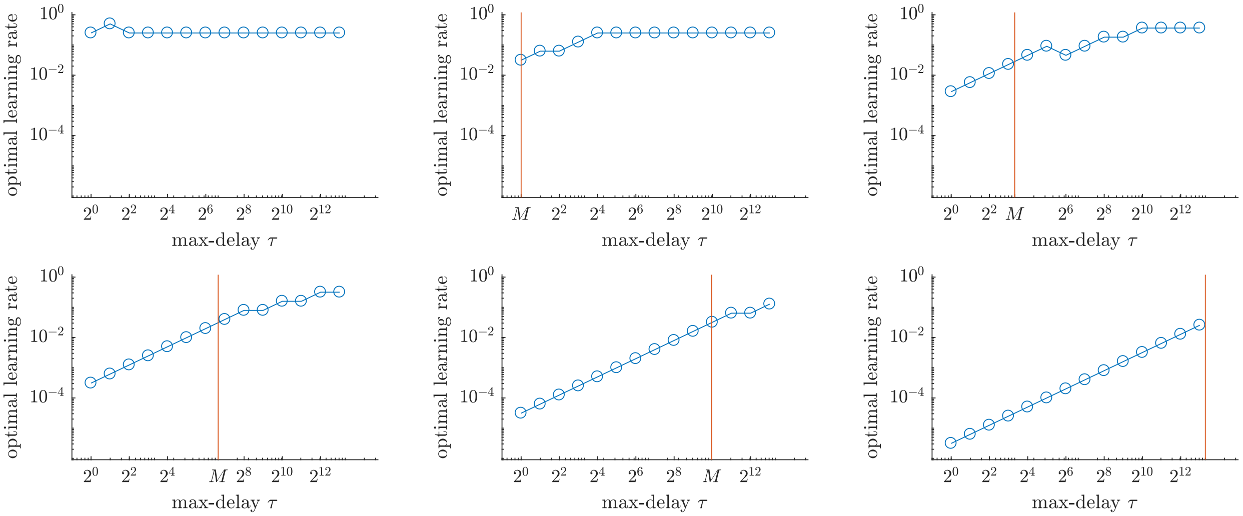

Figure 14: **Optimal learning rate γ_{d} for delayed SGD** for the results reported in Figure 13.

B.5 Hyperparameters for Deep Learning Experiments

For Figures 7 and 8, we tune the learning rate for each batch size. In particular, for ResNet-8, we use as step size, 0.4 when batch size is 32, 0.2 when batch size is 64 and 0.05 for all other batch sizes. The step size was chosen from the set $\{0.005, 0.05, 0.02, 0.1, 0.2, 0.4\}$.

For ResNet-18, we use as step size, 0.02 when batch size is 32, 0.04 when batch size is 64 and 0.1 for all other batch sizes. The step size was chosen from the set $\{0.005, 0.02, 0.05, 0.1\}$.