Supplementary Material

A. Proofs

Proof of Theorem 3.2 relies on Theorem 3.1, which in turn relies on Theorem A.1 and Lemma A.1, both of which are stated below. Proofs of the lemma and theorems follow in the subsequent subsections.

The next result is a standard result from convex optimization (Theorem 2.1.14 in (Nesterov, 2014)) and is used in the proof of Theorem A.1 below.

Next, we introduce the *population gradient AM operator*, $\mathcal{G}_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K)$, where $i = 1, 2, \dots, K$, defined as

$$\mathcal{G}_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K) \coloneqq \boldsymbol{\theta}_i + \eta \nabla_i f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K),$$

where η is the step size.

Lemma A.1. *For any $d = 1, 2, \dots, K$, the gradient operator $\mathcal{G}_d(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \dots, \boldsymbol{\theta}_{K-1}^*, \boldsymbol{\theta}_K^*)$ under Assumption 3 (strong concavity) and Assumption 3 (smoothness) with constant step size choice $0 < \eta \le \frac{2}{\mu_d + \lambda_d}$ is contractive, i.e.*

$$\|\mathcal{G}_d(\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \dots, \boldsymbol{\theta}_K^*) - \boldsymbol{\theta}_d^*\|_2 \le \left(1 - \frac{2\eta\mu_d\lambda_d}{\mu_d + \lambda_d}\right) \|\boldsymbol{\theta}_d - \boldsymbol{\theta}_d^*\|_2 \tag{11}$$

for all $\boldsymbol{\theta}_d \in B_2(r_d, \boldsymbol{\theta}_d^*)$.

The next theorem also holds for any d from 1 to K. Let $r_1, \dots, r_{d-1}, r_{d+1}, \dots, r_K > 0$ and $\boldsymbol{\theta}_1 \in B_2(r_1, \boldsymbol{\theta}_1^*), \dots, \boldsymbol{\theta}_{d-1} \in B_2(r_{d-1}, \boldsymbol{\theta}_{d-1}^*), \boldsymbol{\theta}_{d+1} \in B_2(r_{d+1}, \boldsymbol{\theta}_{d+1}^*), \dots, \boldsymbol{\theta}_K \in B_2(r_k, \boldsymbol{\theta}_K^*)$.

Theorem A.1. *For some radius $r_d > 0$ and a triplet $(\gamma_d, \lambda_d, \mu_d)$ such that $0 \le \gamma_d < \lambda_d \le \mu_d$, suppose that the function $L(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \dots, \boldsymbol{\theta}_{K-1}^*, \boldsymbol{\theta}_K^*)$ is λ_d-strongly concave (Assumption 3) and μ_d-smooth (Assumption 3), and that the GS (γ_d) condition of Assumption 3 holds. Then the population gradient AM operator $\mathcal{G}_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K)$ with step η such that $0 < \eta \le \min_{i=1,2,\dots,K} \frac{2}{\mu_i + \lambda_i}$ is contractive over a ball $B_2(r_d, \boldsymbol{\theta}_d^*)$, i.e.*

$$\|\mathcal{G}_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K) - \boldsymbol{\theta}_d^*\|_2 \le (1 - \xi\eta)\|\boldsymbol{\theta}_d - \boldsymbol{\theta}_d^*\|_2 + \eta\gamma \sum_{\substack{i=1 \\ i \ne d}}^{K} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*\|_2 \tag{12}$$

where $\gamma \coloneqq \max_{i=1,2,\dots,K} \gamma_i$, and $\xi \coloneqq \min_{i=1,2,\dots,K} \frac{2\mu_i\lambda_i}{\mu_i + \lambda_i}$.

A.1. Proof of Theorem A.1

$$\begin{aligned}
&\|\mathcal{G}_d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K) - \boldsymbol{\theta}_d^*\|_2 = \|\boldsymbol{\theta}_d + \eta \nabla_d L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \dots, \boldsymbol{\theta}_K) - \boldsymbol{\theta}_d^*\|_2 \\
&\text{by the triangle inequality we further get} \\
&\le \|\boldsymbol{\theta}_d + \eta \nabla_d L(\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \dots, \boldsymbol{\theta}_K^*) - \boldsymbol{\theta}_d^*\|_2 \\
&+\eta\|\nabla_d L(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d, \dots, \boldsymbol{\theta}_K) \\
&-\nabla_d L(\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \dots, \boldsymbol{\theta}_K^*)\|_2 \\
&\text{by the contractivity of } T \text{ from Equation 11 from Lemma A.1 and GS condition} \\
&\le \left(1 - \frac{2\eta\mu_d\lambda_d}{\mu_d + \lambda_d}\right) \|\theta_d - \theta_d^*\|_2 + \eta\gamma_d \sum_{\substack{i=1 \\ i \ne d}}^{K} \|\theta_i - \theta_i^*\|_2.
\end{aligned}$$

A.2. Proof of Theorem 3.1

Let $\boldsymbol{\theta}_d^{t+1} = \Pi_d(\tilde{\boldsymbol{\theta}}_d^{t+1})$, where $\tilde{\boldsymbol{\theta}}_d^{t+1} \coloneqq \boldsymbol{\theta}_d^t + \eta^t \nabla_d L^1(\boldsymbol{\theta}_1^{t+1}, \boldsymbol{\theta}_2^{t+1}, \dots, \boldsymbol{\theta}_{d-1}^{t+1}, \boldsymbol{\theta}_d^t, \boldsymbol{\theta}_{d+1}^t, \dots, \boldsymbol{\theta}_K^t)$ ($\nabla_d L^1$ is the gradient computed with respect to a single data sample) is the update vector prior to the projection onto a ball $B_2(\frac{r_d}{2}, \boldsymbol{\theta}_d^0)$. Let

$\boldsymbol{\Delta}_d^{t+1} \coloneqq \boldsymbol{\theta}_d^{t+1} - \boldsymbol{\theta}_d^*$ and $\tilde{\boldsymbol{\Delta}}_d^{t+1} \coloneqq \tilde{\boldsymbol{\theta}}_d^{t+1} - \boldsymbol{\theta}_d^*$. Thus

$$\begin{aligned}
\|\boldsymbol{\Delta}_d^{t+1}\|_2^2 - \|\boldsymbol{\Delta}_d^t\|_2^2 &\leq \|\tilde{\boldsymbol{\Delta}}_d^{t+1}\|_2^2 - \|\boldsymbol{\Delta}_d^t\|_2^2 \\
&= \|\tilde{\boldsymbol{\theta}}_d^{t+1} - \boldsymbol{\theta}_d^*\| - \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\| \\
&= \left\langle \tilde{\boldsymbol{\theta}}_d^{t+1} - \boldsymbol{\theta}_d^t, \tilde{\boldsymbol{\theta}}_d^{t+1} + \boldsymbol{\theta}_d^t - 2\boldsymbol{\theta}_d^* \right\rangle .
\end{aligned}$$

Let $\hat{\boldsymbol{W}}_d^t \coloneqq \nabla_d L^1(\boldsymbol{\theta}_1^{t+1}, \boldsymbol{\theta}_2^{t+1}, \ldots, \boldsymbol{\theta}_{d-1}^{t+1}, \boldsymbol{\theta}_d^t, \boldsymbol{\theta}_{d+1}^t, \ldots, \boldsymbol{\theta}_K^t)$. Then we have that $\tilde{\boldsymbol{\theta}}_d^{t+1} - \boldsymbol{\theta}_d^t = \eta^t \hat{\boldsymbol{W}}_d^t$. We combine it with Equation 13 and obtain:

$$\begin{aligned}
& \|\boldsymbol{\Delta}_d^{t+1}\|_2^2 - \|\boldsymbol{\Delta}_d^t\|_2^2 \\
\leq\ & \left\langle \eta^t \hat{\boldsymbol{W}}_d^t, \eta^t \hat{\boldsymbol{W}}_d^t + 2(\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*) \right\rangle \\
=\ & (\eta^t)^2 (\hat{\boldsymbol{W}}_d^t)^\top \hat{\boldsymbol{W}}_d^t + 2\eta^t (\hat{\boldsymbol{W}}_d^t)^\top (\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*) \\
=\ & (\eta^t)^2 \|\hat{\boldsymbol{W}}_d^t\|_2^2 + 2\eta^t \left\langle \hat{\boldsymbol{W}}_d^t, \boldsymbol{\Delta}_d^t \right\rangle .
\end{aligned}$$

Let $\boldsymbol{W}_d^t \coloneqq \nabla_d L(\boldsymbol{\theta}_1^{t+1}, \boldsymbol{\theta}_2^{t+1}, \ldots, \boldsymbol{\theta}_{d-1}^{t+1}, \boldsymbol{\theta}_d^t, \boldsymbol{\theta}_{d+1}^t, \ldots, \boldsymbol{\theta}_K^t)$. Recall that $\mathbb{E}[\hat{\boldsymbol{W}}_d^t] = \boldsymbol{W}_d^t$. By the properties of martingales, i.e. iterated expectations and tower property:

$$\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] \quad \leq \quad \mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + (\eta^t)^2 \mathbb{E}[\|\hat{\boldsymbol{W}}_d^t\|_2^2] + 2\eta^t \mathbb{E}[\langle \boldsymbol{W}_d^t, \boldsymbol{\Delta}_d^t \rangle] \tag{13}$$

Let $\boldsymbol{W}_d^* \coloneqq \nabla_d L(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_K^*)$. By self-consistency, i.e. $\boldsymbol{\theta}_d^* = \arg\max_{\boldsymbol{\theta}_d \in \Omega_d} L(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{d-1}^*, \boldsymbol{\theta}_d, \boldsymbol{\theta}_{d+1}^*, \ldots, \boldsymbol{\theta}_K^*)$ and convexity of Ω_d we have that

$$\langle \boldsymbol{W}_d^*, \boldsymbol{\Delta}_d^t \rangle = \langle \nabla_d L(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_K^*), \boldsymbol{\Delta}_d^t \rangle \leq 0.$$

Combining this with Equation 13 we have

$$\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] \quad \leq \quad \mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + (\eta^t)^2 \mathbb{E}[\|\hat{\boldsymbol{W}}_d^t\|_2^2] + 2\eta^t \mathbb{E}[\langle \boldsymbol{W}_d^t - \boldsymbol{W}_d^*, \boldsymbol{\Delta}_d^t \rangle].$$

Define $\mathcal{G}_d^t \coloneqq \boldsymbol{\theta}_d^t + \eta^t \boldsymbol{W}_d^t$ and $\mathcal{G}_d^{t*} \coloneqq \boldsymbol{\theta}_d^* + \eta^t \boldsymbol{W}_d^*$. Thus

$$\begin{aligned}
& \eta^t \left\langle \boldsymbol{W}_d^t - \boldsymbol{W}_d^*, \boldsymbol{\Delta}_d^t \right\rangle \\
=\ & \left\langle \mathcal{G}_d^t - \mathcal{G}_d^{t*} - (\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*), \boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^* \right\rangle \\
=\ & \left\langle \mathcal{G}_d^t - \mathcal{G}_d^{t*}, \boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^* \right\rangle - \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\|_2^2 \\
& \text{by the fact that } \mathcal{G}_d^{t*} = \boldsymbol{\theta}_d^* + \eta^t \boldsymbol{W}_d^* = \boldsymbol{\theta}_d^* \text{ (since } \boldsymbol{W}_d^* = 0\text{):} \\
=\ & \left\langle \mathcal{G}_d^t - \boldsymbol{\theta}_d^*, \boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^* \right\rangle - \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\|_2^2 \\
& \text{by the contractivity of } \mathcal{G}^t \text{ from Theorem A.1:} \\
\leq\ & \left\{ (1 - \eta^t \xi) \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\| + \eta^t \gamma \left(\sum_{i=1}^{d-1} \|\boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}_i^*\|_2 \ + \sum_{i=d+1}^{K} \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}_i^*\|_2 \right) \right\} \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\|_2 - \|\boldsymbol{\theta}_d^t - \boldsymbol{\theta}_d^*\|_2^2 \\
\leq\ & \left\{ (1 - \eta^t \xi) \|\boldsymbol{\Delta}_d^t\|_2 + \eta^t \gamma \left(\sum_{i=1}^{d-1} \|\boldsymbol{\Delta}_i^{t+1}\|_2 + \sum_{i=d+1}^{K} \|\boldsymbol{\Delta}_i^t\|_2 \right) \right\} \cdot \|\boldsymbol{\Delta}_d^t\|_2 - \|\boldsymbol{\Delta}_d^t\|_2^2
\end{aligned}$$

Combining this result with Equation 14 gives

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] \quad &\leq \quad \mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + (\eta^t)^2\mathbb{E}[\|\hat{\boldsymbol{W}}_d^t\|_2^2] + 2\mathbb{E}\left[\left\{(1-\eta^t\xi)\|\boldsymbol{\Delta}_d^t\|_2 + \eta^t\gamma\left(\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2 + \sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2\right)\right\}\right. \\
&\quad \left.\cdot\|\boldsymbol{\Delta}_d^t\|_2 - \|\boldsymbol{\Delta}_d^t\|_2^2\right] \\
&\leq \mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + (\eta^t)^2\sigma_d^2 + 2\mathbb{E}\left[\left\{(1-\eta^t\xi)\|\boldsymbol{\Delta}_d^t\|_2 + \eta^t\gamma\left(\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2 + \sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2\right)\right\}\right. \\
&\quad \left.\cdot\|\boldsymbol{\Delta}_d^t\|_2 - \|\boldsymbol{\Delta}_d^t\|_2^2\right], \quad \text{where}
\end{aligned}$$

$\sigma_d^2 = \sup_{\substack{\boldsymbol{\theta}_1\in B_2(r_1,\boldsymbol{\theta}_1^*)\\ \cdots \\ \boldsymbol{\theta}_K\in B_2(r_K,\boldsymbol{\theta}_K^*)}} \ \mathbb{E}[\|\nabla_d L^1(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_K)\|_2^2].$

After re-arranging the terms we obtain

$$\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] \leq (\eta^t)^2\sigma_d^2 + (1-2\eta^t\xi)\mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + 2\eta^t\gamma\mathbb{E}\left[\left(\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2 + \sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2\right)\|\boldsymbol{\Delta}_d^t\|_2\right]$$

apply $2ab \leq a^2 + b^2$

$$\begin{aligned}
&\leq \quad (\eta^t)^2\sigma_d^2 + (1-2\eta^t\xi)\mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + \eta^t\gamma\mathbb{E}\left[\sum_{i=1}^{d-1}\left(\|\boldsymbol{\Delta}_i^{t+1}\|_2^2 + \|\boldsymbol{\Delta}_d^t\|_2^2\right)\right] + \eta^t\gamma\mathbb{E}\left[\sum_{i=d+1}^{K}\left(\|\boldsymbol{\Delta}_i^t\|_2^2 + \|\boldsymbol{\Delta}_d^t\|_2^2\right)\right] \\
&= \quad (\eta^t)^2\sigma_d^2 + \mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2]\cdot\left[1-2\eta^t\xi + \eta^t\gamma(K-1)\right] + \eta^t\gamma\mathbb{E}\left[\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2^2\right] + \eta^t\gamma\mathbb{E}\left[\sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2^2\right]
\end{aligned}$$

We obtained

$$\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] \leq (\eta^t)^2\sigma_d^2 + [1-2\eta^t\xi + \eta^t\gamma(K-1)]\mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + \eta^t\gamma\mathbb{E}\left[\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2^2\right] + \eta^t\gamma\mathbb{E}\left[\sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2^2\right]$$

we next re-group the terms as follows

$$\mathbb{E}[\|\boldsymbol{\Delta}_d^{t+1}\|_2^2] - \eta^t\gamma\mathbb{E}\left[\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2^2\right] \leq [1-2\eta^t\xi + \eta^t\gamma(K-1)]\mathbb{E}[\|\boldsymbol{\Delta}_d^t\|_2^2] + \eta^t\gamma\mathbb{E}\left[\sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2^2\right] + (\eta^t)^2\sigma_d^2$$

and then sum over d from 1 to K

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] - \eta^t\gamma\mathbb{E}\left[\sum_{d=1}^{K}\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2^2\right] \\
\leq \quad &[1-2\eta^t\xi + \eta^t\gamma(K-1)]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right] + \eta^t\gamma\mathbb{E}\left[\sum_{d=1}^{K}\sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2^2\right] + (\eta^t)^2\sum_{d=1}^{K}\sigma_d^2
\end{aligned}$$

Let $\sigma = \sqrt{\sum_{d=1}^K \sigma_d^2}$. Also, note that

$$\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] - \eta^t\gamma(K-1)\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \leq \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] - \eta^t\gamma\mathbb{E}\left[\sum_{d=1}^{K}\sum_{i=1}^{d-1}\|\boldsymbol{\Delta}_i^{t+1}\|_2^2\right]$$

and

$$
\begin{aligned}
&[1-2\eta^t\xi+\eta^t\gamma(K-1)]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+\eta^t\gamma\mathbb{E}\left[\sum_{d=1}^{K}\sum_{i=d+1}^{K}\|\boldsymbol{\Delta}_i^t\|_2^2\right]+(\eta^t)^2\sigma^2 \\
\leq\quad &[1-2\eta^t\xi+\eta^t\gamma(K-1)]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+\eta^t\gamma(K-1)\mathbb{E}\left[\sum_{d=1}^{K}\|\Delta_d^t\|_2^2\right]+(\eta^t)^2\sigma^2
\end{aligned}
$$

Combining these two facts with our previous results yields:

$$
\begin{aligned}
&[1-(K-1)\eta^t\gamma]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
&\leq[1-2\eta^t\xi+\eta^t\gamma(K-1)]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+\eta^t\gamma(K-1)\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+(\eta^t)^2\sigma^2 \\
&=[1-2\eta^t\xi+2\eta^t\gamma(K-1)]\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+(\eta^t)^2\sigma^2
\end{aligned}
$$

Thus:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right]\leq&\frac{1-2\eta^t\xi+2\eta^t\gamma(K-1)}{1-(K-1)\eta^t\gamma}\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right] \\
&+\frac{(\eta^t)^2}{1-(K-1)\eta^t\gamma}\sigma^2.
\end{aligned}
$$

Since $\gamma<\frac{2\xi}{3(K-1)}$, $\frac{1-2\eta^t\xi+2\eta^t\gamma(K-1)}{1-(K-1)\eta^t\gamma}<1$.

A.3. Proof of Theorem 3.2

To obtain the final theorem we need to expand the recursion from Theorem 3.1. We obtained

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
&\leq\frac{1-2\eta^t[\xi-\gamma(K-1)]}{1-(K-1)\eta^t\gamma}\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+\frac{(\eta^t)^2}{1-(K-1)\eta^t\gamma}\sigma^2 \\
&=\left(1-\frac{\eta^t[2\xi-3\gamma(K-1)]}{1-(K-1)\eta^t\gamma}\right)\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^t\|_2^2\right]+\frac{(\eta^t)^2}{1-(K-1)\eta^t\gamma}\sigma^2
\end{aligned}
$$

Recall that we defined q^t in Theorem 3.1 as

$$
q^t=1-\frac{1-2\eta^t\xi+2\eta^t\gamma(K-1)}{1-(K-1)\eta^t\gamma}=\frac{\eta^t[2\xi-3\gamma(K-1)]}{1-(K-1)\eta^t\gamma}
$$

and denote

$$
\beta^t=\frac{(\eta^t)^2}{1-(K-1)\eta^t\gamma}.
$$

Thus we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \leq (1-q^t)\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t}\|_2^2\right] + \beta^t\sigma^2 \\
\leq\ & (1-q^t)\left\{(1-q^{t-1})\mathbb{E}\left[\sum_{d=1}^{K}\|\Delta_d^{t-1}\|_2^2\right] + \beta^{t-1}\sigma^2\right\} + \beta^t\sigma^2 \\
=\ & (1-q^t)(1-q^{t-1})\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t-1}\|_2^2\right] + (1-q^t)\beta^{t-1}\sigma^2 + \beta^t\sigma^2 \\
\leq\ & (1-q^t)(1-q^{t-1})\left\{(1-q^{t-2})\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t-2}\|_2^2\right] + \beta^{t-2}\sigma^2\right\} + (1-q^t)\beta^{t-1}\sigma^2 + \beta^t\sigma^2 \\
=\ & (1-q^t)(1-q^{t-1})(1-q^{t-2})\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t-2}\|_2^2\right] \\
& +(1-q^t)(1-q^{t-1})\beta^{t-2}\sigma^2 + (1-q^t)\beta^{t-1}\sigma^2 + \beta^t\sigma^2
\end{aligned}$$

We end-up with the following

$$\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \leq \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=0}^{t}(1-q^i) + \sigma^2\sum_{i=0}^{t-1}\beta^i\prod_{j=i+1}^{t}(1-q^j) + \beta^t\sigma^2.$$

Set $q^t = \frac{\frac{3}{2}}{t+2}$ and

$$\begin{aligned}
\eta^t &= \frac{q^t}{2\xi - 3\gamma(K-1) + q^t(K-1)\gamma} \\
&= \frac{\frac{3}{2}}{[2\xi - 3\gamma(K-1)](t+2) + \frac{3}{2}(K-1)\gamma}.
\end{aligned}$$

Denote $A = 2\xi - 3\gamma(K-1)$ and $B = \frac{3}{2}(K-1)\gamma$. Thus

$$\eta^t = \frac{\frac{3}{2}}{A(t+2)+B}$$

and

$$\beta^t = \frac{(\eta^t)^2}{1-\frac{2}{3}B\eta^t} = \frac{\frac{9}{4}}{A(t+2)[A(t+2)+B]}.$$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=0}^{t}\left(1-\frac{\frac{3}{2}}{i+2}\right) + \sigma^2\sum_{i=0}^{t-1}\frac{\frac{9}{4}}{A(i+2)[A(i+2)+B]}\prod_{j=i+1}^{t}\left(1-\frac{\frac{3}{2}}{j+2}\right) \\
& +\sigma^2\frac{\frac{9}{4}}{A(t+2)[A(t+2)+B]} \\
=\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=2}^{t+2}\left(1-\frac{\frac{3}{2}}{i}\right) + \sigma^2\sum_{i=2}^{t+1}\frac{\frac{9}{4}}{Ai[Ai+B]}\prod_{j=i+1}^{t+2}\left(1-\frac{\frac{3}{2}}{j}\right) + \sigma^2\frac{\frac{9}{4}}{A(t+2)[A(t+2)+B]}
\end{aligned}$$

Since $A > 0$ and $B > 0$ thus

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=2}^{t+2}\left(1-\frac{\frac{3}{2}}{i}\right)+\sigma^2\sum_{i=2}^{t+1}\frac{\frac{9}{4}}{Ai[Ai+B]}\prod_{j=i+1}^{t+2}\left(1-\frac{\frac{3}{2}}{j}\right)+\sigma^2\frac{\frac{9}{4}}{A(t+2)[A(t+2)+B]} \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=2}^{t+2}\left(1-\frac{\frac{3}{2}}{i}\right)+\sigma^2\sum_{i=2}^{t+1}\frac{\frac{9}{4}}{(Ai)^2}\prod_{j=i+1}^{t+2}\left(1-\frac{\frac{3}{2}}{j}\right)+\sigma^2\frac{\frac{9}{4}}{[A(t+2)]^2}
\end{aligned}$$

We can next use the fact that for any $a \in (1,2)$:

$$\prod_{i=\tau+1}^{t+2}\left(1-\frac{a}{i}\right) \leq \left(\frac{\tau+1}{t+3}\right)^{a}.$$

The bound then becomes

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\prod_{i=2}^{t+2}\left(1-\frac{\frac{3}{2}}{i}\right)+\sigma^2\sum_{i=2}^{t+1}\frac{\frac{9}{4}}{(Ai)^2}\prod_{j=i+1}^{t+2}\left(1-\frac{\frac{3}{2}}{j}\right)+\sigma^2\frac{\frac{9}{4}}{[A(t+2)]^2} \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\sum_{i=2}^{t+1}\frac{\frac{9}{4}}{(Ai)^2}\left(\frac{i+1}{t+3}\right)^{\frac{3}{2}}+\sigma^2\frac{\frac{9}{4}}{[A(t+2)]^2} \\
=\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\sum_{i=2}^{t+2}\frac{\frac{9}{4}}{(Ai)^2}\left(\frac{i+1}{t+3}\right)^{\frac{3}{2}}
\end{aligned}$$

Note that $(i+1)^{\frac{3}{2}} \leq 2i$ for $i = 2, 3, \ldots$, thus

$$\begin{aligned}
&\mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{t+1}\|_2^2\right] \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\frac{\frac{9}{4}}{A^2(t+3)^{\frac{3}{2}}}\sum_{i=2}^{t+2}\frac{(i+1)^{\frac{3}{2}}}{i^2} \\
\leq\ & \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\frac{\frac{9}{2}}{A^2(t+3)^{\frac{3}{2}}}\sum_{i=2}^{t+2}\frac{1}{i^{\frac{1}{2}}}
\end{aligned}$$

finally note that $\sum_{i=2}^{t+2}\frac{1}{i^{\frac{1}{2}}} \leq \int_{1}^{t+2}\frac{1}{x^{\frac{1}{2}}}dx \leq 2(t+3)^{\frac{1}{2}}$. Thus

$$\leq\ \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\frac{9}{A^2(t+3)}$$

substituting $A = 2\xi - 3\gamma(K-1)$ gives

$$=\ \mathbb{E}\left[\sum_{d=1}^{K}\|\boldsymbol{\Delta}_d^{0}\|_2^2\right]\left(\frac{2}{t+3}\right)^{\frac{3}{2}}+\sigma^2\frac{9}{[2\xi-3\gamma(K-1)]^2(t+3)}$$

This leads us to the final theorem.

B. CNNs experiments: details

We compare SGD, Adam, and AM-Adam on the LeNet-5(LeCun et al., 1998) architecture on both MNIST and Fashion-MNIST (Xiao et al., 2017) datasets.

Fashion-MNIST is a dataset of Zalando's article images, consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. We intend Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

We fix the batchsize to 128, and run a hyperparameter grid search for each algorithm and dataset using the following values: weight-learning rates of 2e-M for M=2,3,4,5; batch-wise mu-increments of 1e-2,1e-5, 1e-7; epoch-wise mu-multipliers of 1, 1,1; code learning-rates of 0.1, 1 (note: only weight learning rates are varied for SGD and Adam). SGD was allowed a standard epoch-wise learning rate decay of 0.9. AM-Adam used only one subproblem iteration (both codes and weights) for each minibatch, an initial μ value of 0.01, and a maximum μ value of 1.5. In total, six total grid searches were performed.

For each hyperparameter combination, each algorithm was run on at least 5 initializations, training for 10 epochs on 5/6 of the training dataset. The mean final accuracy on the validation set (the remaining 1/6 of the training dataset) was used to select the best hyperparameters.

Finally, each algorithm with its best hyperparameters on each dataset was used to re-train Lenet-? with N intializations, this time evaluated on the test set. The mean performances are plotted in Figures ?? for MNIST and Figures ?? for Fashion-MNIST.

The winning hyperparameters for Fashion-MNIST are: Adam: LR=0.002 SGD: LR=0.02 AM: weight-LR= 0.002; code-LR= 1.0; batchwise μ-increment=1e-5; epochwise μ-multiplier=1.1

The winning hyperparameters for MNIST are: Adam: LR=0.002 SGD: LR=0.02 AM: weight-LR= 0.002; code-LR= 1.0; batchwise μ-increment=1e-7; epochwise μ-multiplier=1.1

C. RNN experiments: details

C.1. Architecture and AM Adaptation

We also compare SGD, Adam, and AM-Adam on a standard Elman RNN architecture. That is a recurrent unit that, at time t, yields an output z^t and hidden state h^t based on a combination of input x^t and the previous hidden state h^{t-1}, for $t = 1, ..., T$. The equations for the unit are:

$$h^t = \sigma\{\mathbf{U}x^t + \mathbf{W}h^{t-1} + \mathbf{b}\} \tag{14}$$

$$z^t = \mathbf{V}h^t, \tag{15}$$

where $\mathbf{b}$ is a bias, σ is a *tanh* activation function, and $\mathbf{U} \in \mathbb{R}^{d\times 1}, \mathbf{W} \in \mathbb{R}^{d\times d}$, and $\mathbf{V} \in \mathbb{R}^{1\times d}$ are learnable parameter matrices that do not vary with t. Denote with m the length of one sequence element, so $x^t, z^t \in \mathbb{R}^m$. Then let d be the number of hidden units, so $h^t \in \mathbb{R}^d$.

We train this architecture to classify MNIST digits where each image is vectorized and fed to the RNN as a sequence of $T = 784$ pixels (termed "Sequential MNIST" in (Le et al., 2015)). Thus for each t, the input x^t is a single pixel. A final matrix $\mathbf{C}$ is then used to classify the output sequence z^t using the same multinomial loss function as before:

$$\sum_n \mathcal{L}(y_n, ReLU(\mathbf{z}_n), \mathbf{C}), \tag{16}$$

where $\mathbf{z}_n = [z_n^1, ..., z_n^{784}]^{\mathrm{T}}$ is the output sequence for the n^{th} training sample, and $\mathbf{C} \in \mathbb{R}^{10\times 784}$. In summary, the prediction is made only after processing all 784 pixels.

To train this family of architectures using Alt-Min, we introduce two sets of auxiliary variables (codes). First, we introduce a code for each element of the sequence just before input to the activation function:

$$c^t = \mathbf{U}x^t + \mathbf{W}h^{t-1} + \mathbf{b} \tag{17}$$

where c^t is the internal RNN code at time t. Using the "unfolded" interpretation of an RNN, we have introduced a code between each repeated "layer". Second, we treat the output sequence $\mathbf{z}$ as an auxiliary variable in order to break the gradient chain between the loss function and the recurrent unit.

C.2. Experiments

We compare SGD, Adam, and AM-Adam on the Elman RNN architecture with hidden sizes $d = 15$ and $d = 50$ on the Sequential MNIST dataset. We fix the batchsize to 1024, and run a hyperparameter grid search for each algorithm using the following values: weight-learning rates of 5e-M, for M=1,2,3,4,5 (all methods); weight sparsity = 0, 0.01, 0.1 (SGD and Adam); batch-wise mu-increment 1e-M for M=2,3,4; epoch-wise mu-multiplier for 1, 1.1, 1.25, 1.5; mu-max=1, 5. SGD was allowed a standard learning-rate-decay of 0.9. AM-Adam used an initial μ value of 0.01, and used 5 subproblem iterations for both code and weight optimization subproblems.

Note: in an offline hand-tuning search, we determined that weight-sparsity only hurt Alt-Min, so it was not included in official the grid search. Also note that a larger batchsize is used for the RNN experiments because of the relatively strong dependence of the training time on batchsize. This dependence is because for each minibatch, a series of loops though $t = 1, ..., 784$ are required.

For each hyperparameter combination, each algorithm was run on at least 3 initializations, training for 10 epochs on 5/6 of the training dataset. The mean final accuracy on the validation set (the remaining 1/6 of the training dataset) was used to select the best hyperparameters.

Finally, each algorithm with its best hyperparameters on each dataset was used to re-train the Elman RNN with N intializations, this time evaluated on the test set.

The winning hyperparameters for *d=15* are: Adam: learning rate = 0.005, L1=0; SGD: learning rate = 0.05, L1=0; AM-Adam: learning rate = 0.005, max-mu=1, mu-multiplier=1.1, mu-increment=0.01. Results are depicted in Figure 6.

The winning hyperparameters for *d=50* are: Adam: learning rate = 0.005, L1=0.01; SGD: learning rate = 0.005, L1=0; AM-Adam: learning rate = 0.005, max-mu=1, mu-multiplier=1.0, mu-increment=0.0001. Results are depicted in Figure 9.

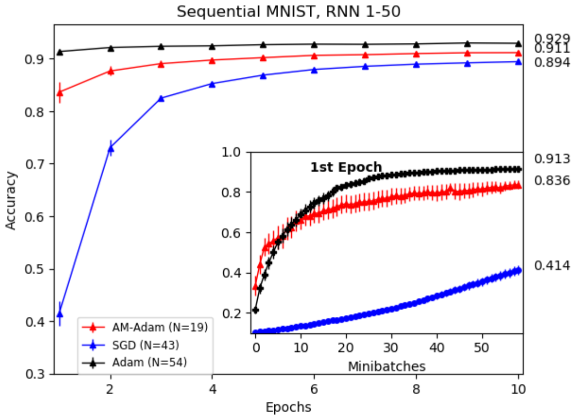

Figure 9. RNN-50, Sequential MNIST.

D. Fully connected networks: details

Performance of the online (i.e., SGD, Adam, AM-Adam, AM-mem) and offline (i.e., AM-Adam-off, AM-mem-off, Taylor) methods are compared on the MNIST and CIFAR-10 datasets for two fully connected network architectures with two identical hidden layers of 100 and 500 units each. We also consider a different architecture with one hidden layer of 300 units for the larger HIGGS dataset. Optimal hyperparameters are reported below for each set of experiments.

D.1. MNIST Experiments

The standard MNIST training dataset is split into a reduced training set (first 50,000 samples) and a validation set (last 10,000 samples) for hyperparameter optimization. More specifically, an iterative bayesian optimization scheme is used to find the optimal learning rates (lr) maximizing classification accuracy on the validation set after 50 epochs of training.

Rather than learning rates, for Taylor's method we optimize the γ_{prod} and γ_{nonlin} parameters. The procedure is repeated for five different weight initializations and for both architectures considered. Table 1 reports hyperparameters yielding the highest accuracy among the 5 weight initializations.

Algorithm	Hidden units per layer	lr	γ_{prod}	γ_{nonlin}
Adam	100	0.0210		
Adam	500	0.0005		
SGD	100	0.2030		
SGD	500	0.1497		
AM-Adam	100	0.1973		
AM-Adam	500	0.1171		
AM-mem	100	0.1737		
AM-mem	500	0.1376		
AM-Adam-off	100	0.5003		
AM-Adam-off	500	0.4834		
AM-mem-off	100	0.4664		
AM-mem-off	500	0.2503		
Taylor	100		582.8	54.15
Taylor	500		444.2	111.7

Table 1. Optimal hyperparameters for fully connected networks on MNIST

D.2. CIFAR-10 Experiments

Similary to what done for the MNIST dataset, we split the standard CIFAR-10 training dataset into a reduced training set (first 40,000 samples) and a validation set (last 10,000 samples) used to evaluate accuracy for hyperparameter optimization. Table 2 reports hyperparameters for all the methods yielding the highest accuracy among the 5 weight initializations. Since not included in the original publication, we do not consider Taylor's method on this dataset.

Algorithm	Hidden units per layer	lr
Adam	100	0.0029
Adam	500	0.0002
SGD	100	0.1500
SGD	500	0.1428
AM-Adam	100	0.1974
AM-Adam	500	0.1011
AM-mem	100	0.1746
AM-mem	500	0.1016
AM-Adam-off	100	0.5000
AM-Adam-off	500	0.4844
AM-mem-off	100	0.2343
AM-mem-off	500	0.2277

Table 2. Optimal hyperparameters for fully connected networks on CIFAR-10

D.3. HIGGS Experiments

For the Higgs experiment, we compare only our best performing AM-Adam online method to Adam and SGD. Also, due to the increased computational costs associated to this dataset, we consider only one weight initialization and replace the bayesian optimization scheme with a simpler grid search. Table 3 reports the hyperparameters yielding the highest accuracy.

Algorithm	Hidden units per layer	lr
Adam	300	0.001
SGD	300	0.050
AM-Adam	300	0.001

Table 3. Hyperparameters used for fully connected networks on HIGGS

D.4. Related Work: ProxProp

As we mentioned in the introduction, a closely related auxiliary methods, called ProxProp, was recently proposed in (Thomas Frerix, 2018). However, there are several importnant differences between ProxProp and our approach. ProxProp only analyzes and experimentally evaluates a batch version, only briefly mentioning in section 4.2.3 that theory is extendable to mini-batch setting, without explicit convergence rates/formal proofs/experiments. Also, an assumption on eigenvalues (from eq. 14 in (Thomas Frerix, 2018)) bounded away from zero is mentioned; however, in flat regions of optimization landscape (often found by solvers like SGD) this condition is not met, as most eigenvalues are close to zero (see, e.g. Chaudhari et al 2016). We believe that our assumptions are less restrictive from that perspective (and convergence in mini-batch setting is formally proven). Further differences include: (1) our formulation involves only one set of auxiliary variables/"codes" (linear z in ProxProp) rather than two (linear and nonlinear), reducing memory footprint (and potentially computing time); (2) ProxProp experiments are limited to batch mode, while we compare batch vs mini-batch vs SGD; (3) ProxProp processes both auxiliary variables and weights sequentially, layer by layer (we process auxiliary variables first, then weights in all layers independently/in parallel), which is important for ProxProp. (4) Finally, we also propose two different mini-batch methods, AM-SGD (closer to ProxProp) and AM-mem, which is very different from ProxProp as. it exploits surrogate objective method of online dictionary learning in (Mairal et al., 2009).

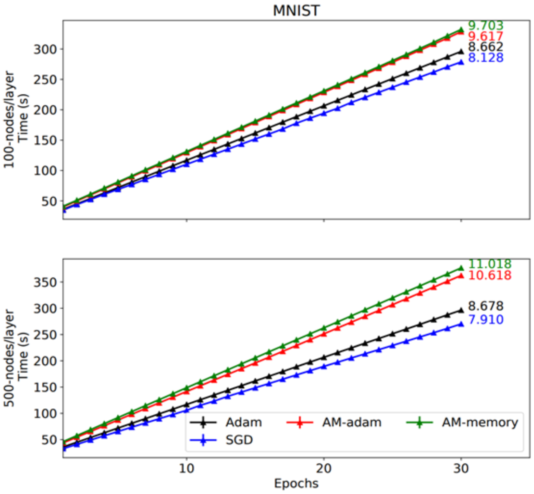

Figure 10. Runtimes on MNIST, fully-connected architecture

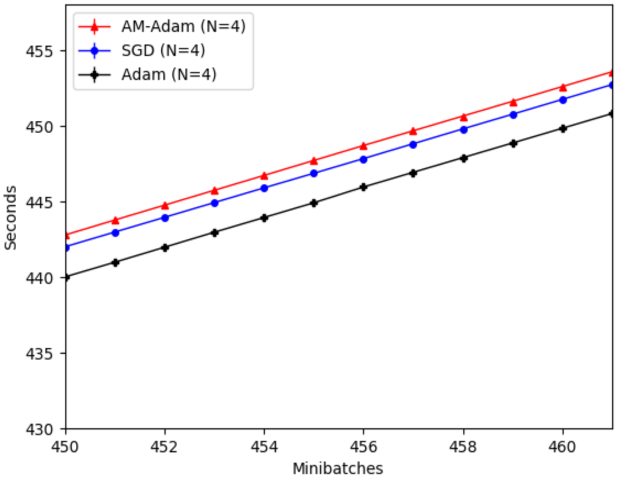

Figure 11. Runtimes on MNIST, LeNet5.

D.5. Computational Efficiency: Runtimes

Runtime results for AM-Adam were quite comparable in most experiments to those of Adam and SGD (see Figures 10 and 11). Runtimes of all methods grew linearly with mini-batches/epochs, and were similar to each other: e.g., for LeNet/MNIST (Figure 11), practically same slope was observed for all methods, and the runtimes were really close (e.g. 440, 442 and 443 seconds for 450 mini-batches for Adam, SGD and AM, respectively). On MNIST, using fully-connected networks (Figure 10), slight increase was observed in the slope of AM versus SGD and Adam, but the times were quite comparable: e.g., at 30 epochs, Adam took 8.7 seconds, while AM-SGD and AM-mem took 9.6 and 9.7 seconds, respectively. Note that we are comparing an implementation of AM which does not yet exploit parallelization; the latter is likely to provide a considerable speedup, similar to the one presented in (Carreira-Perpinan & Wang, 2014).