Supplementary Material

A Proof of Lemma 1

Proof. We know that $-\nabla \ell_{\mathcal{B}}(x)$ is a descent direction iff the following condition holds:

$$\nabla \ell_{\mathcal{B}}(x)^T \nabla \ell(x) > 0. \tag{10}$$

Expanding $\|\nabla \ell_{\mathcal{B}}(x) - \nabla \ell(x)\|^2$ we get

$$\begin{aligned} \|\nabla \ell_{\mathcal{B}}(x)\|^2 + \|\nabla \ell(x)\|^2 - 2\nabla \ell_{\mathcal{B}}(x)^T \nabla \ell(x) &< \|\nabla \ell_{\mathcal{B}}(x)\|^2, \\ \implies -2\nabla \ell_{\mathcal{B}}(x)^T \nabla \ell(x) &< -\|\nabla \ell(x)\|_2^2 \leq 0, \end{aligned}$$

which is always true for a descent direction (10). □

B Proof of Theorem 1

Proof. Let $\bar{z} = \mathbb{E}[z]$ be the mean of z. Given the current iterate x, we assume that the batch $\mathcal{B}$ is sampled uniformly with replacement from p. We then have the following bound:

$$\begin{aligned} \|\nabla f(x; z) - \nabla \ell(x)\|^2 &\leq 2\|\nabla f(x; z) - \nabla f(x, \bar{z})\|^2 + 2\|\nabla f(x, \bar{z}) - \nabla \ell(x)\|^2 \\ &\leq 2L_z^2 \|z - \bar{z}\|^2 + 2\|\nabla f(x, \bar{z}) - \nabla \ell(x)\|^2 \\ &= 2L_z^2 \|z - \bar{z}\|^2 + 2\|\mathbb{E}_z[\nabla f(x, \bar{z}) - \nabla f(x, z)]\|^2 \\ &\leq 2L_z^2 \|z - \bar{z}\|^2 + 2\mathbb{E}_z\|\nabla f(x, \bar{z}) - \nabla f(x, z)\|^2 \\ &\leq 2L_z^2 \|z - \bar{z}\|^2 + 2L_z^2 \mathbb{E}_z\|\bar{z} - z\|^2 \\ &= 2L_z^2 \|z - \bar{z}\|^2 + 2L_z^2 \operatorname{Tr} \operatorname{Var}_z(z), \end{aligned}$$

where the first inequality uses the property $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the second and fourth inequalities use Assumption 1, and the third line uses Jensen's inequality. This bound is *uniform* in x. We then have

$$\begin{aligned} \mathbb{E}_z\|\nabla f(x; z) - \nabla \ell(x)\|^2 &\leq 2L_z^2 \mathbb{E}_z\|z - \bar{z}\|^2 + 2L_z^2 \operatorname{Tr} \operatorname{Var}_z(z) \\ &= 4L_z^2 \operatorname{Tr} \operatorname{Var}_z(z) \end{aligned}$$

uniformly for all x. The result follows from the observation that

$$\mathbb{E}_{\mathcal{B}}\|\nabla f_{\mathcal{B}}(x) - \nabla \ell(x)\|^2 = \frac{1}{|\mathcal{B}|} E_z\|\nabla f(x; z) - \nabla \ell(x)\|^2.$$

□

C Proof of Lemma 2

Proof. From (5) and Assumption 2 we get

$$\ell(x_{t+1}) \leq \ell(x_t) - \alpha {g_t}^T \nabla \ell(x_t) + \frac{L\alpha^2}{2}\|g_t\|^2.$$

Taking expectation with respect to the batch $\mathcal{B}_t$ and conditioning on x_t, we get

$$\begin{aligned} \mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq& \ell(x_t) - \ell(x^\star) - \alpha \mathbb{E}[g_t]^T \nabla \ell(x_t) + \frac{L\alpha^2}{2}\mathbb{E}\|g_t\|^2 \\ =& \ell(x_t) - \ell(x^\star) - \alpha\|\nabla \ell(x_t)\|^2 + \frac{L\alpha^2}{2}(\|\nabla \ell(x_t)\|^2 + \mathbb{E}\|e_t\|^2 + \mathbb{E}[e_t]^T \nabla \ell(x_t)) \\ =& \ell(x_t) - \ell(x^\star) - \big(\alpha - \frac{L\alpha^2}{2}\big)\|\nabla \ell(x_t)\|^2 + \frac{L\alpha^2}{2}\mathbb{E}\|e_t\|^2 \\ \leq& \Big(1 - 2\mu\big(\alpha - \frac{L\alpha^2}{2}\big)\Big)(\ell(x_t) - \ell(x^\star)) + \frac{L\alpha^2}{2}\mathbb{E}\|e_t\|^2, \end{aligned}$$

where the second inequality follows from Assumption 3. Taking expectation, the result follows. □

D Proof of Theorem 2

Proof. We begin by applying the reverse triangle inequality to (4) to get

$$(1-\theta)\mathbb{E}\|\nabla \ell_{\mathcal{B}}(x)\| \leq \mathbb{E}\|\nabla \ell(x)\|$$

which applied to (4) yields

$$\frac{\theta^2}{(1-\theta)^2}\mathbb{E}\|\nabla \ell(x_t)\|^2 \geq \mathbb{E}\|\nabla \ell_{\mathcal{B}}(x_t) - \nabla \ell(x_t)\|^2 = \mathbb{E}\|e_t\|^2. \tag{11}$$

Now, we apply (11) to the result in Lemma 2 to get

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \mathbb{E}[\ell(x_t) - \ell(x^\star)] - \big(\alpha - \frac{L\alpha^2\beta}{2}\big)\mathbb{E}\|\nabla \ell(x_t)\|^2,$$

where $\beta = \frac{\theta^2+(1-\theta)^2}{(1-\theta)^2} \geq 1$. Assuming $\alpha - \frac{L\alpha^2\beta}{2} \geq 0$, we can apply Assumption 3 to write

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \Big(1 - 2\mu\big(\alpha - \frac{L\alpha^2\beta}{2}\big)\Big)\mathbb{E}[\ell(x_t) - \ell(x^\star)],$$

which proves the theorem. Note that $\max_\alpha\{\alpha - \frac{L\alpha^2\beta}{2}\} = \frac{1}{2L\beta}$, and $\mu \leq L$. It follows that

$$0 \leq \Big(1 - 2\mu\big(\alpha - \frac{L\alpha^2\beta}{2}\big)\Big) < 1.$$

The second result follows immediately. □

E Proof of Theorem 3

Proof. Applying the reverse triangle inequality to (4) and using Lemma 2 we get, as in Theorem 2:

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \mathbb{E}[\ell(x_t) - \ell(x^\star)] - \big(\alpha - \frac{L\alpha^2\beta}{2}\big)\mathbb{E}\|\nabla \ell(x_t)\|^2, \tag{12}$$

where $\beta = \frac{\theta^2+(1-\theta)^2}{(1-\theta)^2} \geq 1$.

We will show that the backtracking condition in (7) is satisfied whenever $0 < \alpha_t \leq \frac{1}{\beta L}$. First notice that:

$$0 < \alpha_t \leq \frac{1}{\beta L} \quad \implies \quad -\alpha_t + \frac{L\alpha_t^2\beta}{2} \leq -\frac{\alpha_t}{2}.$$

Thus, we can rewrite (12) as

$$\begin{aligned}\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] &\leq \mathbb{E}[\ell(x_t) - \ell(x^\star)] - \frac{\alpha_t}{2}\mathbb{E}\|\nabla \ell(x_t)\|^2 \\ &\leq \mathbb{E}[\ell(x_t) - \ell(x^\star)] - c\alpha_t\mathbb{E}\|\nabla \ell(x_t)\|^2,\end{aligned}$$

where $0 < c \leq 0.5$. Thus, the backtracking line search condition (7) is satisfied whenever $0 < \alpha_t \leq \frac{1}{L\beta}$.

Now we know that either $\alpha_t = \alpha_0$ (the initial stepsize), or $\alpha_t \geq \frac{1}{2\beta L}$, where the stepsize is decreased by a factor of 2 each time the backtracking condition fails. Thus, we can rewrite the above as

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \mathbb{E}[\ell(x_t) - \ell(x^\star)] - c\min\Big(\alpha_0, \frac{1}{2\beta L}\Big)\mathbb{E}\|\nabla \ell(x_t)\|^2.$$

Using Assumption 3 we get

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \left(1 - 2c\mu\min\left(\alpha_0, \frac{1}{2\beta L}\right)\right)\mathbb{E}[\ell(x_t) - \ell(x^\star)].$$

Assuming we start off the stepsize at a large value such that $\min(\alpha_0, \frac{1}{2\beta L}) = \frac{1}{2\beta L}$, we can rewrite this as:

$$\mathbb{E}[\ell(x_{t+1}) - \ell(x^\star)] \leq \Big(1 - \frac{c\mu}{\beta L}\Big)\mathbb{E}[\ell(x_t) - \ell(x^\star)].$$

□

F Algorithmic Details for Automated Big Batch Methods

The complete details of the backtracking Armijo line search we used with big batch SGD are explained in detail in Algorithm 2. The adaptive stepsize method using Barzilai-Borwein curvature estimates with big batch SGD is presented in Algorithm 3.

Algorithm 2 Big batch SGD: backtracking line search

initialize starting pt. x_0, initial stepsize α, initial batch size $K > 1$, batch size increment δ_k, backtracking line search parameter c, flag $F = 0$
while not converged **do**
  Draw random batch with size $|\mathcal{B}| = K$
  Calculate $V_{\mathcal{B}}$ and $\nabla \ell_{\mathcal{B}}(x_t)$ using (6)
  while $\|\nabla \ell_{\mathcal{B}}(x_t)\|^2 \leq V_{\mathcal{B}}/K$ **do**
    Increase batch size $K \leftarrow K + \delta_K$
    Sample more gradients
    Update $V_{\mathcal{B}}$ and $\nabla \ell_{\mathcal{B}}(x_t)$
    Set flag $F = 1$
  end while
  if flag $F == 1$ **then**
    $\alpha \leftarrow \alpha * 2$
    Reset flag $F = 0$
  end if
  while $\ell_{\mathcal{B}}(x_t - \alpha \nabla \ell_{\mathcal{B}}(x_t)) > \ell_{\mathcal{B}}(x_t) - c\alpha_t \|\nabla \ell_{\mathcal{B}}(x_t)\|^2$ **do**
    $\alpha \leftarrow \alpha/2$
  end while
  $x_{t+1} = x_t - \alpha \nabla \ell_{\mathcal{B}}(x_t)$
end while

Algorithm 3 Big batch SGD: with BB stepsizes

initialize starting pt. x, initial stepsize α, initial batch size $K > 1$, batch size increment δ_k, backtracking line search parameter c
while not converged **do**
  Draw random batch with size $|\mathcal{B}| = K$
  Calculate $V_{\mathcal{B}}$ and $G_{\mathcal{B}} = \nabla \ell_{\mathcal{B}}(x)$ using (6)
  while $\|G_{\mathcal{B}}\|^2 \leq V_{\mathcal{B}}/K$ **do**
    Increase batch size $K \leftarrow K + \delta_K$
    Sample more gradients
    Update $V_{\mathcal{B}}$ and $G_{\mathcal{B}}$
  end while
  while $\ell_{\mathcal{B}}(x - \alpha \nabla \ell_{\mathcal{B}}(x)) > \ell_{\mathcal{B}}(x) - c\alpha \|\nabla \ell_{\mathcal{B}}(x)\|^2$ **do**
    $\alpha \leftarrow \alpha/2$
  end while
  $x \leftarrow x - \alpha \nabla \ell_{\mathcal{B}}(x)$
  if $K < N$ **then**
    Calculate $\tilde{\alpha} = (1 - V_B/(K\|G_{\mathcal{B}}\|^2))/\nu$ using (8) and (9)
  else
    Calculate $\tilde{\alpha} = 1/\nu$ using (9)
  end if
  Stepsize smoothing: $\alpha \leftarrow \alpha(1 - K/N) + \tilde{\alpha} K/N$
  while $\ell_{\mathcal{B}}(x - \alpha \nabla \ell_{\mathcal{B}}(x)) > \ell_{\mathcal{B}}(x) - c\alpha \|\nabla \ell_{\mathcal{B}}(x)\|^2$ **do**
    $\alpha \leftarrow \alpha/2$
  end while
  $x \leftarrow x - \alpha \nabla \ell_{\mathcal{B}}(x)$
end while

G Derivation of Adaptive Step Size

Here we present the complete derivation of the adaptive stepsizes presented in Section 5. Our derivation follows the classical adaptive Barzilai and Borwein (1988) (BB) method. The BB methods fits a quadratic model to the objective on each iteration, and a stepsize is proposed that is optimal for the local quadratic model (Goldstein et al., 2014). To derive the analog of the BB method for stochastic problems, we consider quadratic approximations of the form $\ell(x) = \mathbb{E}_\theta f(x,\theta)$, where we define $f(x,\theta) = \frac{\nu}{2}\|x-\theta\|^2$ with $\theta \sim \mathcal{N}(x^\star, \sigma^2 I)$.

We derive the optimal stepsize for this. We can rewrite the quadratic approximation as

$$\ell(x) = \mathbb{E}_\theta f(x,\theta) = \frac{\nu}{2}\mathbb{E}_\theta\|x-\theta\|^2 = \frac{\nu}{2}[x^Tx - 2x^Tx^\star - \mathbb{E}(\theta^T\theta)] = \frac{\nu}{2}\big(\|x-x^\star\|^2 + d\sigma^2\big),$$

since we can write

$$\mathbb{E}(\theta^T\theta) = \mathbb{E}\sum_{i=1}^d \theta_i^2 = \sum_{i=1}^d \mathbb{E}\theta_i^2 = \sum_{i=1}^d (x_i^\star)^2 + \sigma^2 = \|x^\star\|^2 + d\sigma^2.$$

Further, notice that:

$$\begin{aligned}\mathbb{E}_\theta[\nabla\ell(x)] &= \mathbb{E}_\theta[\nu(x-\theta)] = \nu(x-x^\star), \quad \text{and}\\ \operatorname{Tr}\operatorname{Var}_\theta[\nabla\ell(x)] &= \mathbb{E}_\theta[\nu^2(x-\theta)^T(x-\theta)] - \nu^2(x-x^\star)^T(x-x^\star) = d\nu^2\sigma^2.\end{aligned}$$

Using the quadratic approximation, we can rewrite the update for big batch SGD as follows:

$$x_{t+1} = x_t - \alpha_t\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\nu(x_t-\theta_i) = (1-\nu\alpha_t)x_t + \frac{\nu\alpha_t}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\theta_i = (1-\nu\alpha_t)x_t + \nu\alpha_t x^\star + \frac{\nu\sigma\alpha_t}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\xi_i,$$

where we write $\theta_i = x^\star + \sigma\xi_i$ with $\xi_i \sim \mathcal{N}(0,1)$. Thus, the expected value of the function is:

$$\begin{aligned}\mathbb{E}[\ell(x_{t+1})] &= \mathbb{E}_\xi\left[\ell\left((1-\nu\alpha_t)x_t + \nu\alpha_t x^\star + \frac{\nu\sigma\alpha_t}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\xi_i\right)\right]\\ &= \frac{\nu}{2}\mathbb{E}_\xi\left[\left\|(1-\nu\alpha_t)(x_t-x^\star) + \frac{\nu\sigma\alpha_t}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\xi_i\right\|^2 + d\sigma^2\right]\\ &= \frac{\nu}{2}\left(\|(1-\nu\alpha_t)(x_t-x^\star)\|^2 + \mathbb{E}_\xi\left\|\frac{\nu\sigma\alpha_t}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\xi_i\right\|^2 + d\sigma^2\right)\\ &= \frac{\nu}{2}\left(\|(1-\nu\alpha_t)(x_t-x^\star)\|^2 + (1+\frac{\nu^2\alpha_t^2}{|\mathcal{B}|})d\sigma^2\right).\end{aligned}$$

Minimizing $\mathbb{E}[\ell(x_{t+1})]$ w.r.t. α_t we get:

$$\begin{aligned}\alpha_t &= \frac{1}{\nu}\cdot\frac{\big\|\mathbb{E}[\nabla\ell_{\mathcal{B}_t}(x_t)]\big\|^2}{\big\|\mathbb{E}[\nabla\ell_{\mathcal{B}_t}(x_t)]\big\|^2 + \frac{1}{|\mathcal{B}_t|}\operatorname{Tr}\operatorname{Var}[\nabla f(x_t)]}\\ &= \frac{1}{\nu}\cdot\frac{\mathbb{E}\big\|\nabla\ell_{\mathcal{B}_t}(x_t)\big\|^2 - \frac{1}{|\mathcal{B}_t|}\operatorname{Tr}\operatorname{Var}[\nabla f(x_t)]}{\mathbb{E}\big\|\nabla\ell_{\mathcal{B}_t}(x_t)\big\|^2}\\ &= \frac{1}{\nu}\cdot\left(1 - \frac{\frac{1}{|\mathcal{B}_t|}\operatorname{Tr}\operatorname{Var}[\nabla f(x_t)]}{\mathbb{E}\big\|\nabla\ell_{\mathcal{B}_t}(x_t)\big\|^2}\right)\\ &\geq \frac{1-\theta^2}{\nu}.\end{aligned}$$

Here ν denotes the curvature of the quadratic approximation. Thus, the optimal stepsize for big batch SGD is the optimal stepsize for deterministic gradient descent scaled down by at most $1-\theta^2$.

H Details of Neural Network Experiments and Additional Results

Here we present details of the ConvNet and exact hyper-parameters used for training the neural network models for our experiments.

We train a convolutional neural network (LeCun et al., 1998) (ConvNet) to classify three benchmark image datasets: CIFAR-10 (Krizhevsky and Hinton, 2009), SVHN (Netzer et al., 2011) and MNIST (LeCun et al., 1998). The ConvNet used in our experiments is composed of 4 layers, excluding the input layer. We use 32×32 pixel images as input. The first layer of the ConvNet contains $16 \times 3 \times 3$, and the second layer contains $256 \times 3 \times 3$ filters. The third and fourth layers are fully connected (LeCun et al., 1998) with 256 and 10 outputs respectively. Each layer except the last one is followed by a ReLu non-linearity (Krizhevsky et al., 2012) and a max pooling stage (Ranzato et al., 2007) of size 2×2. This ConvNet has over 4.3 million weights.

To compare against fine-tuned SGD, we used a comprehensive grid search on the stepsize schedule to identify optimal parameters (up to a factor of 2 accuracy). For CIFAR10, the stepsize starts from 0.5 and is divided by 2 every 5 epochs with 0 stepsize decay. For SVHN, the stepsize starts from 0.5 and is divided by 2 every 5 epochs with $1e{-}05$ learning rate decay. For MNIST, the learning rate starts from 1 and is divided by 2 every 3 epochs with 0 stepsize decay. All algorithms use a momentum parameter of 0.9, and SGD and AdaDelta use mini-batches of size 128.

Fixed stepsize methods use the default decay rule of the *Torch* library: $\alpha_t = \alpha_0/(1 + 10^{-7}t)$, where α_0 was chosen to be the stepsize used in the fine-tuned experiments. We also tune the hyper-parameter ρ in the Adadelta algorithm, and we found 0.9, 0.9 and 0.8 to be best-performing parameters for CIFAR10, SVHN and MNIST respectively.

Figure 3 shows the change in the loss function over time for the same neural network experiments shown in the main paper (on CIFAR-10, SVHN and MNIST).

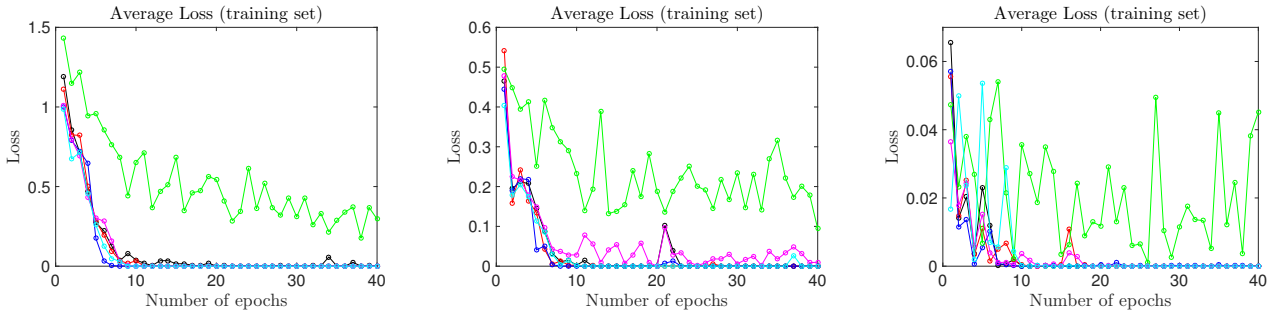

Figure 3: Neural Network Experiments. Figure shows the change in the loss function for CIFAR-10, SVHN, and MNIST (left to right).