Appendix: Stochastic Variance Reduction for Nonconvex Optimization

A. Nonconvex SGD: Convergence Rate

Proof of Theorem 1

Theorem. *Suppose $f \in \mathcal{F}_n$ has σ-bounded gradients; let $\eta_t = \eta = c/\sqrt{T}$ where $c = \sqrt{\frac{2(f(x^0)-f(x^*))}{L\sigma^2}}$, and x^* is an optimal solution to (1). Then, the iterates of* SGD *satisfy*

$$\min_{0\le t\le T-1} \mathbb{E}[\|\nabla f(x^t)\|^2] \le \sqrt{\frac{2(f(x^0)-f(x^*))L}{T}}\sigma.$$

Proof. We include the proof here for completeness. Please refer to (Ghadimi & Lan, 2013) for a more general result.

The iterates of SGD satisfy the following bound:

$$\begin{aligned}\mathbb{E}[f(x^{t+1})] &\le \mathbb{E}[f(x^t) + \langle \nabla f(x^t), x^{t+1}-x^t\rangle \\ &\qquad + \tfrac{L}{2}\|x^{t+1}-x^t\|^2] \qquad (4)\\ &\le \mathbb{E}[f(x^t)] - \eta_t\mathbb{E}[\|\nabla f(x^t)\|^2] + \tfrac{L\eta_t^2}{2}\mathbb{E}[\|\nabla f_{i_t}(x^t)\|^2]\\ &\le \mathbb{E}[f(x^t)] - \eta_t\mathbb{E}[\|\nabla f(x^t)\|^2] + \tfrac{L\eta_t^2}{2}\sigma^2. \qquad (5)\end{aligned}$$

The first inequality follows from Lipschitz continuity of ∇f. The second inequality follows from the update of SGD and since $\mathbb{E}_{i_t}[\nabla f_{i_t}(x^t)] = \nabla f(x^t)$ (unbiasedness of the stochastic gradient). The last step uses our assumption on gradient boundedness. Rearranging Equation (5) we obtain

$$\mathbb{E}[\|\nabla f(x^t)\|^2] \le \tfrac{1}{\eta_t}\mathbb{E}[f(x^t) - f(x^{t+1})] + \tfrac{L\eta_t}{2}\sigma^2. \quad (6)$$

Summing Equation (6) from $t = 0$ to $T-1$ and using that η_t is fixed η we obtain

$$\begin{aligned}\min_t \mathbb{E}[\|\nabla f(x^t)\|^2] &\le \tfrac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|f(x^t)\|^2]\\ &\le \tfrac{1}{T\eta}\mathbb{E}[f(x^0) - f(x^T)] + \tfrac{L\eta}{2}\sigma^2\\ &\le \tfrac{1}{T\eta}(f(x^0) - f(x^*)) + \tfrac{L\eta}{2}\sigma^2\\ &\le \tfrac{1}{\sqrt{T}}\Big(\tfrac{1}{c}\big(f(x^0) - f(x^*)\big) + \tfrac{Lc}{2}\sigma^2\Big).\end{aligned}$$

The first step holds because the minimum is less than the average. The second and third steps are obtained from Equation (6) and the fact that $f(x^*) \le f(x^T)$, respectively. The final inequality follows upon using $\eta = c/\sqrt{T}$. By setting

$$c = \sqrt{\frac{2(f(x^0) - f(x^*))}{L\sigma^2}}$$

in the above inequality, we get the desired result. □

B. Nonconvex SVRG

In this section, we provide the proofs of the results for nonconvex SVRG. We first start with few useful lemmas and then proceed towards the main results.

Lemma 1. *Let $f \in \mathcal{F}_n$. For $c_t, c_{t+1}, \beta_t > 0$, suppose we have the following:*

$$c_t = c_{t+1}(1 + \eta_t\beta_t + 2\eta_t^2L^2) + \eta_t^2L^3.$$

Let η_t, β_t and c_{t+1} be chosen such that $\Gamma_t > 0$ (in Equation (3)). The iterate x_t^{s+1} in Alg. 1 satisfy the bound:

$$\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \le \frac{R_t^{s+1} - R_{t+1}^{s+1}}{\Gamma_t},$$

where $R_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_t\|x_t^{s+1} - \tilde{x}^s\|^2]$ for all $0 \le s \le S-1$.

Proof. Since f is L-smooth we have

$$\begin{aligned}\mathbb{E}[f(x_{t+1}^{s+1})] \le \mathbb{E}[f(x_t^{s+1}) &+ \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1}\rangle\\ &+ \tfrac{L}{2}\|x_{t+1}^{s+1} - x_t^{s+1}\|^2].\end{aligned}$$

Using the SVRG update in Alg. 1 and its unbiasedness, the right hand side above is further upper bounded by

$$\mathbb{E}[f(x_t^{s+1}) - \eta_t\|\nabla f(x_t^{s+1})\|^2 + \tfrac{L\eta_t^2}{2}\|v_t^{s+1}\|^2]. \quad (7)$$

Consider now the Lyapunov function

$$R_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + c_t\|x_t^{s+1} - \tilde{x}^s\|^2].$$

For bounding it we will require the following:

$$\begin{aligned}&\mathbb{E}[\|x_{t+1}^{s+1} - \tilde{x}^s\|^2] = \mathbb{E}[\|x_{t+1}^{s+1} - x_t^{s+1} + x_t^{s+1} - \tilde{x}^s\|^2]\\ &= \mathbb{E}[\|x_{t+1}^{s+1} - x_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2\\ &\qquad + 2\langle x_{t+1}^{s+1} - x_t^{s+1}, x_t^{s+1} - \tilde{x}^s\rangle]\\ &= \mathbb{E}[\eta_t^2\|v_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2]\\ &\qquad - 2\eta_t\mathbb{E}[\langle \nabla f(x_t^{s+1}), x_t^{s+1} - \tilde{x}^s\rangle]\\ &\le \mathbb{E}[\eta_t^2\|v_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2]\\ &+ 2\eta_t\mathbb{E}\left[\tfrac{1}{2\beta_t}\|\nabla f(x_t^{s+1})\|^2 + \tfrac{1}{2}\beta_t\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \qquad (8)\end{aligned}$$

The second equality follows from the unbiasedness of the update of SVRG. The last inequality follows from a simple application of Cauchy-Schwarz and Young's inequality. Plugging Equation (7) and Equation (8) into R_{t+1}^{s+1}, we obtain the following bound:

$$\begin{aligned}R_{t+1}^{s+1} &\le \mathbb{E}[f(x_t^{s+1}) - \eta_t\|\nabla f(x_t^{s+1})\|^2 + \tfrac{L\eta_t^2}{2}\|v_t^{s+1}\|^2]\\ &\quad + \mathbb{E}[c_{t+1}\eta_t^2\|v_t^{s+1}\|^2 + c_{t+1}\|x_t^{s+1} - \tilde{x}^s\|^2]\\ &\quad + 2c_{t+1}\eta_t\mathbb{E}\left[\tfrac{1}{2\beta_t}\|\nabla f(x_t^{s+1})\|^2 + \tfrac{1}{2}\beta_t\|x_t^{s+1} - \tilde{x}^s\|^2\right]\\ &\le \mathbb{E}[f(x_t^{s+1}) - \left(\eta_t - \tfrac{c_{t+1}\eta_t}{\beta_t}\right)\|\nabla f(x_t^{s+1})\|^2\\ &\qquad + \left(\tfrac{L\eta_t^2}{2} + c_{t+1}\eta_t^2\right)\mathbb{E}[\|v_t^{s+1}\|^2]\\ &\qquad + (c_{t+1} + c_{t+1}\eta_t\beta_t)\,\mathbb{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \qquad (9)\end{aligned}$$

To further bound this quantity, we use Lemma 3 to bound $\mathbb{E}[\|v_t^{s+1}\|^2]$, so that upon substituting it in Equation (9), we see that

$$\begin{aligned}
&R_{t+1}^{s+1} \leq \mathbb{E}[f(x_t^{s+1})] \\
&- \left(\eta_t - \tfrac{c_{t+1}\eta_t}{\beta_t} - \eta_t^2 L - 2c_{t+1}\eta_t^2\right) \mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \\
&+ \left[c_{t+1}\left(1 + \eta_t\beta_t + 2\eta_t^2 L^2\right) + \eta_t^2 L^3\right] \mathbb{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\
&\leq R_t^{s+1} - \left(\eta_t - \tfrac{c_{t+1}\eta_t}{\beta_t} - \eta_t^2 L - 2c_{t+1}\eta_t^2\right)\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2].
\end{aligned}$$

The second inequality follows from the definition of c_t and R_t^{s+1}, thus concluding the proof. □

Proof of Theorem 2

Theorem. *Let $f \in \mathcal{F}_n$. Let $c_m = 0$, $\eta_t = \eta > 0$, $\beta_t = \beta > 0$, and $c_t = c_{t+1}(1 + \eta\beta + 2\eta^2 L^2) + \eta^2 L^3$ such that $\Gamma_t > 0$ for $0 \leq t \leq m-1$. Define the quantity $\gamma_n := \min_t \Gamma_t$. Further, let $p_i = 0$ for $0 \leq i < m$ and $p_m = 1$, and let T be a multiple of m. Then for the output x_a of Algorithm 1 we have*

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \leq \frac{f(x^0) - f(x^*)}{T\gamma_n},$$

where x^* is an optimal solution to (1).

Proof. Since $\eta_t = \eta$ for $t \in \{0, \ldots, m-1\}$, using Lemma 1 and telescoping the sum, we obtain

$$\sum\nolimits_{t=0}^{m-1} \mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{R_0^{s+1} - R_m^{s+1}}{\gamma_n}.$$

This inequality in turn implies that

$$\sum\nolimits_{t=0}^{m-1} \mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{\mathbb{E}[f(\tilde{x}^s) - f(\tilde{x}^{s+1})]}{\gamma_n}, \quad (10)$$

where we used that $R_m^{s+1} = \mathbb{E}[f(x_m^{s+1})] = \mathbb{E}[f(\tilde{x}^{s+1})]$ (since $c_m = 0$, $p_m = 1$, and $p_i = 0$ for $i < m$), and that $R_0^{s+1} = \mathbb{E}[f(\tilde{x}^s)]$ (since $x_0^{s+1} = \tilde{x}^s$, as $p_m = 1$ and $p_i = 0$ for $i < m$). Now sum over all epochs to obtain

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1} \mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \leq \frac{f(x^0) - f(x^*)}{T\gamma_n}. \quad (11)$$

The above inequality used the fact that $\tilde{x}^0 = x^0$. Using the above inequality and the definition of x_a in Algorithm 1, we obtain the desired result. □

Proof of Theorem 3

Theorem. *Suppose $f \in \mathcal{F}_n$. Let $\eta = \mu_0/(Ln^\alpha)$ ($0 < \mu_0 < 1$ and $0 < \alpha \leq 1$), $\beta = L/n^{\alpha/2}$, $m = \lfloor n^{3\alpha/2}/(3\mu_0)\rfloor$ and T is some multiple of m. Then there exists universal constants $\mu_0, \nu > 0$ such that we have the following: $\gamma_n \geq \frac{\nu}{Ln^\alpha}$ in Theorem 2 and*

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \leq \frac{Ln^\alpha[f(x^0) - f(x^*)]}{T\nu},$$

where x^* is an optimal solution to the problem in (1) and x_a is the output of Algorithm 1.

Proof. For our analysis, we will require an upper bound on c_0. We observe that $c_0 = \frac{\mu_0^2 L}{n^{2\alpha}} \frac{(1+\theta)^m - 1}{\theta}$ where $\theta = 2\eta^2 L^2 + \eta\beta$. This is obtained using the relation $c_t = c_{t+1}(1 + \eta\beta + 2\eta^2 L^2) + \eta^2 L^3$ and the fact that $c_m = 0$. Using the specified values of β and η we have

$$\theta = 2\eta^2 L^2 + \eta\beta = \frac{2\mu_0^2}{n^{2\alpha}} + \frac{\mu_0}{n^{3\alpha/2}} \leq \frac{3\mu_0}{n^{3\alpha/2}}.$$

The above inequality follows since $\mu_0 \leq 1$ and $n \geq 1$. Using the above bound on θ, we get

$$\begin{aligned}
c_0 &= \frac{\mu_0^2 L}{n^{2\alpha}} \frac{(1+\theta)^m - 1}{\theta} = \frac{\mu_0 L((1+\theta)^m - 1)}{2\mu_0 + n^{\alpha/2}} \\
&\leq \frac{\mu_0 L((1 + \frac{3\mu_0}{n^{3\alpha/2}})^{\lfloor n^{3\alpha/2}/(3\mu_0)\rfloor} - 1)}{2\mu_0 + n^{\alpha/2}} \\
&\leq n^{-\alpha/2}(\mu_0 L(e-1)), \qquad (12)
\end{aligned}$$

wherein the second inequality follows upon noting that $(1 + \frac{1}{l})^l$ is increasing for $l > 0$ and $\lim_{l\to\infty}(1 + \frac{1}{l})^l = e$ (here e is the Euler's number). Now we can lower bound γ_n, as

$$\begin{aligned}
\gamma_n &= \min_t\left(\eta - \tfrac{c_{t+1}\eta}{\beta} - \eta^2 L - 2c_{t+1}\eta^2\right) \\
&\geq \left(\eta - \tfrac{c_0\eta}{\beta} - \eta^2 L - 2c_0\eta^2\right) \geq \frac{\nu}{Ln^\alpha},
\end{aligned}$$

where $\nu > 0$ is a universal constant. The first inequality holds since c_t decreases with t. The second inequality holds since (a) c_0/β is upper bounded by $\mu_0(e-1)$ (follows from Equation (12)), (b) $\eta^2 L \leq \mu_0\eta$ and (c) $2c_0\eta^2 \leq 2\mu_0^2(e-1)\eta$ (follows from Equation (12)). By choosing a universal constant μ_0 appropriately, one can ensure that $\gamma_n \geq \nu/(Ln^\alpha)$ for some universal constant ν. For example, choosing $\mu_0 = 1/4$, we have $\gamma_n \geq \nu/(Ln^\alpha)$ with $\nu = 1/40$. Substituting the above lower bound in Equation (11), we obtain the desired result. □

Proof of Corollary 2

Corollary. *Suppose $f \in \mathcal{F}_n$. Then the IFO complexity of Alg. 1 (with parameters from Thm. 3) for achieving an ϵ-accurate solution is:*

$$\textit{IFO calls} = \begin{cases} O\left(n + (n^{1-\frac{\alpha}{2}}/\epsilon)\right), & \text{if } \alpha < 2/3, \\ O\left(n + (n^\alpha/\epsilon)\right), & \text{if } \alpha \geq 2/3. \end{cases}$$

Proof. This result follows from Theorem 3 and the fact that $m = \lfloor n^{3\alpha/2}/(3\mu_0) \rfloor$. Suppose $\alpha < 2/3$, then $m = o(n)$ (little-o notation). However, n IFO calls are invested in calculating the average gradient at the end of each epoch. Thus, $O(n)$ IFO calls are made for every m (inner) iterations of the algorithm. Using this relationship, we get an IFO complexity of $O\big(n + (n^{1-\frac{\alpha}{2}}/\epsilon)\big)$ in this case.

On the other hand, when $\alpha \geq 2/3$, the total number of IFO calls made by Alg. 1 in each epoch is $\Omega(n)$ since $m = \lfloor n^{3\alpha/2}/(3\mu_0) \rfloor$. Hence, the oracle calls required for calculating the average gradient (per epoch) is of lower order, leading to $O\big(n + (n^{\alpha}/\epsilon)\big)$ IFO calls. □

C. GD-SVRG: Convergence Rate

Proof of Theorem 4

Theorem. *Suppose $f \in \mathcal{F}_n$ is τ-gradient dominated ($\tau > n^{1/3}$). Then, the iterates of Algorithm 2 with $T = \lceil 2L\tau n^{2/3}/\nu_1 \rceil$, $m = \lfloor n/(3\mu_1) \rfloor$, $\eta_t = \mu_1/(Ln^{2/3})$ for $0 \leq t < m$, $p_m = 1$ and $p_i = 0$ for $0 \leq i < m$ satisfy*

$$\mathbb{E}[\|\nabla f(x^k)\|^2] \leq 2^{-k}[\|\nabla f(x^0)\|^2],$$
$$\mathbb{E}[f(x^k) - f(x^*)] \leq 2^{-k}[f(x^0) - f(x^*)].$$

Here μ_1 and ν_1 are the constants used in Corollary 3.

Proof. Note that Algorithm 2 uses SVRG as a subroutine. Using Corollary 3 (SVRG result), we observe that the iterates of Algorithm 2 satisfy the following:

$$\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{Ln^{2/3}\mathbb{E}[f(x^{k-1}) - f(x^*)]}{T\nu_1}.$$

Substituting the specified value of T in the above inequality, we have the following:

$$\begin{aligned}\mathbb{E}[\|\nabla f(x^k)\|^2] &\leq \frac{1}{2\tau}\big(\mathbb{E}[f(x^{k-1}) - f(x^*)]\big) \\ &\leq \tfrac{1}{2}\mathbb{E}[\|\nabla f(x^{k-1})\|^2].\end{aligned}$$

The second inequality follows from τ-gradient dominance of the function f. This completes the proof for the first part.

The proof of second part mimics that of the first part. Now we have the following condition on the iterates of Algorithm 2:

$$\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{\mathbb{E}[f(x^{k-1}) - f(x^*)]}{2\tau}. \tag{13}$$

However, f is τ-gradient dominated, so $\mathbb{E}[\|\nabla f(x^k)\|^2] \geq \mathbb{E}[f(x^k) - f(x^*)]/\tau$, which combined with Equation (13) concludes the proof. □

D. Convex SVRG: Convergence Rate

Proof of Theorem 5

Theorem. *If $f \in \mathcal{F}_n$ and f_i is convex ($i \in [n]$), $p_i = 1/m$ for $0 \leq i < m$, and $p_m = 0$. Then for Alg. 1, we have*

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \leq \frac{L\|x^0 - x^*\|^2 + 4mL^2\eta^2[f(x^0) - f(x^*)]}{T\eta(1 - 4L\eta)},$$

where x^* is optimal for (1) and x_a is the output of Alg. 1.

Proof. Consider the following sequence of inequalities:

$$\begin{aligned}&\mathbb{E}[\|x_{t+1}^{s+1} - x^*\|^2] = \mathbb{E}[\|x_t^{s+1} - \eta v_t^{s+1} - x^*\|^2] \\ &\leq \mathbb{E}[\|x_t^{s+1} - x^*\|^2] + \eta^2\mathbb{E}[\|v_t^{s+1}\|^2] \\ &\qquad - 2\eta\mathbb{E}[\langle v_t^{s+1}, x_t^{s+1} - x^*\rangle] \\ &\leq \mathbb{E}[\|x_t^{s+1} - x^*\|^2] + \eta^2\mathbb{E}[\|v_t^{s+1}\|^2] \\ &\qquad - 2\eta\mathbb{E}[f(x_t^{s+1}) - f(x^*)] \\ &\leq \mathbb{E}[\|x_t^{s+1} - x^*\|^2] - 2\eta(1 - 2L\eta)\mathbb{E}[f(x_t^{s+1}) - f(x^*)] \\ &\qquad + 4L\eta^2\mathbb{E}[f(\tilde{x}^s) - f(x^*)] \\ &= \mathbb{E}[\|x_t^{s+1} - x^*\|^2] - 2\eta(1 - 4L\eta)\mathbb{E}[f(x_t^{s+1}) - f(x^*)] \\ &+ 4L\eta^2\mathbb{E}[f(\tilde{x}^s) - f(x^*)] - 4L\eta^2\mathbb{E}[f(x_t^{s+1}) - f(x^*)].\end{aligned}$$

The second inequality uses unbiasedness of the SVRG update and convexity of f. The third inequality follows from Lemma 8. Defining the Lyapunov function

$$P^s := \mathbb{E}[\|x_m^s - x^*\|^2] + 4mL\eta^2\mathbb{E}[f(\tilde{x}^s) - f(x^*)],$$

and summing the above inequality over t, we get

$$2\eta(1 - 4L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_t^{s+1}) - f(x^*)] \leq P^s - P^{s+1}.$$

This due is to the fact that

$$\begin{aligned}P^{s+1} &= \mathbb{E}[\|x_m^{s+1} - x^*\|^2] + 4mL\eta^2\mathbb{E}[f(\tilde{x}^{s+1}) - f(x^*)] \\ &= \mathbb{E}[\|x_m^{s+1} - x^*\|^2] + 4L\eta^2\sum_{t=0}^{m-1}\mathbb{E}[f(x_t^{s+1}) - f(x^*)].\end{aligned}$$

The above equality uses the fact that $p_m = 0$ and $p_i = 1/m$ for $0 \leq i < m$. Summing over all epochs and telescoping we then obtain

$$\mathbb{E}[f(x_a) - f(x^*)] \leq P^0\big(2T\eta(1 - 4L\eta)\big)^{-1}.$$

The inequality also uses the definition of x_a given in Alg 1. On this inequality we use Lemma 7, which yields

$$\begin{aligned}&\mathbb{E}[\|\nabla f(x_a)\|^2] \leq 2L\mathbb{E}[f(x_a) - f(x^*)] \\ &\leq \frac{L\|x^0 - x^*\|^2 + 4mL^2\eta^2[f(x^0) - f(x^*)]}{T\eta(1 - 4L\eta)}.\end{aligned}$$ □

It is easy to see that we can obtain convergence rates for $E[f(x_a) - f(x^*)]$ from the above reasoning. This leads to a *direct* analysis of SVRG for convex functions.

Algorithm 3 Mini-batch SVRG

Input: $\tilde{x}^0 = x_m^0 = x^0 \in \mathbb{R}^d$, epoch length m, step sizes $\{\eta_i > 0\}_{i=0}^{m-1}$, $S = \lceil T/m \rceil$, discrete probability distribution $\{p_i\}_{i=0}^m$, mini-batch size b
for $s = 0$ **to** $S-1$ **do**
  $x_0^{s+1} = x_m^s$
  $g^{s+1} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x}^s)$
  for $t = 0$ **to** $m-1$ **do**
    Choose a mini-batch (uniformly random with replacement) $I_t \subset [n]$ of size b
    $u_t^{s+1} = \frac{1}{b}\sum_{i_t \in I_t}(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)) + g^{s+1}$
    $x_{t+1}^{s+1} = x_t^{s+1} - \eta_t u_t^{s+1}$
  end for
  $\tilde{x}^{s+1} = \sum_{i=0}^m p_i x_i^{s+1}$
end for
Output: Iterate x_a chosen uniformly random from $\{\{x_t^{s+1}\}_{t=0}^{m-1}\}_{s=0}^{S-1}$.

E. Mini-batch Nonconvex SVRG

Proof of Theorem 6

The proofs essentially follow along the lines of Lem. 1, Theorem 2 and Theorem 3 with the added complexity of mini-batch. We first prove few intermediate results before proceeding to the proof of Theorem 6.

Lemma 2. *Let $f \in \mathcal{F}_n$. Suppose we have*

$$\overline{R}_t^{s+1} := \mathbb{E}[f(x_t^{s+1}) + \bar{c}_t\|x_t^{s+1} - \tilde{x}^s\|^2],$$
$$\bar{c}_t = \bar{c}_{t+1}(1 + \eta_t\beta_t + \tfrac{2\eta_t^2L^2}{b}) + \frac{\eta_t^2L^3}{b},$$

for $0 \le s \le S-1$ and $0 \le t \le m-1$ and the parameters $\eta_t, \beta_t > 0$ and $\bar{c}_{t+1}$ are chosen such that

$$\left(\eta_t - \frac{\bar{c}_{t+1}\eta_t}{\beta_t} - \eta_t^2L - 2\bar{c}_{t+1}\eta_t^2\right) > 0.$$

Then the iterates x_t^{s+1} in the mini-batch version of Alg. 1 i.e., Alg. 3 with mini-batch size b satisfy the bound:

$$\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \le \frac{\overline{R}_t^{s+1} - \overline{R}_{t+1}^{s+1}}{\left(\eta_t - \frac{\bar{c}_{t+1}\eta_t}{\beta_t} - \eta_t^2L - 2\bar{c}_{t+1}\eta_t^2\right)},$$

Proof. Using essentially the same argument as the proof of Lemma. 1 until Equation (9), we have

$$\begin{aligned}\overline{R}_{t+1}^{s+1} &\le \mathbb{E}[f(x_t^{s+1}) - \left(\eta_t - \tfrac{\bar{c}_{t+1}\eta_t}{\beta_t}\right)\|\nabla f(x_t^{s+1})\|^2 \\ &+ \left(\tfrac{L\eta_t^2}{2} + \bar{c}_{t+1}\eta_t^2\right)\mathbb{E}[\|u_t^{s+1}\|^2] \\ &+ (\bar{c}_{t+1} + \bar{c}_{t+1}\eta_t\beta_t)\,\mathbb{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right]. \qquad (14)\end{aligned}$$

We use Lem. 4 in order to bound $\mathbb{E}[\|u_t^{s+1}\|^2]$ in the above inequality. Substituting it in Equation (14), we see that

$$\begin{aligned}&\overline{R}_{t+1}^{s+1} \le \mathbb{E}[f(x_t^{s+1})] \\ &- \left(\eta_t - \tfrac{\bar{c}_{t+1}\eta_t}{\beta_t} - \eta_t^2L - 2\bar{c}_{t+1}\eta_t^2\right)\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \\ &+ \left[\bar{c}_{t+1}\big(1 + \eta_t\beta_t + \tfrac{2\eta_t^2L^2}{b}\big) + \tfrac{\eta_t^2L^3}{b}\right]\mathbb{E}\left[\|x_t^{s+1} - \tilde{x}^s\|^2\right] \\ &\le \overline{R}_t^{s+1} - \big(\eta_t - \tfrac{\bar{c}_{t+1}\eta_t}{\beta_t} - \eta_t^2L - 2\bar{c}_{t+1}\eta_t^2\big)\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2].\end{aligned}$$

The second inequality follows from the definition of $\bar{c}_t$ and $\overline{R}_t^{s+1}$, thus concluding the proof. □

Our intermediate key result is the following theorem that provides convergence rate of mini-batch SVRG.

Theorem 8. *Let $f \in \mathcal{F}_n$ and $\overline{\gamma}_n$ denote the following:*

$$\overline{\gamma}_n := \min_{0 \le t \le m-1} \big(\eta - \tfrac{\bar{c}_{t+1}\eta}{\beta} - \eta^2L - 2\bar{c}_{t+1}\eta^2\big).$$

Let $\eta_t = \eta > 0$ and $\beta_t = \beta > 0$ for all $t \in \{0, \ldots, m-1\}$, $\bar{c}_m = 0$, $\bar{c}_t = \bar{c}_{t+1}(1 + \eta_t\beta_t + \frac{2\eta_t^2L^2}{b}) + \frac{\eta_t^2L^3}{b}$ for $t \in \{0, \ldots, m-1\}$ such that $\overline{\gamma}_n > 0$. Further, let $p_m = 1$ and $p_i = 0$ for $0 \le i < m$. Then for the output x_a of mini-batch version of Alg. 1 i.e., Alg. 3 with mini-batch size b, we have

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \le \frac{f(x^0) - f(x^*)}{T\overline{\gamma}_n},$$

where x^* is an optimal solution to (1).

Proof. Since $\eta_t = \eta$ for $t \in \{0, \ldots, m-1\}$, using Lem. 2 and telescoping the sum, we obtain

$$\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \le \frac{\overline{R}_0^{s+1} - \overline{R}_m^{s+1}}{\overline{\gamma}_n}.$$

This inequality in turn implies that

$$\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \le \frac{\mathbb{E}[f(\tilde{x}^s) - f(\tilde{x}^{s+1})]}{\overline{\gamma}_n},$$

where we used that $\overline{R}_m^{s+1} = \mathbb{E}[f(x_m^{s+1})] = \mathbb{E}[f(\tilde{x}^{s+1})]$ (since $\bar{c}_m = 0$, $p_m = 1$, and $p_i = 0$ for $i < m$), and that $\overline{R}_0^{s+1} = \mathbb{E}[f(\tilde{x}^s)]$ (since $x_0^{s+1} = \tilde{x}^s$, as $p_m = 1$ and $p_i = 0$ for $i < m$). Now sum over all epochs and using the fact that $\tilde{x}^0 = x^0$, we get the desired result. □

We now present the proof of Theorem 6.

Theorem. *Let $f \in \mathcal{F}_n$ and $\overline{\gamma}_n$ denote the following:*

$$\overline{\gamma}_n := \min_{0 \le t \le m-1} \big(\eta - \tfrac{\bar{c}_{t+1}\eta}{\beta} - \eta^2L - 2\bar{c}_{t+1}\eta^2\big),$$

where $\bar{c}_m = 0$, $\bar{c}_t = \bar{c}_{t+1}(1 + \eta\beta + 2\eta^2L^2/b) + \eta_t^2L^3/b$ for $0 \le t < m$. Suppose $\eta = \mu_2 b/(Ln^{2/3})$ $(0 < \mu_2 < 1)$,

$\beta = L/n^{1/3}$, $m = \lfloor n/(3b\mu_2) \rfloor$ and T is some multiple of m. Then for $b < n^{2/3}$, there exists universal constants $\mu_2, \nu_2 > 0$ such that: $\overline{\gamma}_n \geq \frac{\nu_2 b}{Ln^{2/3}}$ and

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \leq \frac{Ln^{2/3}[f(x^0) - f(x^*)]}{bT\nu_2},$$

where x^* is optimal for (1) and x_a is the output of the mini-batch version of Alg. 1.

Proof of Theorem 6. We first observe that using the specified values of β and η we obtain

$$\overline{\theta} := \frac{2\eta^2 L^2}{b} + \eta\beta = \frac{2\mu_2^2 b}{n^{4/3}} + \frac{\mu_2 b}{n} \leq \frac{3\mu_2 b}{n}.$$

The above inequality follows since $\mu_2 \leq 1$ and $n \geq 1$. For our analysis, we will require the following bound on $\overline{c}_0$:

$$\begin{aligned}\overline{c}_0 &= \frac{\mu_2^2 b^2 L}{bn^{4/3}} \frac{(1+\overline{\theta})^m - 1}{\overline{\theta}} = \frac{\mu_2 bL((1+\overline{\theta})^m - 1)}{2b\mu_2 + bn^{1/3}} \\ &\leq n^{-1/3}(\mu_2 L(e-1)), \end{aligned} \tag{15}$$

wherein the first equality holds due to the relation $\overline{c}_t = \overline{c}_{t+1}(1 + \eta_t\beta_t + \frac{2\eta_t^2 L^2}{b}) + \frac{\eta_t^2 L^3}{b}$, and the inequality follows upon again noting that $(1 + 1/l)^l$ is increasing for $l > 0$ and $\lim_{l\to\infty}(1 + \frac{1}{l})^l = e$. Now we can lower bound $\overline{\gamma}_n$, as

$$\begin{aligned}\overline{\gamma}_n &= \min_t \big(\eta - \tfrac{\overline{c}_{t+1}\eta}{\beta} - \eta^2 L - 2\overline{c}_{t+1}\eta^2\big) \\ &\geq \big(\eta - \tfrac{\overline{c}_0\eta}{\beta} - \eta^2 L - 2\overline{c}_0\eta^2\big) \geq \frac{b\nu_2}{Ln^{2/3}},\end{aligned}$$

where $\nu_2 > 0$ is a universal constant. The first inequality holds since $\overline{c}_t$ decreases with t. The second one holds since (a) $\overline{c}_0/\beta$ is upper bounded by $\mu_2(e-1)$ (due to Equation (15)), (b) $\eta^2 L \leq \mu_2\eta$ (as $b < n^{2/3}$) and (c) $2\overline{c}_0\eta^2 \leq 2\mu_2^2(e-1)\eta$ (again due to Equation (15) and the fact $b < n^{2/3}$). By choosing an appropriately small universal constant μ_2, one can ensure that $\overline{\gamma}_n \geq b\nu_2/(Ln^{2/3})$ for some universal constant ν_2. For example, choosing $\mu_2 = 1/4$, we have $\overline{\gamma}_n \geq b\nu_2/(Ln^{2/3})$ with $\nu_2 = 1/40$. Substituting the above lower bound in Theorem 8, we get the desired result. □

F. MSVRG: Convergence Rate

Proof of Theorem 7

Theorem. *Suppose $f \in \mathcal{F}_n$ has σ-bounded gradients. Let $\eta_t = \eta = \max\{c/\sqrt{T}, \mu_1/(Ln^{2/3})\}$ (μ_1 is the constant from Corr. 3), $m = \lfloor n/(3\mu_1) \rfloor$, and $c = \sqrt{\frac{f(x^0)-f(x^*)}{2L\sigma^2}}$. Further, let T be a multiple of m, $p_m = 1$, and $p_i = 0$ for $0 \leq i < m$. Then, the output x_a of Alg. 1 satisfies*

$$\begin{aligned}&\mathbb{E}[\|\nabla f(x_a)\|^2] \\ &\leq \bar{\nu}\min\Big\{2\sqrt{\frac{2(f(x^0)-f(x^*))L}{T}}\sigma, \frac{Ln^{2/3}[f(x^0)-f(x^*)]}{T\nu_1}\Big\},\end{aligned}$$

where $\bar{\nu} > 0$ is a universal constant, ν_1 is the universal constant from Corr. 3 and x^* is an optimal solution to (1).

Proof. First, we observe that the step size η is chosen to be $\max\{c/\sqrt{T}, \mu_1/(Ln^{2/3})\}$ where

$$c = \sqrt{\frac{f(x^0) - f(x^*)}{2L\sigma^2}}.$$

Suppose $\eta = \mu_1/(Ln^{2/3})$, we obtain the convergence rate in Corollary 3. Now, lets consider the case where $\eta = c/\sqrt{T}$. In this case, we have the following bound:

$$\begin{aligned}&\mathbb{E}[\|v_t^{s+1}\|^2] = \mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s) + \nabla f(\tilde{x}^s)\|^2] \\ &\leq 2\left(\mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1})\|^2 + \|\nabla f_{i_t}(\tilde{x}^s) - \nabla f(\tilde{x}^s)\|^2]\right) \\ &\leq 2\left(\mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1})\|^2 + \|\nabla f_{i_t}(\tilde{x}^s)\|^2]\right) \leq 4\sigma^2.\end{aligned}$$

The first inequality follows from Lemma 6 with $r = 2$. The second inequality follows from (a) σ-bounded gradient property of f and (b) the fact that for a random variable ζ, $\mathbb{E}[\|\zeta - \mathbb{E}[\zeta]\|^2] \leq \mathbb{E}[\|\zeta\|^2]$. The rest of the proof is along exactly the lines as in Theorem 1. This provides a convergence rate similar to Theorem 1. More specifically, using step size $c/\sqrt{T}$, we get

$$\mathbb{E}[\|f(x_a)\|^2] \leq 2\sqrt{\frac{2(f(x^0) - f(x^*))L}{T}}\sigma. \tag{16}$$

The only thing that remains to be proved is that with the step size choice of $\max\{c/\sqrt{T}, \mu_1/(Ln^{2/3})\}$, the minimum of two bounds hold. Consider the case $c/\sqrt{T} > \mu_1/(Ln^{2/3})$. In this case, we have the following:

$$\begin{aligned}\frac{2\sqrt{\frac{2(f(x^0)-f(x^*))L}{T}}\sigma}{\frac{Ln^{2/3}[f(x^0)-f(x^*)]}{T\nu_1}} &= \frac{2\nu_1\sigma\sqrt{2LT}}{Ln^{2/3}\sqrt{f(x^0)-f(x^*)}} \\ &\leq 2\nu_1/\mu_1 \leq \bar{\nu} := \max\left\{\frac{2\nu_1}{\mu_1}, \frac{\mu_1}{2\nu_1}\right\},\end{aligned}$$

where ν_1 is the constant in Corollary 3. This inequality holds since $c/\sqrt{T} > \mu_1/(Ln^{2/3})$. Rearranging the above inequality, we have

$$2\sqrt{\frac{2(f(x^0) - f(x^*))L}{T}}\sigma \leq \frac{\bar{\nu}Ln^{2/3}[f(x^0) - f(x^*)]}{T}$$

in this case. Note that the left hand side of the above inequality is precisely the bound obtained by using step size $c/\sqrt{T}$ (see Equation (16)). Similarly, when $c/\sqrt{T} \leq \mu_1/(Ln^{2/3})$, the inequality holds in the other direction. Using these two observations, we have the desired result. □

G. Key Lemmatta

Lemma 3. *For the intermediate iterates v_t^{s+1} computed by Alg. 1, we have the following:*

$$\mathbb{E}[\|v_t^{s+1}\|^2] \leq 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + 2L^2\mathbb{E}[\|x_t^{s+1} - \tilde{x}^s\|^2].$$

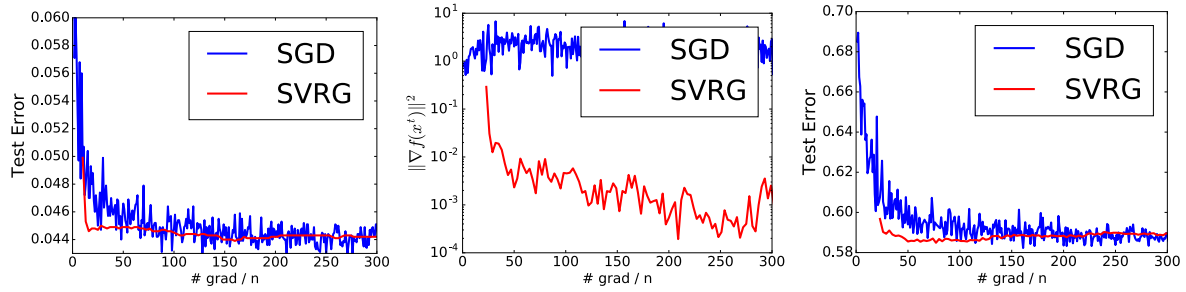

Figure 2. Neural network results for MNIST and STL-10. The leftmost result is for MNIST. The remaining two plots are of STL-10.

Proof. The proof simply follows from the proof of Lemma 4 with $I_t = \{i_t\}$. □

We now present a result to bound the variance of mini-batch SVRG.

Lemma 4. *Let u_t^{s+1} be computed by the mini-batch version of Alg. 1 i.e., Alg. 3 with mini-batch size b. Then,*

$$\mathbb{E}[\|u_t^{s+1}\|^2] \leq 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + \tfrac{2L^2}{b}\mathbb{E}[\|x_t^{s+1} - \tilde{x}^s\|^2].$$

Proof. For the ease of exposition, we use the following notation:

$$\zeta_t^{s+1} = \frac{1}{|I_t|}\sum_{i_t \in I_t}\left(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)\right).$$

We use the definition of u_t^{s+1} to get

$$\begin{aligned}
&\mathbb{E}[\|u_t^{s+1}\|^2] = \mathbb{E}[\|\zeta_t^{s+1} + \nabla f(\tilde{x}^s)\|^2] \\
&= \mathbb{E}[\|\zeta_t^{s+1} + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1}) + \nabla f(x_t^{s+1})\|^2] \\
&\leq 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + 2\mathbb{E}[\|\zeta_t^{s+1} - \mathbb{E}[\zeta_t^{s+1}]\|^2] \\
&= 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \\
&+ \frac{2}{b^2}\mathbb{E}\left[\left\|\sum_{i_t \in I_t}\left(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s) - \mathbb{E}[\zeta_t^{s+1}]\right)\right\|^2\right]
\end{aligned}$$

The first inequality follows from Lemma 6 (with $r = 2$) and the fact that $\mathbb{E}[\zeta_t^{s+1}] = \nabla f(x_t^{s+1}) - \nabla f(\tilde{x}^s)$. From the above inequality, we get

$$\begin{aligned}
\mathbb{E}[\|u_t^{s+1}\|^2] &\leq 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] \\
&+ \frac{2}{b^2}\mathbb{E}\left[\sum_{i_t \in I_t}\|\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)\|^2\right] \\
&\leq 2\mathbb{E}[\|\nabla f(x_t^{s+1})\|^2] + \frac{2L^2}{b}\mathbb{E}[\|x_t^{s+1} - \tilde{x}^s\|^2]
\end{aligned}$$

The first inequality follows from Lemma 5 and noting that for a random variable ζ, $\mathbb{E}[\|\zeta - \mathbb{E}[\zeta]\|^2] \leq \mathbb{E}[\|\zeta\|^2]$. The last inequality follows from L-smoothness of f_{i_t}. □

H. Experiments

Figure 2 shows the remaining plots for MNIST and STL-10 datasets. As seen in the plots, there is no significant difference in the test error of SVRG and SGD for these datasets.

I. Other Lemmas

Lemma 5. *For random variables $z_1, \ldots, z_r$ are independent and mean 0, we have*

$$\mathbb{E}\left[\|z_1 + \ldots + z_r\|^2\right] = \mathbb{E}\left[\|z_1\|^2 + \ldots + \|z_r\|^2\right].$$

Proof. We have the following:

$$\begin{aligned}
&\mathbb{E}\left[\|z_1 + \ldots + z_r\|^2\right] \\
&= \sum_{i,j=1}^{r}\mathbb{E}\left[z_i z_j\right] = \mathbb{E}\left[\|z_1\|^2 + \ldots + \|z_r\|^2\right].
\end{aligned}$$

The second equality follows from the fact that z_i's are independent and mean 0. □

Lemma 6. *For random variables $z_1, \ldots, z_r$, we have*

$$\mathbb{E}\left[\|z_1 + \ldots + z_r\|^2\right] \leq r\mathbb{E}\left[\|z_1\|^2 + \ldots + \|z_r\|^2\right].$$

We need the next lemma (Lemma 7) for our results in the convex case.

Lemma 7 (Johnson & Zhang (2013))**.** *Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex with L-Lipschitz continuous gradient. Then,*

$$\|\nabla g(x) - \nabla g(y)\|^2 \leq 2L[g(x) - g(y) - \langle \nabla g(y), x - y\rangle],$$

for all $x, y \in \mathbb{R}^d$.

Proof. Consider $h(x) := g(x) - g(y) - \langle \nabla g(y), x - y\rangle$ for arbitrary $y \in \mathbb{R}^d$. Observe that ∇h is also L-Lipschitz continuous. Note that $h(x) \geq 0$ (since $h(y) = 0$ and $\nabla h(y) = 0$, or alternatively since h defines a Bregman

divergence), from which it follows that

$$\begin{aligned}0 &\leq \min_{\rho}[h(x-\rho\nabla h(x))]\\ &\leq \min_{\rho}[h(x)-\rho\|\nabla h(x)\|^2+\tfrac{L\rho^2}{2}\|\nabla h(x)\|^2]\\ &= h(x)-\tfrac{1}{2L}\|\nabla h(x)\|^2.\end{aligned}$$

Rewriting in terms of g we obtain the required result. □

Lemma 8 bounds the variance of SVRG for the convex case. Please refer to (Johnson & Zhang, 2013) for more details.

Lemma 8 ((Johnson & Zhang, 2013))**.** *Suppose f_i is convex for all $i \in [n]$. For the updates in Alg. 1 we have the following inequality:*

$$\mathbb{E}[\|v_t^{s+1}\|^2] \leq 4L[f(x_t^{s+1})-f(x^*)+f(\tilde{x}^s)-f(x^*)].$$

Proof. The proof follows upon observing the following:

$$\begin{aligned}&\mathbb{E}[\|v_t^{s+1}\|^2 = \mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1})-\nabla f_{i_t}(x_0^{s+1})+\nabla f(\tilde{x}^s)\|^2]\\ &\leq 2\mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1})-\nabla f_{i_t}(x^*)\|^2]\\ &\quad + 2\mathbb{E}[\|\nabla f_{i_t}(\tilde{x}^s)-\nabla f_{i_t}(x^*)-(\nabla f(\tilde{x}^s)-\nabla f(x^*))\|^2]\\ &\leq 2\mathbb{E}[\|\nabla f_{i_t}(x_t^{s+1})-\nabla f_{i_t}(x^*)\|^2]\\ &\qquad + 2\mathbb{E}[\|\nabla f_{i_t}(\tilde{x}^s)-\nabla f_{i_t}(x^*)\|^2]\\ &\leq 4L[f(x_t^{s+1}-f(x^*)+f(\tilde{x}^s)-f(x^*)].\end{aligned}$$

The first inequality follows from Cauchy-Schwarz and Young inequality; the second one from $\mathbb{E}[\|\xi-\mathbb{E}[\xi]\|^2] \leq \mathbb{E}[\|\xi\|^2]$, and the third one from Lemma 7. □