Simple Stochastic Gradient Methods for Non-Smooth Non-Convex Regularized Optimization: Supplementary Material

1. Proof of Lemma 2

Lemma 2. *For an initial value $w_1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, and $\alpha, \theta \in \mathbb{R}$, MBSGA generates w^R satisfying the following bound.*

$$\mathbb{E}||\nabla E^R_\lambda(w^R)||_2^2 \leq \frac{\tilde{\Delta}}{N}(L + N^\theta) + \frac{\sigma}{\sqrt{N}}\left(\tilde{\Delta} + \frac{L + N^\theta}{\lceil N^\alpha \rceil}\right),$$

where $\tilde{\Delta} = 2(\tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w^*_\lambda))$ and w^*_λ is a global minimizer of $\tilde{h}_\lambda(\cdot)$.

In order to prove this result, we require the following property.

Property 13.

$$\mathbb{E}||\nabla A^k_{\lambda M}(w^k, \xi^k) - \nabla E^k_\lambda(w^k)||_2^2 \leq \frac{\sigma^2}{M}$$

Proof. From the definition of $\nabla A^k_{\lambda M}(w^k, \xi^k)$ found in Algorithm 1 and (11), $\nabla A^k_{\lambda M}(w^k, \xi^k) - \nabla E^k_\lambda(w^k) = \frac{1}{M}\sum_{j=1}^M \nabla F(w^k, \xi^k_j) - \nabla f(w^k)$. Taking the expectation of its squared norm,

$$\begin{aligned}\mathbb{E}||\nabla A^k_{\lambda M}(w^k, \xi^k) - \nabla E^k_\lambda(w^k)||_2^2 &= \mathbb{E}||\frac{1}{M}\sum_{j=1}^M(\nabla F(w^k, \xi^k_j) - \nabla f(w^k))||_2^2 \\ &= \frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\left(\sum_{j=1}^M \nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i\right)^2.\end{aligned}$$

For $j \neq l$, $\nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i$ and $\nabla F(w^k, \xi^k_l)_i - \nabla f(w^k)_i$ are independent random variables with zero mean. It follows that

$$\begin{aligned}&\mathbb{E}[(\nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i)(\nabla F(w^k, \xi^k_l)_i - \nabla f(w^k)_i)] = \\ &\mathbb{E}[(\nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i)]\mathbb{E}[(\nabla F(w^k, \xi^k_l)_i - \nabla f(w^k)_i)] = 0,\end{aligned}$$

and

$$\begin{aligned}\frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\left(\sum_{j=1}^M \nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i\right)^2 &= \frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\sum_{j=1}^M(\nabla F(w^k, \xi^k_j)_i - \nabla f(w^k)_i)^2 \\ &= \frac{1}{M^2}\sum_{j=1}^M \mathbb{E}||\nabla F(w^k, \xi^k_j) - \nabla f(w^k)||_2^2 \leq \frac{\sigma^2}{M}\end{aligned}$$

using (5). □

Proof of Lemma 2. Given the smoothness of $E^k_\lambda(w)$ as shown in Property 1,

$$\begin{aligned}E^k_\lambda(w^{k+1}) &\leq E^k_\lambda(w^k) + \langle \nabla E^k_\lambda(w^k), w^{k+1} - w^k\rangle + \frac{L_{E\lambda}}{2}||w^{k+1} - w^k||_2^2 \\ &= E^k_\lambda(w^k) + \langle \nabla E^k_\lambda(w^k), -\gamma\nabla A^k_{\lambda M}(w^k, \xi^k)\rangle + \frac{L_{E\lambda}}{2}|| - \gamma\nabla A^k_{\lambda M}(w^k, \xi^k)||_2^2.\end{aligned}$$

Using (12) and (13),

$$\tilde{h}(w^{k+1}) \leq \tilde{h}(w^k) - \gamma\langle \nabla E_\lambda^k(w^k), \nabla A_{\lambda M}^k(w^k, \xi^k)\rangle + \frac{L_{E\lambda}}{2}\gamma^2 ||\nabla A_{\lambda M}^k(w^k, \xi^k)||_2^2.$$

Setting $\delta_k = \nabla A_{\lambda M}^k(w^k, \xi^k) - \nabla E_\lambda^k(w^k)$,

$$\begin{aligned}\tilde{h}(w^{k+1}) \leq& \tilde{h}(w^k) - \gamma\left(||\nabla E_\lambda^k(w^k)||_2^2 + \langle \nabla E_\lambda^k(w^k), \delta_k\rangle\right) + \frac{L_{E\lambda}}{2}\gamma^2\left(||\nabla E_\lambda^k(w^k)||_2^2 + 2\langle \nabla E_\lambda^k(w^k), \delta_k\rangle + ||\delta_k||_2^2\right)\\ =& \tilde{h}(w^k) + \left(\frac{L_{E\lambda}}{2}\gamma^2 - \gamma\right)||\nabla E_\lambda^k(w^k)||_2^2 + (L_{E\lambda}\gamma^2 - \gamma)\langle \nabla E_\lambda^k(w^k), \delta_k\rangle + \frac{L_{E\lambda}}{2}\gamma^2||\delta_k||_2^2,\end{aligned}$$

as

$$\langle \nabla E_\lambda^k(w^k), \nabla A_{\lambda M}^k(w^k, \xi^k)\rangle = ||\nabla E_\lambda^k(w^k)||_2^2 + \langle \nabla E_\lambda^k(w^k), \delta_k\rangle$$

and

$$||\nabla A_{\lambda M}^k(w^k, \xi^k)||_2^2 = ||\nabla E_\lambda^k(w^k)||_2^2 + 2\langle \nabla E_\lambda^k(w^k), \delta_k\rangle + ||\delta_k||_2^2.$$

After N iterations,

$$\begin{aligned}\left(\gamma - \frac{L_{E\lambda}}{2}\gamma^2\right)\sum_{k=1}^N ||\nabla E_\lambda^k(w^k)||_2^2 \leq& \tilde{h}(w^1) - \tilde{h}(w^{N+1}) + (L_{E\lambda}\gamma^2 - \gamma)\sum_{k=1}^N\langle \nabla E_\lambda^k(w^k), \delta_k\rangle + \frac{L_{E\lambda}}{2}\gamma^2\sum_{k=1}^N||\delta_k||_2^2\\ \leq& \tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w_\lambda^*) + (L_{E\lambda}\gamma^2 - \gamma)\sum_{k=1}^N\langle \nabla E_\lambda^k(w^k), \delta_k\rangle + \frac{L_{E\lambda}}{2}\gamma^2\sum_{k=1}^N||\delta_k||_2^2.\end{aligned}$$

It follows from (4) that for w independent of ξ^k, $\mathbb{E}\nabla A_{\lambda M}^k(w, \xi^k) = \nabla E_\lambda^k(w)$, and so $\mathbb{E}[\delta_k] = 0$. Taking the expectation of both sides,

$$\begin{aligned}\left(\gamma - \frac{L_{E\lambda}}{2}\gamma^2\right)\sum_{k=1}^N \mathbb{E}||\nabla E_\lambda^k(w^k)||_2^2 \leq& \tilde{h}(w^1) - \tilde{h}(w_\lambda^*) + \frac{L_{E\lambda}}{2}\gamma^2\sum_{k=1}^N\mathbb{E}||\delta_k||_2^2\\ \leq& \tilde{h}(w^1) - \tilde{h}(w_\lambda^*) + \frac{L_{E\lambda}}{2}\gamma^2\frac{N}{M}\sigma^2,\end{aligned}$$

where the second inequality uses Property 13. Choosing R uniformly over $\{1, ..., N\}$,

$$\begin{aligned}\mathbb{E}||\nabla E_\lambda^R(w^R)||_2^2 =& \frac{1}{N}\sum_{k=1}^N \mathbb{E}||\nabla E_\lambda^k(w^k)||_2^2\\ \leq& \frac{1}{N\left(\gamma - \frac{L_{E\lambda}}{2}\gamma^2\right)}\left(\tilde{h}(w^1) - \tilde{h}(w^*) + \frac{L_{E\lambda}}{2}\gamma^2\frac{N}{M}\sigma^2\right).\end{aligned}$$

Since $\gamma \leq \frac{1}{L_{E\lambda}}$, it holds that $\gamma - \frac{L_{E\lambda}}{2}\gamma^2 \geq \frac{1}{2}\gamma$, and

$$\begin{aligned}\frac{1}{N\left(\gamma - \frac{L_{E\lambda}}{2}\gamma^2\right)}\left(\frac{\tilde{\Delta}}{2} + \frac{L_{E\lambda}}{2}\gamma^2\frac{N}{M}\sigma^2\right) \leq& \frac{1}{N\gamma}\left(\tilde{\Delta} + L_{E\lambda}\gamma^2\frac{N}{M}\sigma^2\right)\\ =& \frac{\tilde{\Delta}}{N\gamma} + L_{E\lambda}\frac{\gamma}{M}\sigma^2\\ \leq& \frac{\tilde{\Delta}}{N}\max\left\{L_{E\lambda}, \sigma\sqrt{N}\right\} + L_{E\lambda}\frac{\sigma}{M\sqrt{N}}\\ \leq& \frac{\tilde{\Delta}L_{E\lambda}}{N} + \frac{\sigma}{\sqrt{N}}\left(\tilde{\Delta} + \frac{L_{E\lambda}}{M}\right)\\ =& \frac{\tilde{\Delta}}{N}(L + N^\theta) + \frac{\sigma}{\sqrt{N}}\left(\tilde{\Delta} + \frac{L + N^\theta}{\lceil N^\alpha\rceil}\right)\end{aligned}$$

□

2. Proof of Lemma 7

Lemma 7. *For an initial value $\tilde{w}_1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, $\alpha, \theta \in \mathbb{R}$, VRSGA generates w_T^R satisfying the following bound.*

$$\mathbb{E}\left[||\nabla E_{T\lambda}^R(w_T^R)||_2^2\right] \leq \tilde{\Delta}\frac{L+(Sm)^\theta}{Sm},$$

where $\tilde{\Delta} = 36(\tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^*))$ and w_λ^* is a global minimizer of $\tilde{h}_\lambda(\cdot)$.

In order to prove this result, we require the following lemmas.

Lemma 14. *Consider arbitrary $w, V, z \in \mathbb{R}^d$, $\gamma \in \mathbb{R}$, and $w^+ = w - \gamma V$,*

$$E_{t\lambda}^k(w^+) \leq E_{t\lambda}^k(z) + \langle \nabla E_{t\lambda}^k(w) - V, w^+ - z\rangle + \frac{L_{E\lambda}}{2}||w^+ - w||_2^2 + \frac{L_{E\lambda}}{2}||z - w||_2^2 - \frac{1}{\gamma}\langle w^+ - w, w^+ - z\rangle.$$

Proof. Adding the following three inequalities proves the result, where the first two come from the smoothness of $E_{t\lambda}^k(w)$ and $-E_{t\lambda}^k(w)$, see Property 1, and the third is due to $V + \frac{1}{\gamma}(w^+ - w) = 0$.

$$\begin{aligned}
E_{t\lambda}^k(w^+) &\leq E_{t\lambda}^k(w) + \langle \nabla E_{t\lambda}^k(w), w^+ - w\rangle + \frac{L_{E\lambda}}{2}||w^+ - w||_2^2 \\
-E_{t\lambda}^k(z) &\leq -E_{t\lambda}^k(w) + \langle -\nabla E_{t\lambda}^k(w), z - w\rangle + \frac{L_{E\lambda}}{2}||z - w||_2^2 \\
0 &= -\langle V + \frac{1}{\gamma}(w^+ - w), w^+ - z\rangle
\end{aligned}$$

□

Lemma 15. *For vectors w, x, z, and $\beta > 0$,*

$$||w - x||_2^2 \leq (1+\beta)||w - z||_2^2 + \left(1 + \frac{1}{\beta}\right)||z - x||_2^2.$$

Proof.

$$\begin{aligned}
||w - x||_2^2 &= ||w - z + z - x||_2^2 \\
&\leq (||w - z||_2 + ||z - x||_2)^2 \\
&= ||w - z||_2^2 + 2||w - z||_2||z - x||_2 + ||z - x||_2^2 \\
&\leq ||w - z||_2^2 + \left(\beta||w - z||_2^2 + \frac{1}{\beta}||z - x||_2^2\right) + ||z - x||_2^2 \\
&= (1+\beta)||w - z||_2^2 + \left(1 + \frac{1}{\beta}\right)||z - x||_2^2,
\end{aligned}$$

where the second inequality uses Young's inequality.

□

Proof of Lemma 7. Let $\hat{w}_{t+1}^k = w_t^k - \gamma\nabla E_{t\lambda}^k(w_t^k)$, with $w^+ = w_{t+1}^k$, $w = w_t^k$, $V = V_t^k$, and $z = \hat{w}_{t+1}^k$ in Lemma 14 to get the inequality

$$\begin{aligned}
E_{t\lambda}^k(w_{t+1}^k) \leq E_{t\lambda}^k(\hat{w}_{t+1}^k) &+ \langle \nabla E_{t\lambda}^k(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k\rangle + \frac{L_{E\lambda}}{2}||w_{t+1}^k - w_t^k||_2^2 \\
&+ \frac{L_{E\lambda}}{2}||\hat{w}_{t+1}^k - w_t^k||_2^2 - \frac{1}{\gamma}\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k\rangle. \qquad (16)
\end{aligned}$$

In addition, let $w^+ = \hat{w}_{t+1}^k$, $w = w_t^k$, $V = \nabla E_{t\lambda}^k(w_t^k)$, and $z = w_t^k$ in Lemma 14 to get

$$\begin{aligned} E_{t\lambda}^k(\hat{w}_{t+1}^k) \leq & E_{t\lambda}^k(w_t^k) + \langle \nabla E_{t\lambda}^k(w_t^k) - \nabla E_{t\lambda}^k(w_t^k), \hat{w}_{t+1}^k - w_{t+1}^k \rangle + \frac{L_{E\lambda}}{2} ||\hat{w}_{t+1}^k - w_t^k||_2^2 \\ & + \frac{L_{E\lambda}}{2} ||w_t^k - w_t^k||_2^2 - \frac{1}{\gamma} \langle \hat{w}_{t+1}^k - w_t^k, \hat{w}_{t+1}^k - w_t^k \rangle \\ = & E_{t\lambda}^k(w_t^k) + \left(\frac{L_{E\lambda}}{2} - \frac{1}{\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{17}$$

Adding (16) and (17),

$$\begin{aligned} E_{t\lambda}^k(w_{t+1}^k) \leq & E_{t\lambda}^k(w_t^k) + \langle \nabla E_{t\lambda}^k(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \frac{L_{E\lambda}}{2} ||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{\gamma} \langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(L_{E\lambda} - \frac{1}{\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{18}$$

Plugging $\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle = \frac{1}{2} \left(||w_{t+1}^k - w_t^k||_2^2 + ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 - ||\hat{w}_{t+1}^k - w_t^k||_2^2 \right)$ into (18) and rearranging,

$$\begin{aligned} E_{t\lambda}^k(w_{t+1}^k) \leq & E_{t\lambda}^k(w_t^k) + \langle \nabla E_{t\lambda}^k(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_{E\lambda}}{2} - \frac{1}{2\gamma} \right) ||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{2\gamma} ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 + \left(L_{E\lambda} - \frac{1}{2\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{19}$$

Focusing on the term $-\frac{1}{2\gamma} ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2$, we apply Lemma 15 with $w = w_{t+1}^k$, $x = w_t^k$, and $z = \hat{w}_{t+1}^k$. Rearranging,

$$-(1+\beta) ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -||w_{t+1}^k - w_t^k||_2^2 + \left(1 + \frac{1}{\beta} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2$$

$$-\frac{1}{2\gamma} ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -\frac{1}{(1+\beta)2\gamma} ||w_{t+1}^k - w_t^k||_2^2 + \frac{\left(1 + \frac{1}{\beta} \right)}{(1+\beta)2\gamma} ||\hat{w}_{t+1}^k - w_t^k||_2^2.$$

Choosing $\beta = 3$,

$$-\frac{1}{2\gamma} ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -\frac{1}{8\gamma} ||w_{t+1}^k - w_t^k||_2^2 + \frac{1}{6\gamma} ||\hat{w}_{t+1}^k - w_t^k||_2^2.$$

Using this inequality in (19),

$$\begin{aligned} E_{t\lambda}^k(w_{t+1}^k) \leq & E_{t\lambda}^k(w_t^k) + \langle \nabla E_{t\lambda}^k(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_{E\lambda}}{2} - \frac{1}{2\gamma} \right) ||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{8\gamma} ||w_{t+1}^k - w_t^k||_2^2 + \frac{1}{6\gamma} ||\hat{w}_{t+1}^k - w_t^k||_2^2 + \left(L_{E\lambda} - \frac{1}{2\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2 \\ = & E_{t\lambda}^k(w_t^k) + \langle \nabla E_{t\lambda}^k(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma} \right) ||w_{t+1}^k - w_t^k||_2^2 \\ & + \left(L_{E\lambda} - \frac{1}{3\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2 \\ = & E_{t\lambda}^k(w_t^k) + \gamma ||\nabla E_{t\lambda}^k(w_t^k) - V_t^k||_2^2 + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma} \right) ||w_{t+1}^k - w_t^k||_2^2 + \left(L_{E\lambda} - \frac{1}{3\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2, \end{aligned}$$

where the last equality holds since $w_{t+1}^k - \hat{w}_{t+1}^k = \gamma(\nabla E_{t\lambda}^k(w_t^k) - V_t^k)$. Using (12) and (13), and taking the expectation of both sides,

$$\mathbb{E}\tilde{h}_\lambda(w_{t+1}^k) \leq \mathbb{E} \left[\tilde{h}_\lambda(w_t^k) + \gamma ||\nabla E_{t\lambda}^k(w_t^k) - V_t^k||_2^2 + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma} \right) ||w_{t+1}^k - w_t^k||_2^2 + \left(L_{E\lambda} - \frac{1}{3\gamma} \right) ||\hat{w}_{t+1}^k - w_t^k||_2^2 \right]. \tag{20}$$

Focusing on $\mathbb{E}\left[||\nabla E^k_{t\lambda}(w^k_t) - V^k_t||^2_2\right]$, from (11) and the definition of V^k_t found in Algorithm 2, $\nabla E^k_{t\lambda}(w^k_t) - V^k_t = \nabla f(w^k_t) - (\frac{1}{b}\sum_{j\in I}\left(\nabla f_j(w^k_t) - \nabla f_j(\tilde{w}^k)\right) + G^k)$. Rearranging, and taking the expectation of its squared norm,

$$\begin{aligned}
\mathbb{E}||\nabla E^k_{t\lambda}(w^k_t) - V^k_t||^2_2 &= \mathbb{E}||\frac{1}{b}\sum_{j\in I}\left(\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t)\right) - \left(G^k - \nabla f(w^k_t)\right)||^2_2 \\
&= \frac{1}{b^2}\mathbb{E}\sum_{j\in I}||\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t) - \left(G^k - \nabla f(w^k_t)\right)||^2_2 \\
&\leq \frac{1}{b^2}\mathbb{E}\sum_{j\in I}||\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t)||^2_2 \\
&\leq \frac{L^2}{b}\mathbb{E}||\tilde{w}^k - w^k_t||^2_2.
\end{aligned}$$

As the squared norm of a sum of independent random variables with zero mean, the second equality holds using the same reasoning as found in Property 13, and the first inequality holds since $\mathbb{E}||x - \mathbb{E}[x]||^2_2 \leq \mathbb{E}||x||^2_2$ for any random variable x. Using this bound in (20),

$$\begin{aligned}
\mathbb{E}\tilde{h}_\lambda(w^k_{t+1}) &\leq \mathbb{E}\left[\tilde{h}_\lambda(w^k_t) + \gamma\frac{L^2}{b}||\tilde{w}^k - w^k_t||^2_2 + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma}\right)||w^k_{t+1} - w^k_t||^2_2 + \left(L_{E\lambda} - \frac{1}{3\gamma}\right)||\hat{w}^k_{t+1} - w^k_t||^2_2\right] \\
&\leq \mathbb{E}\left[\tilde{h}_\lambda(w^k_t) + \frac{L_{E\lambda}}{6b}||\tilde{w}^k - w^k_t||^2_2 - \frac{13L_{E\lambda}}{4}||w^k_{t+1} - w^k_t||^2_2 - L_{E\lambda}||\hat{w}^k_{t+1} - w^k_t||^2_2\right] \\
&= \mathbb{E}\left[\tilde{h}_\lambda(w^k_t) + \frac{L_{E\lambda}}{6b}||\tilde{w}^k - w^k_t||^2_2 - \frac{13L_{E\lambda}}{4}||w^k_{t+1} - w^k_t||^2_2 - \frac{1}{36L_{E\lambda}}||\nabla E^k_{t\lambda}(w^k_t)||^2_2\right], \qquad (21)
\end{aligned}$$

where the last two lines use the fact that $\gamma = \frac{1}{6L_{E\lambda}}$. Focusing on $-\frac{13L_{E\lambda}}{4}||w^k_{t+1} - w^k_t||^2_2$, we apply Lemma 15 with $w = w^k_{t+1}$, $x = \tilde{w}^k$, and $z = w^k_t$,

$$\begin{aligned}
(1+\beta)||w^k_{t+1} - w^k_t||^2_2 &\geq ||w^k_{t+1} - \tilde{w}^k||^2_2 - \left(1 + \frac{1}{\beta}\right)||w^k_t - \tilde{w}^k||^2_2 \\
-\frac{13L_{E\lambda}}{4}||w^k_{t+1} - w^k_t||^2_2 &\leq -\frac{13L_{E\lambda}}{4(1+\beta)}||w^k_{t+1} - \tilde{w}^k||^2_2 + \frac{13L_{E\lambda}\left(1 + \frac{1}{\beta}\right)}{4(1+\beta)}||w^k_t - \tilde{w}^k||^2_2.
\end{aligned}$$

Setting $\beta = 2t - 1$,

$$-\frac{13L_{E\lambda}}{4}||w^k_{t+1} - w^k_t||^2_2 \leq -\frac{13L_{E\lambda}}{8t}||w^k_{t+1} - \tilde{w}^k||^2_2 + \frac{13L_{E\lambda}}{8t-4}||w^k_t - \tilde{w}^k||^2_2.$$

Applying this bound in (21),

$$\mathbb{E}\tilde{h}_\lambda(w^k_{t+1}) \leq \mathbb{E}\left[\tilde{h}_\lambda(w^k_t) + \left(\frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t-4}\right)||\tilde{w}^k - w^k_t||^2_2 - \frac{13L_{E\lambda}}{8t}||w^k_{t+1} - \tilde{w}^k||^2_2 - \frac{1}{36L_{E\lambda}}||\nabla E^k_{t\lambda}(w^k_t)||^2_2\right].$$

Summing over t,

$$\begin{aligned}
\mathbb{E}\tilde{h}_\lambda(w^k_{m+1}) \leq \mathbb{E}\Bigg[&\tilde{h}_\lambda(w^k_1) + \sum_{t=1}^{m}\left(\frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t-4}\right)||\tilde{w}^k - w^k_t||^2_2 \\
&- \sum_{t=1}^{m}\frac{13L_{E\lambda}}{8t}||w^k_{t+1} - \tilde{w}^k||^2_2 - \frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E^k_{t\lambda}(w^k_t)||^2_2\Bigg].
\end{aligned}$$

Considering that $\tilde{w}^k = w_1^k$ and $||w_{m+1}^k - \tilde{w}^k||_2^2 \geq 0$,

$$\begin{aligned}
\mathbb{E}\tilde{h}_\lambda(w_{m+1}^k) \leq& \mathbb{E}\left[\tilde{h}_\lambda(w_1^k) + \sum_{t=2}^{m}\left(\frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t-4}\right)||\tilde{w}^k - w_t^k||_2^2\right. \\
&\left. - \sum_{t=1}^{m-1}\frac{13L_{E\lambda}}{8t}||w_{t+1}^k - \tilde{w}^k||_2^2 - \frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] \\
=& \mathbb{E}\left[\tilde{h}_\lambda(w_1^k) + \sum_{t=1}^{m-1}\left(\frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t+4} - \frac{13L_{E\lambda}}{8t}\right)||w_{t+1}^k - \tilde{w}^k||_2^2 - \frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] \\
\leq& \mathbb{E}\left[\tilde{h}_\lambda(w_1^k) + \sum_{t=1}^{m-1}\left(\frac{L_{E\lambda}}{6b} - \frac{L_{E\lambda}}{2t^2}\right)||w_{t+1}^k - \tilde{w}^k||_2^2 - \frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] \\
\leq& \mathbb{E}\left[\tilde{h}_\lambda(w_1^k) - \frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right],
\end{aligned}$$

where the last inequality holds since $6b = 6m^2 > 2(m-1)^2 \geq 2t^2$ for $t = 1, ..., m-1$. This summation can be equivalently written as

$$\begin{aligned}
\mathbb{E}\tilde{h}_\lambda(\tilde{w}^{k+1}) &\leq \mathbb{E}\tilde{h}_\lambda(\tilde{w}^k) - \mathbb{E}\left[\frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] \\
\mathbb{E}\left[\frac{1}{36L_{E\lambda}}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] &\leq \mathbb{E}\tilde{h}_\lambda(\tilde{w}^k) - \mathbb{E}\tilde{h}_\lambda(\tilde{w}^{k+1}) \\
\mathbb{E}\left[\frac{1}{36L_{E\lambda}}\sum_{k=1}^{S}\sum_{t=1}^{m}||\nabla E_{t\lambda}^k(w_t^k)||_2^2\right] &\leq \tilde{h}_\lambda(\tilde{w}^1) - \mathbb{E}\tilde{h}_\lambda(\tilde{w}^{S+1}) \\
&\leq \tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^*) \\
\mathbb{E}\left[||\nabla E_{T\lambda}^R(w_T^R)||_2^2\right] &\leq \frac{36L_{E\lambda}\left(\tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^*)\right)}{Sm}. \\
&= \tilde{\Delta}\frac{L + (Sm)^\theta}{Sm}.
\end{aligned}$$

□

3. Implementation details of SSD-SPG and SSD-SVRG

In this section we describe all chosen parameter values using the notation found in (Xu et al., 2018). The algorithm SSDC-SPG calls a stochastic proximal gradient (SPG) algorithm K times. For the k^{th} iteration, the number of iterations of SPG equals $T_k = 4k$. Each iteration of SPG uses one gradient call. We used the minimum K which ensured at least en gradient calls were used. The convex majorant parameter $\gamma = 3L$, and the step size $\eta_t = 1/(L(t+1))$. The Moreau envelope parameter $\mu = \epsilon$, where $K = O(1/\epsilon^4)$, is the only non-explicitly given parameter, which we set to $\mu = 1/\left(K^{\frac{1}{4}}\right)$. SSDC-SVRG calls a stochastic variance reduced gradient (SVRG) algorithm K times. We set the inner loop length $T_k = \max(2, 200L/\gamma)$, and the outer loop length $S_k = \lceil \log_2(k) \rceil$. The step size $\eta_k = 0.05/L$. Two parameters are not explicitly given, similar to in SSDC-SPG, we set $\mu = 1/\left(K^{\frac{1}{4}}\right)$. For these parameter settings, there seems to be no restriction on γ. Their SVRG algorithm is based off of the work of Xiao & Zhang (2014), where empirical testing of different sizes of T_k was done for a binary classification problem. The best performance was found with a choice of $T_k = 2n$, from which we were able to determine γ. Given γ, we were then able to solve for K, ensuring at least en gradient calls were used.

References

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xu, Y., Qi, Q., Lin, Q., Jin, R., and Yang, T. Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829v1,* Access date: December 3, 2018.