Supplementary material for
“Variance-Reduced and Projection-Free Stochastic Optimization”

A. Proof of Property (1)

Proof. We drop the subscript $i$ for conciseness. Define $g(w) = f(w) - \nabla f(v)\trans w$, which is clearly also convex and $L$-smooth on $\Omega$. Since $\nabla g(v) = 0$, $v$ is one of the minimizers of $g(w)$. Therefore we have

$$g(v) - g(w) \leq g(w - \frac{1}{L} \nabla g(w)) - g(w)$$

$$\leq \nabla g(w)\trans (w - \frac{1}{L} \nabla g(w) - w) + \frac{L}{2} \|w - \frac{1}{L} \nabla g(w) - w\|^2 \quad \text{(by smoothness of $g$)}$$

$$= -\frac{1}{2L} \|\nabla g(w)\|^2 = -\frac{1}{2L} \|\nabla f(w) - \nabla f(v)\|^2$$

Rearranging and plugging in the definition of $g$ concludes the proof.

B. Analysis for SFW

The concrete update of SFW is

$$v_k = \arg\min_{v \in \Omega} \tilde{\nabla}_k^\trans v$$

$$w_k = (1 - \gamma_k)w_{k-1} + \gamma_k v_k$$

where $\tilde{\nabla}_k$ is the average of $m_k$ iid samples of stochastic gradient $\nabla f_i(w_{k-1})$. The convergence rate of SFW is presented below.

**Theorem 3.** If each $f_i$ is $G$-Lipschitz, then with $\gamma_k = \frac{2}{k+1}$ and $m_k = \left(\frac{G(k+1)}{LD}\right)^2$, SFW ensures for any $k$,

$$\mathbb{E}[f(w_k) - f(w^*)] \leq \frac{4LD^2}{k+2}.$$  

Proof. Similar to the proof of Lemma 2, we first proceed as follows,

$$f(w_k) \leq f(w_{k-1}) + \nabla f(w_{k-1})\trans (w_k - w_{k-1}) + \frac{L}{2} \|w_k - w_{k-1}\|^2 \quad \text{(smoothness)}$$

$$= f(w_{k-1}) + \gamma_k \nabla f(w_{k-1})\trans (v_k - w_{k-1}) + \frac{L\gamma_k^2}{2} \|v_k - x_{k-1}\|^2$$

$$\leq f(w_{k-1}) + \gamma_k \tilde{\nabla}_k^\trans (v_k - w_{k-1}) + \gamma_k (\nabla f(w_{k-1}) - \tilde{\nabla}_k)\trans (v_k - w_{k-1}) + \frac{LD^2\gamma_k^2}{2}$$

$$\leq f(w_{k-1}) + \gamma_k \tilde{\nabla}_k^\trans (w^* - w_{k-1}) + \gamma_k (\nabla f(w_{k-1}) - \tilde{\nabla}_k)\trans (v_k - w_{k-1}) + \frac{LD^2\gamma_k^2}{2} \quad \text{(by optimality of $v_k$)}$$

$$= f(w_{k-1}) + \gamma_k \nabla f(w_{k-1})\trans (w^* - w_{k-1}) + \gamma_k (\nabla f(w_{k-1}) - \tilde{\nabla}_k)\trans (v_k - w^*) + \frac{LD^2\gamma_k^2}{2}$$

$$\leq f(w_{k-1}) + \gamma_k (f(w^*) - f(w_{k-1})) + \gamma_k D\|\tilde{\nabla}_k - \nabla f(w_{k-1})\| + \frac{LD^2\gamma_k^2}{2},$$

where the last step is by convexity and Cauchy-Schwarz inequality. Since $f_i$ is $G$-Lipschitz, with Jensen’s inequality, we further have $\mathbb{E}[\|\tilde{\nabla}_k - \nabla f(w_{k-1})\|] \leq \sqrt{\mathbb{E}[\|\tilde{\nabla}_k - \nabla f(w_{k-1})\|^2]} \leq \frac{G}{\sqrt{m_k}}$, which is at most $\frac{LD^2}{\epsilon}$ with the choice of $\gamma_k$ and $m_k$. So we arrive at $\mathbb{E}[f(w_k) - f(w^*)] \leq (1 - \gamma_k)\mathbb{E}[f(w_{k-1}) - f(w^*)] + LD^2\gamma_k^2$. It remains to use a simple induction to conclude the proof.

Now it is clear that to achieve $1 - \epsilon$ accuracy, SFW needs $\mathcal{O}\left(\frac{LD^2}{\epsilon}\right)$ iterations, and in total $\mathcal{O}\left(\frac{G^2}{\epsilon^2}\left(\frac{LD^2}{\epsilon}\right)^3\right) = \mathcal{O}\left(\frac{G^2LD^4}{\epsilon^4}\right)$ stochastic gradients.
C. Proof of Lemma 3

Proof. Let $\delta_s = \nabla f(z_s) - \nabla f(z_s)$. For any $s \leq k$, we proceed as follows:

$$
\begin{align*}
    f(y_s) &\leq f(z_s) + \nabla f(z_s)^T (y_s - z_s) + \frac{L}{2} \|y_s - z_s\|^2 \\
    &= (1 - \gamma_s) (f(z_s) + \nabla f(z_s)^T (y_{s-1} - z_s)) + \gamma_s (f(z_s) + \nabla f(z_s)^T (w_s - z_s)) + \gamma_s \nabla f(z_s)^T (x_s - w_s) \\
    &\quad + \frac{L\gamma_s^2}{2} \|x_s - x_{s-1}\|^2 \\
    &\leq (1 - \gamma_s) f(y_{s-1}) + \gamma_s f(w_s) + \gamma_s \nabla f(z_s)^T (x_s - w_s) + \frac{L\gamma_s^2}{2} \|x_s - x_{s-1}\|^2 \\
    &\leq (1 - \gamma_s) f(y_{s-1}) + \gamma_s f(w_s) + \gamma_s \eta_s + \beta_s (x_s - x_{s-1})^T (x_s - w_s) + \frac{L\gamma_s^2}{2} \|x_s - x_{s-1}\|^2 + \gamma_s \delta_s^T (w_s - x_s) \\
    &\quad + \frac{\gamma_s}{2} \left( (L\gamma_s - \beta_s) \|x_s - x_{s-1}\|^2 + 2\delta_s^T (x_s - x_{s-1} + 2\delta_s^T (w_s - x_{s-1})) \right) \\
    &\leq (1 - \gamma_s) f(y_{s-1}) + \gamma_s f(w_s) + \gamma_s \eta_s + \beta_s (x_s - x_{s-1})^T (x_s - w_s) + \frac{L\gamma_s^2}{2} \|x_s - x_{s-1}\|^2 + \gamma_s \left( \frac{\|\delta_s\|^2}{\beta_s - L\gamma_s} + 2\delta_s^T (w_s - x_{s-1}) \right),
\end{align*}
$$

where the last inequality is by the fact $\beta_s \geq L\gamma_s$ and thus

$$(L\gamma_s - \beta_s) \|x_s - x_{s-1}\|^2 + 2\delta_s^T (x_s - x_{s-1}) = \frac{\|\delta_s\|^2}{\beta_s - L\gamma_s} - (\beta_s - L\gamma_s) \|x_s - x_{s-1} - \frac{\delta_s}{\beta_s - L\gamma_s}\|^2 \leq \frac{\|\delta_s\|^2}{\beta_s - L\gamma_s}.$$

Note that $E[\delta_s^T (w_s - x_{s-1})] = 0$. So with the condition $E[\|\delta_s\|^2] \leq \frac{L^2D^2}{N^2(s+1)^2} \equiv \sigma_s^2$ we arrive at

$$E[f(y_s) - f(w_s)] \leq (1 - \gamma_s) E[f(y_{s-1}) - f(w_s)] + \gamma_s \left( \eta_s + \beta_s \left( E[\|x_{s-1} - w_s\|^2] - E[\|x_{s-1} - x_{s-1}\|^2] \right) + \frac{\sigma_s^2}{2(\beta_s - L\gamma_s)} \right).$$

Now define $\Gamma_s = \Gamma_{s-1} - (1 - \gamma_s)$ when $s > 1$ and $\Gamma_1 = 1$. By induction, one can verify $\Gamma_s = \frac{2}{s(s+1)}$ and the following:

$$E[f(y_k) - f(w_s)] \leq \Gamma_k \sum_{s=1}^{k} \frac{\gamma_s}{\Gamma_s} \left( \eta_s + \frac{\sigma_s^2}{2(\beta_s - L\gamma_s)} \right) + \frac{\Gamma_k}{2} \left( \frac{\gamma_1 \beta_1}{\Gamma_1} E[\|x_0 - w_s\|^2] + \sum_{s=2}^{k} \left( \frac{\gamma_s \beta_s}{\Gamma_s} - \frac{\gamma_{s-1} \beta_{s-1}}{\Gamma_{s-1}} \right) E[\|x_{s-1} - w_s\|^2] \right).$$

Finally plugging in the parameters $\gamma_s$, $\beta_s$, $\eta_s$, $\Gamma_s$ and the bound $E[\|x_0 - w_s\|^2] \leq D^2$ concludes the proof:

$$E[f(y_k) - f(w_s)] \leq \frac{2}{k(k+1)} \sum_{s=1}^{k} \frac{2LD^2}{N_t k} \left( + \frac{LD^2}{2N_t(k+1)} \right) + \frac{3LD^2}{k(k+1)} \leq \frac{8LD^2}{k(k+1)}.$$

\[\square\]