Supplementary material for "Variance-Reduced and Projection-Free Stochastic Optimization"

A. Proof of Property (1)

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

$$\begin{aligned} g(\boldsymbol{v}) - g(\boldsymbol{w}) &\le g(\boldsymbol{w} - \frac{1}{L}\nabla g(\boldsymbol{w})) - g(\boldsymbol{w}) \\ &\le \nabla g(\boldsymbol{w})^\top(\boldsymbol{w} - \frac{1}{L}\nabla g(\boldsymbol{w}) - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{w} - \frac{1}{L}\nabla g(\boldsymbol{w}) - \boldsymbol{w}\|^2 && \text{(by smoothness of } g) \\ &= -\frac{1}{2L}\|\nabla g(\boldsymbol{w})\|^2 = -\frac{1}{2L}\|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\|^2 \end{aligned}$$

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

B. Analysis for SFW

The concrete update of SFW is

$$\begin{aligned} \boldsymbol{v}_k &= \operatorname*{argmin}_{\boldsymbol{v}\in\Omega} \tilde{\nabla}_k^\top \boldsymbol{v} \\ \boldsymbol{w}_k &= (1-\gamma_k)\boldsymbol{w}_{k-1} + \gamma_k \boldsymbol{v}_k \end{aligned}$$

where $\tilde{\nabla}_k$ is the average of m_k iid samples of stochastic gradient $\nabla f_i(\boldsymbol{w}_{k-1})$. 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,*

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

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

$$\begin{aligned} f(\boldsymbol{w}_k) &\le f(\boldsymbol{w}_{k-1}) + \nabla f(\boldsymbol{w}_{k-1})^\top(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}) + \frac{L}{2}\|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|^2 && \text{(smoothness)} \\ &= f(\boldsymbol{w}_{k-1}) + \gamma_k \nabla f(\boldsymbol{w}_{k-1})^\top(\boldsymbol{v}_k - \boldsymbol{w}_{k-1}) + \frac{L\gamma_k^2}{2}\|\boldsymbol{v}_k - \boldsymbol{x}_{k-1}\|^2 && (\boldsymbol{w}_k - \boldsymbol{w}_{k-1} = \gamma_k(\boldsymbol{v}_k - \boldsymbol{w}_{k-1})) \\ &\le f(\boldsymbol{w}_{k-1}) + \gamma_k \tilde{\nabla}_k^\top(\boldsymbol{v}_k - \boldsymbol{w}_{k-1}) + \gamma_k(\nabla f(\boldsymbol{w}_{k-1}) - \tilde{\nabla}_k)^\top(\boldsymbol{v}_k - \boldsymbol{w}_{k-1}) + \frac{LD^2\gamma_k^2}{2} && (\|\boldsymbol{v}_k - \boldsymbol{w}_{k-1}\| \le D) \\ &\le f(\boldsymbol{w}_{k-1}) + \gamma_k \tilde{\nabla}_k^\top(\boldsymbol{w}^* - \boldsymbol{w}_{k-1}) + \gamma_k(\nabla f(\boldsymbol{w}_{k-1}) - \tilde{\nabla}_k)^\top(\boldsymbol{v}_k - \boldsymbol{w}_{k-1}) + \frac{LD^2\gamma_k^2}{2} && \text{(by optimality of } \boldsymbol{v}_k) \\ &= f(\boldsymbol{w}_{k-1}) + \gamma_k \nabla f(\boldsymbol{w}_{k-1})^\top(\boldsymbol{w}^* - \boldsymbol{w}_{k-1}) + \gamma_k(\nabla f(\boldsymbol{w}_{k-1}) - \tilde{\nabla}_k)^\top(\boldsymbol{v}_k - \boldsymbol{w}^*) + \frac{LD^2\gamma_k^2}{2} \\ &\le f(\boldsymbol{w}_{k-1}) + \gamma_k(f(\boldsymbol{w}^*) - f(\boldsymbol{w}_{k-1})) + \gamma_k D\|\tilde{\nabla}_k - \nabla f(\boldsymbol{w}_{k-1})\| + \frac{LD^2\gamma_k^2}{2}, \end{aligned}$$

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

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

C. Proof of Lemma 3

Proof. Let $\boldsymbol{\delta}_s = \tilde{\nabla}_s - \nabla f(\boldsymbol{z}_s)$. For any $s \le k$, we proceed as follows:

$$\begin{aligned}
f(\boldsymbol{y}_s) &\le f(\boldsymbol{z}_s) + \nabla f(\boldsymbol{z}_s)^\top (\boldsymbol{y}_s - \boldsymbol{z}_s) + \frac{L}{2}\|\boldsymbol{y}_s - \boldsymbol{z}_s\|^2 && \text{(by smoothness)} \\
&= (1-\gamma_s)(f(\boldsymbol{z}_s) + \nabla f(\boldsymbol{z}_s)^\top(\boldsymbol{y}_{s-1} - \boldsymbol{z}_s)) + \gamma_s(f(\boldsymbol{z}_s) + \nabla f(\boldsymbol{z}_s)^\top(\boldsymbol{w}^* - \boldsymbol{z}_s)) + \gamma_s \nabla f(\boldsymbol{z}_s)^\top(\boldsymbol{x}_s - \boldsymbol{w}^*) \\
&\quad + \frac{L\gamma_s^2}{2}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 && \text{(by definition of } \boldsymbol{y}_s \text{ and } \boldsymbol{z}_s) \\
&\le (1-\gamma_s) f(\boldsymbol{y}_{s-1}) + \gamma_s f(\boldsymbol{w}^*) + \gamma_s \nabla f(\boldsymbol{z}_s)^\top(\boldsymbol{x}_s - \boldsymbol{w}^*) + \frac{L\gamma_s^2}{2}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 && \text{(by convexity)} \\
&= (1-\gamma_s) f(\boldsymbol{y}_{s-1}) + \gamma_s f(\boldsymbol{w}^*) + \gamma_s \tilde{\nabla}_s^\top(\boldsymbol{x}_s - \boldsymbol{w}^*) + \frac{L\gamma_s^2}{2}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 + \gamma_s \boldsymbol{\delta}_s^\top(\boldsymbol{w}^* - \boldsymbol{x}_s) \\
&\le (1-\gamma_s) f(\boldsymbol{y}_{s-1}) + \gamma_s f(\boldsymbol{w}^*) + \gamma_s \eta_{t,s} - \gamma_s\beta_s(\boldsymbol{x}_s - \boldsymbol{x}_{s-1})^\top(\boldsymbol{x}_s - \boldsymbol{w}^*) + \frac{L\gamma_s^2}{2}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 + \gamma_s \boldsymbol{\delta}_s^\top(\boldsymbol{w}^* - \boldsymbol{x}_s) \\
& && \text{(by Eq. (4))} \\
&= (1-\gamma_s) f(\boldsymbol{y}_{s-1}) + \gamma_s f(\boldsymbol{w}^*) + \gamma_s \eta_{t,s} + \frac{\beta_s\gamma_s}{2}(\|\boldsymbol{x}_{s-1} - \boldsymbol{w}^*\|^2 - \|\boldsymbol{x}_s - \boldsymbol{w}^*\|^2) + \\
&\quad \frac{\gamma_s}{2}\Big((L\gamma_s - \beta_s)\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 + 2\boldsymbol{\delta}_s^\top(\boldsymbol{x}_{s-1} - \boldsymbol{x}_s) + 2\boldsymbol{\delta}_s^\top(\boldsymbol{w}^* - \boldsymbol{x}_{s-1})\Big) \\
&\le (1-\gamma_s) f(\boldsymbol{y}_{s-1}) + \gamma_s f(\boldsymbol{w}^*) + \gamma_s \eta_{t,s} + \frac{\beta_s\gamma_s}{2}(\|\boldsymbol{x}_{s-1} - \boldsymbol{w}^*\|^2 - \|\boldsymbol{x}_s - \boldsymbol{w}^*\|^2) + \frac{\gamma_s}{2}\left(\frac{\|\boldsymbol{\delta}_s\|^2}{\beta_s - L\gamma_s} + 2\boldsymbol{\delta}_s^\top(\boldsymbol{w}^* - \boldsymbol{x}_{s-1})\right),
\end{aligned}$$

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

$$(L\gamma_s - \beta_s)\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 + 2\boldsymbol{\delta}_s^\top(\boldsymbol{x}_{s-1} - \boldsymbol{x}_s) = \frac{\|\boldsymbol{\delta}_s\|^2}{\beta_s - L\gamma_s} - (\beta_s - L\gamma_s)\left\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1} - \frac{\boldsymbol{\delta}_s}{\beta_s - L\gamma_s}\right\|^2 \le \frac{\|\boldsymbol{\delta}_s\|^2}{\beta_s - L\gamma_s}.$$

Note that $\mathbb{E}[\boldsymbol{\delta}_s^\top(\boldsymbol{w}^* - \boldsymbol{x}_{s-1})] = \mathbf{0}$. So with the condition $\mathbb{E}[\|\boldsymbol{\delta}_s\|^2] \le \frac{L^2 D_t^2}{N_t(s+1)^2} \stackrel{\text{def}}{=} \sigma_s^2$ we arrive at

$$\mathbb{E}[f(\boldsymbol{y}_s) - f(\boldsymbol{w}^*)] \le (1-\gamma_s)\mathbb{E}[f(\boldsymbol{y}_{s-1}) - f(\boldsymbol{w}^*)] + \gamma_s\left(\eta_{t,s} + \frac{\beta_s}{2}(\mathbb{E}[\|\boldsymbol{x}_{s-1} - \boldsymbol{w}^*\|^2] - \mathbb{E}[\|\boldsymbol{x}_s - \boldsymbol{w}^*\|^2]) + \frac{\sigma_s^2}{2(\beta_s - L\gamma_s)}\right).$$

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:

$$\mathbb{E}[f(\boldsymbol{y}_k) - f(\boldsymbol{w}^*)] \le \Gamma_k \sum_{s=1}^k \frac{\gamma_s}{\Gamma_s}\left(\eta_{t,s} + \frac{\beta_s}{2}(\mathbb{E}[\|\boldsymbol{x}_{s-1} - \boldsymbol{w}^*\|^2] - \mathbb{E}[\|\boldsymbol{x}_s - \boldsymbol{w}^*\|^2]) + \frac{\sigma_s^2}{2(\beta_s - L\gamma_s)}\right),$$

which is at most

$$\Gamma_k \sum_{s=1}^k \frac{\gamma_s}{\Gamma_s}\left(\eta_s + \frac{\sigma_s^2}{2(\beta_s - L\gamma_s)}\right) + \frac{\Gamma_k}{2}\left(\frac{\gamma_1\beta_1}{\Gamma_1}\mathbb{E}[\|\boldsymbol{x}_0 - \boldsymbol{w}^*\|^2] + \sum_{s=2}^k\left(\frac{\gamma_s\beta_s}{\Gamma_s} - \frac{\gamma_{s-1}\beta_{s-1}}{\Gamma_{s-1}}\right)\mathbb{E}[\|\boldsymbol{x}_{s-1} - \boldsymbol{w}^*\|^2]\right).$$

Finally plugging in the parameters $\gamma_s, \beta_s, \eta_{t,s}, \Gamma_s$ and the bound $\mathbb{E}[\|\boldsymbol{x}_0 - \boldsymbol{w}^*\|^2] \le D_t^2$ concludes the proof:

$$\mathbb{E}[f(\boldsymbol{y}_k) - f(\boldsymbol{w}^*)] \le \frac{2}{k(k+1)}\sum_{s=1}^k k\left(\frac{2LD_t^2}{N_t k} + \frac{LD_t^2}{2N_t(k+1)}\right) + \frac{3LD_t^2}{k(k+1)} \le \frac{8LD_t^2}{k(k+1)}.$$

□