A Incorporating additive error for Nesterov acceleration

For this section, we assume an additive error in the the strong growth condition implying that the following equation is satisfied for all w, z.

$$\mathbb{E}_z \left\|\nabla f(w,z)\right\|^2 \leq \rho \left\|\nabla f(w)\right\|^2 + \sigma^2$$

In this case, we have the counterparts of Theorems 1 and 2 as follows:

Theorem 7 (Strongly convex)**.** *Under L-smoothness and μ strongly-convexity, if f satisfies SGC with constant ρ and an additive error σ, then SGD with Nesterov acceleration with the following choice of parameters,*

$$\gamma_k = \frac{1}{\sqrt{\mu\eta\rho}} \quad ; \quad \beta_k = 1 - \sqrt{\frac{\mu\eta}{\rho}}$$
$$b_{k+1} = \frac{\sqrt{\mu}}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$
$$a_{k+1} = \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$
$$\alpha_k = \frac{\gamma_k \beta_k b_{k+1}^2 \eta}{\gamma_k \beta_k b_{k+1}^2 \eta + a_k^2}; \quad \eta = \frac{1}{\rho L}$$

results in the following convergence rate:

$$[\mathbb{E}[f(w_{k+1})] - f(w^*)] \leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k \left[f(x_0) - f(w^*) + \frac{\mu}{2}\left\|x_0 - w^*\right\|^2\right] + \frac{\sigma^2\sqrt{\eta}}{\sqrt{\rho\mu}}$$

Theorem 8 (Convex)**.** *Under L-smoothness and convexity, if f satisfies SGC with constant ρ and an additive error σ, then SGD with Nesterov acceleration with the following choice of parameters,*

$$\gamma_k = \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2}$$
$$a_{k+1} = \gamma_k \sqrt{\eta\rho}$$
$$\alpha_k = \frac{\gamma_k \eta}{\gamma_k \eta + a_k^2}; \quad \eta = \frac{1}{\rho L}$$

results in the following convergence rate:

$$[\mathbb{E} f(w_{k+1}) - f(w^*)] \leq \frac{2\rho}{k^2\eta}\left\|x_0 - w^*\right\|^2 + \frac{k\sigma^2\eta}{\rho}$$

The above theorems are proved in appendices B.1.1 and B.1.3

B Proofs

B.1 Proofs for SGD with Nesterov Acceleration

Recall the update equations for SGD with Nesterov acceleration as follows:

$$w_{k+1} = \zeta_k - \eta \nabla f(\zeta_k, z_k)$$
$$\zeta_k = \alpha_k v_k + (1 - \alpha_k) w_k$$
$$v_{k+1} = \beta_k v_k + (1 - \beta_k)\zeta_k - \gamma_k \eta \nabla f(\zeta_k, z_k)$$

Since the stochastic gradients are unbiased, we obtain the following equation,

$$\mathbb{E}_z \left[\nabla f(y, z)\right] = \nabla f(y) \tag{9}$$

For the proof, we consider the more general strong-growth condition with an additive error σ^2.

$$\mathbb{E}_z \left\|\nabla f(w, z)\right\|^2 \leq \rho \left\|\nabla f(w)\right\|^2 + \sigma^2 \tag{10}$$

We choose the parameters γ_k, α_k, β_k, a_k, b_k such that the following equations are satisfied:

$$\gamma_k = \frac{1}{\rho} \cdot \left[1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right] \tag{11}$$

$$\alpha_k = \frac{\gamma_k \beta_k b_{k+1}^2 \eta}{\gamma_k \beta_k b_{k+1}^2 \eta + a_k^2} \tag{12}$$

$$\beta_k \geq 1 - \gamma_k \mu \eta \tag{13}$$

$$a_{k+1} = \gamma_k \sqrt{\eta \rho} b_{k+1} \tag{14}$$

$$b_{k+1} \leq \frac{b_k}{\sqrt{\beta_k}} \tag{15}$$

We now prove the following lemma assuming that the function $f(\cdot)$ is L-smooth and μ strongly-convex.

Lemma 3. *Assume that the function is L-smooth and μ strongly-convex and satisfies the strong-growth condition in Equation 10. Then, using the updates in Equation 3-5 and setting the parameters according to Equations 11-15, if $\eta \leq \frac{1}{\rho L}$, then the following relation holds:*

$$b_{k+1}^2 \gamma_k^2 \left[\mathbb{E} f(w_{k+1}) - f^*\right] \leq \frac{a_0^2}{\rho \eta}\left[f(x_0) - f^*\right] + \frac{b_0^2}{2\rho\eta}\left\|x_0 - w^*\right\|^2 + \frac{\sigma^2 \eta}{\rho} \sum_{i=0}^{k} [\gamma_i^2 b_{i+1}^2]$$

Proof.

Let $r_{k+1} = \|v_{k+1} - w^*\|$, then using equation 5

$$r_{k+1}^2 = \left\|\beta_k v_k + (1-\beta_k)\zeta_k - w^* - \gamma_k \eta \nabla f(\zeta_k, z_k)\right\|^2$$

$$r_{k+1}^2 = \left\|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k \eta \langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle$$

Taking expecation wrt to z_k,

$$\begin{aligned}
\mathbb{E}[r_{k+1}^2] &= \mathbb{E}[\|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2] + \gamma_k^2 \eta^2 \mathbb{E}\left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k \eta \left[\mathbb{E}\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle\right] \\
&\leq \|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&= \|\beta_k (v_k - w^*) + (1-\beta_k)(\zeta_k - w^*)\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&\leq \beta_k \left\|v_k - w^*\right\|^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&\qquad \text{(By convexity of } \|\cdot\|^2\text{)} \\
&= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle \beta_k(\zeta_k - v_k) + w^* - \zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\langle \frac{\beta_k(1-\alpha_k)}{\alpha_k}(w_k - \zeta_k) + w^* - \zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&\qquad \text{(From equation 4)} \\
&= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\langle \nabla f(\zeta_k), (w_k - \zeta_k)\rangle + \langle \nabla f(\zeta_k), w^* - \zeta_k\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k \eta \left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + \langle \nabla f(\zeta_k), w^* - \zeta_k\rangle\right] + \gamma_k^2 \eta^2 \sigma^2 \\
&\qquad \text{(By convexity)}
\end{aligned}$$

By strong convexity,

$$
\begin{aligned}
\mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k) \left\|\zeta_k - w^*\right\|^2 + \gamma_k^2 \eta^2 \rho \left\|\nabla f(\zeta_k)\right\|^2 \\
&+ 2\gamma_k \eta \left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\left(f(w_k) - f(\zeta_k)\right) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2 \eta^2 \sigma^2
\end{aligned} \tag{16}
$$

By Lipschitz continuity of the gradient,

$$
\begin{aligned}
f(w_{k+1}) - f(\zeta_k) &\leq \langle \nabla f(\zeta_k), w_{k+1} - \zeta_k \rangle + \frac{L}{2}\left\|w_{k+1} - \zeta_k\right\|^2 \\
&\leq -\eta \langle \nabla f(\zeta_k), \nabla f(\zeta_k, z_k) \rangle + \frac{L\eta^2}{2}\left\|\nabla f(\zeta_k, z_k)\right\|^2
\end{aligned}
$$

Taking expectation wrt z_k and using equations 9, 10

$$
\begin{aligned}
\mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq -\eta \left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\rho\eta^2}{2}\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\
\mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq \left[-\eta + \frac{L\rho\eta^2}{2}\right]\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2}
\end{aligned}
$$

If $\eta \leq \frac{1}{\rho L}$,

$$
\begin{aligned}
\mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq \left(\frac{-\eta}{2}\right)\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\
\implies \left\|\nabla f(\zeta_k)\right\|^2 &\leq \left(\frac{2}{\eta}\right)\mathbb{E}[f(\zeta_k) - f(w_{k+1})] + L\eta\sigma^2
\end{aligned} \tag{17}
$$

From equations 16 and 17,

$$
\begin{aligned}
\mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2 \rho \eta \mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\
&+ 2\gamma_k \eta \left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\left(f(w_k) - f(\zeta_k)\right) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2 \eta^2 \sigma^2 + L\gamma_k^2 \eta^3 \rho \sigma^2 \\
&\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2 \eta \rho \mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\
&+ 2\gamma_k \eta \left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\left(f(w_k) - f(\zeta_k)\right) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + 2\gamma_k^2 \eta^2 \sigma^2 \qquad \left(\text{Since } \eta \leq \tfrac{1}{\rho L}\right) \\
&= \beta_k r_k^2 + \left\|\zeta_k - w^*\right\|^2 \left[(1-\beta_k) - \gamma_k \mu \eta\right] + f(\zeta_k)\left[2\gamma_k^2 \eta \rho - 2\gamma_k \eta \cdot \frac{\beta_k(1-\alpha_k)}{\alpha_k} - 2\gamma_k \eta\right] \\
&- 2\gamma_k^2 \eta \rho \mathbb{E} f(w_{k+1}) + 2\gamma_k \eta f^* + \left[2\gamma_k \eta \cdot \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right] f(w_k) + 2\gamma_k^2 \eta^2 \sigma^2
\end{aligned}
$$

Since $\beta_k \geq 1 - \gamma_k \mu \eta$ and $\gamma_k = \frac{1}{\rho} \cdot \left(1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right)$,

$$
\mathbb{E}[r_{k+1}^2] \leq \beta_k r_k^2 - 2\gamma_k^2 \eta \rho \mathbb{E} f(w_{k+1}) + 2\gamma_k \eta f^* + \left[2\gamma_k \eta \cdot \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right] f(w_k) + 2\gamma_k^2 \eta^2 \sigma^2
$$

Multiplying by b_{k+1}^2,

$$
b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq b_{k+1}^2 \beta_k r_k^2 - 2b_{k+1}^2 \gamma_k^2 \eta \rho \mathbb{E} f(w_{k+1}) + 2b_{k+1}^2 \gamma_k \eta f^* + \left[2b_{k+1}^2 \gamma_k \eta \cdot \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right] f(w_k) + 2b_{k+1}^2 \gamma_k^2 \eta^2 \sigma^2
$$

Since $b_{k+1}^2 \beta_k \leq b_k^2$, $b_{k+1}^2 \gamma_k^2 \eta \rho = a_{k+1}^2$, $\frac{\gamma_k \eta \beta_k (1-\alpha_k)}{\alpha_k} = \frac{a_k^2}{b_{k+1}^2}$

$$
\begin{aligned}
b_{k+1}^2 \mathbb{E}[r_{k+1}^2] &\leq b_k^2 r_k^2 - 2a_{k+1}^2 \mathbb{E} f(w_{k+1}) + 2b_{k+1}^2 \gamma_k \eta f^* + 2a_k^2 f(w_k) + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho} \\
&= b_k^2 r_k^2 - 2a_{k+1}^2 \left[\mathbb{E} f(w_{k+1}) - f^*\right] + 2a_k^2 \left[f(w_k) - f^*\right] + 2\left[b_{k+1}^2 \gamma_k \eta - a_{k+1}^2 + a_k^2\right] f^* + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho}
\end{aligned}
$$

Since $\left[b_{k+1}^2 \gamma_k \eta - a_{k+1}^2 + a_k^2\right] = 0$,

$$
b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq b_k^2 r_k^2 - 2a_{k+1}^2 \left[\mathbb{E} f(w_{k+1}) - f^*\right] + 2a_k^2 \left[f(w_k) - f^*\right] + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho}
$$

Denoting $\mathbb{E} f(w_{k+1})$ as ϕ_k,

$$
2a_{k+1}^2 \left[\phi_{k+1} - f^*\right] + b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq 2a_k^2 \left[\phi_k - f^*\right] + b_k^2 \mathbb{E}[r_k^2] + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho}
$$

By recursion,

$$
\begin{aligned}
2a_{k+1}^2 \left[\phi_{k+1} - f^*\right] + b_{k+1}^2 \mathbb{E}[r_{k+1}^2] &\leq 2a_0^2 \left[f(x_0) - f^*\right] + b_0^2 \left\|x_0 - w^*\right\|^2 + \frac{2\sigma^2 \eta}{\rho} \sum_{i=0}^{k} [a_{i+1}^2] \\
2a_{k+1}^2 \left[\phi_{k+1} - f^*\right] &\leq 2a_0^2 \left[f(x_0) - f^*\right] + b_0^2 \left\|x_0 - w^*\right\|^2 + \frac{2\sigma^2 \eta}{\rho} \sum_{i=0}^{k} [a_{i+1}^2] \\
2b_{k+1}^2 \gamma_k^2 \rho \eta \left[\phi_{k+1} - f^*\right] &\leq 2a_0^2 \left[f(x_0) - f^*\right] + b_0^2 \left\|x_0 - w^*\right\|^2 + 2\sigma^2 \eta^2 \rho \sum_{i=0}^{k} [\gamma_i^2 b_{i+1}^2] \\
b_{k+1}^2 \gamma_k^2 \left[\mathbb{E} f(w_{k+1}) - f^*\right] &\leq \frac{a_0^2}{\rho \eta} \left[f(x_0) - f^*\right] + \frac{b_0^2}{2\rho \eta} \left\|x_0 - w^*\right\|^2 + \frac{\sigma^2 \eta}{\rho} \sum_{i=0}^{k} [\gamma_i^2 b_{i+1}^2]
\end{aligned}
$$

□

Lemma 4. *Under the parameter setting according to Equations 11- 15, the following relation is true:*

$$
\gamma_k^2 - \gamma_k \left[\frac{1}{\rho} - \mu \eta \gamma_{k-1}^2\right] = \gamma_{k-1}^2
$$

Proof.

$$
\begin{aligned}
\gamma_k &= \frac{1}{\rho} \left[1 + \frac{\beta_k (1 - \alpha_k)}{\alpha_k}\right] && \text{(From equation 11)} \\
\gamma_k^2 - \frac{\gamma_k}{\rho} &= \frac{\gamma_k \beta_k (1 - \alpha_k)}{\rho \alpha_k} \\
&= \frac{1}{\eta \rho} \frac{a_k^2}{b_{k+1}^2} && \text{(From equation 12)} \\
&= \frac{\beta_k}{\eta \rho} \frac{a_k^2}{b_k^2} && \text{(From equation 15)} \\
&= \frac{1 - \gamma_k \mu \eta}{\eta \rho} \frac{a_k^2}{b_k^2} && \text{(From equation 13)} \\
&= \frac{1 - \gamma_k \mu \eta}{\eta \rho} \left(\gamma_{k-1} \sqrt{\eta \rho}\right)^2 && \text{(From equation 13)} \\
&= \left(1 - \gamma_k \mu \eta\right) \gamma_{k-1}^2
\end{aligned}
$$

$$
\implies \gamma_k^2 - \gamma_k \left[\frac{1}{\rho} - \mu \eta \gamma_{k-1}^2\right] = \gamma_{k-1}^2 \tag{18}
$$

□

B.1.1 Strongly-convex case

We now consider the strongly-convex case,

Using Lemma 4,

$$\gamma_k^2 - \gamma_k \left[\frac{1}{\rho} - \mu\eta\gamma_{k-1}^2\right] = \gamma_{k-1}^2$$

If $\gamma_k = C$, then

$$\gamma_k = \frac{1}{\sqrt{\mu\eta\rho}}$$
$$\beta_k = 1 - \sqrt{\frac{\mu\eta}{\rho}}$$
$$b_{k+1} = \frac{b_0}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$
$$a_{k+1} = \frac{1}{\sqrt{\mu\eta\rho}} \cdot \sqrt{\eta\rho} \cdot \frac{b_0}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} = \frac{b_0}{\sqrt{\mu}} \cdot \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$

If $b_0 = \sqrt{\mu}$,

$$a_{k+1} = \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$

The above equation implies that $a_0 = 1$. This gives us the parameter settings used in Theorem 1.

Using the result of Lemma 3 and the above relations, we obtain the following inequality. Note that $\phi_{k+1} = \mathbb{E}[f(w_{k+1})]$.

$$\frac{\mu}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)}} \cdot \frac{1}{\mu\eta\rho} [\phi_{k+1} - f^*] \leq \frac{1}{\rho\eta} [f(x_0) - f^*] + \frac{\mu}{2\rho\eta} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho} \cdot \frac{1}{\mu\eta\rho} \sum_{i=0}^{k} \frac{\mu}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(i+1)}}$$

$$\frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{k}} [\phi_{k+1} - f^*] \leq [f(x_0) - f^*] + \frac{\mu}{2} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho} \sum_{i=0}^{k} \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(i+1)}}$$

$$\frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{k}} [\phi_{k+1} - f^*] \leq \left[f(x_0) - f^* + \frac{\mu}{2} \|x_0 - w^*\|^2\right] + \frac{\sigma^2\sqrt{\eta}}{\sqrt{\rho\mu}} \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{-k}$$

$$[\phi_{k+1} - f^*] \leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{k} \left[f(x_0) - f^* + \frac{\mu}{2} \|x_0 - w^*\|^2\right] + \frac{\sigma^2\sqrt{\eta}}{\sqrt{\rho\mu}}$$

B.1.2 Proof of Theorem 1

We use the above relation to complete the proof for Theorem 1. Substituting $\eta = \frac{1}{\rho L}$ and $\sigma = 0$, we obtain the following:

$$[\mathbb{E}[f(w_{k+1})] - f^*] \leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k \left[f(x_0) - f^* + \frac{\mu}{2}\|x_0 - w^*\|^2\right]$$

B.1.3 Convex case

We now use the above lemmas to first prove the convergence rate in the convex case. In this case, $\mu = 0$ and the result of Lemma 4 can be written as:

$$\gamma_k^2 - \frac{\gamma_k}{\rho} - \gamma_{k-1}^2 = 0$$
$$\implies \gamma_k = \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2}$$

Let $\gamma_0 = 0$. From equation 13, for all k,

$$\beta_k = 1$$
$$b_{k+1} = b_k = b_0 = 1 \qquad \text{(From equation 15)}$$
$$a_{k+1} = \gamma_k\sqrt{\eta\rho}b_0 \implies a_{k+1} = \gamma_k\sqrt{\eta\rho} \qquad \text{(From equation 14)}$$

The above equation implies that $a_0 = 0$. This gives us the parameter settings used in Theorem 2.

Using the result of Lemma 3 by setting $\mu = 0$ and the above relations, we obtain the following inequality. Note that $\phi_{k+1} = \mathbb{E}[f(w_{k+1})]$.

$$\gamma_k^2 [\phi_{k+1} - f^*] \leq \frac{1}{2\rho\eta}\|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho}\sum_{i=1}^{k-1}[\gamma_i^2]$$

By induction, $\gamma_i \geq \frac{i}{2\rho}$,

$$\frac{k^2}{4\rho^2}[\phi_{k+1} - f^*] \leq \frac{1}{2\rho\eta}\|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{4\rho^3}\sum_{i=1}^{k-1}[i^2]$$
$$[\phi_{k+1} - f^*] \leq \frac{2\rho}{k^2\eta}\|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{k^2\rho}\sum_{i=1}^{k-1}[i^2]$$
$$[\phi_{k+1} - f^*] \leq \frac{2\rho}{k^2\eta}\|x_0 - w^*\|^2 + \frac{k\sigma^2\eta}{\rho}$$

B.1.4 Proof of Theorem 2

We use the above relation to complete the proof for Theorem 2. Substituting $\eta = \frac{1}{\rho L}$ and $\sigma = 0$, we obtain the following:

$$[\mathbb{E}[f(w_{k+1})] - f^*] \leq \frac{2\rho^2 L}{k^2}\|x_0 - w^*\|^2$$

B.2 Proof of Theorem 3

Proof. Recall the stochastic gradient descent update,

$$w_{k+1} = w_k - \eta \nabla f(w_k, z_k) \tag{19}$$

By Lipschitz continuity of the gradient,

$$\begin{aligned}
f(w_{k+1}) - f(w_k) &\leq \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \left\|w_{k+1} - w_k\right\|^2 \\
&\leq -\eta \langle \nabla f(w_k), \nabla f(w_k, z_k) \rangle + \frac{L\eta^2}{2} \left\|\nabla f(w_k, z_k)\right\|^2
\end{aligned}$$

Taking expectation wrt z_k and using equations 9, 10

$$\begin{aligned}
\mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq -\eta \left\|\nabla f(w_k)\right\|^2 + \frac{L\rho\eta^2}{2} \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\
\mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq \left[-\eta + \frac{L\rho\eta^2}{2}\right] \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2}
\end{aligned}$$

If $\eta \leq \frac{1}{\rho L}$,

$$\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq \left(\frac{-\eta}{2}\right) \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2}$$

$$\implies \left\|\nabla f(w_k)\right\|^2 \leq \left(\frac{2}{\eta}\right) \mathbb{E}[f(w_k) - f(w_{k+1})] + L\eta\sigma^2 \tag{20}$$

Taking expectation wrt $z_0, z_1, \dots z_{t-1}$ and summing from $k = 0$ to $t - 1$,

$$\begin{aligned}
\sum_{k=0}^{t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2}{\eta}\right) \sum_{k=0}^{t-1} \mathbb{E}\left[f(w_k) - f(w_{k+1})\right] + L\eta t\sigma^2 \\
\implies \sum_{k=0}^{t-1} \min_{k=0,1,\dots t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2}{\eta}\right) \sum_{k=0}^{t-1} \mathbb{E}\left[f(w_k) - f(w_{k+1})\right] + L\eta\sigma^2 \\
\min_{k=0,1,\dots t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2}{\eta t}\right) \left[f(w_0) - \mathbb{E}[f(w_t)]\right] + L\eta\sigma^2 \\
\min_{k=0,1,\dots t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2}{\eta t}\right) \left[f(w_0) - f(w^*)\right] + L\eta\sigma^2
\end{aligned}$$

If $\sigma = 0$,

$$\begin{aligned}
\min_{k=0,1,\dots t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2}{\eta t}\right) \left[f(w_0) - f(w^*)\right] \\
\implies \min_{k=0,1,\dots t-1} \mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] &\leq \left(\frac{2\rho L}{t}\right) \left[f(w_0) - f(w^*)\right] && \left(\text{Setting } \eta = \tfrac{1}{\rho L}\right)
\end{aligned}$$

□

B.3 Proof of Theorem 4

Proof. Similar to the proof of Theorem 3, we can use the SGD update and Lipschitz continuity of the gradient to obtain the following equation for the stepsize $\eta \leq \frac{1}{\rho L}$:

$$\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq \left(\frac{-\eta}{2}\right) \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2}$$

We now use the PL inequality with constant μ as follows:

$$\|\nabla f(w_k)\|^2 \geq 2\mu \left[f(w_k) - f^*\right]$$

Combining the above two inequalities,

$$\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq -\eta\mu \left[f(w_k) - f^*\right] + \frac{L\eta^2\sigma^2}{2}$$

If $\sigma = 0$,

$$\begin{aligned} &\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq -\eta\mu \left[f(w_k) - f^*\right] \\ \implies &\mathbb{E}[f(w_{k+1}) - f^*] \leq (1 - \eta\mu) \left[f(w_k) - f^*\right] \end{aligned}$$

Substituting $\eta = \frac{1}{\rho L}$,

$$\begin{aligned} &\mathbb{E}[f(w_{k+1}) - f^*] \leq \left(1 - \frac{\mu}{\rho L}\right) \left[f(w_k) - f^*\right] \\ \implies &\mathbb{E}[f(w_{k+1}) - f^*] \leq \left(1 - \frac{\mu}{\rho L}\right)^k \left[f(w_0) - f^*\right] \end{aligned} \tag{21}$$

□

B.4 Proof of Theorem 5

Proof.

$$\begin{aligned} \|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla f(w_k, z) - w^*\|^2 \\ &= \|w_k - w^*\|^2 - 2\eta\langle\nabla f(w_k, z), w_k - w^*\rangle + \eta^2 \|\nabla f(w_k, z)\|^2 \\ \mathbb{E}_z[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\eta\mathbb{E}[\langle\nabla f(w_k, z), w_k - w^*\rangle] + \eta^2\mathbb{E}[\|\nabla f(w_k, z)\|^2] \\ &= \|w_k - w^*\|^2 - 2\eta\langle\nabla f(w_k), w_k - w^*\rangle + \eta^2\mathbb{E}[\|\nabla f(w_k, z)\|^2] \\ &\qquad \text{(From the unbiasedness of stochastic gradients.)} \\ &\leq \|w_k - w^*\|^2 - 2\eta\langle\nabla f(w_k), w_k - w^*\rangle + 2\rho\eta^2 L[f(w_k) - f^*] \qquad \text{(From equation 6)} \\ &\leq \|w_k - w^*\|^2 + 2\eta\left[f^* - f(w_k) - \frac{\mu}{2}\|w_k - w^*\|^2\right] + 2\rho\eta^2 L[f(w_k) - f^*] \\ &\qquad \text{(By strong convexity)} \\ &= (1 - \mu\eta)\|w_k - w^*\|^2 + \left(2\eta^2\rho L - 2\eta\right)[f(w_k) - f^*] \\ \|w_{k+1} - w^*\|^2 &\leq \left(1 - \frac{\mu}{\rho L}\right)\|w_k - w^*\|^2 \qquad \text{(Setting } \eta = \tfrac{1}{\rho L}\text{)} \\ \implies \|w_{k+1} - w^*\|^2 &\leq \left(1 - \frac{\mu}{\rho L}\right)^k \|x_0 - w^*\|^2 \end{aligned}$$

□

B.5 Proof of Theorem 6

Proof.

By convexity,

$$f(w_k) \leq f(w^*) + \langle\nabla f(w_k), w_k - w^*\rangle$$

For any $\beta \leq 1$,

$$f(w_k) \leq \beta f(w_k) + (1-\beta) f(w^*) + (1-\beta)\langle \nabla f(w_k), w_k - w^* \rangle$$

By Lipschitz continuity of $\nabla f(f)$,

$$\begin{aligned}
& f(w_{k+1}) \leq f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \left\| w_{k+1} - w_k \right\|^2 \\
\implies & f(w_{k+1}) \leq f(w_k) - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \left\| \nabla f(w_k, z) \right\|^2
\end{aligned}$$

From the above equations,

$$f(w_{k+1}) \leq \beta f(w_k) + (1-\beta) f(w^*) + (1-\beta)\langle \nabla f(w_k), w_k - w^* \rangle - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \left\| \nabla f(w_k, z) \right\|^2$$

Note that,

$$\begin{aligned}
\frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \left\| w_{k+1} - w^* \right\|^2 \right) &= \frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \left\| w_k - \eta \nabla f(w_k, z) - w^* \right\|^2 \right) \\
&= \frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \left\| w_k - w^* \right\|^2 - \eta^2 \left\| \nabla f(w_k, z) \right\|^2 + 2\eta \langle w_k - w^*, \nabla f(w_k, z) \rangle \right) \\
\frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \left\| w_{k+1} - w^* \right\|^2 \right) &= \frac{-\eta}{2} \left\| \nabla f(w_k, z) \right\|^2 + \langle w_k - w^*, \nabla f(w_k, z) \rangle \\
\implies \langle w_k - w^*, \nabla f(w_k, z) \rangle &= \frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \left\| w_{k+1} - w^* \right\|^2 \right) + \frac{\eta}{2} \left\| \nabla f(w_k, z) \right\|^2
\end{aligned}$$

Taking expectation

$$\begin{aligned}
\mathbb{E}\left[\langle w_k - w^*, \nabla f(w_k, z) \rangle \right] &= \frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \mathbb{E}\left[\left\| w_{k+1} - w^* \right\|^2 \right] \right) + \frac{\eta}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right] \\
\implies \langle w_k - w^*, \nabla f(w_k) \rangle &= \frac{1}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \mathbb{E}\left[\left\| w_{k+1} - w^* \right\|^2 \right] \right) + \frac{\eta}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right]
\end{aligned}$$

Using the above equations,

$$\begin{aligned}
f(w_{k+1}) \leq\ & \beta f(w_k) + (1-\beta) f(w^*) + \frac{1-\beta}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \mathbb{E}\left[\left\| w_{k+1} - w^* \right\|^2 \right] \right) + \frac{(1-\beta)(\eta)}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right] \\
& - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \left\| \nabla f(w_k, z) \right\|^2
\end{aligned}$$

Taking expectation,

$$\begin{aligned}
\mathbb{E}[f(w_{k+1})] \leq\ & \beta f(w_k) + (1-\beta) f(w^*) + \frac{1-\beta}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \mathbb{E}\left[\left\| w_{k+1} - w^* \right\|^2 \right] \right) + \frac{(1-\beta)(\eta)}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right] \\
& - \eta \langle \nabla f(w_k), \mathbb{E}\left[\nabla f(w_k, z) \right] \rangle + \frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right] \\
=\ & \beta f(w_k) + (1-\beta) f(w^*) + \frac{1-\beta}{2\eta}\left(\left\| w_k - w^* \right\|^2 - \mathbb{E}\left[\left\| w_{k+1} - w^* \right\|^2 \right] \right) + \frac{(1-\beta)(\eta)}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right] \\
& - \eta \left\| \nabla f(w_k) \right\|^2 + \frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \nabla f(w_k, z) \right\|^2 \right]
\end{aligned}$$

The term $-\eta \left\|\nabla f(w_k)\right\|^2 \leq 0$

$$\begin{aligned}
\implies \mathbb{E}[f(w_{k+1})] &\leq \beta f(w_k) + (1-\beta) f(w^*) + \frac{1-\beta}{2\eta}\left(\|w_k - w^*\|^2 - \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]\right) \\
&+ \frac{(1-\beta)(\eta)}{2}\mathbb{E}\left[\|\nabla f(w_k, z)\|^2\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\|\nabla f(w_k, z)\|^2\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[f(w_{k+1})] - f(w^*) &\leq \beta\left(f(w_k) - f(w^*)\right) + \frac{1-\beta}{2\eta}\left(\|w_k - w^*\|^2 - \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]\right) \\
&+ \left(\frac{(1-\beta)(\eta)}{2} + \frac{\eta^2 L}{2}\right)\mathbb{E}\left[\|\nabla f(w_k, z)\|^2\right]
\end{aligned}$$

From equation 6,

$$\begin{aligned}
\mathbb{E}[f(w_{k+1})] - f(w^*) &\leq \beta\left(f(w_k) - f(w^*)\right) + \frac{1-\beta}{2\eta}\left(\|w_k - w^*\|^2 - \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]\right) \\
&+ \left(\rho(1-\beta)\eta L + \eta^2 \rho L^2\right)\left(f(w_k) - f(w^*)\right)
\end{aligned}$$

Let us choose $1 - \beta = \eta L$,

$$\mathbb{E}[f(w_{k+1})] - f(w^*) \leq \beta\left(f(w_k) - f(w^*)\right) + \frac{1-\beta}{2\eta}\left(\|w_k - w^*\|^2 - \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]\right) + 2\rho\eta^2 L^2\left(f(w_k) - f(w^*)\right)$$

$$\mathbb{E}[f(w_{k+1})] - f(w^*) \leq \left(\beta + 2\rho\eta^2 L^2\right)\left(f(w_k) - f(w^*)\right) + \frac{L}{2}\left(\|w_k - w^*\|^2 - \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]\right)$$

Let $\delta_{k+1} = \mathbb{E}[f(w_{k+1})] - f(w^*)$ and $\Delta_{k+1} = \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right]$

$$\implies \delta_{k+1} \leq \left(\beta + 2\rho\eta^2 L^2\right)\delta_k + \frac{L}{2}\left[\Delta_k - \Delta_{k+1}\right]$$

Summing from $i = 0$ to $k - 1$,

$$\begin{aligned}
\sum_{i=0}^{k-1}\delta_{i+1} &\leq \left(\beta + 2\rho\eta^2 L^2\right)\sum_{i=0}^{k-1}\delta_i + \frac{L}{2}\sum_{i=0}^{k-1}\left[\Delta_i - \Delta_{i+1}\right] \\
\implies \sum_{i=0}^{k-1}\delta_{i+1} &\leq \left(\beta + 2\rho\eta^2 L^2\right)\sum_{i=0}^{k-1}\delta_i + \frac{L}{2}\Delta_0 \\
\implies \sum_{i=1}^{k}\delta_i &\leq \frac{\left(\beta + 2\rho\eta^2 L^2\right)\delta_0 + \frac{L}{2}\Delta_0}{\left(1 - \beta - 2\rho\eta^2 L^2\right)}
\end{aligned}$$

Let $\bar{w}_k = \frac{\left[\sum_{i=1}^{k} w_i\right]}{k}$. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[f(\bar{w}_k)] &\leq \frac{\sum_{i=1}^{k}\mathbb{E}[f(w_i)]}{k} \\
\implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \sum_{i=1}^{k}\delta_i \\
\implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{\left(\beta + 2\rho\eta^2 L^2\right)\delta_0 + \frac{L}{2}\Delta_0}{\left(1 - \beta - 2\rho\eta^2 L^2\right)k} \\
\mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{\left(1 - \eta L + 2\rho\eta^2 L^2\right)\left[f(w_0) - f(w^*)\right] + \frac{L}{2}\left\|w_0 - w^*\right\|^2}{\left(\eta L - 2\rho\eta^2 L^2\right)k} \qquad \text{(Since } 1 - \beta = \eta L\text{)}
\end{aligned}$$

If $\eta = \frac{1}{4\rho L}$,

$$\mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{\frac{7}{8\rho}\left[f(w_0) - f(w^*)\right] + \frac{L}{2}\left\|w_0 - w^*\right\|^2}{\frac{1}{8\rho}k}$$

$$\mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{7\left[f(w_0) - f(w^*)\right] + 4\rho L\left\|w_0 - w^*\right\|^2}{k}$$

$$\mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{(7L/2)\left\|w_0 - w^*\right\|^2 + 4\rho L\left\|w_0 - w^*\right\|^2}{k}$$

$$\implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{4(1+\rho)\left\|w_0 - w^*\right\|^2}{k}$$

□

B.6 Proof for Proposition 1

Proof.

For the first part, we use the PL inequality which states the for all w,

$$2\left[f(w) - f(w^*)\right] \leq \frac{1}{\mu}\left\|\nabla f(w)\right\|^2$$

Combining this with the WGC gives us the desired result

For the converse, we use smoothness and the convexity of $f(\cdot)$. Specifically, for all points a, b,

$$f(a) - f(b) \geq \langle f(b), a - b\rangle + \frac{1}{2L}\left\|\nabla f(a) - \nabla f(b)\right\|^2$$

Substituting $a = w$ and $b = w^*$ and rearranging,

$$\left\|\nabla f(w)\right\|^2 \leq 2L \cdot \left[f(w) - f(w^*)\right]$$

Combining this with the SGC gives us the desired result.

□

B.7 Proof for Proposition 2

Proof.

$$\mathbb{E}_i\left\|\nabla f_i(w)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w)\right\|^2 \tag{22}$$

By Lipschitz continuity of $\nabla f_i(w)$ and convexity,

$$f_i(w) - f_i(w^*) \geq \langle \nabla f_i(w^*), w - w^*\rangle + \frac{1}{2L_i}\left\|\nabla f_i(w) - \nabla f_i(w^*)\right\|^2$$

For all i, $\nabla f_i(w^*) = \nabla f(w^*) = 0$. Hence,

$$
\begin{aligned}
& f_i(w) - f_i(w^*) \geq \frac{1}{2L_i} \left\|\nabla f_i(w)\right\|^2 \\
\implies & \left\|\nabla f_i(w)\right\|^2 \leq 2L_i \left[f_i(w) - f_i(w^*)\right]
\end{aligned}
$$

Using Equation 22,

$$
\begin{aligned}
\mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 &\leq \sum_{i=1}^{n} \left[\frac{2L_i}{n} \left[f_i(w) - f_i(w^*)\right]\right] \\
&\leq \frac{2L_{max}}{n} \sum_{i=1}^{n} \left[f_i(w) - f_i(w^*)\right] \\
\mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 &\leq 2L_{max} \left[f(w) - f(w^*)\right] \qquad (23)
\end{aligned}
$$

□

B.8 Proof for Lemma 1

Proof. Let $a = y \cdot x$. For the squared-hinge loss, the strong growth condition is equivalent to

$$
\begin{aligned}
\mathbb{E}\big[(1 - w^\top a)_+^2\big] &\leqslant \rho \big\|\mathbb{E}\big[(1 - w^\top a)_+ a\big]\|^2 \\
\big\|\mathbb{E}\big[(1 - w^\top a)_+ a\big]\big\| &\geqslant \frac{1}{\|w_*\|} \mathbb{E}\big[(1 - w^\top a)_+ a^\top w_*\big] \\
&\geqslant \tau \mathbb{E}\big[(1 - w^\top a)_+\big]
\end{aligned}
$$

We thus need to upper bound $\mathbb{E}\big[(1 - w^\top a)_+^2\big]$ by a constant c times $\big(\mathbb{E}\big[(1 - w^\top a)_+\big]\big)^2$. We must have $c \geqslant 1$ (as a consequence of Jensen's inequality). Then we have $\rho = c/\tau^2$. Next, we prove that if the distribution of a is uniform over κ values, then $c = \kappa$.

Consider a random variable $A \in \mathbb{R}+$ taking κ values $a_1, \ldots, a_\kappa$ with probabilities $p_1, \ldots, p_\kappa$. Then $(\mathbb{E}A)^2 = \sum_{i,j} p_i p_j a_i a_j \geqslant \sum_i a_i^2 p_i^2 \geqslant \min_i p_i \sum_i a_i^2 p_i$, □

B.9 Proof for Lemma 2

Proof. Let $a = y \cdot x$.

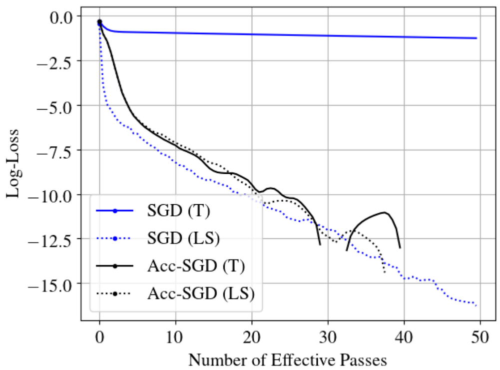
$$
\begin{aligned}
\mathbb{P}(a^\top w \leqslant 0) &\leqslant \mathbb{P}((1 - a^\top w)_+^2 \geqslant 1) \\
&\leqslant \mathbb{E}(1 - a^\top w)_+^2 \\
\implies \mathbb{P}(a^\top w \leqslant 0) &\leq \mathbb{E} f(w, a)
\end{aligned}
$$

□

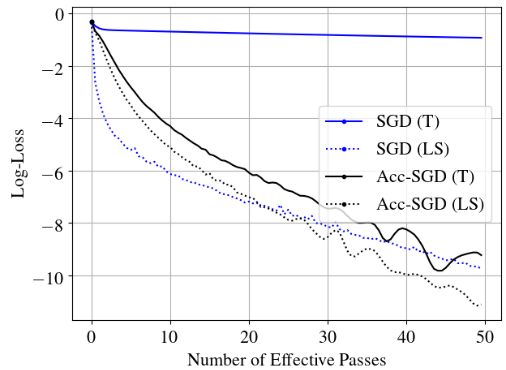
C Additional experimental results

In this section, we propose to use a line-search heuristic for both constant step-size SGD and its accelerated variant. For SGD, we use the line-search proposed in SAG [31]: start with an initial estimate $\hat{L} = 1$ and in each iteration, we double the estimate when the condition $f_k\left(w_k - \frac{1}{\hat{L}} \nabla f_k(w_k)\right) \leq f_k(w_k) - \frac{1}{2\hat{L}} \left\|\nabla f_k(w_k)\right\|^2$ is not satisfied. We denote this variant as SGD(LS) and the corresponding variant that uses a $1/L$ step-size as SGD(T). For the accelerated case, we use the same line-search procedure as above, but search for an appropriate value of ρL. We denote the accelerated variant with and without line-search as Acc-SGD(LS) and Acc-SGD(T) respectively.

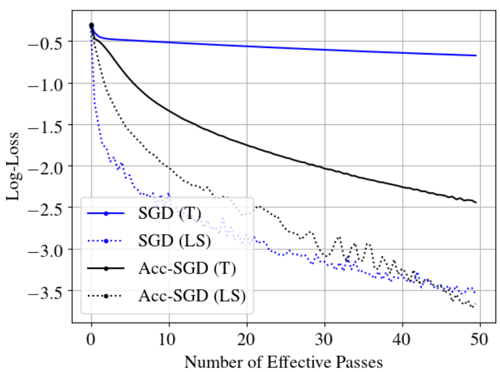
We make the following observations: (i) Accelerated SGD in conjunction with our line-search heuristic is stable across datasets. (ii) Acc-SGD(LS) either matches or outperforms Acc-SGD(T). (iii) In some cases, SGD(LS) can result in faster empirical convergence as compared to the accelerated variants. We plan to investigate better line-search methods for both SGD [31] and Acc-SGD [21] in the future.



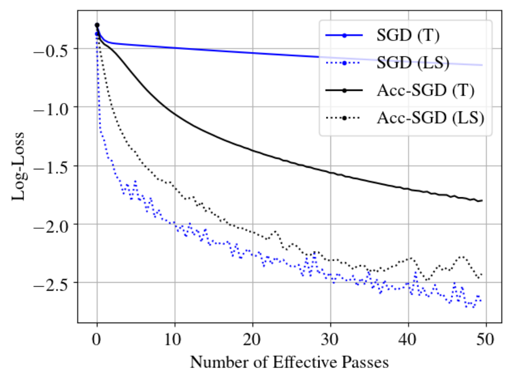
(a) $\tau = 0.1$



(b) $\tau = 0.05$

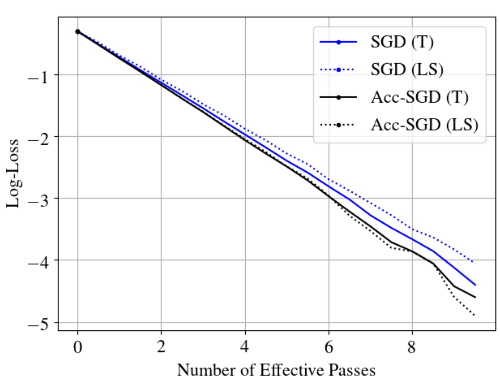


(c) $\tau = 0.01$

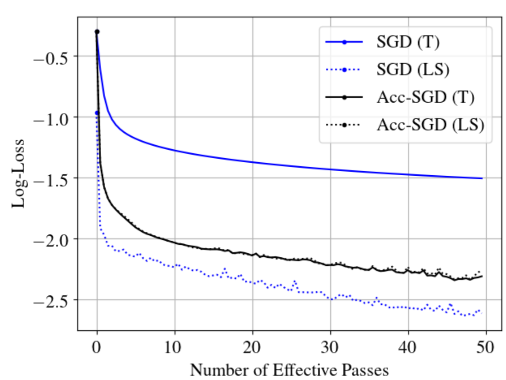


(d) $\tau = 0.005$

Figure 3: Comparison of SGD and variants of accelerated SGD on a synthetic linearly separable dataset with margin τ.



(a) CovType



(b) Protein

Figure 4: Comparison of SGD and accelerated SGD for learning a linear classifier with RBF features on the (a) CovType and (b) Protein datasets.