Supplementary Materials for “Exploring Fast and Communication-Efficient Algorithms in Large-scale Distributed Networks”

1 Proof of Unbiased Quantization Variance

Lemma 1. *If $x \in \mathbb{R}$ is in the convex hull of $\mathrm{dom}(\delta, b)$, then the quantization variance can be bounded as*

$$\mathbf{E}[\big(Q_{(\delta,b)}(x) - x\big)^2] \leq \frac{\delta^2}{4}. \tag{1}$$

Proof. From the manuscript we know that if x is in the convex hull of $\mathrm{dom}(\delta, b)$, then it will be stochastically rounded up or down. Without loss of generality, let $z + \delta$ and z be the up and down quantization values respectively, then

$$Q_{(\delta,b)}(x) = \begin{cases} z & \text{with probability } \frac{z+\delta-x}{\delta}, \\ z+\delta & \text{otherwise}. \end{cases}$$

Note that if x equals to the smallest or largest value in $\mathrm{dom}(\delta, b)$, then $Q_{(\delta,b)}(x) = x$ from the above definition of function Q. Firstly, it can be verified that $\mathbf{E}\big[Q_{(\delta,b)}(x)\big] = x$, then we have

$$\mathbf{E}[\big(Q_{(\delta,b)}(x) - x\big)^2] = \frac{z+\delta-x}{\delta}(z-x)^2 + \frac{x-z}{\delta}(z+\delta-x)^2 = (x-z)(z+\delta-x) \leq \frac{\delta^2}{4}. \tag{2}$$

□

2 Proof of Theorem 2

Lemma 2. *For $\omega \in \mathbb{R}^d$, $\lambda \in (0, 1]$, if $\delta = \frac{\lambda||\omega||_\infty}{2^{b-1}-1}$, then*

$$\mathbf{E}||Q_{(\delta,b)}(\omega) - \omega||^2 \leq \frac{(d-d_\lambda)\delta^2}{4} + d_\lambda(1-\lambda)^2||\omega||^2, \tag{3}$$

where d_λ is the number of coordinates in ω exceeding $\mathrm{dom}(\delta, b)$.

Proof. Since the squared norm $||Q_{(\delta,b)}(\omega) - \omega||^2$ separates along dimensions, it suffices to consider a single coordinate ω_i. If ω_i is in the convex hull of $\mathrm{dom}(\delta, b)$, then according to Lemma 1 we have

$$\mathbf{E}[\big(Q_{(\delta,b)}(\omega_i) - \omega_i\big)^2] \leq \frac{\delta^2}{4}. \tag{4}$$

On the other hand, if ω_i is not in the convex hull of $\mathrm{dom}(\delta, b)$, then $Q_{(\delta,b)}(\omega_i)$ is either the smallest or the largest value of $\mathrm{dom}(\delta, b)$. Therefore, $\big(Q_{(\delta,b)}(\omega_i) - \omega_i\big)^2 \leq (\lambda||\omega||_\infty - ||\omega||_\infty)^2 = (1-\lambda)^2||\omega||_\infty^2 \leq (1-\lambda)^2||\omega||^2$. Summing up over all dimensions we get (3). □

Lemma 3. *For the iterates x_t^{s+1}, $\tilde{x}^s$ in Algorithm 1, define $g_t \triangleq \frac{1}{B}\sum_{j=1}^{B}\left[\nabla f_j(x_t^{s+1}) - \nabla f_j(\tilde{x}^s)\right] + \nabla f(\tilde{x}^s)$, where each element j is uniformly and independently sampled from {1,...,n}, we have*

$$\mathbf{E}||g_t - \nabla f(x_t^{s+1})||^2 \leq \frac{L^2}{B}\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2. \tag{5}$$

Proof.

$$\begin{aligned}\mathbf{E}||g_t - \nabla f(x_t^{s+1})||^2 &= \mathbf{E}||\frac{1}{B}\sum_{j=1}^{B}\left[\nabla f_j(x_t^{s+1}) - \nabla f_j(\tilde{x}^s) + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})\right]||^2 \\ &= \frac{1}{B^2}\sum_{j=1}^{B}\mathbf{E}||\nabla f_j(x_t^{s+1}) - \nabla f_j(\tilde{x}^s) + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})||^2 \\ &\leq \frac{1}{B^2}\sum_{j=1}^{B}\mathbf{E}||\nabla f_j(x_t^{s+1}) - \nabla f_j(\tilde{x}^s)||^2 \\ &\leq \frac{L^2}{B}\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2,\end{aligned} \tag{6}$$

where the first inequality uses $\mathbf{E}||x - \mathbf{E}x||^2 \leq \mathbf{E}||x||^2$ and the last inequality follows from the Lipschitz smooth property of $f_j(x)$. □

Lemma 4. *Denote $d_\lambda = \max_i\{d_\lambda^i\}$, where d_λ^i is the number of coordinates in u_t^i exceeding $\mathrm{dom}(\delta_t^i, b)$. Under communication scheme **(a)** or **(b)**, i.e., $\tilde{u}_t = \frac{1}{N}\sum_{i=1}^{N}\tilde{u}_t^i$, we have*

$$\mathbf{E}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 \leq 2L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2. \tag{7}$$

*Note that δ_t^i has different value for **(a)** and **(b)**.*

Proof. **Case 1.** First of all, we consider communication scheme **(a)**, i.e., $\tilde{u}_t = \frac{1}{N}\sum_{i=1}^{N}\tilde{u}_t^i = \frac{1}{N}\sum_{i=1}^{N}Q_{(\delta_t^i,b)}(u_t^i)$, where $\delta_t^i = \frac{\lambda||u_t^i||_\infty}{2^{b-1}-1}$. Therefore

$$\begin{aligned}&\mathbf{E}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 \\ &= \mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}Q_{(\delta_t^i,b)}(u_t^i) - \frac{1}{N}\sum_{i=1}^{N}u_t^i + \frac{1}{N}\sum_{i=1}^{N}u_t^i + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})||^2 \\ &\leq 2\,\underbrace{\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}Q_{(\delta_t^i,b)}(u_t^i) - \frac{1}{N}\sum_{i=1}^{N}u_t^i||^2}_{L_1} + 2\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}u_t^i + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})||^2,\end{aligned} \tag{8}$$

where L_1 can be bounded as follows.

$$\begin{aligned}L_1 &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}||Q_{(\delta_t^i,b)}(u_t^i) - u_t^i||^2 \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda^i(1-\lambda)^2\Big]||u_t^i||^2 \\ &\leq \frac{1}{N}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2\Big]\sum_{i=1}^{N}\mathbf{E}||u_t^i||^2 \\ &\leq L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2\Big]\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2,\end{aligned} \tag{9}$$

where the second inequality uses Lemma 2 and the last inequality is due to the smoothness of $f_i(x)$. Substituting (9) into (8) and using Lemma 3, we get (7).

Case 2. If employing communication scheme **(b)**, we have $\tilde{u}_t = \frac{1}{N}\sum\limits_{i=1}^{N}\tilde{u}_t^i = \frac{1}{N}\sum\limits_{i=1}^{N} Q_{(\delta_t,b)}(u_t^i)$, where $\delta_t = \max_i\{\delta_t^i\}$. Denote $j = \arg\max_i ||u_t^i||_\infty$, therefore $\delta_t = \frac{\lambda||u_t^j||_\infty}{2^{b-1}-1}$. Then let $\delta_t^i = \delta_t$ for all i, putting back into (8) we obtain

$$\begin{aligned}
&\mathbf{E}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 \\
&\leq 2\underbrace{\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N} Q_{(\delta_t,b)}(u_t^i) - \frac{1}{N}\sum_{i=1}^{N} u_t^i||^2}_{L_1'} + 2\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N} u_t^i + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})||^2,
\end{aligned} \tag{10}$$

where L_1' has a bound of

$$\begin{aligned}
L_1' &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}||Q_{(\delta_t,b)}(u_t^i) - u_t^i||^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda^i(1-\lambda)^2\Big]||u_t^j||^2 \\
&\leq \frac{1}{N}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2\Big]\sum_{i=1}^{N}\mathbf{E}||u_t^j||^2 \\
&\leq L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2\Big]\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2.
\end{aligned} \tag{11}$$

We adopt

$$\mathbf{E}||Q_{(\delta_t,b)}(u_t^i) - u_t^i||^2 \leq \mathbf{E}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda^i(1-\lambda)^2\Big]||u_t^j||^2 \tag{12}$$

in the second inequality in (11), which can be verified following the proof of Lemma 2 with $\delta_t = \frac{\lambda||u_t^j||_\infty}{2^{b-1}-1}$. Putting the above inequalities together we obtain (7). $\square$

Lemma 5. *Under communication scheme **(c)** with $\tilde{u}_t = Q_{(\delta_t,b)}(\frac{1}{N}\sum\limits_{i=1}^{N}\tilde{u}_t^i)$, $\tilde{u}_t^i = Q_{(\delta_t,b)}(u_t^i)$, $\delta_t = \max_i\{\delta_t^i\}$, we obtain*

$$\mathbf{E}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 \leq 2L^2\Big[\frac{3d\lambda^2}{8(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2, \tag{13}$$

where $d_\lambda = \max_i\{d_\lambda^i\}$, d_λ^i is the number of coordinates in u_t^i exceeding $\mathrm{dom}(\delta_t^i, b)$.

Proof.

$$\mathbf{E}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 = \underbrace{\mathbf{E}||Q_{(\delta_t,b)}(\frac{1}{N}\sum_{i=1}^{N}\tilde{u}_t^i) - \frac{1}{N}\sum_{i=1}^{N}\tilde{u}_t^i||^2}_{L_2} + \underbrace{\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}\tilde{u}_t^i + \nabla f(\tilde{x}^s) - \nabla f(x_t^{s+1})||^2}_{L_3} \tag{14}$$

where the equality holds because $Q_{(\delta_t,b)}(\frac{1}{N}\sum\limits_{i=1}^{N}\tilde{u}_t^i)$ is an unbiased quantization (note that each u_t^i is quantized using δ_t, therefore, all coordinates of $\frac{1}{N}\sum\limits_{i=1}^{N}\tilde{u}_t^i$ are in the convex hull of $\mathrm{dom}(\delta_t, b)$).

From Lemma 1 and **Case 2** in Lemma 4 we obtain

$$
\begin{aligned}
L_2 &\leq \frac{d\delta_t^2}{4} \\
&\leq \frac{d\lambda^2 ||u_t^j||_\infty^2}{4(2^{b-1}-1)^2}, \quad j = \arg\max_i ||u_t^i||_\infty \\
&\leq \frac{L^2 d\lambda^2}{4(2^{b-1}-1)^2}\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2
\end{aligned} \tag{15}
$$

and

$$
L_3 \leq 2L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]\mathbf{E}||x_t^{s+1} - \tilde{x}^s||^2. \tag{16}
$$

Putting them together, we get (13). □

Proof of Theorem 2. Define $\bar{x}_{t+1}^{s+1} = \text{prox}_{\eta h}(x_t^{s+1} - \eta\nabla f(x_t^{s+1}))$. Following the proof of Theorem 5 in [2] (equations (8)-(12)), we get

$$
\mathbf{E}\Big[P(x_{t+1}^{s+1})\Big] \leq \mathbf{E}\Big[P(x_t^{s+1}) + \frac{\eta}{2}||v_t^{s+1} - \nabla f(x_t^{s+1})||^2 + (L - \frac{1}{2\eta})||\bar{x}_{t+1}^{s+1} - x_t^{s+1}||^2 + (\frac{L}{2} - \frac{1}{2\eta})||x_{t+1}^{s+1} - x_t^{s+1}||^2\Big]. \tag{17}
$$

If adopting communication scheme **(a)** or **(b)**, combining Lemma 4, we have

$$
\begin{aligned}
\mathbf{E}\Big[P(x_{t+1}^{s+1})\Big] \leq &\mathbf{E}\Big[P(x_t^{s+1}) + \eta L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]||x_t^{s+1} - \tilde{x}^s||^2 \\
&+ (L - \frac{1}{2\eta})||\bar{x}_{t+1}^{s+1} - x_t^{s+1}||^2 + (\frac{L}{2} - \frac{1}{2\eta})||x_{t+1}^{s+1} - x_t^{s+1}||^2\Big].
\end{aligned} \tag{18}
$$

Define $R_t^{s+1} \triangleq \mathbf{E}\Big[P(x_t^{s+1}) + c_t||x_t^{s+1} - \tilde{x}^s||^2\Big]$ and a sequence $\{c_t\}_{t=0}^m$ with $c_m = 0$ and $c_t = c_{t+1}(1+\beta) + \eta L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]$, where $\beta = \frac{1}{m}$. Therefore $\{c_t\}$ is a decreasing sequence. We first derive the bound of c_0 in the following.

$$
\begin{aligned}
c_0 &\leq \eta L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big] \cdot \frac{(1+\beta)^m - 1}{\beta} \\
&\leq 2m\eta L^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]
\end{aligned} \tag{19}
$$

where the second inequality uses $\beta = \frac{1}{m}$. Denote $\eta = \frac{\rho}{L}$. Then inequality (19) can be simplified as

$$
c_0 \leq 2m\rho L\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]. \tag{20}
$$

On the other hand,

$$
\begin{aligned}
R_{t+1}^{s+1} &= \mathbf{E}\Big[P(x_{t+1}^{s+1}) + c_{t+1}||x_{t+1}^{s+1} - \tilde{x}^s||^2\Big] \\
&\leq \mathbf{E}\Big[P(x_{t+1}^{s+1}) + c_{t+1}(1 + \frac{1}{\beta})||x_{t+1}^{s+1} - x_t^{s+1}||^2 + c_{t+1}(1+\beta)||x_t^{s+1} - \tilde{x}^s||^2\Big] \\
&\leq \mathbf{E}\Big[P(x_t^{s+1}) + c_t||x_t^{s+1} - \tilde{x}^s||^2 + (c_{t+1}(1 + \frac{1}{\beta}) + \frac{L}{2} - \frac{1}{2\eta})||x_{t+1}^{s+1} - x_t^{s+1}||^2 + (L - \frac{1}{2\eta})||\bar{x}_{t+1}^{s+1} - x_t^{s+1}||^2\Big].
\end{aligned} \tag{21}
$$

Now we derive the bound for ρ and b to make sure $(c_{t+1}(1+\frac{1}{\beta}) + \frac{L}{2} - \frac{1}{2\eta}) \leq 0$, and it suffices to let $c_0(1+\frac{1}{\beta}) + \frac{L}{2} \leq \frac{1}{2\eta}$. Combining (20) and $\beta = \frac{1}{m}$, $\eta = \frac{\rho}{L}$, we only need to guarantee

$$
8m^2\rho^2\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big] + \rho \leq 1. \tag{22}
$$

If the above constraint holds, then

$$R_{t+1}^{s+1} \leq R_t^{s+1} + (L - \frac{1}{2\eta})\mathbf{E}||\bar{x}_{t+1}^{s+1} - x_t^{s+1}||^2\Big]. \tag{23}$$

Summing it up over $t = 0$ to $m-1$ and $s = 0$ to $S-1$, using $c_m = 0$, $x_0^{s+1} = \tilde{x}^s$ and $x_m^{s+1} = \tilde{x}^{s+1}$ we get

$$(\frac{1}{2\eta} - L)\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbf{E}||\bar{x}_{t+1}^{s+1} - x_t^{s+1}||^2 \leq P(x^0) - P(x^*). \tag{24}$$

Applying the definition of $G_\eta(x_t^{s+1})$, we obtain results in Theorem 2. Moreover, the analysis of communication scheme **(c)** can be similarly obtained using the above proof steps. □

3 Proof of ALPC-SVRG

Lemma 6. *For $\hat{v}_{k+1} = \frac{1}{B}\sum_{j=1}^{B}\big[\nabla f_j(x_{k+1}) - \nabla f_j(\tilde{x}^s)\big] + \nabla f(\tilde{x}^s)$, where each element j is uniformly and independently sampled from {1,...,n}, we have*

$$\mathbf{E}||\hat{v}_{k+1} - \nabla f(x_{k+1})||^2 \leq \frac{2L}{B}\mathbf{E}\big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\big]. \tag{25}$$

Proof.

$$\begin{aligned}\mathbf{E}||\hat{v}_{k+1} - \nabla f(x_{k+1})||^2 &\leq \frac{1}{B^2}\sum_{j=1}^{B}\mathbf{E}||\nabla f_j(x_{k+1}) - \nabla f_j(\tilde{x}^s)||^2 \\ &\leq \frac{2L}{B}\mathbf{E}\big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\big],\end{aligned} \tag{26}$$

where the first inequality follows from Lemma 3 and the last inequality adopts the Lipschitz smooth property of $f_j(x)$. □

Lemma 7. *Denote $d_\lambda = \max_i\{d_\lambda^i\}$, where d_λ^i is the number of coordinates in u_{k+1}^i exceeding $\mathrm{dom}(\delta_{k+1}^i, b)$, then we have*

$$\mathbf{E}||v_{k+1} - \nabla f(x_{k+1})||^2 \leq 4L\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2} + d_\lambda(1-\lambda)^2 + \frac{1}{NB}\Big]\mathbf{E}\big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\big]. \tag{27}$$

Proof.

$$\begin{aligned}&\mathbf{E}||v_{k+1} - \nabla f(x_{k+1})||^2 \\ &\leq 2\underbrace{\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}Q_{(\delta_{k+1}^i, b)}(u_{k+1}^i) - \frac{1}{N}\sum_{i=1}^{N}u_{k+1}^i||^2}_{A_1} + 2\underbrace{\mathbf{E}||\frac{1}{N}\sum_{i=1}^{N}u_{k+1}^i + \nabla f(\tilde{x}^s) - \nabla f(x_{k+1})||^2}_{A_2}.\end{aligned} \tag{28}$$

Using the same arguments of Lemma 6 we obtain

$$A_2 \leq \frac{2L}{NB}\mathbf{E}\big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\big]. \tag{29}$$

Moreover,

$$
\begin{aligned}
A_1 &\le \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}||Q_{(\delta^i_{k+1},b)}(u^i_{k+1})-u^i_{k+1}||^2\\
&\le \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2}+d^i_\lambda(1-\lambda)^2\Big]||u^i_{k+1}||^2\\
&\le \frac{1}{N}\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2}+d_\lambda(1-\lambda)^2\Big]\sum_{i=1}^{N}\mathbf{E}||u^i_{k+1}||^2\\
&\le 2L\Big[\frac{d\lambda^2}{4(2^{b-1}-1)^2}+d_\lambda(1-\lambda)^2\Big]\mathbf{E}\big[f(\tilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\tilde{x}^s-x_{k+1}\rangle\big],
\end{aligned}
\tag{30}
$$

where the second inequality follows from Lemma 2. Putting them together, we obtain (27). □

Lemma 8.

$$
\mathbf{E}||\hat{v}_{k+1}-v_{k+1}||^2 \le 4L\Big[\frac{1}{2B}+\frac{d\lambda^2}{4(2^{b-1}-1)^2}+d_\lambda(1-\lambda)^2+\frac{1}{NB}\Big]\mathbf{E}\big[f(\tilde{x}^s)-f(x_{k+1})-\langle\nabla f(x_{k+1}),\tilde{x}^s-x_{k+1}\rangle\big].
\tag{31}
$$

Proof.

$$
\begin{aligned}
\mathbf{E}||\hat{v}_{k+1}-v_{k+1}||^2 &= \mathbf{E}||\hat{v}_{k+1}-\nabla f(x_{k+1})+\nabla f(x_{k+1})-v_{k+1}||^2\\
&= \mathbf{E}||\hat{v}_{k+1}-\nabla f(x_{k+1})||^2+\mathbf{E}||v_{k+1}-\nabla f(x_{k+1})||^2.
\end{aligned}
\tag{32}
$$

The second equality holds because the mini-batches for calculating $\hat{v}_{k+1}$ and v_{k+1} are independent and $\mathbf{E}\hat{v}_{k+1} = \nabla f(x_{k+1})$. Combining Lemma 6 and Lemma 7, we obtain (31). □

Lemma 9. *Define*

$$
\mathrm{Prog}(x_{k+1}) \triangleq -min_y\{2L||y-x_{k+1}||^2+\langle v_{k+1},y-x_{k+1}\rangle+h(y)-h(x_{k+1})\},
\tag{33}
$$

then from the update rule of y, we obtain

$$
\mathbf{E}\big[P(x_{k+1})-P(y_{k+1})\big] \ge \mathbf{E}\big[\,\mathrm{Prog}(x_{k+1})-\frac{1}{6L}||\nabla f(x_{k+1})-v_{k+1}||^2\big].
\tag{34}
$$

Proof Sketch. (34) follows from the proof of Lemma 3.3 in [1] with different coefficients. □

Lemma 10. *If $h(x)$ is σ-strongly convex, then for any $u\in\mathbb{R}^d$, we have (Lemma 3.5 in [1])*

$$
\alpha\langle\hat{v}_{k+1},z_{k+1}-u\rangle+\alpha h(z_{k+1})-\alpha h(u) \le -\frac{1}{2}||z_k-z_{k+1}||^2+\frac{1}{2}||z_k-u||^2-\frac{1+\alpha\sigma}{2}||z_{k+1}-u||^2.
\tag{35}
$$

Lemma 11. *Let $\tau_2 = \frac{5}{3}\zeta+\frac{1}{2B}$, $\zeta = \frac{d\lambda^2}{4(2^{b-1}-1)^2}+d_\lambda(1-\lambda)^2+\frac{1}{NB}$, $\alpha = \frac{1}{6\tau_1 L}$. Suppose with proper choice of parameters B, b, λ, we have $\tau_2 \le \frac{1}{2}$, then*

$$
\begin{aligned}
&\mathbf{E}\Big[\alpha\langle\nabla f(x_{k+1}),z_k-u\rangle-\alpha h(u)\Big]\\
&\le \mathbf{E}\Big[\frac{\alpha}{\tau_1}\big[P(x_{k+1})-P(y_{k+1})+\tau_2 P(\tilde{x}^s)-\tau_2 f(x_{k+1})-\tau_2\langle\nabla f(x_{k+1}),\tilde{x}^s-x_{k+1}\rangle\big]\\
&+\frac{\alpha}{\tau_1}(1-\tau_1-\tau_2)h(y_k)-\frac{\alpha}{\tau_1}h(x_{k+1})+\frac{1}{2}||z_k-u||^2-\frac{1+\alpha\sigma}{2}||z_{k+1}-u||^2\Big].
\end{aligned}
\tag{36}
$$

Proof.

$$
\begin{aligned}
&\mathbf{E}\Big[\alpha\langle\hat{v}_{k+1}, z_k - u\rangle + \alpha h(z_{k+1}) - \alpha h(u)\Big]\\
=&\mathbf{E}\Big[\alpha\langle v_{k+1}, z_k - z_{k+1}\rangle + \alpha\langle\hat{v}_{k+1} - v_{k+1}, z_k - z_{k+1}\rangle + \alpha\langle\hat{v}_{k+1}, z_{k+1} - u\rangle + \alpha h(z_{k+1}) - \alpha h(u)\Big]\\
\leq&\mathbf{E}\Big[\alpha\langle v_{k+1}, z_k - z_{k+1}\rangle + \frac{\alpha}{\tau_1}\frac{1}{4L}||\hat{v}_{k+1} - v_{k+1}||^2 + \frac{1}{6}||z_k - z_{k+1}||^2 + \alpha\langle\hat{v}_{k+1}, z_{k+1} - u\rangle + \alpha h(z_{k+1}) - \alpha h(u)\Big]\\
\leq&\mathbf{E}\Big[\alpha\langle v_{k+1}, z_k - z_{k+1}\rangle + \frac{\alpha}{\tau_1}\frac{1}{4L}||\hat{v}_{k+1} - v_{k+1}||^2 - \frac{1}{3}||z_k - z_{k+1}||^2 + \frac{1}{2}||z_k - u||^2 - \frac{1+\alpha\sigma}{2}||z_{k+1} - u||^2\Big]\\
\leq&\mathbf{E}\Big[\alpha\langle v_{k+1}, z_k - z_{k+1}\rangle + \frac{\alpha}{\tau_1}(\zeta + \frac{1}{2B})\mathbf{E}\big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\big] - \frac{1}{3}||z_k - z_{k+1}||^2\\
&+ \frac{1}{2}||z_k - u||^2 - \frac{1+\alpha\sigma}{2}||z_{k+1} - u||^2\Big],
\end{aligned}
\tag{37}
$$

where the first inequality uses Young's inequality and $\alpha = \frac{1}{6\tau_1 L}$, the last two inequalities follow from Lemma 10 and Lemma 8 respectively. Define $v \triangleq \tau_1 z_{k+1} + \tau_2\tilde{x}^s + (1-\tau_1-\tau_2)y_k$, therefore $x_{k+1} - v = \tau_1(z_k - z_{k+1})$, then we obtain

$$
\begin{aligned}
&\mathbf{E}\Big[\alpha\langle v_{k+1}, z_k - z_{k+1}\rangle - \frac{1}{3}||z_k - z_{k+1}||^2\Big]\\
=&\mathbf{E}\Big[\frac{\alpha}{\tau_1}\langle v_{k+1}, x_{k+1} - v\rangle - \frac{1}{3\tau_1^2}||x_{k+1} - v||^2\Big]\\
=&\mathbf{E}\Big[\frac{\alpha}{\tau_1}\Big(\langle v_{k+1}, x_{k+1} - v\rangle - \frac{1}{3\alpha\tau_1}||x_{k+1} - v||^2 - h(v) + h(x_{k+1})\Big) + \frac{\alpha}{\tau_1}\Big(h(v) - h(x_{k+1})\Big)\Big]\\
=&\mathbf{E}\Big[\frac{\alpha}{\tau_1}\Big(\langle v_{k+1}, x_{k+1} - v\rangle - 2L||x_{k+1} - v||^2 - h(v) + h(x_{k+1})\Big) + \frac{\alpha}{\tau_1}\Big(h(v) - h(x_{k+1})\Big)\Big]\\
\leq&\mathbf{E}\Big[\frac{\alpha}{\tau_1}\Big(P(x_{k+1}) - P(y_{k+1}) + \frac{1}{6L}||v_{k+1} - \nabla f(x_{k+1})||^2\Big) + \frac{\alpha}{\tau_1}\Big(h(v) - h(x_{k+1})\Big)\Big]\\
\leq&\frac{\alpha}{\tau_1}\mathbf{E}\Big[P(x_{k+1}) - P(y_{k+1})\Big] + \frac{\alpha}{\tau_1}\frac{2}{3}\zeta\mathbf{E}\Big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\Big] + \frac{\alpha}{\tau_1}\mathbf{E}\Big[h(v) - h(x_{k+1})\Big],
\end{aligned}
\tag{38}
$$

where the third equality uses $\alpha = \frac{1}{6\tau_1 L}$, the first inequality follows from Lemma 9 and the last inequality adopts Lemma 7. Substituting (38) into (37) we get

$$
\begin{aligned}
&\mathbf{E}\Big[\alpha\langle\hat{v}_{k+1}, z_k - u\rangle + \alpha h(z_{k+1}) - \alpha h(u)\Big]\\
&\leq \mathbf{E}\Big[\frac{\alpha}{\tau_1}\big[P(x_{k+1}) - P(y_{k+1})\big] + \frac{\alpha}{\tau_1}\tau_2\Big[f(\tilde{x}^s) - f(x_{k+1}) - \langle\nabla f(x_{k+1}), \tilde{x}^s - x_{k+1}\rangle\Big]\\
&+ \frac{1}{2}||z_k - u||^2 - \frac{1+\alpha\sigma}{2}||z_{k+1} - u||^2 + \frac{\alpha}{\tau_1}\big[\tau_1 h(z_{k+1}) + \tau_2 h(\tilde{x}^s) + (1-\tau_1-\tau_2)h(y_k) - h(x_{k+1})\big]\Big].
\end{aligned}
\tag{39}
$$

Because $\hat{v}_{k+1}$ is unbiased, we get (36) after rearranging terms. □

***Proof of Theorem* 3.** Starting form Lemma 11, following the proof of ([1], Lemma 3.7, Theorem 3.1), we get

$$
\begin{aligned}
&\mathbf{E}\Big[\frac{\tau_1 + \tau_2 - (1-\frac{1}{\theta})}{\tau_1}\theta\tilde{D}^{s+1}\cdot\sum_{t=0}^{m-1}\theta^t\Big]\\
&\leq \mathbf{E}\Big[\frac{1-\tau_1-\tau_2}{\tau_1}(D_{sm} - \theta^m D_{(s+1)m}) + \frac{\tau_2}{\tau_1}\tilde{D}^s\sum_{t=0}^{m-1}\theta^t + \frac{1}{2\alpha}||z_{sm} - x^*||^2 - \frac{\theta^m}{2\alpha}||z_{(s+1)m} - x^*||^2\Big],
\end{aligned}
\tag{40}
$$

where $\theta = (1+\alpha\sigma)$, $D_k \triangleq P(y_k) - P(x^*)$, $\tilde{D}^s \triangleq P(\tilde{x}^s) - P(x^*)$.

If $\frac{m\sigma}{L} \leq \frac{3}{2}$, then $\sqrt{\frac{m\sigma}{6L}} \leq \frac{1}{2}$. Choosing $\alpha = \frac{1}{\sqrt{6m\sigma L}}$, then $\tau_1 = \frac{1}{6\alpha L} = m\sigma\alpha = \sqrt{\frac{m\sigma}{6L}} \leq \frac{1}{2}$ and $\alpha\sigma \leq \frac{1}{2m}$.

It can be verified that the above parameter settings guarantee **Case** 1. in ([1], Theorem 1), therefore, with the same arguments we arrive at

$$\mathbf{E}\big[P(\tilde{x}^S) - P(x^*)\big] \leq O((1+\alpha\sigma)^{-Sm})\big[P(x_0) - P(x^*)\big]. \tag{41}$$

$\square$

Proof Sketch of Theorem 4. Let $\alpha_s = \frac{1}{6L\tau_{1,s}}, \tau_{1,s} = \frac{2}{s+4}$ and τ_2 unchanged. It can be verified that Lemma 11 also holds in the current parameter setting (with $\sigma = 0$), then plug Lemma 11 into the proof of Theorem 4.1 in [1], we get

$$\mathbf{E}\Big[P(\tilde{x}^S) - P(x^*)\Big] \leq O(\frac{1}{mS^2})\Big[m(P(x_0) - P(x^*)) + L||x_0 - x^*||^2\Big]. \tag{42}$$

$\square$

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, pages 1200–1205, 2017.

[2] S. J. Reddi, S. Sra, B. Poczos, and A. Smola. Fast stochastic methods for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2016.