We show that FedAv with Accelerated SGD has $O(1/T)$ rate under μ-strong convexity and L-smoothness. The proof follows the framework of the ICLR paper. The FedAv algorithm with Nesterov Accelerated SGD (NASGD) follows the updates

$$
\begin{aligned}
y_{t+1}^k &= w_t^k - \alpha_t g_{t,k} \\
w_{t+1}^k &= \begin{cases} y_{t+1}^k + \beta_t (y_{t+1}^k - y_t^k) & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^N p_k \left[y_{t+1}^k + \beta_t (y_{t+1}^k - y_t^k) \right] & \text{if } t+1 \in \mathcal{I}_E \end{cases}
\end{aligned}
$$

and define the virtual sequences $\overline{y}_t = \sum_{k=1}^N p_k y_t^k$, $\overline{w}_t = \sum_{k=1}^N p_k w_t^k$, and $\overline{g}_t = \sum_{k=1}^N p_k \mathbb{E} g_{t,k}$. We have $\mathbb{E} g_t = \overline{g}_t$ and $\overline{y}_{t+1} = \overline{w}_t - \alpha_t g_t$, and $\overline{w}_{t+1} = \overline{y}_{t+1} + \beta_t (\overline{y}_{t+1} - \overline{y}_t)$.

Theorem 1. *Let the parameters satisfy the assumptions in the ICLR paper and learning rate* $\alpha_t = \frac{2}{\mu(\gamma+t)}$, β_t *such that* $\alpha_t^2 + \beta_{t-1}^2 \leq \frac{1}{2}$, $\beta_t \leq \alpha_t$ *for all* t. *Then with full device participation,*

$$
\begin{aligned}
\mathbb{E} F(w_T) - F^* &\leq \frac{2\kappa}{\gamma + T} \left(\frac{B}{\mu} + 2L(\|w_0 - w^*\|^2 \right) \\
B &= \sum_{k=1}^N p_k^2 \sigma_k^2 + 9L\Gamma + 32(E-1)^2 G^2 + 2 + G^2 + GK
\end{aligned}
$$

and K *is such that*

$$
\alpha_0 B + 2\sqrt{K} \cdot G \leq \mu K
$$

and

$$
\|w_0 - w^*\|^2 \leq K
$$

Proof. We have the recursion

$$
\begin{aligned}
y_{t+1}^k - y_t^k &= w_t^k - w_{t-1}^k - (\alpha_t g_{t,k} - \alpha_{t-1} g_{t-1,k}) \\
w_{t+1}^k - w_t^k &= -\alpha_t g_{t,k} + \beta_t (y_{t+1}^k - y_t^k)
\end{aligned}
$$

so that

$$
\begin{aligned}
y_{t+1}^k - y_t^k &= -\alpha_{t-1} g_{t-1,k} + \beta_{t-1}(y_t^k - y_{t-1}^k) - (\alpha_t g_{t,k} - \alpha_{t-1} g_{t-1,k}) \\
&= \beta_{t-1}(y_t^k - y_{t-1}^k) - \alpha_t g_{t,k}
\end{aligned}
$$

First, we derive a bound on $\mathbb{E}\|\overline{y}_{t+1} - \overline{y}_t\|^2$ that is useful in the proof. Since the identity $y_{t+1}^k - y_t^k = \beta_{t-1}(y_t^k - y_{t-1}^k) - \alpha_t g_{t,k}$ implies

$$
\mathbb{E}\|y_{t+1}^k - y_t^k\|^2 \leq 2\beta_{t-1}^2 \mathbb{E}\|y_t^k - y_{t-1}^k\|^2 + 2\alpha_t^2 G^2
$$

as long as α_t, β_t **satisfy** $2\beta_{t-1}^2 + 2\alpha_t^2 \leq 1$**, and** $\mathbb{E}\|w_0 - \alpha_t g_{t,k}\|^2 \leq G^2$**,** we can guarantee that $\mathbb{E}\|y_t^k - y_{t-1}^k\|^2 \leq G^2$. This together with Jensen implies $\mathbb{E}\|\overline{y}_t - \overline{y}_{t-1}\|^2 \leq G^2$.

Now we turn to $\|\overline{w}_{t+1} - w^*\|^2$. We have

$$
\begin{aligned}
\|\overline{w}_{t+1} - w^*\|^2 &= \|(\overline{w}_t - \alpha_t g_t) + \beta_t(\overline{y}_{t+1} - \overline{y}_t) - w^*\|^2 \\
&= \|(\overline{w}_t - \alpha_t \overline{g}_t - w^*) + \beta_t(\overline{y}_{t+1} - \overline{y}_t) - \alpha_t(\overline{g}_t - g_t)\|^2 \\
&= A_1 + A_2 + \alpha_t^2 \|g_t - \overline{g}_t\|^2
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= \|\overline{w}_t - w^* - \alpha_t \overline{g}_t + \beta_t(\overline{y}_{t+1} - \overline{y}_t)\|^2 \\
A_2 &= 2\alpha_t \langle \overline{w}_t - w^* - \alpha_t \overline{g}_t + \beta_t(\overline{y}_{t+1} - \overline{y}_t), \overline{g}_t - g_t \rangle
\end{aligned}
$$

and $\mathbb{E}A_2 = 0$ by definition of g_t and $\overline{g}_t$. Next we bound A_1:

$$
\begin{aligned}
\|\overline{w}_t - w^* - \alpha_t \overline{g}_t + \beta_t(\overline{y}_{t+1} - \overline{y}_t)\|^2 &= \|\overline{w}_t - w^*\|^2 + 2\langle \overline{w}_t - w^*, \beta_t(\overline{y}_{t+1} - \overline{y}_t) - \alpha_t \overline{g}_t \rangle + \|\beta_t(\overline{y}_{t+1} - \overline{y}_t) - \alpha_t \overline{g} \\
&\leq \|\overline{w}_t - w^*\|^2 + 2\langle \overline{w}_t - w^*, \beta_t(\overline{y}_{t+1} - \overline{y}_t) - \alpha_t \overline{g}_t \rangle + 2\|\beta_t(\overline{y}_{t+1} - \overline{y}_t)\|^2 +
\end{aligned}
$$

and by the convexity of $\|\cdot\|^2$ and L-smoothness of F_k,

$$
\alpha_t^2 \|\overline{g}_t\|^2 \leq \alpha_t^2 \sum_{k=1}^N p_k \|\nabla F_k(w_t^k)\|^2 \leq 2L\alpha_t^2 \sum_{k=1}^N p_k (F_k(w_t^k) - F_k^*)
$$

and if $\beta_t = \alpha_t$,

$$
\begin{aligned}
2\|\beta_t(\overline{y}_{t+1} - \overline{y}_t)\|^2 &= 2\beta_t^2 \|\sum_{k=1}^N p_k (y_{t+1}^k - y_t^k)\|^2 \\
&\leq 2\beta_t^2 \sum_{k=1}^N p_k \|y_{t+1}^k - y_t^k\|^2 \\
&= 2\alpha_t^2 \sum_{k=1}^N p_k \|y_{t+1}^k - y_t^k\|^2
\end{aligned}
$$

and taking expectation we get

$$
2\mathbb{E}\|\beta_t(\overline{y}_{t+1} - \overline{y}_t)\|^2 \leq 2\alpha_t^2 G^2
$$

Now

$$
2\mathbb{E}\langle \overline{w}_t - w^*, \beta_t(\overline{y}_{t+1} - \overline{y}_t) - \alpha_t \overline{g}_t \rangle = 2\beta_t \mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle - 2\alpha_t \langle \overline{w}_t - w^*, \overline{g}_t \rangle
$$

and so

$$
\begin{aligned}
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 \leq \mathbb{E}\|\overline{w}_t - w^*\|^2 - 2\alpha_t \langle \overline{w}_t - w^*, \overline{g}_t \rangle + 4L\alpha_t^2 \sum_{k=1}^N p_k (F_k(w_t^k) - F_k^*) + \alpha_t^2 \mathbb{E}\|g_t - \overline{g}_t\|^2 \\
+ 2\alpha_t^2 G^2 + 2\beta_t \mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle
\end{aligned}
$$

At this point, the exact same argument in the ICLR paper implies that

$$\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 \leq (1 - \mu\alpha_t) + 9L\alpha_t^2\Gamma + \alpha_t^2\mathbb{E}\|g_t - \overline{g}_t\|^2 + 2\mathbb{E}\sum_{k=1}^{N} p_k\|\overline{w}_t - w_k^t\|^2$$
$$+ 2\alpha_t^2 G^2 + 2\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle$$

Now we bound $\mathbb{E}\sum_{k=1}^{N} p_k\|\overline{w}_t - w_t^k\|^2$. Since communication is done every E steps, for any $t \geq 0$, we can find a $t_0 \leq t$ such that $t - t_0 \leq E - 1$ and $w_{t_0}^k = \overline{w}_{t_0}$for all k. Moreover, using η_t is non-increasing and $\eta_{t_0} \leq 2\eta_t$ for any $t - t_0 \leq E - 1$, we have

$$\begin{aligned}
\mathbb{E}\sum_{k=1}^{N} p_k\|\overline{w}_t - w_t^k\|^2 &= \mathbb{E}\sum_{k=1}^{N} p_k\|w_t^k - \overline{w}_{t_0} - (\overline{w}_t - \overline{w}_{t_0})\|^2 \\
&\leq \mathbb{E}\sum_{k=1}^{N} p_k\|w_t^k - \overline{w}_{t_0}\|^2 \\
&= \mathbb{E}\sum_{k=1}^{N} p_k\|w_t^k - w_{t_0}^k\|^2 \\
&= \mathbb{E}\sum_{k=1}^{N} p_k\|\sum_{i=t_0}^{t-1}\beta_i(y_{i+1}^k - y_i^k) - \sum_{i=t_0}^{t-1}\alpha_i g_{i,k}\|^2 \\
&\leq 2\sum_{k=1}^{N} p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\alpha_i^2\|g_{i,k}\|^2 + 2\sum_{k=1}^{N} p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\beta_i^2\|(y_{i+1}^k - y_i^k)\|^2
\end{aligned}$$

where we recall that

$$y_{t+1}^k = w_t^k - \alpha_t g_{t,k}$$
$$w_{t+1}^k = \begin{cases} y_{t+1}^k + \beta_t(y_{t+1}^k - y_t^k) & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^{N} p_k\left[y_{t+1}^k + \beta_t(y_{t+1}^k - y_t^k)\right] & \text{if } t+1 \in \mathcal{I}_E \end{cases}$$

The first term $2\sum_{k=1}^{N} p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\alpha_i^2\|g_{i,k}\|^2$ is bounded above by $8\alpha_t^2(E-1)^2G^2$ following the ICLR paper. The term $\mathbb{E}\|(y_{i+1}^k - y_i^k)\|^2$ is bounded above by G^2 as well, as proved earlier. It follows that

$$\mathbb{E}\sum_{k=1}^{N} p_k\|\overline{w}_t - w_t^k\|^2 \leq 16\alpha_t^2(E-1)^2G^2$$

Using the bound on $\mathbb{E}\sum_{k=1}^{N} p_k\|\overline{w}_t - w_t^k\|^2$, we can conclude that

$$
\begin{aligned}
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 &\le (1-\mu\alpha_t)\mathbb{E}\|\overline{w}_t - w^*\|^2 + 9L\alpha_t^2\Gamma + \alpha_t^2\sum_{k=1}^N p_k^2\sigma_k^2 + 32\alpha_t^2(E-1)^2G^2 \\
&\quad + 2\alpha_t^2G^2 + 2\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle \\
&= (1-\mu\alpha_t)\mathbb{E}\|\overline{w}_t - w^*\|^2 + \alpha_t^2 B + 2\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle
\end{aligned}
$$

where

$$
B = \sum_{k=1}^N p_k^2\sigma_k^2 + 9L\Gamma + 32(E-1)^2G^2 + 2
$$

Our next step is to show that $2\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle = O(\alpha_t^2)$.

With appropriate choice of constant K depending on the other constants(to be detailed), we first show that

$$
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 \le K^2
$$

for all t, i.e. the updates always stay in a large ball around the optimum during the Nesterov accelerated gradient descent. Note that

$$
\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle \le \beta_t\sqrt{\mathbb{E}\|\overline{w}_t - w^*\|^2}\cdot\sqrt{\mathbb{E}\|\overline{y}_{t+1} - \overline{y}_t\|^2}
$$

so that

$$
\begin{aligned}
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 &\le (1-\alpha_t\mu)\mathbb{E}\|\overline{w}_t - w^*\|^2 + \alpha_t^2 B + 2\beta_t\sqrt{\mathbb{E}\|\overline{w}_t - w^*\|^2}\cdot\sqrt{\mathbb{E}\|\overline{y}_{t+1} - \overline{y}_t\|^2} \\
&\le (1-\alpha_t\mu)\mathbb{E}\|\overline{w}_t - w^*\|^2 + \alpha_t^2 B + 2\beta_t\sqrt{\mathbb{E}\|\overline{w}_t - w^*\|^2}\cdot G
\end{aligned}
$$

where $B = \sum_{k=1}^N p_k^2\sigma_k^2 + 9L\Gamma + 32(E-1)^2G^2 + 2$. Suppose $\mathbb{E}\|\overline{w}_t - w^*\|^2 \le K^2$ for $t \ge 0$, then

$$
\begin{aligned}
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 &\le (1-\alpha_t\mu)K^2 + \alpha_t^2 B + 2\beta_t K\cdot G \\
&\le K^2 + (\alpha_t^2 B + 2\alpha_t K\cdot G - \alpha_t\mu K^2)
\end{aligned}
$$

as long as α_0 and K are chosen so that

$$
\alpha_0 B + 2\sqrt{K}\cdot G \le \mu K
$$

and

$$
\|w_0 - w^*\|^2 \le K
$$

then since $\alpha_t \le \alpha_0$, we get $\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 \le K$, where K only depends on $G, \sigma_k, L, \mu, \Gamma, \|w_0 - w^*\|^2, E$.

Now we can finally bound $\beta_t\mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t)\rangle$.

Using the recursive relations

$$
\begin{aligned}
y_{t+1}^k - y_t^k &= w_t^k - w_{t-1}^k - (\alpha_t g_{t,k} - \alpha_{t-1} g_{t-1,k}) \\
w_{t+1}^k - w_t^k &= -\alpha_t g_{t,k} + \beta_t (y_{t+1}^k - y_t^k)
\end{aligned}
$$

so that $y_{t+1}^k - y_t^k = \beta_{t-1}(y_t^k - y_{t-1}^k) - \alpha_t g_{t,k}$, we have

$$
\overline{y}_{t+1} - \overline{y}_t = \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) - \alpha_t g_t
$$

and so

$$
\begin{aligned}
\beta_t \langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle &= \beta_t \langle \overline{w}_t - w^*, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) - \alpha_t g_{t,k} \rangle \\
&= \beta_t \langle \overline{w}_t - w^*, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) \rangle - \beta_t \langle \overline{w}_t - w^*, \alpha_t g_t \rangle
\end{aligned}
$$

and we further expand the first term:

$$
\begin{aligned}
\beta_t \langle \overline{w}_t - w^*, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) \rangle &= \beta_t \langle \overline{w}_t - \overline{w}_{t-1} + \overline{w}_{t-1} - w^*, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) \rangle \\
&= \beta_t \langle \overline{w}_t - \overline{w}_{t-1}, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) \rangle + \beta_t \langle \overline{w}_{t-1} - w^*, \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}) \rangle \\
&= \beta_t \beta_{t-1} \langle -\alpha_{t-1} g_{t-1} + \beta_{t-1}(\overline{y}_t - \overline{y}_{t-1}), (\overline{y}_t - \overline{y}_{t-1}) \rangle + \beta_t \langle \overline{w}_{t-1} - w^*, \beta_{t-1}(\overline{y}_t - \\
&= \beta_t \beta_{t-1} \langle -\alpha_{t-1} g_{t-1}, (\overline{y}_t - \overline{y}_{t-1}) \rangle + \beta_t \beta_{t-1}^2 \|\overline{y}_t - \overline{y}_{t-1}\|^2 + \beta_t \beta_{t-1} \langle \overline{w}_{t-1} - w^*,
\end{aligned}
$$

and so

$$
\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle = -\beta_{t-1}\alpha_{t-1} \langle g_{t-1}, \overline{y}_t - \overline{y}_{t-1} \rangle + \beta_{t-1}^2 \|\overline{y}_t - \overline{y}_{t-1}\|^2 + \beta_{t-1} \langle \overline{w}_{t-1} - w^*, (\overline{y}_t - \overline{y}_{t-1}) \rangle -
$$

from which we can conclude that $|\beta_t \langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle| \leq \alpha_t^2 (G^2 + GK)$ and so

$$
\begin{aligned}
\mathbb{E}\|\overline{w}_{t+1} - w^*\|^2 \leq{} & (1 - \mu\alpha_t)\mathbb{E}\|\overline{w}_t - w^*\|^2 + 9L\alpha_t^2\Gamma + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 + 32\alpha_t^2(E-1)^2 G^2 \\
& + 2\alpha_t^2 G^2 + 2\beta_t \mathbb{E}\langle \overline{w}_t - w^*, (\overline{y}_{t+1} - \overline{y}_t) \rangle \\
={} & (1 - \mu\alpha_t)\mathbb{E}\|\overline{w}_t - w^*\|^2 + \alpha_t^2 B'
\end{aligned}
$$

where

$$
\begin{aligned}
B' &= B + G^2 + GK \\
&= \sum_{k=1}^N p_k^2 \sigma_k^2 + 9L\Gamma + 32(E-1)^2 G^2 + 2 + G^2 + GK
\end{aligned}
$$

and the rest of the proof follows as ICLR paper. □