# Supplementary Materials

## A Proofs

**Lemma 4**

*If*

$$r_{k+1} \leqslant (1 - C_1 k^{-a}) r_k + C_2 (k^{-2a+\varepsilon} + k^{-2\gamma+a+\varepsilon}), \tag{10}$$

*for any constant* $1 > C_1 > 0, C_2 > 0, 1 > a > 0, 1 - a > \varepsilon > 0$, *there exists a constant* $\mathcal{E}$ *such that*

$$r_{k+1} \leqslant (k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon})\mathcal{E}.$$

**Proof to Lemma 4** First we expand the recursion relation (10) till $k = 1$:

$$\begin{aligned}
r_{k+1} \leqslant &(1 - C_1 k^{-a}) r_k + C_2 (k^{-2a+\varepsilon} + k^{-2\gamma+a+\varepsilon}) \\
\leqslant &\left(\prod_{\varkappa=1}^{k} (1 - C_1 \varkappa^{-a})\right) r_1 + C_2 \left(\sum_{\varkappa=1}^{k} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \left(\prod_{\hbar=\kappa+1}^{k} (1 - C_1 \hbar^{-a})\right)\right).
\end{aligned} \tag{11}$$

As the first step to bound (11), we use the monotonicity of the $\ln(\cdot)$ function to bound the form $\prod_{\hbar=m}^{k}(1 - C_1 \hbar^{-a})$ for some $m$. Notice that

$$\ln\left(\prod_{\hbar=m}^{k} (1 - C_1 \hbar^{-a})\right) = \sum_{\hbar=m}^{k} \ln(1 - C_1 \hbar^{-a}) \leqslant -C_1 \sum_{\hbar=m}^{k} \hbar^{-a}.$$

Then it follows from the monotonicity of $\ln(\cdot)$ that:

$$\prod_{\hbar=m}^{k} (1 - C_1 \hbar^{-a}) \leqslant \exp\left(-C_1 \sum_{\hbar=m}^{k} \hbar^{-a}\right) \leqslant \exp(-C_1 (k - m + 1) k^{-a}). \tag{12}$$

Using this inequality (11) can be bounded by the following inequality:

$$\begin{aligned}
r_{k+1} \leqslant &\left(\prod_{\varkappa=1}^{k} (1 - C_1 \varkappa^{-a})\right) r_1 + C_2 \left(\sum_{\varkappa=1}^{k} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \left(\prod_{\hbar=\kappa+1}^{k} (1 - C_1 \hbar^{-a})\right)\right) \\
\overset{(12)}{\leqslant} &\exp(-k^{1-a}) r_1 + \underbrace{C_2 \left(\sum_{\varkappa=1}^{k} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \exp(-C_1 (k - \varkappa) k^{-a})\right)}_{=:T_k}.
\end{aligned} \tag{13}$$

The final step will be bounding $T_k$ above. The term can actually be bounded by a power of reciprocal function. Notice that for any $1 - a > \varepsilon > 0$, we have the following inequality:

$$\begin{aligned}
T_k = &C_2 \sum_{\varkappa=1}^{k} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \exp(-C_1 (k - \varkappa) k^{-a}) \\
\leqslant &C_2 \sum_{\varkappa=1}^{k - \lfloor k^{a+\varepsilon} \rfloor} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \exp(-C_1 \lfloor k^{a+\varepsilon} \rfloor k^{-a}) + C_2 \sum_{\varkappa=k-\lfloor k^{a+\varepsilon} \rfloor + 1}^{k} (\varkappa^{-2a+\varepsilon} + \varkappa^{-2\gamma+a+\varepsilon}) \\
= &O(\exp(-C_1 \lfloor k^{a+\varepsilon} \rfloor k^{-a})) + O(k^{a+\varepsilon} (k^{-2a+\varepsilon} + k^{-2\gamma+a+\varepsilon})) = O(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}),
\end{aligned}$$

where $O(\cdot)$ notation is used to ignore any constant terms to simplify the notation. It immediately follows from (13) that there exists a constant $\mathcal{E}$ such that for any $1 - a > \varepsilon > 0$, the following inequality holds:

$$r_{k+1} \leqslant (k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon})\mathcal{E}.$$

**Proof to Lemma 1** First note that for any $\alpha_k \in [0,1]$, the following identity always holds for the difference between the estimation and the true expectation. The first step comes from the update rule in Algorithm 2.

$$
\begin{aligned}
&\hat{e}_n^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) \\
=&(1-\alpha_k)\hat{e}_n^{(k)} + \alpha_k \hat{\mathbf{e}}_n(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) \\
=&(1-\alpha_k)(\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)) + \alpha_k(\hat{\mathbf{e}}_n(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)) \\
=&(1-\alpha_k)(\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) + \alpha_k(\hat{\mathbf{e}}_n(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)) \\
&- (1-\alpha_k)(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)).
\end{aligned} \tag{14}
$$

Then take $\ell_2$ norm on both sides. The second step comes from the fact that $n$ is the last layer's index, so $\mathbf{e}_n(x;\xi) = \hat{\mathbf{e}}_n(x^{(k)};\xi;\hat{e}^{(k)})$.

$$
\begin{aligned}
&\mathbb{E}\|\hat{e}_n^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)\|^2 \\
=&\mathbb{E}\left\| \begin{array}{c} (1-\alpha_k)(\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \\ +\alpha_k(\hat{\mathbf{e}}_n(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)) \\ -(1-\alpha_k)(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2 \\
=&\mathbb{E}\left\| \begin{array}{c} (1-\alpha_k)(\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \\ +\alpha_k(\mathbf{e}_n(x^{(k)};\xi_k) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)) \\ -(1-\alpha_k)(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2 \\
=&(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} \hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2 + \alpha_k^2\mathbb{E}\|\mathbf{e}_n(x^{(k)};\xi_k) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)\|^2 \\
&+ 2(1-\alpha_k)\alpha_k\mathbb{E}\left\langle \begin{array}{c} \hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array}, \mathbf{e}_n(x^{(k)};\xi_k) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) \right\rangle \\
=&(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} \hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2 + \alpha_k^2\mathbb{E}\|\mathbf{e}_n(x^{(k)};\xi_k) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)\|^2 \\
&+ 2(1-\alpha_k)\alpha_k\mathbb{E}\left\langle \begin{array}{c} \hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array}, \underbrace{\mathbb{E}_{\xi_k} \mathbf{e}_n(x^{(k)};\xi_k) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)}_{=0} \right\rangle \\
\overset{\text{Assumption 1-2}}{\leqslant}& \underbrace{(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} \hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2}_{=:\mathcal{T}_1} + \alpha_k^2\sigma^2.
\end{aligned} \tag{15}
$$

The second term is bounded by the bounded variance assumption. We denote the first term as $\mathcal{T}_1$ and bound it separately. With the fact that $\|x+y\|^2 \leqslant (1+\mathscr{d})\|x\|^2 + (1+\frac{1}{\mathscr{d}})\|y\|^2, \forall \mathscr{d} \in (0,1), \forall x, \forall y$ we can bound $\mathcal{T}_1$ as following:

$$
\begin{aligned}
\mathcal{T}_1 =&(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} (\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \\ -(\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)) \end{array} \right\|^2 \\
\leqslant&(1-\alpha_k)^2\left((1+\mathscr{d})\mathbb{E}\|\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 + \left(1+\frac{1}{\mathscr{d}}\right)\mathbb{E}\|\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2\right), \forall \mathscr{d} \in (0,1) \\
\overset{\mathscr{d}\leftarrow\alpha_k}{=}&(1-\alpha_k)^2(1+\alpha_k)\mathbb{E}\|\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 + \left(\frac{(1+\alpha_k)(1-\alpha_k)^2}{\alpha_k}\right)\mathbb{E}\|\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi e_n(x^{(k-1)};\xi)\|^2 \\
\leqslant&(1-\alpha_k)\mathbb{E}\|\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 + \frac{1}{\alpha_k}\mathbb{E}\|\mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi) - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 \\
\overset{\text{Assumption 1-4}}{\leqslant}&(1-\alpha_k)\mathbb{E}\|\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 + \frac{1}{\alpha_k}L\mathbb{E}\|x^{(k)} - x^{(k-1)}\|^2.
\end{aligned} \tag{16}
$$

The last step comes from the Lipschitzian assumption on the functions.

Putting (16) back into (15) we obtain

$$\mathbb{E}\|\hat{e}_n^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k)};\xi)\|^2 \leqslant (1-\alpha_k)\mathbb{E}\|\hat{e}_n^{(k)} - \mathbb{E}_\xi \mathbf{e}_n(x^{(k-1)};\xi)\|^2 + \frac{1}{\alpha_k}\underbrace{L\mathbb{E}\|x^{(k)} - x^{(k-1)}\|^2}_{=O(\gamma_{k-1}^2)} + \alpha_k^2\sigma^2, \tag{17}$$

where the second term is of order $O(\gamma_{k-1}^2)$ because the step length is $\gamma_{k-1}$ for the $(k-1)$-th step and the gradient is bounded according to Assumption 1-1. For the case $i = n$, (8) directly follows from combining (17), Lemma 4 and Assumption 1-6.

We then prove (8) for all $i$ by induction. Assuming for all $p > i$ and for any $1 - a > \epsilon > 0$ there exists a constant $\mathcal{E}$ (8) holds such that

$$\mathbb{E}\|\hat{e}_p^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_p(x^{(k)};\xi)\|^2 \leqslant \mathcal{E}(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}). \tag{18}$$

Similar to (14) we can split the difference between the estimation and the expectation into three parts:

$$\begin{aligned}
&\hat{e}_i^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\\
=&(1-\alpha_k)\hat{e}_i^{(k)} + \alpha_k \hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\\
=&(1-\alpha_k)(\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)) + (1-\alpha_k)\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) + \alpha_k \hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\\
=&(1-\alpha_k)(\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)) + (1-\alpha_k)(\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi))\\
&+ \alpha_k(\hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)),
\end{aligned}$$

Taking the $\ell_2$ norm on both sides we obtain

$$\begin{aligned}
&\mathbb{E}\|\hat{e}_i^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2\\
\leqslant&\mathbb{E}\left\| \begin{array}{c} (1-\alpha_k)(\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)) \\ +(1-\alpha_k)(\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)) \\ +\alpha_k(\hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)) \end{array} \right\|^2\\
=&\underbrace{(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) \\ +(\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)) \end{array} \right\|^2}_{\mathcal{T}_2}\\
&+ 2(1-\alpha_k)\alpha_k \underbrace{\left\langle \begin{array}{c} \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) \\ +\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi) \end{array}, \hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi) \right\rangle}_{\mathcal{T}_4}\\
&+ \underbrace{\alpha_k^2\mathbb{E}\|\hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2}_{\mathcal{T}_3},
\end{aligned}$$

We define the there parts in the above inequality as $\mathcal{T}_2, \mathcal{T}_3$ and $\mathcal{T}_4$ so that we can bound them one by one. Firstly for $\mathcal{T}_3$, it can be easily bounded using (18), Assumption 1-2 and Assumption 1-4:

$$\begin{aligned}
\mathcal{T}_3 =&\alpha_k^2\mathbb{E}\|\hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2\\
\leqslant&\alpha_k^2\mathbb{E}\|\hat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)})\| + \alpha_k^2\mathbb{E}\|\mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|\\
\overset{\text{Assumption 1-2}}{\leqslant}&\alpha_k^2\sigma^2 + \alpha_k^2\mathbb{E}\|\mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|\\
\overset{\text{Assumption 1-4}}{\leqslant}&\alpha_k^2\sigma^2 + \alpha_k^2 L \sum_{j=i+1}^{n} \mathbb{E}\|\hat{e}_j^{(k)} - \mathbb{E}_\xi \mathbf{e}_j(x^{(k)};\xi)\|^2\\
\overset{(18)}{\leqslant}&\alpha_k^2\sigma^2 + n\alpha_k^2 L\mathcal{E}.
\end{aligned} \tag{19}$$

We then take a look at $\mathcal{T}_2$. It can be bounded similar to what we did in (16) to bound $\mathcal{T}_1$.

$$\mathcal{T}_2 =(1-\alpha_k)^2\mathbb{E}\left\| \begin{array}{c} \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) \\ +(\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)) \end{array} \right\|^2$$

$$
\begin{aligned}
&\leqslant (1-\alpha_k)^2 \begin{pmatrix} (1+\mathscr{d})\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 \\ + \left(1+\frac{1}{\mathscr{d}}\right)\mathbb{E}\|\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2 \end{pmatrix}, \quad \forall \mathscr{d} \in (0,1)\\
&\overset{\mathscr{d}\leftarrow\alpha_k}{\leqslant} (1-\alpha_k)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \frac{1}{\alpha_k}\mathbb{E}\|\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2\\
&\overset{\text{Assumption 1-4}}{\leqslant} (1-\alpha_k)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \frac{L}{\alpha_k}(\mathbb{E}\|x^{(k)} - x^{(k-1)}\|^2). \qquad (20)
\end{aligned}
$$

Finally we need to bound $\mathscr{T}_4$. This one is a litter harder than $\mathscr{T}_2$ and $\mathscr{T}_3$. Different from what we were doing in (15), we no longer have the nice property $\mathbf{e}_i(x;\xi) = \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)})$ since here we are dealing with $i \neq n$. We need to further split it and bound each part separately.

$$
\begin{aligned}
\mathscr{T}_4 =& \left\langle \begin{matrix} \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) \\ +\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi) \end{matrix}, \mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi) \right\rangle\\
=& \underbrace{\langle \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi), \mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\rangle}_{\mathscr{T}_6}\\
&+ \underbrace{\langle \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi), \mathbb{E}_\xi \hat{\mathbf{e}}_i(x^{(k)};\xi;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\rangle}_{\mathscr{T}_7}. \qquad (21)
\end{aligned}
$$

We first bound $\mathscr{T}_7$:

$$
\begin{aligned}
\mathscr{T}_7 =& \langle \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi), \mathbb{E}\widehat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\rangle\\
\leqslant& \underbrace{\|\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|}_{=O(\gamma_{k-1})} \|\mathbb{E}\widehat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|\\
=& O\left(\gamma_{k-1}\|\mathbb{E}\widehat{\mathbf{e}}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|\right) \qquad (22)\\
\overset{\text{Assumption 1-4}}{\leqslant}& O\left(\gamma_{k-1}\sum_{j=i+1}^{n}\|\hat{e}_j^{(k)} - \mathbb{E}_\xi \mathbf{e}_j(x^{(k)};\xi)\|\right)\\
\overset{(18)}{=}& O\left(\gamma_{k-1}\sqrt{k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}}\right) = O(\gamma_{k-1}(k^{-\gamma+a+\varepsilon/2} + k^{-a/2+\varepsilon/2}))\\
=& O(k^{-2\gamma+a+\varepsilon/2} + k^{-\gamma-a/2+\varepsilon/2}), \qquad (23)
\end{aligned}
$$

where in the second step we have $\|\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\| = O(\gamma_{k-1})$ since by Assumption 1-4 we obtain $\|\mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\| \leqslant L\|x^{(k-1)} - x^{(k)}\|$. Then it follows from the same argument as in (17).

After bounding $\mathscr{T}_7$, we start investigating $\mathscr{T}_6$. The last step follows the same procedure as in (22).

$$
\begin{aligned}
\mathscr{T}_6 =& \langle \hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi), \mathbb{E}\mathbf{e}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\rangle\\
\leqslant& \frac{1}{2\mathscr{d}_k}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \mathscr{d}_k\|\mathbb{E}\mathbf{e}_i(x^{(k)};\xi_k;\hat{e}^{(k)}) - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2, \quad \forall \mathscr{d}_k > 0\\
=& \frac{1}{2\mathscr{d}_k}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \mathscr{d}_k O(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}), \quad \forall \mathscr{d}_k > 0. \qquad (24)
\end{aligned}
$$

Finally plug (19), (20), (24) and (23) back into (21). By choosing $\mathscr{d}_k$ in (24) to be 1 we obtain:

$$
\begin{aligned}
&\mathbb{E}\|\hat{e}_i^{(k+1)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k)};\xi)\|^2\\
\leqslant& \alpha_k^2\sigma^2 + n\alpha_k^2 L\mathscr{E}\\
&+ (1-\alpha_k)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \frac{L}{\alpha_k}(\mathbb{E}\|x^{(k)} - x^{(k-1)}\|^2)\\
&+ 2(1-\alpha_k)\alpha_k O(k^{-2\gamma+a+\varepsilon/2} + k^{-\gamma-a/2+\varepsilon/2})\\
&+ (1-\alpha_k)\alpha_k \frac{1}{\mathscr{d}_k}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + 2(1-\alpha_k)\alpha_k\mathscr{d}_k O(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon})
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\mathscr{I}_k \leftarrow 1}{\leqslant} \alpha_k^2\sigma^2 + n\alpha_k^2 L\mathcal{E} \\
&\quad + (1-\alpha_k)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \frac{L}{\alpha_k}(\mathbb{E}\|x^{(k)} - x^{(k-1)}\|^2) \\
&\quad + 2(1-\alpha_k)\alpha_k O(k^{-2\gamma+a+\varepsilon/2} + k^{-\gamma-a/2+\varepsilon/2}) \\
&\quad + (1-\alpha_k)\alpha_k \frac{1}{2}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + 4(1-\alpha_k)\alpha_k O(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}) \\
&\leqslant \left(1 - \frac{\alpha_k}{2}\right)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \alpha_k^2\sigma^2 + n\alpha_k^2 L\mathcal{E} + \frac{L\gamma_{k-1}^2}{\alpha_k}\mathcal{G}^2 \\
&\quad + 2(1-\alpha_k)\alpha_k O(k^{-2\gamma+a+\varepsilon/2} + k^{-\gamma-a/2+\varepsilon/2}) + 4(1-\alpha_k)\alpha_k O(k^{-2\gamma+2a+\varepsilon} + k^{-a+\varepsilon}) \\
&= \left(1 - \frac{\alpha_k}{2}\right)\mathbb{E}\|\hat{e}_i^{(k)} - \mathbb{E}_\xi \mathbf{e}_i(x^{(k-1)};\xi)\|^2 + \alpha_k^2\sigma^2 \\
&\quad + n\alpha_k^2 L\mathcal{E} + \frac{L\gamma_{k-1}^2}{\alpha_k}\mathcal{G}^2 + O(k^{-2\gamma+a+\varepsilon} + k^{-2a+\varepsilon}).
\end{aligned} \tag{25}
$$

By combining (25) and Lemma 4 we obtain (8). (9) follows from (17) and (25).

**Proof to Theorem 2** It directly follows from the definition of Lipschitz condition of $f$ that,

$$
\begin{aligned}
f(x^{(k+1)}) &\overset{\text{Assumption 1-4}}{\leqslant} f(x^{(k)}) + \langle \partial f(x^{(k)}), x^{(k+1)} - x^{(k)}\rangle + L\|x^{(k+1)} - x^{(k)}\|^2 \\
&= f(x^{(k)}) + \langle \partial f(x^{(k)}), -\gamma_k g^{(k)}\rangle + L\|x^{(k+1)} - x^{(k)}\|^2 \\
&= f(x^{(k)}) - \gamma_k\|\partial f(x^{(k)})\|^2 + \underbrace{\langle \partial f(x^{(k)}), -\gamma_k(g^{(k)} - \partial f(x^{(k)}))\rangle}_{\mathcal{T}\text{cross}} + L\underbrace{\|x^{(k+1)} - x^{(k)}\|^2}_{\mathcal{T}\text{progress}}.
\end{aligned} \tag{26}
$$

Here we define two new terms and try to bound the expectation of $\mathcal{T}$cross and $\mathcal{T}$progress separately as shown below:

$$
\begin{aligned}
\mathbb{E}\mathcal{T}\text{cross} &= \mathbb{E}\langle \partial f(x^{(k)}), -\gamma_k(g^{(k)} - \partial f(x^{(k)}))\rangle \\
&\leqslant \frac{2\gamma_k^2 L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2 + \frac{\alpha_{k+1}}{2L_g}\mathbb{E}\|g^{(k)} - \partial f(x^{(k)})\|^2 \\
&\overset{\text{Assumption 1-3}}{\leqslant} \frac{2\gamma_k^2 L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2 + \frac{\alpha_{k+1}}{2}\sum_{i=1}^n \|\hat{e}_i^{(k+1)} - \mathbb{E}_\xi[\mathbf{e}_i(x^{(k)};\xi)]\|^2,
\end{aligned} \tag{27}
$$

and

$$
\mathbb{E}\mathcal{T}\text{progress} = \mathbb{E}\|x_{k+1} - x_k\|^2 \overset{\text{Assumption 1-1}}{\leqslant} \gamma_k^2\mathcal{G}^2. \tag{28}
$$

With the help of (27) and (28), (26) becomes

$$
\begin{aligned}
\mathbb{E}f(x^{(k+1)}) \leqslant& \mathbb{E}f(x^{(k)}) - \gamma_k\mathbb{E}\|\partial f(x^{(k)})\|^2 \\
&+ \frac{2\gamma_k^2 L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2 + \frac{\alpha_{k+1}}{2}\sum_{i=1}^{n-1}\mathbb{E}\|\hat{e}_i^{(k+1)} - \mathbb{E}_{\xi_k}[e_i(x^{(k)};\xi_k)]\|^2 + \gamma_k^2 L\mathcal{G}^2.
\end{aligned} \tag{29}
$$

To show how this converges, we define the following $\mathscr{I}_k$ to derive a recursive relation for (29):

$$
\mathscr{I}_k := f(x^{(k)}) + \sum_{i=1}^{n-1}\|\hat{e}_i^{(k+1)} - \mathbb{E}_{\xi_k}[e_i(x^{(k)};\xi_k)]\|^2.
$$

Then the recursive relation is derived by observing

$$
\mathbb{E}\mathscr{I}_{k+1} \leqslant \mathbb{E}f(x^{(k)}) - \gamma_k\mathbb{E}\|\partial f(x^{(k)})\|^2 + \frac{2\gamma_k^2 L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2 + \frac{\alpha_{k+1}}{2}\sum_{i=1}^{n-1}\|\hat{e}_i^{(k+1)} - \mathbb{E}_{\xi_k}[e_i(x^{(k)};\xi_k)]\|^2 + \gamma_k^2 L\mathcal{G}^2
$$

$$
\begin{aligned}
&+\sum_{i=1}^{n-1}\|\hat{e}_i^{(k+2)}-\mathbb{E}_{\xi_{k+1}}[e_i(x^{(k+1)};\xi_{k+1})]\|^2\\
\overset{\text{Lemma 1}}{\leqslant}\;&\mathbb{E}f(x^{(k)})-\gamma_k\mathbb{E}\|\partial f(x^{(k)})\|^2\\
&+\frac{2\gamma_k^2L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2+\frac{\alpha_{k+1}}{2}\sum_{i=1}^{n-1}\mathbb{E}\|\hat{e}_i^{(k+1)}-\mathbb{E}_{\xi_k}[e_i(x^{(k)};\xi_k)]\|^2+\gamma_k^2L\mathcal{G}^2\\
&+\sum_{i=1}^{n-1}\left(\left(1-\frac{\alpha_{k+1}}{2}\right)\mathbb{E}\|\hat{e}_i^{(k+1)}-\mathbb{E}_{\xi}\mathbf{e}_i(x^{(k+1)};\xi)\|^2+\mathcal{C}((k+1)^{-2\gamma+a+\varepsilon}+(k+1)^{-2a+\varepsilon})\right)\\
\leqslant&\mathbb{E}f(x^{(k)})-\gamma_k\mathbb{E}\|\partial f(x^{(k)})\|^2+\frac{2\gamma_k^2L_g}{\alpha_{k+1}}\mathbb{E}\|\partial f(x^{(k)})\|^2+\sum_{i=1}^{n-1}\mathbb{E}\|\hat{e}_i^{(k+1)}-\mathbb{E}_{\xi}\mathbf{e}_i(x^{(k)};\xi)\|^2+O(k^{-2\gamma+a+\varepsilon}+k^{-2a+\varepsilon})\\
=&\mathbb{E}\mathcal{d}_k-\left(\gamma_k-\frac{2\gamma_k^2L_g}{\alpha_{k+1}}\right)\mathbb{E}\|\partial f(x^{(k)})\|^2+O(k^{-2\gamma+a+\varepsilon}+k^{-2a+\varepsilon}).
\end{aligned}
$$

Thus as long as $a-2\gamma<-1, a<1/2$ we can choose $\epsilon$ such that there exists a constant $\mathcal{R}$:

$$
\begin{aligned}
\mathbb{E}\mathcal{d}_{K+1}\leqslant&\mathbb{E}\mathcal{d}_0-\sum_{k=0}^{K}\left(\gamma_k-\frac{2\gamma_k^2L_g}{\alpha_{k+1}}\right)\mathbb{E}\|\partial f(x^{(k)})\|^2+\mathcal{R}\\
\sum_{k=0}^{K}\left(\gamma_k-\frac{2\gamma_k^2L_g}{\alpha_{k+1}}\right)\mathbb{E}\|\partial f(x^{(k)})\|^2\leqslant&\mathcal{R}+\mathbb{E}\mathcal{d}_0-\mathbb{E}\mathcal{d}_{K+1}\\
\xrightarrow{\frac{\gamma_kL_g}{\alpha_{k+1}}\leqslant\frac{1}{2}}\ \frac{\sum_{k=0}^{K}\gamma_k\mathbb{E}\|\partial f(x^{(k)})\|^2}{\sum_{k=0}^{K}\gamma_k}\leqslant&\frac{2(\mathcal{R}+\mathbb{E}\mathcal{d}_0-\mathbb{E}\mathcal{d}_{K+1})}{\sum_{k=0}^{K}\gamma_k}.
\end{aligned}
$$

By Assumption 1-5 we complete the proof.

**Proof to Corollary 3** With the given choice of $\alpha_k$ and $\gamma_k$ the prerequisites in Theorem 2 are satisfied. Then by Theorem 2 there exists a constant $\mathcal{H}$ (which may differ from the constant in Theorem 2) such that

$$
\begin{aligned}
\sum_{k=0}^{K}(k+2)^{-4/5}\mathbb{E}\|\partial f(x^{(k)})\|^2\leqslant&\mathcal{H}\\
\sum_{k=0}^{K}\mathbb{E}\|\partial f(x^{(k)})\|^2\leqslant&\frac{\mathcal{H}}{(K+2)^{-4/5}}\\
\frac{\sum_{k=0}^{K}\mathbb{E}\|\partial f(x^{(k)})\|^2}{K+2}\leqslant&\frac{\mathcal{H}}{(K+2)^{1/5}},
\end{aligned}
$$

completing the proof.