Supplementary Materials for "Leveraging First and Zeroth-Order Gradient to Address Imbalanced Black-box Prompt Tuning via Minimax Optimization"

Proofs for Technical Lemma

Lemma 3. *For Algorithm 1, let $\Delta_t = \mathbb{E}[\Phi(\boldsymbol{v}_t) - f(\boldsymbol{v}_t, \alpha_t)]$ and $\eta_{\boldsymbol{v}} > \frac{c}{\ell}$ (c is a constant value), the following statement holds true,*

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_t)] \leq & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\Delta_{t-1} - \frac{1}{4}(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})||^2] \\ & + \frac{\eta_{\boldsymbol{v}}^2\ell}{c}\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\|^2] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2].\end{aligned} \tag{10}$$

Proof. Let $\hat{\boldsymbol{v}}_{t-1} = \mathrm{prox}_{\Phi/2\ell}(\boldsymbol{v}_{t-1})$, we have

$$\Phi_{1/2\ell}(\boldsymbol{v}_t) = \min_{\mathbf{w}} \Phi(\mathbf{w}) + \ell||\mathbf{w} - \boldsymbol{v}_t||^2 \leq \Phi(\hat{\boldsymbol{v}}_{t-1}) + \ell||\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_t||^2 \tag{11}$$

For the update of $\boldsymbol{v}_t$, we have

$$\begin{aligned}\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_t\|^2 &= \|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1} + \eta_{\boldsymbol{v}}\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2 \\ &= \|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2 + 2\eta_{\boldsymbol{v}}\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\rangle + \eta_{\boldsymbol{v}}^2\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2\end{aligned} \tag{12}$$

Combining with Equation (11) and (12) yields that

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{v}_t) &\leq \Phi(\hat{\boldsymbol{v}}_{t-1}) + \ell\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2 + 2\eta_{\boldsymbol{v}}\ell\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\rangle + \eta_{\boldsymbol{v}}^2\ell\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2 \\ &= \Phi_{1/2\ell}(\boldsymbol{v}_{t-1}) + 2\eta_{\boldsymbol{v}}\ell\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\rangle + \eta_{\boldsymbol{v}}^2\ell\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2\end{aligned} \tag{13}$$

Taking the expectation of both sides together yields that

$$\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_t)] \leq \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\mathbb{E}[\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\rangle] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \tag{14}$$

Since f is ℓ-smooth, we have the lower bound for the function value

$$f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1}) \geq f(\boldsymbol{v}_{t-1}, \alpha_{t-1}) + \langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle - \frac{\ell}{2}\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2 \tag{15}$$

That is

$$\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle \leq f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1}) + \frac{\ell}{2}\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2 \tag{16}$$

Taking the expectation of both sides together yields that

$$\mathbb{E}[\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle] \leq \mathbb{E}[f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] + \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] \tag{17}$$

Since $\Phi(\hat{\boldsymbol{v}}_{t-1}) = \max_{\alpha\in\mathcal{A}} f(\hat{\boldsymbol{v}}_{t-1}, \alpha) > f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1})$, we have

$$\mathbb{E}[f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] \leq \mathbb{E}[\Phi(\hat{\boldsymbol{v}}_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] \tag{18}$$

Furthermore, by the definition of $\Delta_{t-1} = \mathbb{E}[\Phi(\boldsymbol{v}_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})]$ and $\hat{\boldsymbol{v}}_{t-1} = \mathrm{prox}_{\Phi/2\ell}(\boldsymbol{v}_{t-1}) = \mathrm{argmin}_{\boldsymbol{y}}\{\Phi(\boldsymbol{y}) + \ell\|\boldsymbol{y} - \boldsymbol{v}_{t-1}\|^2\}$, we have

$$\mathbb{E}[\Phi(\hat{\boldsymbol{v}}_{t-1})] + \ell\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] \leq \mathbb{E}[\Phi(\boldsymbol{v}_{t-1})] = \Delta_{t-1} + \mathbb{E}[f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] \tag{19}$$

Thus, we have

$$\mathbb{E}[f(\hat{\boldsymbol{v}}_{t-1}, \alpha_{t-1}) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] \leq \Delta_{t-1} - \ell\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] \tag{20}$$

Combining with Equation (17) and (20) yields that

$$\mathbb{E}[\langle\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle] \leq \Delta_{t-1} - \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] \tag{21}$$

Combining with Equation (14) and (21) yields that

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_t)] \leq & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\mathbb{E}[\langle \hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\rangle] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
= & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\mathbb{E}[\langle \hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle] \\
& + 2\eta_{\boldsymbol{v}}\ell\mathbb{E}[\langle \hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
\leq & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell(\Delta_{t-1} - \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2]) \\
& + 2\eta_{\boldsymbol{v}}\ell\mathbb{E}[\langle \hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}, \hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\rangle] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
= & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell(\Delta_{t-1} - \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2]) \\
& + 2\ell\mathbb{E}[\langle \sqrt{c}(\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}), \frac{\eta_{\boldsymbol{v}}}{\sqrt{c}}(\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1}))\rangle] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
\overset{(a)}{\leq} & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell(\Delta_{t-1} - \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2]) \\
& + 2\ell(\frac{1}{2}\mathbb{E}[\|\sqrt{c}(\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1})\|^2] + \frac{1}{2}\mathbb{E}[\|\frac{\eta_{\boldsymbol{v}}}{\sqrt{c}}(\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1}))\|^2]) + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
= & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell(\Delta_{t-1} - \frac{\ell}{2}\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2]) \\
& + c\ell\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] + \frac{\eta_{\boldsymbol{v}}^2\ell}{c}\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\|^2] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
= & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\Delta_{t-1} - \ell^2(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\mathbb{E}[\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\|^2] \\
& + \frac{\eta_{\boldsymbol{v}}^2\ell}{c}\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\|^2] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2] \\
\overset{(b)}{=} & \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})] + 2\eta_{\boldsymbol{v}}\ell\Delta_{t-1} - \frac{1}{4}(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})||^2] \\
& + \frac{\eta_{\boldsymbol{v}}^2\ell}{c}\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}_{t-1}, \alpha_{t-1})\|^2] + \eta_{\boldsymbol{v}}^2\ell\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}^{(t-1)}\|^2]
\end{aligned} \tag{22}$$

where the inequality (a) uses Young's inequality $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2$ and the equality (b) uses $\|\hat{\boldsymbol{v}}_{t-1} - \boldsymbol{v}_{t-1}\| = \|\nabla\Phi_{1/2\ell}(\boldsymbol{v}_{t-1})\|/2\ell$ (Lemma 3.6 in (Lin, Jin, and Jordan 2020)). □

Lemma 4. *For Algorithm 1, let $\Delta_t = \mathbb{E}[\Phi(\boldsymbol{v}_t) - f(\boldsymbol{v}_t, \alpha_t)]$, B_1 denote as the bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|]$ and $\eta_\alpha \leq 1/2\ell$, the following statement holds true for $\forall s \leq t-1$,*

$$\begin{aligned}
\Delta_{t-1} \leq & \eta_{\boldsymbol{v}} L(2t - 2s - 1)B_1 + \frac{1}{2\eta_\alpha}(\mathbb{E}[\|\alpha_{t-1} - \alpha^*(\boldsymbol{v}_s)\|^2] - \mathbb{E}[\|\alpha_t - \alpha^*(\boldsymbol{v}_s)\|^2]) \\
& + \mathbb{E}[f(\boldsymbol{v}_t, \alpha_t) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] + \eta_\alpha\sigma^2.
\end{aligned} \tag{23}$$

Proof. Since α is updated using the first-order gradient rather than the zeroth-order gradient, we can use Lemm D.4 in (Lin, Jin, and Jordan 2020) directly to get the following inequalities for $\eta_\alpha \leq 1/2\ell$,

$$\begin{aligned}
\Delta_{t-1} \leq & \mathbb{E}[f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_{t-1})) - f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_s)) + (f(\boldsymbol{v}_t, \alpha_t) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})) + (f(\boldsymbol{v}_{t-1}, \alpha_t) - f(\boldsymbol{v}_t, \alpha_t))] \\
& + \eta_\alpha\sigma^2 + \frac{1}{2\eta_\alpha}\left(\mathbb{E}[\|\alpha_{t-1} - \alpha^*(\boldsymbol{v}_s)\|^2] - \mathbb{E}[\|\alpha_t - \alpha^*(\boldsymbol{v}_s)\|^2]\right)
\end{aligned} \tag{24}$$

$$f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_{t-1})) - f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_s)) \leq f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_{t-1})) - f(\boldsymbol{v}_s, \alpha^*(\boldsymbol{v}_{t-1})) + f(\boldsymbol{v}_s, \alpha^*(\boldsymbol{v}_s)) - f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_s)) \tag{25}$$

Since $f(\cdot, \alpha)$ is L-Lipschitz for any $\alpha \in \mathcal{A}$, we have

$$\begin{gathered}
\mathbb{E}[f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_{t-1})) - f(\boldsymbol{v}_s, \alpha^*(\boldsymbol{v}_{t-1}))] \leq L\mathbb{E}[\|\boldsymbol{v}_{t-1} - \boldsymbol{v}_s\|] \leq \eta_{\boldsymbol{v}} L(t-1-s)B_1, \\
\mathbb{E}[f(\boldsymbol{v}_s, \alpha^*(\boldsymbol{v}_s)) - f(\boldsymbol{v}_{t-1}, \alpha^*(\boldsymbol{v}_s))] \leq L\mathbb{E}[\|\boldsymbol{v}_{t-1} - \boldsymbol{v}_s\|] \leq \eta_{\boldsymbol{v}} L(t-1-s)B_1, \\
\mathbb{E}[f(\boldsymbol{v}_{t-1}, \alpha_t) - f(\boldsymbol{v}_t, \alpha_t)] \leq L\mathbb{E}[\|\boldsymbol{v}_t - \boldsymbol{v}_{t-1}\|] \leq \eta_{\boldsymbol{v}} LB_1
\end{gathered} \tag{26}$$

Putting Equations (24) - (26) together, we have

$$\begin{aligned}
\Delta_{t-1} \leq & \eta_{\boldsymbol{v}} L(2t - 2s - 1)B_1 + \frac{1}{2\eta_\alpha}(\mathbb{E}[\|\alpha_{t-1} - \alpha^*(\boldsymbol{v}_s)\|^2] - \mathbb{E}[\|\alpha_t - \alpha^*(\boldsymbol{v}_s)\|^2]) \\
& + \mathbb{E}[f(\boldsymbol{v}_t, \alpha_t) - f(\boldsymbol{v}_{t-1}, \alpha_{t-1})] + \eta_\alpha\sigma^2
\end{aligned} \tag{27}$$

□

Lemma 5. *For Algorithm 1, let* $\Delta_t = \mathbb{E}[\Phi(\boldsymbol{v}_t) - f(\boldsymbol{v}_t, \alpha_t)]$, $\widehat{\Delta}_0 = \mathbb{E}[\Phi(\boldsymbol{v}_0) - f(\boldsymbol{v}_0, \alpha_0)]$, B_1 *denote as the bound of* $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|]$ *and* $\eta_\alpha \leq 1/2\ell$*, the following statement holds true,*

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\Delta_t\right) \leq \eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2 + \frac{\widehat{\Delta}_0}{T+1}. \tag{28}$$

Proof. Similar to Lemma D.5 in (Lin, Jin, and Jordan 2020), we divide $\{\Delta_t\}_{t=0}^T$ into several blocks where each block contains at most B terms. Then we have,

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\Delta_t\right) = \frac{B}{T+1}\left[\sum_{j=0}^{(T+1)/B-1}\left(\frac{1}{B}\sum_{t=jB}^{(j+1)B-1}\Delta_t\right)\right] \tag{29}$$

Letting $s = jB$ in the inequality (23) in Lemma 4 yields that

$$\sum_{t=jB}^{(j+1)B-1}\Delta_t \leq \eta_{\boldsymbol{v}} LB^2B_1 + \frac{D_{\mathcal{A}}^2}{2\eta_\alpha} + B\eta_\alpha\sigma^2 + \mathbb{E}[f(\boldsymbol{v}_{jB+B}, \alpha_{jB+B}) - f(\boldsymbol{v}_{jB}, \alpha_{jB})] \tag{30}$$

Then we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\Delta_t\right) \leq \eta_{\boldsymbol{v}} LBB_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2 + \frac{\mathbb{E}[f(\boldsymbol{v}_{T+1}, \alpha_{T+1}) - f(\boldsymbol{v}_0, \alpha_0)]}{T+1} \tag{31}$$

Since $f(\cdot, \alpha)$ is L-Lipschitz for $\forall\alpha \in \mathcal{A}$, we have

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{v}_{T+1}, \alpha_{T+1}) - f(\boldsymbol{v}_0, \alpha_0)] =& \mathbb{E}[f(\boldsymbol{v}_{T+1}, \alpha_{T+1}) - f(\boldsymbol{v}_0, \alpha_{T+1})] + \mathbb{E}[f(\boldsymbol{v}_0, \alpha_{T+1}) - f(\boldsymbol{v}_0, \alpha_0)] \\
\leq& \mathbb{E}[f(\boldsymbol{v}_{T+1}, \alpha_{T+1}) - f(\boldsymbol{v}_0, \alpha_{T+1})] + \mathbb{E}[\Phi(\boldsymbol{v}_0) - f(\boldsymbol{v}_0, \alpha_0)] \\
\leq& L\mathbb{E}[\|\boldsymbol{v}_{T+1} - \boldsymbol{v}_0\|] + \widehat{\Delta}_0 \\
\leq& \eta_{\boldsymbol{v}} LB_1(T+1) + \widehat{\Delta}_0
\end{aligned} \tag{32}$$

Putting these pieces together, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\Delta_t\right) \leq \eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2 + \frac{\widehat{\Delta}_0}{T+1} \tag{33}$$

□

Proofs for Theorem 1

Proof. Summing up the inequality in Lemma 3 over $t = 1, 2, \ldots, T+1$ yields that

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{T+1})] \leq& \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_0)] + 2\eta_{\boldsymbol{v}}\ell\sum_{t=0}^{T}\Delta_t - \frac{1}{4}(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \\
&+ \frac{\eta_{\boldsymbol{v}}^2\ell B_2(T+1)}{c} + \eta_{\boldsymbol{v}}^2\ell B_3(T+1)
\end{aligned} \tag{34}$$

Combining the above inequality with the inequality (28) in Lemma 5 yields that

$$\begin{aligned}
\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{T+1})] \leq& \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_0)] + 2\eta_{\boldsymbol{v}}\ell(T+1)\big(\eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2\big) + 2\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0 \\
&- \frac{1}{4}(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] + \frac{\eta_{\boldsymbol{v}}^2\ell B_2(T+1)}{c} + \eta_{\boldsymbol{v}}^2\ell B_3(T+1)
\end{aligned} \tag{35}$$

Thus we have

$$\begin{aligned}
\frac{1}{4}(\eta_{\boldsymbol{v}} - \frac{c}{\ell})\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq& \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_0)] - \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{v}_{T+1})] + 2\eta_{\boldsymbol{v}}\ell(T+1)(\eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
&+ 2\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0 + \frac{\eta_{\boldsymbol{v}}^2\ell B_2(T+1)}{c} + \eta_{\boldsymbol{v}}^2\ell B_3(T+1). \\
\overset{(a)}{\leq}& \widehat{\Delta}_\Phi + 2\eta_{\boldsymbol{v}}\ell(T+1)(\eta_{\boldsymbol{v}} L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
&+ 2\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0 + \frac{\eta_{\boldsymbol{v}}^2\ell B_2(T+1)}{c} + \eta_{\boldsymbol{v}}^2\ell B_3(T+1)
\end{aligned} \tag{36}$$

The inequality (a) uses the definition of $\widehat{\Delta}_\Phi$. By rearranging the above inequality, we have

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq & \frac{4\widehat{\Delta}_\Phi}{(\eta_{\boldsymbol{v}}-c/\ell)(T+1)} + \frac{8\eta_{\boldsymbol{v}}\ell}{\eta_{\boldsymbol{v}}-c/\ell}(\eta_{\boldsymbol{v}}L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\ & + \frac{8\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0}{(\eta_{\boldsymbol{v}}-c/\ell)(T+1)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_2}{c(\eta_{\boldsymbol{v}}-c/\ell)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_3}{\eta_{\boldsymbol{v}}-c/\ell}\end{aligned} \tag{37}$$

Bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|]$ (B_1): By the definition of $\hat{\mathbf{g}}_{\boldsymbol{v}}$ we have

$$\|\hat{\mathbf{g}}_{\boldsymbol{v}}\| = (\|\hat{\mathbf{g}}_{\mathbf{z}}\|^2 + \|\mathbf{g}_{a,b}\|^2)^{\frac{1}{2}} \leq \|\hat{\mathbf{g}}_{\mathbf{z}}\| + \|\mathbf{g}_{a,b}\| \overset{(a)}{\leq} \|\hat{\mathbf{g}}_{\mathbf{z}}\| + L \tag{38}$$

The inequality (a) uses the fact that $F(\cdot,\alpha;\xi)$ is L-Lipschitz. By the definition of $\hat{\mathbf{g}}_{\mathbf{z}}$ and since $F(\cdot,\alpha;\xi)$ is L-Lipschitz, we have

$$\begin{aligned}\|\hat{\mathbf{g}}_{\mathbf{z}}\| &= \|\frac{1}{m}\sum_{i=1}^{m}(\frac{F(\mathbf{z}+\mu\mathbf{u}_i,a,b,\alpha;\xi_i)-F(\mathbf{z}-\mu\mathbf{u}_i,a,b,\alpha;\xi_i)}{2\mu}\mathbf{u}_i)\| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\|\frac{F(\mathbf{z}+\mu\mathbf{u}_i,a,b,\alpha;\xi_i)-F(\mathbf{z}-\mu\mathbf{u}_i,a,b,\alpha;\xi_i)}{2\mu}\mathbf{u}_i\| \\ &= \frac{1}{m}\sum_{i=1}^{m}\|\frac{F(\mathbf{z}+\mu\mathbf{u}_i,a,b,\alpha;\xi_i)-F(\mathbf{z}-\mu\mathbf{u}_i,a,b,\alpha;\xi_i)}{2\mu}\|\|\mathbf{u}_i\| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}L\|\mathbf{u}_i\|^2\end{aligned} \tag{39}$$

Thus we have

$$\|\hat{\mathbf{g}}_{\boldsymbol{v}}\| \leq \frac{1}{m}\sum_{i=1}^{m}L\|\mathbf{u}_i\|^2 + L \tag{40}$$

Taking the expectation of both sides together yields that

$$\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|] \leq \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m}L\|\mathbf{u}_i\|^2] + L = L\mathbb{E}[\|\mathbf{u}\|^2] + L = (d+1)L \tag{41}$$

where the last equality uses the fact that the expectation of a chi-square distribution with d degrees of freedom is d. Finally, we have $B_1 = (d+1)L$

Bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)\|^2]$ (B_2): By the definition of $\hat{\mathbf{g}}_{\boldsymbol{v}}$, we have

$$
\begin{aligned}
\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}} - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)\|^2] =& \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}} - \nabla_{\mathbf{z}} f(\boldsymbol{v}, \alpha)\|^2] + \mathbb{E}[\|\mathbf{g}_a - \nabla_a f(\boldsymbol{v}, \alpha)\|^2 + \|\mathbf{g}_b - \nabla_b f(\boldsymbol{v}, \alpha)\|^2] \\
\overset{(a)}{\le}& \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}} - \nabla_{\mathbf{z}} f(\boldsymbol{v}, \alpha)\|^2] + \frac{\sigma^2}{m} \\
=& \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}} - \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha) + \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha) - \nabla_{\mathbf{z}} f(\boldsymbol{v}, \alpha)\|^2] + \frac{\sigma^2}{m} \\
=& \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}} - \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2] + \mathbb{E}[\|\nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha) - \nabla_{\mathbf{z}} f(\boldsymbol{v}, \alpha)\|^2] + \frac{\sigma^2}{m} \\
\overset{(b)}{\le}& \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}} - \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2] + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
=& \mathbb{E}[\|\frac{1}{m}\sum_{i=1}^{m} \hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi_i) - \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2] + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
=& \frac{1}{m}\mathbb{E}[\|\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi) - \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2] + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
=& \frac{1}{m}\mathbb{E}[\|\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi)\|^2] - \frac{2}{m}\mathbb{E}[\langle \hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi), \nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\rangle] + \frac{1}{m}\|\nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2 \\
&+ \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
\overset{(c)}{=}& \frac{1}{m}\mathbb{E}[\|\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi)\|^2] - \frac{1}{m}\|\nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)\|^2 + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
\le& \frac{1}{m}\mathbb{E}[\|\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi)\|^2] + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
\overset{(d)}{\le}& \frac{1}{m}\mathbb{E}[2(d+4)\|\nabla_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi)\|^2 + \frac{\mu^2 \ell^2 (d+6)^3}{2}] + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
\le& \frac{1}{m}(2(d+4)L^2 + \frac{\mu^2 \ell^2 (d+6)^3}{2}) + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m} \\
=& \frac{2(d+4)L^2}{m} + \frac{\mu^2 \ell^2 (d+6)^3}{2m} + \frac{\mu^2 \ell^2 (d+3)^3}{4} + \frac{\sigma^2}{m}
\end{aligned}
\tag{42}
$$

where $f_\mu(\boldsymbol{v}, \alpha) = \mathbb{E}_{\mathbf{u}}[f(\mathbf{z} + \mu\mathbf{u}, a, b, \alpha)]$.

The inequality (a) uses the fact below,

$$
\mathbb{E}[||\frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{v}} F(\boldsymbol{v}, \alpha; \xi_i) - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)||^2] = \frac{\sum_{i=1}^{m} \mathbb{E}[||\nabla_{\boldsymbol{v}} F(\boldsymbol{v}, \alpha; \xi_i) - \nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \alpha)||^2]}{m^2} \le \frac{\sigma^2}{m}
\tag{43}
$$

The inequalities (b) and (d) use Lemma 6 in (Huang et al. 2019). The equation (c) uses the fact that $\hat{\nabla}_{\mathbf{z}} F(\mathbf{z}, a, b, \alpha; \xi)$ is unbiased for $\nabla_{\mathbf{z}} f_\mu(\boldsymbol{v}, \alpha)$. Finally we have $B_2 = \frac{2(d+4)L^2}{m} + \frac{\mu^2\ell^2(d+6)^3}{2m} + \frac{\mu^2\ell^2(d+3)^3}{4} + \frac{\sigma^2}{m}$

Bound of $\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|^2]$ (B_3): By the definition of $\hat{\mathbf{g}}_{\boldsymbol{v}}$, we have

$$
\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|^2 = \|\hat{\mathbf{g}}_{\mathbf{z}}\|^2 + \|\mathbf{g}_{a,b}\|^2
\tag{44}
$$

Taking the expectation of both sides together yields that

$$
\begin{aligned}
\mathbb{E}[\|\hat{\mathbf{g}}_{\boldsymbol{v}}\|^2] &= \mathbb{E}[\|\hat{\mathbf{g}}_{\mathbf{z}}\|^2] + \mathbb{E}[\|\mathbf{g}_{a,b}\|^2] \\
&= \mathbb{E}[\|\frac{1}{m}\sum_{i=1}^{m}\hat{\nabla}_{\mathbf{z}}F(\mathbf{z},a,b,\alpha;\xi_i)\|^2] + \mathbb{E}[\|\frac{1}{m}\sum_{i=1}^{m}\nabla_{a,b}F(\mathbf{z},a,b,\alpha;\xi_i)\|^2] \\
&\overset{(a)}{\leq} \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\hat{\nabla}_{\mathbf{z}}F(\mathbf{z},a,b,\alpha;\xi_i)\|^2] + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla_{a,b}F(\mathbf{z},a,b,\alpha;\xi_i)\|^2] \\
&= \mathbb{E}[\|\hat{\nabla}_{\mathbf{z}}F(\mathbf{z},a,b,\alpha;\xi)\|^2] + \mathbb{E}[\|\nabla_{a,b}F(\mathbf{z},a,b,\alpha;\xi)\|^2] \\
&\overset{(b)}{\leq} \mathbb{E}[\|\hat{\nabla}_{\mathbf{z}}F(\mathbf{z},a,b,\alpha;\xi)\|^2] + L^2 \\
&\overset{(c)}{\leq} \mathbb{E}[2(d+4)\|\nabla_{\mathbf{z}}F(\mathbf{z},a,b,\alpha;\xi)\|^2 + \frac{\mu^2\ell^2(d+6)^3}{2}] + L^2 \\
&\overset{(d)}{\leq} 2(d+4)L^2 + \frac{\mu^2\ell^2(d+6)^3}{2} + L^2 \\
&= (2d+9)L^2 + \frac{\mu^2\ell^2(d+6)^3}{2}
\end{aligned}
\tag{45}
$$

where the inequality (a) uses $\|\sum_{i=1}^m \mathbf{x}_i\|^2 \leq m\sum_{i=1}^m \|\mathbf{x}_i\|^2$. The inequality (b) uses the fact that $F(\cdot,\alpha;\xi)$ is L-Lipschitz. The inequalities (c) and (d) use the Lemma 6 in (Huang et al. 2019). Finally we have $B_3 = (2d+9)L^2 + \frac{\mu^2\ell^2(d+6)^3}{2}$ □

Proof of Corollary 1

Proof. Since Theorem 1, we have

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq & \frac{4\widehat{\Delta}_\Phi}{(\eta_{\boldsymbol{v}}-c/\ell)(T+1)} + \frac{8\eta_{\boldsymbol{v}}\ell}{\eta_{\boldsymbol{v}}-c/\ell}(\eta_{\boldsymbol{v}}L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
& + \frac{8\eta_{\boldsymbol{v}}\ell\widehat{\Delta}_0}{(\eta_{\boldsymbol{v}}-c/\ell)(T+1)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_2}{c(\eta_{\boldsymbol{v}}-c/\ell)} + \frac{4\eta_{\boldsymbol{v}}^2\ell B_3}{\eta_{\boldsymbol{v}}-c/\ell}
\end{aligned}
\tag{46}
$$

Letting $c = \frac{\eta_{\boldsymbol{v}}\ell}{2}$, we have

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq & \frac{8\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}(T+1)} + 16\ell(\eta_{\boldsymbol{v}}L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
& + \frac{16\ell\widehat{\Delta}_0}{(T+1)} + 16B_2 + 8\eta_{\boldsymbol{v}}\ell B_3
\end{aligned}
\tag{47}
$$

Letting $B = \frac{D_{\mathcal{A}}}{2}\sqrt{\frac{1}{\eta_{\boldsymbol{v}}\eta_\alpha LB_1}}$, we have

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[||\nabla\Phi_{1/2\ell}(\boldsymbol{v}_t)||^2] \leq & \frac{8\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}(T+1)} + 16\ell(\eta_{\boldsymbol{v}}L(B+1)B_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
& + \frac{16\ell\widehat{\Delta}_0}{(T+1)} + 16B_2 + 8\eta_{\boldsymbol{v}}\ell B_3 \\
\leq & \frac{8\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}(T+1)} + 16\ell(2\eta_{\boldsymbol{v}}LBB_1 + \frac{D_{\mathcal{A}}^2}{2B\eta_\alpha} + \eta_\alpha\sigma^2) \\
& + \frac{16\ell\widehat{\Delta}_0}{(T+1)} + 16B_2 + 8\eta_{\boldsymbol{v}}\ell B_3 \\
\leq & \frac{8\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}(T+1)} + 32\ell D_{\mathcal{A}}\sqrt{\frac{\eta_{\boldsymbol{v}}LB_1}{\eta_\alpha}} + 16\ell\eta_\alpha\sigma^2 \\
& + \frac{16\ell\widehat{\Delta}_0}{(T+1)} + 16B_2 + 8\eta_{\boldsymbol{v}}\ell B_3
\end{aligned}
\tag{48}
$$

For $16\ell\eta_\alpha\sigma^2 \leq \frac{\epsilon^2}{8}$, we get $\eta_\alpha = \min\{\frac{1}{2\ell}, \frac{\epsilon^2}{128\ell\sigma^2}\}$. For $32\ell D_{\mathcal{A}}\sqrt{\frac{\eta_{\boldsymbol{v}} LB_1}{\eta_\alpha}} \leq \frac{\epsilon^2}{8}$ and $8\eta_{\boldsymbol{v}}\ell B_3 \leq \frac{\epsilon^2}{8}$, we get $\eta_{\boldsymbol{v}} = \min\{\frac{\epsilon^2}{64\ell B_3}, \frac{\eta_\alpha\epsilon^4}{65536\ell^2 D_{\mathcal{A}}{}^2 LB_1}\}$. For $16B_2 \leq \frac{\epsilon^2}{8}$, we get $\mu = \frac{\epsilon}{8\ell(d+3)^{\frac{3}{2}}}$ and $m = 128\frac{4(d+4)L^2+\mu^2\ell^2(d+6)^3+2\sigma^2}{\epsilon^2}$. Finally, we can get $T = \max\{\frac{32\widehat{\Delta}_\Phi}{\eta_{\boldsymbol{v}}\epsilon^2}, \frac{64\ell\widehat{\Delta}_0}{\epsilon^2}\}$. $\square$