Supplementary material

A Specifications of Experiments

A.1 Hyper-representation

The hyper-representation problem follows the same problem setup as in [58], and is given by

$$\min_{\lambda} L_\nu(\lambda) := \frac{1}{|S_\nu|} \sum_{(x_i,y_i)\in S_\nu} L_{CE}(w^*(\lambda) f(\lambda; x_i), y_i)$$

$$s.t.\ \ w^*(\lambda) = \arg\min_{w} L_{in}(\lambda, w), \quad L_{in}(\lambda, w) := \frac{1}{|S_\tau|} \sum_{(\tau,y_i)\in S_\tau} L_{CE}(w f(\lambda, x_i), y_i),$$

where L_{CE} denotes the cross-entropy loss, S_ν and S_τ denote the training data and the validation data, and $f(\lambda; x_i)$ donates the features extracted from the data x_i. We perform the hyper-representation with the 7-layer LeNet network to solve the bilevel problem. We take the last two layers as the lower-level parameters w, and all remaining layers as the upper-level parameters λ. In our experiments, the dimension of w is 850, and the dimension of λ is 60856. For the choices of hyperparameters in Figure 1, all η in eq. (6), eq. (7), and eq. (10) are all chosen as 0.9. The stepsize β_t for the lower-level update is chosen as 0.8 and the stepsize α_t for the upper-level update is chosen as 0.008. The stepsize λ_t is 0.05 and all the δ_ϵ used are 0.1. The batch size is 256 for both lower- and upper-level processes and the number of lower iterations is 1. For PZOBO-S, F^2SA, and F^3SA, we use the hyperparameter configurations suggested in their papers and implementations. We run the experiment 3 times with different seeds where the solid lines show the average accuracy or loss and the transparent area indicates the variance filled by the max and min values. We run the comparison algorithms using their repositories. The repository of PZOBO-S is available at https://github.com/sowmaster/esjacobians.

A.2 Data Hyper-Cleaning

The formulation of the data hyper-cleaning problem is given as follows:

$$\min_{\lambda} L_\nu(\lambda, w^*) = \frac{1}{|S_\nu|} \sum_{(x_i,y_i)\in S_\nu} L_{CE}((w^*)^T x_i, y_i)$$

$$s.t. \quad w^* = \arg\min_{w} L(\lambda, w) := \frac{1}{|S_\tau|} \sum_{(x_i,y_i)\in S_\tau} \sigma(\lambda_i) L_{CE}(w^T x_i, y_i) + C\|w\|^2,$$

where L_{CE} denotes the cross-entropy loss, S_ν and S_τ denote the training data and the validation data, whose sizes are set to 20000 and 5000, respectively, $\lambda = \{\lambda_i\}_{i\in S_\tau}$ and C are the regularization parameters, and $\sigma(\cdot)$ is the sigmoid function. In the experiments, we set $C = 0.001$ and use 10000 images for testing. In addition, we set the number of iterations in solving the linear system to 3 and set $\eta = 0.5$ in the hypergradient estimation. For our FMBO, we set both inner stepsize (which is the stepsize to update w) and outer stepsize (which is the stepsize to update λ) to 0.3 and set the batchsize to 100. We also use a decaying η for more stable training. Following the hyperparameter configurations suggested in the codes and papers of other comparison methods, we set their batchsizes to 1000 for MRBO, stocBiO and set the batchsize to 500 for VRBO. We also set the period number to 3 for VRBO. We set the number of inner-loop iterations to 200 for stocBiO, AID-FP and BSA, and to 20 for VRBO, to achieve the best performance. Furthermore, we choose 0.1 as the both inner and outer stepsize for all algorithms except FMBO and SUSTAIN. For SUSTAIN, we set the inner stepsize to 0.03 and the outer stepsize to 0.1 for stability. We run the experiments at two noise rates of 0.1 and 0.15. All results are repeated with 5 random seeds and we use Macbook Pro with a 2.3 GHz Quad-Core Intel Core i5 CPU for training without the requirement of GPU. However, our code also supports GPU.

Technical Proofs

Proof outline. Recall that our proposed FMBO algorithm is a simplification of the proposed Hessian/Jacobian-free FdeHBO method, which uses the Hessian matrix and Jacobian matrix vector

directly without finite-difference approximation. For this reason, we first provide auxiliary lemmas for proving the main theorems in Appendix B. Next, we start the proof of FMBO in Appendix C, and then extend the analysis to the more complex FdeHBO in Appendix D.

B Proofs of Preliminary Lemmas

Lemma 1 (Boundedness of v^*)**.** *Under Assumptions 1 and 2, we have for v^* in eq.* (5)*, $\|v^*\|^2 \leq \frac{C_{f_y}}{\mu_g^2}$.*

Proof. From eq. (5), we have

$$\|v^*\|^2 = \|\big[\nabla^2_{yy} g(x,y)\big]^{-1} \nabla_y f(x,y)\|^2 \leq \|\big[\nabla^2_{yy} g(x,y)\big]^{-1}\|^2 \|\nabla_y f(x,y)\|^2 \overset{(a)}{\leq} \frac{C_{f_y}}{\mu_g^2},$$

where (a) follows Assumptions 1 and 2. Then, the proof is complete. □

Lemma 2. *Under Assumptions 1 and 2, for any $\|v\| \leq r_v$, we have that $\bar{\nabla} f(x,y,v)$ is Lipschitz continuous w.r.t. $(x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ with constant L_F given by*

$$L_F^2 = 2\Big(L_{f_x}^2 + L_{g_{xy}}^2 r_v^2\Big).$$

Proof. Note from eq. (3) that $\bar{\nabla} f(x,y,v) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) v$. Then, we have

$$\begin{aligned}
&\|\bar{\nabla} f(x_1,y_1,v) - \bar{\nabla} f(x_2,y_2,v)\|^2 \\
&\qquad = \|\big[\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\big] - \big[\nabla^2_{xy} g(x_1,y_1) - \nabla^2_{xy} g(x_2,y_2)\big] v\|^2 \\
&\qquad \overset{(a)}{\leq} 2\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\|^2 + 2\|\big[\nabla^2_{xy} g(x_1,y_1) - \nabla^2_{xy} g(x_2,y_2)\big] v\|^2 \\
&\qquad \overset{(b)}{\leq} 2\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\|^2 + 2r_v^2 \|\nabla^2_{xy} g(x_1,y_1) - \nabla^2_{xy} g(x_2,y_2)\|^2 \\
&\qquad \overset{(c)}{\leq} 2\Big(L_{f_x}^2 + L_{g_{xy}}^2 r_v^2\Big)\big(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2\big) \\
&\qquad = L_F^2\big(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2\big)
\end{aligned}$$

where (a) uses Cauchy–Schwartz inequality; (b) follows from $\|v\| \leq r_v$; (c) follows from Assumption 1 and Assumption 2. Then the proof is complete. □

Lemma 3. *Under Assumption 2 and 3, the estimation $\bar{\nabla} f(x,y,v;\xi)$ is unbiased. And the gradient estimate of the upper-level objective satisfies*

$$\mathbb{E}_{\bar{\xi}} \|\bar{\nabla} f(x,y,v;\bar{\xi}) - \bar{\nabla} f(x,y,v)\|^2 \leq \sigma_f^2,$$

where $\sigma_f^2 = 2(\sigma_{f_x}^2 + r_v^2 \sigma_{g_{xy}}^2)$.

Proof. Based on the definition of $\bar{\nabla} f(x,y,v;\xi)$ in eq. (3), we have

$$\begin{aligned}
&\mathbb{E}\left[\bar{\nabla} f(x,y,v;\xi)\right] - \bar{\nabla} f(x,y,v) \\
&\quad = \mathbb{E}\left[\nabla_x f(x,y;\xi) - \nabla_x \nabla_y g(x,y;\xi) v\right] - \left[\nabla_x f(x,y) - \nabla_x \nabla_y g(x,y) v\right] \\
&\quad = \mathbb{E}\left[\nabla_x f(x,y;\xi) - \nabla_x f(x,y)\right] - \mathbb{E}\left[\nabla_y g(x,y;\xi) - \nabla_y g(x,y)\right] v \\
&\quad = 0.
\end{aligned}$$

And

$$\begin{aligned}
&\mathbb{E}\|\bar{\nabla} f(x,y,v;\xi) - \bar{\nabla} f(x,y,v)\|^2 \\
&\qquad = \mathbb{E}\|\left[\nabla_x f(x,y;\xi) - \nabla_x \nabla_y g(x,y;\xi) v\right] - \left[\nabla_x f(x,y) - \nabla_x \nabla_y g(x,y) v\right]\|^2 \\
&\qquad \leq 2\mathbb{E}\|\nabla_x f(x,y;\xi) - \nabla_x f(x,y)\|^2 + 2\mathbb{E}\|\nabla_x \nabla_y g(x,y;\xi) v - \nabla_x \nabla_y g(x,y) v\|^2 \\
&\qquad \overset{(a)}{\leq} 2\mathbb{E}\|\nabla_x f(x,y;\xi) - \nabla_x f(x,y)\|^2 + 2r_v^2 \mathbb{E}\|\nabla_x \nabla_y g(x,y;\xi) - \nabla_x \nabla_y g(x,y)\|^2
\end{aligned}$$

$$\overset{(b)}{\leq} 2\sigma_{f_x}^2 + 2r_v^2\sigma_{g_{xy}}^2.$$

where (a) follows from step 6 in algorithm 2, (b) follows from Assumption 3. Then, the proof is complete. □

Lemma 4 ([19] lemma 2.2)**.** *Under Assumptions 1, 2 and 3, we have, for all* $x, x_1, x_2 \in \mathbb{R}^{d_y}$ *and* $y \in \mathbb{R}^{d_y}$,

$$\begin{aligned}\|\bar{\nabla} f(x,y,v) - \nabla\Phi(x)\| &\leq L\|y^*(x) - y\| \\ \|y^*(x_1) - y^*(x_2)\| &\leq L_y\|x_1 - x_2\| \\ \|\nabla\Phi(x_1) - \nabla\Phi(x_2)\| &\leq L_f\|x_1 - x_2\|,\end{aligned}$$

where the Lipschitz constants are given by

$$L = L_{f_x} + \frac{L_{f_y}C_{g_{xy}}}{\mu_g} + C_y\left(\frac{L_{g_{xy}}}{\mu_g} + \frac{L_{g_{yy}}C_{g_{xy}}}{\mu_g^2}\right), \quad L_f = L + \frac{LC_{g_{xy}}}{\mu_g}, \quad L_y = \frac{C_{g_{xy}}}{\mu_g}.$$

Lemma 5. *Under Assumption 2, for any* ξ *and* ψ*, define* $e_t^J := \widetilde{J}(x_t, y_t, v_t, \delta_\epsilon; \xi) - \nabla_{xy}^2 g(x_t, y_t; \xi)v_t$*,* $e_t^H := \widetilde{H}(x_t, y_t, v_t, \delta_\epsilon; \psi) - \nabla_{yy}^2 g(x_t, y_t; \psi)v_t$*, then we have the bound of* e_t^J *and* e_t^H *that*

$$\|e_t^J\|^2 \leq C_J^2\delta_\epsilon^2, \quad \|e_t^H\|^2 \leq C_H^2\delta_\epsilon^2,$$

where $C_J := L_{g_{xy}}r_v^2$*,* $C_H := L_{g_{yy}}r_v^2$*.*

Proof. According to definition of $\widetilde{H}(x_t, y_t, v_t, \delta_\epsilon; \psi)$ in eq. (9), we have

$$\begin{aligned}\|e_t^H\| &= \|\widetilde{H}(x_t, y_t, v_t, \delta_\epsilon; \psi) - \nabla_{yy}^2 g(x_t, y_t; \psi)v_t\| \\ &= \frac{1}{2\delta_\epsilon}\|\nabla_y g(x_t, y_t + \delta_\epsilon v; \psi) - \nabla_y g(x_t, y_t; \psi) - \nabla_{yy}^2 g(x_t, y_t; \psi)(\delta_\epsilon v_t)\| \\ &\quad + \frac{1}{2\delta_\epsilon}\|\nabla_y g(x_t, y_t; \psi) - \nabla_y g(x_t, y_t - \delta_\epsilon v; \psi) - \nabla_{yy}^2 g(x_t, y_t; \psi)(\delta_\epsilon v_t)\| \\ &\overset{(a)}{\leq} L_{g_{yy}}\delta_\epsilon\|v_t\|^2 \leq L_{g_{yy}}r_v^2\delta_\epsilon.\end{aligned}$$

where (a) uses the Lemma A.1 in [1]. Similarly, according to definition of $\widetilde{J}(x_t, y_t, v_t, \delta_\epsilon; \psi)$ in eq. (11), we have

$$\begin{aligned}\|e_t^J\| &= \|\widetilde{J}(x_t, y_t, v_t, \delta_\epsilon; \psi) - \nabla_{xy}^2 g(x_t, y_t; \psi)v_t\| \\ &= \frac{1}{2\delta_\epsilon}\|\nabla_x g(x_t, y_t + \delta_\epsilon v; \psi) - \nabla_x g(x_t, y_t; \psi) - \nabla_{xy}^2 g(x_t, y_t; \psi)(\delta_\epsilon v_t)\| \\ &\quad + \frac{1}{2\delta_\epsilon}\|\nabla_x g(x_t, y_t; \psi) - \nabla_x g(x_t, y_t - \delta_\epsilon v; \psi) - \nabla_{xy}^2 g(x_t, y_t; \psi)(\delta_\epsilon v_t)\| \\ &\overset{(a)}{\leq} L_{g_{xy}}\delta_\epsilon\|v_t\|^2 \leq L_{g_{xy}}r_v^2\delta_\epsilon.\end{aligned}$$

where (a) uses the Lemma A.1 in [1].

□

C Proof of Theorem 2 (FMBO without first order approximation)

C.1 Some existential lemmas

The following lemmas provide characterizations on descents of (1) function value (by Lemma 6); (2) iterates of the lower level problem (by Lemma 7); and (3) gradient estimation error of the outer-level function (by Lemma 8).

Lemma 6. *For non-convex and smooth $\Phi(\cdot)$,with e_t^f defined as: $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t, v_t)$, the consecutive iterates of Algorithm 2 satisfy:*

$$\mathbb{E}\big[\Phi(x_{t+1})\big] \leq \mathbb{E}\big[\Phi(x_t) - \frac{\alpha_t}{2}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha_t}{2}(1-\alpha_t L_f)\|h_t^f\|^2 + \alpha_t\|e_t^f\|^2 + 4\alpha_t\Big(L_{f_x}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\|y_t - y_t^*\|^2 + 2\alpha_t C_{g_{xy}}\|v_t - v_t^*\|^2\big],$$

for all $t \in \{0, 1, ..., T-1\}$.

Proof. Using the Lipschitz smoothness of the objective function from Lemma 4, we have

$$\begin{aligned}\Phi(x_{t+1}) &\leq \Phi(x_t) + \langle\nabla\Phi(x_t), x_{t+1} - x_t\rangle + \frac{L_f}{2}\|x_{t+1} - x_t\|^2 \\ &= \Phi(x_t) + \langle\nabla\Phi(x_t), h_t^f\rangle + \frac{L_f}{2}\|h_t^f\|^2 \\ &= \Phi(x_t) - \frac{\alpha_t}{2}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha_t}{2}(1-\alpha_t L_f)\|h_t^f\|^2 + \frac{\alpha_t}{2}\|h_t^f - \nabla\Phi(x_t)\|^2, \end{aligned} \tag{15}$$

where the last term of the right-hand side of eq. (15) can be show as

$$\begin{aligned}&\|h_t^f - \nabla\Phi(x_t)\|^2 \\ &= \|h_t^f - \bar{\nabla} f(x_t, y_t, v_t) + \bar{\nabla} f(x_t, y_t, v_t) - \nabla\Phi(x_t)\|^2 \\ &\leq 2\|e_t^f\|^2 + 2\|\bar{\nabla} f(x_t, y_t, v_t) - \bar{\nabla} f(x_t, y_t^*, v_t^*)\|^2 \\ &\leq 2\|e_t^f\|^2 + 4\|\bar{\nabla} f(x_t, y_t, v_t) - \bar{\nabla} f(x_t, y_t, v_t^*)\|^2 + 4\|\bar{\nabla} f(x_t, y_t, v_t^*) - \bar{\nabla} f(x_t, y_t, v_t^*)\|^2 \\ &\overset{(a)}{\leq} 2\|e_t^f\|^2 + 4C_{g_{xy}}\|v_t - v_t^*\|^2 + 8\Big(L_{f_x}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\|y_t - y_t^*\|^2. \end{aligned} \tag{16}$$

By applying eq. (16) to eq. (15), we have

$$\Phi(x_{t+1}) \leq \Phi(x_t) - \frac{\alpha_t}{2}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha_t}{2}(1-\alpha_t L_f)\|h_t^f\|^2 + \alpha_t\|e_t^f\|^2 + 4\alpha_t\Big(L_{f_x}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\|y_t - y_t^*\|^2 + 2\alpha_t C_{g_{xy}}\|v_t - v_t^*\|^2, \tag{17}$$

then we take the expectation of both sides and the proof is complete. □

Lemma 7 ([34] Lemma C.2)**.** *Define $e_t^g := h_t^g - \nabla_y g(x_t, y_t)$. Then the iterates of the inner problem generated by Algorithm 2 satisfy*

$$\begin{aligned}&\mathbb{E}\|y_{t+1} - y_{t+1}^*\|^2 \\ &\leq (1+\gamma_t)(1+\delta_t)\left(1 - 2\beta_t\frac{\mu_g L_g}{\mu_g + L_g}\right)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right)L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\ &\quad - (1+\gamma_t)(1+\delta_t)\left(\frac{2\beta_t}{\mu_g + L_g} - \beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 + (1+\gamma_t)(1+\frac{1}{\delta_t})\beta_t^2\mathbb{E}\|e_t^g\|^2 \end{aligned}$$

for all $t \in \{0, ..., T-1\}$ with some $\gamma_t, \delta_t > 0$.

C.2 Descent in the gradient estimation error of the upper function

Lemma 8. *Define $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t, v_t)$. Under Lemma 3, the iterations of the outer problem generated by Algorithm 2 satisfy*

$$\begin{aligned}\mathbb{E}\|e_{t+1}^f\|^2 \leq{}& (1-\eta_{t+1}^f)^2\mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2\sigma_f^2 \\ &+ 6(1-\eta_{t+1}^f)^2\Bigg[L_F^2\Big(\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\big(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2\big)\Big) \\ &+ 2C_{g_{xy}}\lambda_t^2\big(\mathbb{E}\|e_t^R\|^2 + L_g^2\mathbb{E}\|v_t - v_t^*\|^2\big)\Bigg]\end{aligned}$$

for all $t \in \{0, ..., T-1\}$ with L_F in Lemma 2.

Proof. From the definition of e_t^f, we have

$$
\begin{aligned}
&\mathbb{E}\|e_{t+1}^f\|^2 \\
&= \|h_{t+1}^f - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|\eta_{t+1}^f \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) + (1-\eta_{t+1}^f)\big(h_t^f + \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) \\
&\qquad - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})\big) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\overset{(b)}{=} \mathbb{E}\Big\|(1-\eta_{t+1}^f)e_t^f + \eta_{t+1}^f\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad + (1-\eta_{t+1}^f)\Big(\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\Big)\Big\|^2 \\
&\overset{(c)}{=} (1-\eta_{t+1}^f)^2\mathbb{E}\|e_t^f\|^2 + \mathbb{E}\|\eta_{t+1}^f\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\quad + (1-\eta_{t+1}^f)\Big(\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\Big)\|^2 \\
&\leq (1-\eta_{t+1}^f)^2\mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\quad + 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\|^2 \\
&\overset{(d)}{\leq} (1-\eta_{t+1}^f)^2\mathbb{E}\|e_t^f\|^2 + 2(\eta_{t+1}^f)^2\sigma_f^2 \\
&\quad + 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\|^2
\end{aligned} \tag{18}
$$

where (a) uses the definition of h_{t+1}^f in eq. (13), (b) uses the definition that $e_t^f := h_t^f - \bar{\nabla} f(x_t, y_t, v_t)$, (c) follows because for $\Sigma_{t+1} = \sigma\{y_0, x_0, v_0, ..., y_t, x_t, v_t, y_{t+1}, x_{t+1}, v_{t+1}\}$,

$$
\begin{aligned}
&\mathbb{E}\Big\langle e_t^f, \big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}) \\
&\qquad - (1-\eta_{t+1}^f)\big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\Big\rangle \\
&\quad = \mathbb{E}\Big\langle e_t^f, \mathbb{E}\Big[\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}) \\
&\qquad\qquad - (1-\eta_{t+1}^f)\big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)|\Sigma_{t+1}\Big]\Big\rangle = 0,
\end{aligned} \tag{19}
$$

which follows from the fact that the second term in the inner product of eq. (19) is zero mean as a result of Assumption 3, (d) follows from Assumption 3.

Next, we bound the last term of eq. (18)

$$
\begin{aligned}
&2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\|^2 \\
&\overset{(a)}{\leq} 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})\|^2 \\
&\leq 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_{t+1}, v_{t+1}; \xi_{t+1})\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_t, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_{t+1}; \xi_{t+1})\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_t, y_t, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})\|^2 \\
&\overset{(b)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|x_{t+1} - x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|y_{t+1} - y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\nabla_{xy}^2 g(x_t, y_t)(v_{t+1} - v_t)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2 \mathbb{E}\|x_{t+1}-x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2 \mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2 C_{g_{xy}} \mathbb{E}\|v_{t+1}-v_t\|^2 \\
&\overset{(d)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2 \mathbb{E}\|x_{t+1}-x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2 \mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2 C_{g_{xy}} \mathbb{E}\|w_{t+1}-v_t\|^2 \\
&\overset{(e)}{\leq} 6(1-\eta_{t+1}^f)^2 \Big[L_F^2 \big(\alpha_t^2 \mathbb{E}\|h_t^f\|^2 + \beta_t^2 \mathbb{E}\|h_t^g\|^2 \big) + C_{g_{xy}} \lambda_t^2 \mathbb{E}\|h_t^R\|^2 \Big] \\
&\overset{(f)}{\leq} 6(1-\eta_{t+1}^f)^2 \Big[L_F^2 \big(\alpha_t^2 \mathbb{E}\|h_t^f\|^2 + \beta_t^2 \mathbb{E}\|h_t^g\|^2 \big) + 2C_{g_{xy}} \lambda_t^2 \big(\mathbb{E}\|e_t^R\|^2 + \mathbb{E}\|\nabla R_v(x_t,y_t,v_t)\|^2 \big) \Big] \\
&\overset{(g)}{\leq} 6(1-\eta_{t+1}^f)^2 \Big[L_F^2 \big(\alpha_t^2 \mathbb{E}\|h_t^f\|^2 + \beta_t^2 \mathbb{E}\|h_t^g\|^2 \big) + 2C_{g_{xy}} \lambda_t^2 \big(\mathbb{E}\|e_t^R\|^2 + L_g^2 \mathbb{E}\|v_t - v_t^*\|^2 \big) \Big] \\
&\overset{(h)}{\leq} 6(1-\eta_{t+1}^f)^2 \Bigg[L_F^2 \Big(\alpha_t^2 \mathbb{E}\|h_t^f\|^2 + \beta_t^2 \big(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 \big) \Big) \\
&\quad + 2C_{g_{xy}} \lambda_t^2 \big(\mathbb{E}\|e_t^R\|^2 + L_g^2 \mathbb{E}\|v_t - v_t^*\|^2 \big) \Bigg], \qquad (20)
\end{aligned}
$$

where (a) follows from the mean-variance inequality: For a random variable Z we have $\mathbb{E}\|Z - \mathbb{E}[Z]\|^2 \leq \mathbb{E}\|Z\|^2$ with Z defined as $Z := \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})$, (b) follows from Lemma 2 and eq. (5), (c) uses Assumption 2, (d) follows from the nonexpansiveness of projection that $\|v_{t+1} - v_t^*\| \leq \|\text{Proj}_B(w_{t+1}) - \text{Proj}_B(v_t^*)\| \leq \|w_{t+1} - v_t^*\|$ in convex ball $B(0, r_v^2)$, (e) uses the definition of h_t^f in eq. (13), h_t^g in eq. (6) and h_t^R in eq. (12), (f) follows from the definition that $e_t^R := h_t^R - \nabla_v R(x_t, y_t, v_t)$, (g) uses that result the

$$
\begin{aligned}
\mathbb{E}\|\nabla_v R(x_t,y_t,v_t)\|^2 &= \mathbb{E}\|\nabla_{yy}^2 g(x_t,y_t) v_t - \nabla_y f(x_t,y_t)\|^2 \\
&= \mathbb{E}\|\nabla_{yy}^2 g(x_t,y_t)(v_t - v_t^*)\|^2 \leq L_g^2 \mathbb{E}\|v_t - v_t^*\|^2. \qquad (21)
\end{aligned}
$$

(h) uses the definition that $e_t^g := h_t^g - \nabla_y g(x_t, y_t)$. Finally, substituting eq. (20) in eq. (18), we have the statement in the lemma.

Then, the proof is complete. □

C.3 Descent in the gradient estimation error of the inner function

Lemma 9. *Define $e_t^g := h_t^g - \nabla_y g(x_t, y_t)$. Under Assumption 2 and 3, the iterates generated from Algorithm 2 satisfy*

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 \leq &\big((1-\eta_{t+1}^g)^2 \mathbb{E}\|e_t^g\|^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2 \big) \mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2 \sigma_g^2 \\
&+ 16(1-\eta_{t+1}^g)^2 L_g^2 \alpha_t^2 \mathbb{E}\|h_t^f\|^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2 \mathbb{E}\|\nabla_y g(x_t,y_t)\|^2
\end{aligned}
$$

for all $t \in \{0, 1, ..., T-1\}$.

Proof. From the definition of the e_t^g we have

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 &= \mathbb{E}\|h_{t+1}^g - \nabla_y g(x_{t+1}, y_{t+1})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|\nabla_y g(x_{t+1},y_{t+1};\zeta_{t+1}) + (1-\eta_{t+1}^g)^2 (h_t^g - \nabla_y g(x_t,y_t;\zeta_{t+1})) - \nabla_y g(x_{t+1},y_{t+1})\|^2 \\
&\overset{(b)}{=} \mathbb{E}\|(1-\eta_{t+1}^g) e_t^g + (\nabla_y g(x_{t+1},y_{t+1};\zeta_{t+1}) - \nabla_y g(x_{t+1},y_{t+1})) \\
&\qquad - (1-\eta_{t+1}^g)(\nabla_y g(x_t,y_t;\zeta_{t+1})) - \nabla_y g(x_t,y_t)\|^2 \\
&\overset{(c)}{=} (1-\eta_{t+1}^g)^2 \mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|(\nabla_y g(x_{t+1},y_{t+1};\zeta_{t+1}) - \nabla_y g(x_{t+1},y_{t+1})) \\
&\quad - (1-\eta_{t+1}^g)(\nabla_y g(x_t,y_t;\zeta_{t+1})) - \nabla_y g(x_t,y_t)\|^2 \\
&\overset{(d)}{\leq} (1-\eta_{t+1}^g)^2 \mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2 \sigma_g^2 \\
&\quad + 2(1-\eta_{t+1}^g)^2 \mathbb{E}\|g(x_{t+1},y_{t+1};\zeta_{t+1}) - g(x_{t+1},y_{t+1}) - g(x_t,y_t;\zeta_{t+1}) + g(x_t,y_t)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(e)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 4(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_{t+1}) - g(x_t,y_t)\|^2 \\
&\quad + 4(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_{t+1};\zeta_{t+1}) - g(x_t,y_t;\zeta_{t+1})\|^2 \\
&\overset{(f)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 8(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_{t+1}) - g(x_{t+1},y_t)\|^2 \\
&\quad + 8(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_t) - g(x_t,y_t)\|^2 \\
&\quad + 8(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_{t+1};\zeta_{t+1}) - g(x_{t+1},y_t;\zeta_{t+1})\|^2 \\
&\quad + 8(1-\eta_{t+1}^g)^2\mathbb{E}\|g(x_{t+1},y_t;\zeta_{t+1}) - g(x_t,y_t;\zeta_{t+1})\|^2 \\
&\overset{(g)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 \\
&\quad + 16(1-\eta_{t+1}^g)^2L_g^2\mathbb{E}\|x_{t+1}-x_t\|^2 + 16(1-\eta_{t+1}^g)^2L_g^2\mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\overset{(h)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 \\
&\quad + 16(1-\eta_{t+1}^g)^2L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 16(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\mathbb{E}\|h_t^g\|^2 \\
&\overset{(i)}{\leq} (1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 16(1-\eta_{t+1}^g)^2L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\
&\quad + 32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 \\
&= (1-\eta_{t+1}^g)^2(1+32L_g^2\beta_t^2)\mathbb{E}\|e_t^g\|^2 + 2(\eta_{t+1}^g)^2\sigma_g^2 + 16(1-\eta_{t+1}^g)^2L_g^2\alpha_t^2\mathbb{E}\|h_t^f\|^2 \\
&\quad + 32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2,
\end{aligned}
$$

where (a) uses the definition of h_t^g in eq. (6), (b) uses the definition of e_t^g, (c) follows because for $\Sigma_{t+1} = \sigma\{y_0, x_0, ..., y_t, x_t, y_{t+1}, x_{t+1}\}$,

$$
\begin{aligned}
&\mathbb{E}\Big\langle e_t^g, \big(\nabla_y g(x_{t+1},y_{t+1};\zeta_{t+1}) - \nabla_y g(x_{t+1},y_{t+1}) \\
&\qquad\qquad - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t,y_t;\zeta_{t+1}) - \nabla_y g(x_t,y_t)\big)\Big\rangle \\
&= \mathbb{E}\Big\langle e_t^g, \mathbb{E}\big[\big(\nabla_y g(x_{t+1},y_{t+1};\zeta_{t+1}) - \nabla_y g(x_{t+1},y_{t+1})\big) \\
&\qquad\qquad - (1-\eta_{t+1}^g)\big(\nabla_y g(x_t,y_t;\zeta_{t+1}) - \nabla_y g(x_t,y_t)\big)|\Sigma_{t+1}\big]\Big\rangle = 0,
\end{aligned}
$$

where (d) follows from Cauchy–Schwartz inequality and Assumption 3, (e) and (f) use Cauchy–Schwartz inequality, (g) follows from Assumption 2; (h) follows from Steps 4 and 7 in Algorithm 2; (i) follows from the definition $e_t^g := h_t^g - \nabla_y g(x_t, y_t)$. □

C.4 Descent in the gradient estimation error of the R function

Lemma 10. *Define $e_t^R := h_t^R - \nabla_v R(x_t, y_t, v_t)$. Under Assumption 1, 2, 3, the iterates generated by Algorithm 2 satisfy*

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^R\|^2 \leq\ & (1-\eta_{t+1}^R)^2\left(1+48L_g^2\lambda_t^2\right)\mathbb{E}\|e_t^R\|^2 + 4(\eta_{t+1}^R)^2\left(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2\right) \\
&+ 48(1-\eta_{t+1}^R)^2\left(L_{g_{yy}}^2 r_v^2 + L_{f_y}^2\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t,y_t)\|^2)\right] \\
&+ 48(1-\eta_{t+1}^R)^2L_g^4\lambda_t^2\mathbb{E}\|v_t - v_t^*\|^2
\end{aligned}
$$

for all $t \in \{0, 1, ..., T-1\}$.

Proof. For the gradient estimation error of the R function, we have

$$
\begin{aligned}
&\mathbb{E}\|e_{t+1}^R\|^2 \\
&\quad = \mathbb{E}\|h_{t+1}^R - \nabla_v R(x_t,y_t,v_t)\|^2 \\
&\quad \overset{(a)}{=} \mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) + (1-\eta_{t+1}^R)h_t^R - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_{t+1};\psi_{t+1}) \\
&\qquad - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(b)}{=} \mathbb{E}\|(1-\eta_{t+1}^R)e_t^R + (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) \\
&\quad + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) \\
&\quad - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\overset{(c)}{=} (1-\eta_{t+1}^R)^2\mathbb{E}\|e_t^R\|^2 + \mathbb{E}\|(1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) \\
&\quad + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \qquad (22)
\end{aligned}
$$

where (a) follows from eq. (12), (b) uses the definition of $e_t^R := h_t^R - \nabla_v R(x_t,y_t,v_t)$, (c) follows from the fact that

$$
\begin{aligned}
&\mathbb{E}\Big\langle e_t^R, \big(\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) \\
&\qquad\qquad - (1-\eta_{t+1}^R)\big(\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) - \nabla_v R(x_t,y_t,v_t)\big)\Big\rangle \\
&= \mathbb{E}\Big\langle e_t^R, \mathbb{E}\big[\big(\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla R(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\qquad\qquad - (1-\eta_{t+1}^R)\big(\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) - \nabla_v R(x_t,y_t,v_t)\big)|\Sigma_{t+1}\big]\Big\rangle = 0
\end{aligned}
$$

for $\Sigma_{t+1} = \sigma\{y_0, v_0, x_0, ..., y_t, v_t, x_t, y_{t+1}, v_{t+1}, x_{t+1}\}$. For the second part of the right-hand side of eq. (22),

$$
\begin{aligned}
&\mathbb{E}\|(1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) \\
&\quad + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\qquad \overset{(a)}{\leq} 2(\eta_{t+1}^R)^2\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\qquad\quad + 2(1-\eta_{t+1}^R)^2\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) \\
&\qquad\qquad - \nabla_v R(x_t,y_t,v_t;\psi_{t+1}) + \nabla_v R(x_t,y_t,v_t)\|^2 \qquad (23)
\end{aligned}
$$

where (a) uses Cauchy–Schwartz inequality. For the first term of the right-hand side of eq. (23), we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\Big\| \left[\nabla_{yy}^2 g(x_{t+1},y_{t+1};\psi_{t+1}) - \nabla_{yy}^2 g(x_{t+1},y_{t+1})\right] v_{t+1} \\
&\qquad\qquad - \left[\nabla_y f(x_{t+1},y_{t+1};\psi_{t+1}) - \nabla_y f(x_{t+1},y_{t+1})\right]\Big\|^2 \\
&\leq 2\mathbb{E}\| \left[\nabla_{yy}^2 g(x_{t+1},y_{t+1};\psi_{t+1}) - \nabla_{yy}^2 g(x_{t+1},y_{t+1})\right] v_{t+1}\|^2 \\
&\quad + 2\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1};\psi_{t+1}) - \nabla_y f(x_{t+1},y_{t+1})\|^2 \\
&\overset{(b)}{\leq} 2\Big(r_v^2\sigma_{g_{yy}}^2 + \sigma_{f_y}^2\Big) \qquad (24)
\end{aligned}
$$

where (a) follows the definition of $R(x_t,y_t,v_t)$ in eq. (4), (b) follows from the boundedness of v_t that $\|v_t\| \leq r_v$ (see line 6 in Algorithm 2), Lemma 1 and Assumption 3. Moreover, for the second part of eq. (23), we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) \\
&\qquad - \nabla_v R(x_t,y_t,v_t;\psi_{t+1}) + \nabla_v R(x_t,y_t,v_t)\|^2 \\
&\leq 2\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_t,y_t,v_t;\psi_{t+1})\|^2 \\
&\quad + 2\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1}) - \nabla_v R(x_t,y_t,v_t)\|^2. \qquad (25)
\end{aligned}
$$

Next, we upper bound the second term of the right-hand side of eq. (25). In specific, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1}) - \nabla_v R(x_t,y_t,v_t)\|^2 \\
&\overset{(a)}{\leq} 3\mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1}) - \nabla_v R(x_t,y_{t+1},v_{t+1})\|^2 \\
&\quad + 3\mathbb{E}\|\nabla_v R(x_t,y_{t+1},v_{t+1}) - \nabla_v R(x_t,y_t,v_{t+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\mathbb{E}\|\nabla_v R(x_t, y_t, v_{t+1}) - \nabla_v R(x_t, y_t, v_t)\|^2 \\
&\overset{(b)}{=} 3\mathbb{E}\| \left[\nabla^2_{yy} g(x_{t+1}, y_{t+1}) - \nabla^2_{yy} g(x_t, y_{t+1})\right] v_{t+1} - \left[\nabla_y f(x_{t+1}, y_{t+1}) - \nabla_y f(x_t, y_{t+1})\right] \|^2 \\
&\quad + 3\mathbb{E}\| \left[\nabla^2_{yy} g(x_t, y_{t+1}) - \nabla^2_{yy} g(x_t, y_t)\right] v_{t+1} - \left[\nabla_y f(x_t, y_{t+1}) - \nabla_y f(x_t, y_t)\right] \|^2 \\
&\quad + 3\mathbb{E}\|\nabla_v R(x_t, y_t, v_{t+1}) - \nabla_v R(x_t, y_t, v_t)\|^2 \\
&\overset{(c)}{\leq} 6\left(L^2_{g_{yy}}\|v_{t+1}\|^2 + L^2_{f_y}\right)\mathbb{E}\|x_{t+1} - x_t\|^2 \\
&\quad + 6\left(L^2_{g_{yy}}\|v_{t+1}\|^2 + L^2_{f_y}\right)\mathbb{E}\|y_{t+1} - y_t\|^2 + 3L^2_g\lambda_t^2\mathbb{E}\|h_t^R\|^2 \\
&\overset{(d)}{\leq} 6\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left(\mathbb{E}\|x_{t+1} - x_t\|^2 + \mathbb{E}\|y_{t+1} - y_t\|^2\right) + 3L^2_g\lambda_t^2\mathbb{E}\|h_t^R\|^2 \\
&\overset{(e)}{\leq} 6\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left(\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\mathbb{E}\|h_t^g\|^2\right) + 6L^2_g\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2 + \mathbb{E}\|\nabla_v R(x_t, y_t, v_t)\|^2\right) \\
&\overset{(f)}{\leq} 6\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2)\right] \\
&\quad + 6L^2_g\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2 + L^2_g\mathbb{E}\|v_t - v_t^*\|^2\right), \qquad (26)
\end{aligned}
$$

where (a) uses Cauchy–Schwartz inequality, (b) uses the definition of eq. (5), (c) follows from Assumption 1, 2, and Step 5 and 6 in Algorithm 2, (d) uses Lemma 1; (e) follows from Step 4 and 7 in Algorithm 2 and the definition of h_t^R in eq. (12), and (f) follows from the definition of h_t^g in eq. (6) and the fact that

$$
\begin{aligned}
\mathbb{E}\|\nabla_v R(x_t, y_t, v_t)\|^2 &= \mathbb{E}\|\nabla^2_{yy} g(x_t, y_t)v_t - \nabla_y f(x_t, y_t)\|^2 \\
&= \mathbb{E}\|\nabla^2_{yy} g(x_t, y_t)(v_t - v_t^*)\|^2 \leq L^2_g\mathbb{E}\|v_t - v_t^*\|^2.
\end{aligned}
$$

For the first term of the right-hand side of eq. (25), we follow the same steps and get the same upper bound as in eq. (26). Then, incorporating eq. (26) into eq. (25) yields

$$
\begin{aligned}
\mathbb{E}\|\nabla_v R(x_{t+1}, &y_{t+1}, v_{t+1}; \psi_{t+1}) - \nabla_v R(x_{t+1}, y_{t+1}, v_{t+1}) \\
&\quad - \nabla_v R(x_t, y_t, v_t; \psi_{t+1}) + \nabla_v R(x_t, y_t, v_t)\|^2 \\
&\leq 24\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2)\right] \\
&\quad + 24L^2_g\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2 + L^2_g\mathbb{E}\|v_t - v_t^*\|^2\right). \qquad (27)
\end{aligned}
$$

Then, incorporating eq. (24) and eq. (27) into eq. (23), we have

$$
\begin{aligned}
\mathbb{E}\|(1 - \eta_{t+1}^R)&\nabla_v R(x_t, y_t, v_t) - (1 - \eta_{t+1}^R)\nabla_v R(x_t, y_t, v_t; \psi_{t+1}) \\
&+ \nabla_v R(x_{t+1}, y_{t+1}, v_{t+1}; \psi_{t+1}) - \nabla_v R(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\leq 4(\eta_{t+1}^R)^2\left(r_v^2\sigma^2_{g_{yy}} + \sigma^2_{f_y}\right) \\
&\quad + 48(1 - \eta_{t+1}^R)^2\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2)\right] \\
&\quad + 48(1 - \eta_{t+1}^R)^2 L^2_g\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2 + L^2_g\mathbb{E}\|v_t - v_t^*\|^2\right), \qquad (28)
\end{aligned}
$$

which, incorporated into eq. (22), yields

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^R\|^2 \leq{}& (1 - \eta_{t+1}^R)^2\left(1 + 48L^2_g\lambda_t^2\right)\mathbb{E}\|e_t^R\|^2 + 4(\eta_{t+1}^R)^2\left(r_v^2\sigma^2_{g_{yy}} + \sigma^2_{f_y}\right) \\
&+ 48(1 - \eta_{t+1}^R)^2\left(L^2_{g_{yy}} r_v^2 + L^2_{f_y}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2)\right] \\
&+ 48(1 - \eta_{t+1}^R)^2 L^4_g\lambda_t^2\mathbb{E}\|v_t - v_t^*\|^2,
\end{aligned}
$$

which finishes the proof. □

C.5 Descent in iterates of the LS problem

Lemma 11. *Under the Assumption 1, 2, the iterates of the LS problem generated according to Algorithm 2 satisfy*

$$\mathbb{E}\|v_{t+1} - v_{t+1}^*\|^2$$

$$\leq (1+\gamma_t')\,(1+\delta_t')\left[\left(1-2\lambda_t\frac{(L_g+L_g^3)\mu_g}{\mu_g+L_g}+\lambda_t^2L_g^2\right)\mathbb{E}\|v_t-v_t^*\|^2\right]$$
$$+(1+\gamma_t')\left(1+\frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2$$
$$+(1+\frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2}+\frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2+\beta_t^2\left(2\mathbb{E}\|e_t^g\|^2+2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\right].$$

for all $t\in\{0,...,T-1\}$ with some $\gamma_t'>0$ and $\delta_t'>0$.

Proof. Letting $v_k:=v(x_k,y_k)$ and $v_k^*:=v^*(x_k,y_k)$ and choosing the radius $r_v=\frac{C_{f_y}}{\mu_g}$ (see Step 6 in Algorithm 2), we have

$$\begin{aligned}
\mathbb{E}\|v_{t+1}-v_{t+1}^*\|^2 &\overset{(a)}{\leq}(1+\gamma_t')\mathbb{E}\|v_{t+1}-v_t^*\|^2+(1+\frac{1}{\gamma_t'})\mathbb{E}\|v_t^*-v_{t+1}^*\|^2\\
&\overset{(b)}{\leq}(1+\gamma_t')\mathbb{E}\|w_{t+1}-v_t^*\|^2+(1+\frac{1}{\gamma_t'})\mathbb{E}\|v_t^*-v_{t+1}^*\|^2\\
&\overset{(c)}{\leq}(1+\gamma_t')\mathbb{E}\|v_t-\lambda_t h_t^R-v_t^*\|^2+(1+\frac{1}{\gamma_t'})\mathbb{E}\|v_t^*-v_{t+1}^*\|^2\\
&\overset{(d)}{\leq}(1+\gamma_t')(1+\delta_t')\mathbb{E}\|v_t-\lambda_t\nabla_v R(x_t,y_t,v_t)-v_t^*\|^2\\
&\quad+(1+\gamma_t')(1+\frac{1}{\delta_t'})\mathbb{E}\lambda_t^2\|h_t^R-\nabla_v R(x_t,y_t,v_t)\|^2+(1+\frac{1}{\gamma_t'})\mathbb{E}\|v_t^*-v_{t+1}^*\|^2, \qquad (29)
\end{aligned}$$

where (a) follows from Young's inequality, (b) follows Step 6 in Algorithm 2 and Lemma 1 that v^* is in a ball with radius r_v(define as $B(0,r_v^2)$), then we have $\|v_{t+1}-v_t^*\|=\|Proj_B(w_{t+1})-Proj_B(v_t^*)\|\leq\|w_{t+1}-v_t^*\|$ (this inequality is based on the nonexpansiveness of projection), (c) follows from Step 5 in Algorithm 2, and (d) uses Young's inequality and the definition of h_t^R in eq. (12). For the first term of the above eq. (29), we have

$$\begin{aligned}
&\mathbb{E}\|v_t-\lambda_t\nabla_v R(x_t,y_t,v_t)-v_t^*\|^2\\
&\quad=\mathbb{E}\|v_t-v_t^*\|^2+\lambda_t^2\mathbb{E}\|\nabla_v R(x_t,y_t,v_t)\|^2-2\lambda_t\mathbb{E}\langle\nabla_v R(x_t,y_t,v_t),v_t-v_t^*\rangle\\
&\quad\overset{(a)}{\leq}\left(1-2\lambda_t\frac{\mu_g L_g}{\mu_g+L_g}\right)\mathbb{E}\|v_t-v_t^*\|^2-\left(2\lambda_t\frac{\mu_g L_g}{\mu_g+L_g}-\lambda_t^2\right)\mathbb{E}\|\nabla_v R(x_t,y_t,v_t)\|^2\\
&\quad\overset{(b)}{\leq}\left(1-2\lambda_t\frac{(L_g+L_g^3)\mu_g}{\mu_g+L_g}+\lambda_t^2L_g^2\right)\mathbb{E}\|v_t-v_t^*\|^2 \qquad (30)
\end{aligned}$$

where (a) follows from the strong convexity of R function(in eq. (4)) that

$$\mathbb{E}\langle\nabla_v R(x_t,y_t,v_t),v_t-v_t^*\rangle\geq\frac{\mu_g L_g}{\mu_g+L_g}\mathbb{E}\|v_t-v_t^*\|^2+\frac{1}{\mu_g+L_g}\mathbb{E}\|\nabla_v R(x_t,y_t,v_t)\|^2,$$

and (b) follows from eq. (21). For the last term of eq. (29), we have

$$\begin{aligned}
\mathbb{E}\|v_{t+1}^*-v_t^*\|^2&=\mathbb{E}\|[\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\nabla_y f(x_{t+1},y_{t+1})-[\nabla_{yy}^2g(x_t,y_t)]^{-1}\nabla_y f(x_t,y_t)\|^2\\
&=\mathbb{E}\|[\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\nabla_y f(x_{t+1},y_{t+1})-[\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\nabla_y f(x_t,y_t)\\
&\quad+[\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\nabla_y f(x_t,y_t)-[\nabla_{yy}^2g(x_t,y_t)]^{-1}\nabla_y f(x_t,y_t)\|^2\\
&\overset{(a)}{\leq}2\mathbb{E}\|[\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\left(\nabla_y f(x_{t+1},y_{t+1})-\nabla_y f(x_t,y_t)\right)\|^2\\
&\quad+2\mathbb{E}\|\left([\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}-[\nabla_{yy}^2g(x_t,y_t)]^{-1}\right)\nabla_y f(x_t,y_t)\|^2\\
&\overset{(b)}{\leq}\frac{2L_{f_y}^2}{\mu_g^2}\mathbb{E}\|(x_{t+1},y_{t+1})-(x_t,y_t)\|^2\\
&\quad+2C_{f_y}\mathbb{E}\|[\nabla_{yy}^2g(x_t,y_t)]^{-1}[\nabla_{yy}^2g(x_t,y_t)-\nabla_{yy}^2g(x_{t+1},y_{t+1})][\nabla_{yy}^2g(x_{t+1},y_{t+1})]^{-1}\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\mathbb{E}\|(x_{t+1},y_{t+1})-(x_t,y_t)\|^2\\
&\leq \left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left(\mathbb{E}\|x_{t+1}-x_t\|^2 + \mathbb{E}\|y_{t+1}-y_t\|^2\right)\\
&\overset{(d)}{=} \left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left(\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\mathbb{E}\|h_t^g\|^2\right)\\
&\overset{(e)}{=} \left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\left(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\right], \qquad (31)
\end{aligned}
$$

where (a) follows from Cauchy–Schwartz inequality, (b) follows from Assumption 1 and 2, (c) follows from Assumption 2, (d) follows from Steps 4 and 7 in Algorithm 2, and (e) uses the definition of h_t^g in eq. (6). Finally, incorporating eq. (30) and eq. (31) into eq. (29), we have

$$
\begin{aligned}
&\mathbb{E}\|v_{t+1}-v_{t+1}^*\|^2\\
&\quad\leq (1+\gamma_t')(1+\delta_t')\left[\left(1-2\lambda_t\frac{(L_g+L_g^3)\mu_g}{\mu_g+L_g}+\lambda_t^2L_g^2\right)\mathbb{E}\|v_t-v_t^*\|^2\right]\\
&\qquad+(1+\gamma_t')\left(1+\frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2\\
&\qquad+(1+\frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2}+\frac{2C_{f_y}^2L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2+\beta_t^2\left(2\mathbb{E}\|e_t^g\|^2+2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\right].
\end{aligned}
$$

Then, the proof is complete. □

C.6 Descent in the Potential Function

Define the potential function as

$$
\begin{aligned}
V_t :=& \Phi(x_t) + K_1\|y_t - y^*(x_t)\|^2 + K_2\|v_t - v^*(x_t,y_t)\|^2\\
&+\frac{1}{\bar{c}_{\eta_f}}\frac{\|e_t^f\|^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_g}}\frac{\|e_t^g\|^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_R}}\frac{\|e_t^R\|^2}{\alpha_{t-1}}, \qquad (32)
\end{aligned}
$$

where the coefficients are given by

$$
\begin{aligned}
K_1 &= \frac{8(L_{f_y}^2 + \frac{C_{f_y}L62_{g_{xy}}}{\mu_g^2})}{c_\beta L_{\mu_g}}, \quad K_2 = \frac{4C_{g_{xy}}}{c_\lambda L_{\mu_g}};\\
\bar{c}_{\eta_f} &= \max\left\{96L_F^2,\ 12L_g^2c_\lambda^2,\ \frac{48L_{\mu_g}L_f^2c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}},\ \frac{3}{2}L_{\mu_g}^2c_\lambda^2\right\},\\
\bar{c}_{\eta_g} &= \max\left\{256L_g^2,\ \frac{128L_{\mu_g}L_g^2c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}\right\},\\
\bar{c}_{\eta_R} &= \max\left\{768(L_{g_{yy}}^2r_v^2+L_{f_y}^2),\ \frac{48L_g^4c_\lambda^2}{C_{g_{xy}}},\ \frac{384L_{\mu_g}(L_{g_{yy}}^2r_v^2+L_{f_y}^2)c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}\right\}.
\end{aligned}
$$

Lemma 12. *Suppose Assumptions 1 ,2 and 3 are satisfied. Choose the parameters of Algorithm 2 as*

$$
\alpha_t := \frac{1}{(w+t)^{1/3}},\ \beta_t := c_\beta\alpha_t,\ \lambda_t := c_\lambda\alpha_t;\ \eta_{t+1}^f := c_{\eta_f}\alpha_t^2,\ \eta_{t+1}^g := c_{\eta_g}\alpha_t^2,\ \eta_{t+1}^R := c_{\eta_R}\alpha_t^2,
$$

where the constants are given by

$$
c_\beta \geq \sqrt{\frac{512L_y^2(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}},
$$

$$c_\lambda \geq \sqrt{\max\left\{\frac{1024C_{g_{xy}}}{L_{\mu_g}^2}\Big(\frac{L_{f_y}^2}{\mu_g^2}+\frac{C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\Big), \frac{128(\mu_g+L_g)C_{g_{xy}}}{L_{\mu_g}}c_\beta^2, 128C_{g_{xy}}c_\beta^2\right\}};$$

$$c_{\eta_f}=\frac{1}{3L_f}+\bar{c}_{\eta_f},\ c_{\eta_g}=\frac{1}{3L_f}+32L_g^2c_\beta^2+\left[\frac{17(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\right]\bar{c}_{\eta_g},$$

$$c_{\eta_R}=\frac{1}{3L_f}+48L_g^2c_\lambda^2+\left[\frac{16C_{g_{xy}}}{L_{\mu_g}^2}\right]\bar{c}_{\eta_R};\ \ \sigma_R:=\sqrt{\sigma_{g_{yy}}^2r_v^2+\sigma_{f_y}^2},$$

$$w\geq\left(\max\left\{c_\beta(\mu_g+L_g),\ \frac{c_\lambda(\mu_g+L_g)}{2\mu_gL_g}\right\}\right)^3-1. \tag{33}$$

Then the iterates generated by Algorithm 2 satisfy

$$\mathbb{E}[V_{t+1}-V_t]\leq-\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2+\frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2+\frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2+\frac{4(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2,$$

for all $t\in\{0,1,...,T-1\}$.

Proof. Based on the definition of V_t in eq. (32), it can be seen that V_t contains six parts. We next develop five important inequalities to prove Lemma 12.

Step 1. Bound $\mathbb{E}\|y_t-y^*(x_t)\|^2$ in eq. (32).

Based on Lemma 7, we have

$$\begin{aligned}
&\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2-\mathbb{E}\|y_t-y^*(x_t)\|^2\\
&\quad\leq\left[(1+\gamma_t)(1+\delta_t)\left(1-2\beta\frac{\mu_gL_g}{\mu_g+L_g}\right)-1\right]\mathbb{E}\|y_t-y^*(x_t)]\|^2\\
&\quad\quad-(1+\gamma_t)(1+\delta_t)\left(\frac{2\beta_t}{\mu_g+L_g}-\beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\\
&\quad\quad+(1+\gamma_t)(1+\frac{1}{\delta_t})\beta_t^2\mathbb{E}\|e_t^g\|^2+\left(1+\frac{1}{\gamma_t}\right)L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2.
\end{aligned}\tag{34}$$

Choose $\gamma_t=\frac{\beta_tL_{\mu_g}/2}{1-\beta_tL_{\mu_g}}$ and $\delta_t=\frac{\beta_tL_{\mu_g}}{1-2\beta_tL_{\mu_g}}$. Then, the following equations and inequalities are satisfied.

$$\begin{aligned}
(1+\gamma_t)(1+\delta_t)(1-2\beta_tL_{\mu_g})&=1-\frac{\beta_tL_{\mu_g}}{2},\\
(1+\delta_t)(1-2\beta_tL_{\mu_g})&=1-\beta_tL_{\mu_g},\\
(1+\gamma_t)(1-\beta_tL_{\mu_g})&=1-\frac{\beta_tL_{\mu_g}}{2},\\
1+\frac{1}{\delta_t}\leq\frac{1}{\beta_tL_{\mu_g}},\quad 1+\frac{1}{\gamma_t}&\leq\frac{2}{\beta_tL_{\mu_g}},
\end{aligned}\tag{35}$$

where $L_{\mu_g}=\frac{\mu_gL_g}{\mu_g+L_g}$. Based on the selection of w in eq. (33), we have $(1+\gamma_t)(1+\frac{1}{\delta_t})\leq\frac{2}{\beta_tL_{\mu_g}}$.

Substituting eq. (35)'s bounds in eq. (34), we have

$$\begin{aligned}
&\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2-\mathbb{E}\|y_t-y^*(x_t)\|^2\\
&\quad\leq-\frac{\beta_tL_{\mu_g}}{2}\mathbb{E}\|y_t-y^*(x_t)\|^2-\left(\frac{2\beta_t}{\mu_g+L_g}-\beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\\
&\quad\quad+\frac{2}{\beta_tL_{\mu_g}}\beta_t^2\mathbb{E}\|e_t^g\|^2+\frac{2}{\beta_tL_{\mu_g}}L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2,
\end{aligned}$$

which, in conjunction with our selection in eq. (33) , yields

$$\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2-\mathbb{E}\|y_t-y^*(x_t)\|^2$$

$$\leq -\frac{\beta_t L_{\mu_g}}{2}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{\beta_t}{\mu_g + L_g}\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 + \frac{2\beta_t}{L_{\mu_g}}\mathbb{E}\|e_t^g\|^2 + \frac{2}{\beta_t L_{\mu_g}}L_y^2\alpha_t^2\mathbb{E}\|h_t^f\|^2. \tag{36}$$

Using $\beta_t = c_\beta \alpha_t$, and multiplying both sides of eq. (36) by K_1, we have

$$\begin{aligned} &K_1\mathbb{E}\big[\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2\big] \\ &\leq -4\Big(L_{f_x}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\alpha_t\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{L_{f_y}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{L_{\mu_g}(\mu_g + L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 \\ &\quad + \frac{16(L_{f_y}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2 + \frac{\alpha_t}{32}\mathbb{E}\|h_t^f\|^2. \end{aligned} \tag{37}$$

Step 2. Bound $\mathbb{E}\|v_t - v^*(x_t)\|^2$ in eq. (32).

Next, we deal with $\mathbb{E}\|v_{t+1} - v_{t+1}^*\|^2$ in similar way. Based on the parameter selections in eq. (33), we have $\lambda_t \leq \frac{2\mu_g L_g}{\mu_g + L_g}$, which combined with Lemma 11, yields

$$\begin{aligned} &\mathbb{E}\|v_{t+1} - v_{t+1}^*\|^2 \\ &\leq (1+\gamma_t')(1+\delta_t')\left[\left(1 - 2\lambda_t\frac{(L_g + L_g^3)\mu_g}{\mu_g + L_g} + \lambda_t^2 L_g^2\right)\|v_t - v_t^*\|^2\right] \\ &\quad + (1+\gamma_t')\left(1 + \frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2 \\ &\quad + (1 + \frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\left(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2\right)\right] \\ &\leq (1+\gamma_t')(1+\delta_t')\left[\left(1 - \frac{2\lambda_t L_g\mu_g}{\mu_g + L_g}\right)\mathbb{E}\|v_t - v_t^*\|^2\right] \\ &\quad + (1+\gamma_t')\left(1 + \frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2 \\ &\quad + (1 + \frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\left(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2\right)\right], \end{aligned}$$

Similarly to Step 1, we choose $\delta_t' = \frac{\lambda_t L_{\mu_g}}{1 - 2\lambda_t L_{\mu_g}}$ and $\gamma_t' = \frac{\lambda_t L_{\mu_g}/2}{1 - \lambda_t L_{\mu_g}}$ which implies

$$1 + \frac{1}{\delta_t'} \leq \frac{1}{\lambda_t L_{\mu_g}}, \quad 1 + \frac{1}{\gamma_t'} \leq \frac{2}{\lambda_t L_{\mu_g}}, \quad (1+\gamma_t')(1 + \frac{1}{\delta_t'}) \leq \frac{2}{\lambda_t L_{\mu_g}}.$$

Thus, we have

$$\begin{aligned} &\mathbb{E}\|v_{t+1} - v_{t+1}^*\|^2 \\ &\leq \left(1 - \frac{\lambda_t L_{\mu_g}}{2}\right)\mathbb{E}\|v_t - v_t^*\|^2 + \frac{2}{\lambda_t L_{\mu_g}}\lambda_t^2\mathbb{E}\|e_t^R\|^2 \\ &\quad + \frac{2}{\lambda_t L_{\mu_g}}\left(\frac{2L_{f_y}^2}{\mu_g^2} + \frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|h_t^f\|^2 + \beta_t^2\left(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2\right)\right]. \end{aligned} \tag{38}$$

Rearranging the above eq. (38), we have

$$\begin{aligned} &\mathbb{E}\big[\|v_{t+1} - v_{t+1}^*\|^2 - \|v_t - v_t^*\|^2\big] \\ &\leq -\frac{\lambda_t L_{\mu_g}}{2}\mathbb{E}\|v_t - v_t^*\|^2 + \frac{2}{\lambda_t L_{\mu_g}}\lambda_t^2\mathbb{E}\|e_t^R\|^2 \end{aligned}$$

$$+\frac{2}{\lambda_t L_{\mu_g}}\left(\frac{2L_{f_y}^2}{\mu_g^2}+\frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\Big[\alpha_t^2\mathbb{E}\|h_t^f\|^2+\beta_t^2\left(2\mathbb{E}\|e_t^g\|^2+2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\Big]. \quad (39)$$

Using $\lambda_t = c_\lambda \alpha_t$ with c_λ in eq. (33), and multiplying both sides of eq. (39) by K_2, we have

$$\begin{aligned}
&K_2\mathbb{E}\big[\|v_{t+1}-v_{t+1}^*\|^2-\|v_t-v_t^*\|^2\big]\\
&\quad\le -2C_{g_{xy}}\alpha_t\mathbb{E}\|v_t-v_t^*\|^2+\frac{8C_{g_{xy}}^2}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^R\|^2+\frac{\alpha_t}{32}\mathbb{E}\|h_t^f\|^2\\
&\qquad+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2. \quad (40)
\end{aligned}$$

Step 3. Bound $\mathbb{E}\|e_t^f\|^2$ in eq. (32).

Next, we obtain from Lemma 8 that

$$\begin{aligned}
\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}}&\le\left[\frac{(1-\eta_{t+1})^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\right]\mathbb{E}\|e_t^f\|^2+\frac{2(\eta_{t+1}^f)^2}{\alpha_t}\sigma_f^2+6L_F^2\alpha_t\mathbb{E}\|h_t^f\|^2\\
&\quad+\frac{12L_F^2\beta_t^2}{\alpha_t}\mathbb{E}\|e_t^g\|^2+\frac{12L_F^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\\
&\quad+12C_{g_{xy}}\frac{\lambda_t^2}{\alpha_t}\Big(\mathbb{E}\|e_t^R\|^2+L_g^2\mathbb{E}\|v_t-v_t^*\|^2\Big), \quad (41)
\end{aligned}$$

where the inequality follows from the fact that $0<1-\eta_t<1$ for all $t\in\{0,1,...,T-1\}$. Now considering the coefficient of the first term on the right-hand side of the above eq. (41), we have

$$\frac{(1-\eta_{t+1}^f)^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\le\frac{1}{\alpha_t}-\frac{\eta_{t+1}^f}{\alpha_t}-\frac{1}{\alpha_{t-1}}. \quad (42)$$

Using the definition of α_t in eq. (33), we have

$$\begin{aligned}
\frac{1}{\alpha_t}-\frac{1}{\alpha_{t-1}}&=(w+t)^{1/3}-(w+t-1)^{1/3}\overset{(a)}{\le}\frac{1}{3(w+t-1)^{2/3}}\overset{(b)}{\le}\frac{1}{3(w/2+t)^{2/3}}\\
&=\frac{2^{2/3}}{3(w+2t)^{2/3}}\le\frac{2^{2/3}}{3(w+t)^{2/3}}\overset{(c)}{\le}\frac{2^{2/3}}{3}\alpha_t^2\overset{(d)}{\le}\frac{\alpha_t}{3L_f}, \quad (43)
\end{aligned}$$

where (a) follows from $(x+y)^{1/3}-x^{1/3}\le y/(3x^{2/3})$, (b) follows because we choose $w\ge 2$, (c) follows from the definition of α_t and (d) follows because we choose $\alpha_t\le 1/3L_f$. Substituting eq. (43) into eq. (71) and using $\eta_{t+1}^f=c_{\eta_f}\alpha_t^2$, we have

$$\frac{(1-\eta_{t+1}^f)^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\le\frac{\alpha_t}{3L_f}-c_{\eta_f}\alpha_t\le-\bar{c}_{\eta_f}\alpha_t, \quad (44)$$

where the inequalities follow from $c_{\eta_f}=\frac{1}{3L_f}+\bar{c}_{\eta_f}$ with $\bar{c}_{\eta_f}$ in eq. (33). Then, substituting eq. (44) into eq. (41) yields

$$\begin{aligned}
&\frac{1}{\bar{c}_{\eta_f}}\mathbb{E}\left[\frac{\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\|e_t^f\|^2}{\alpha_{t-1}}\right]\\
&\le-\alpha_t\mathbb{E}\|e_t^f\|^2+\frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2+\frac{\alpha_t}{16}\mathbb{E}\|h_t^f\|^2+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2\\
&\quad+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2+\frac{8C_{g_{xy}}}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^R\|^2+C_{g_{xy}}\alpha_t\mathbb{E}\|v_t-v_t^*\|^2. \quad (45)
\end{aligned}$$

Step 4. Bound $\mathbb{E}\|e_t^g\|^2$ in eq. (32).

Next, from Lemma 9, we have

$$\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}} \leq \left[\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right] \mathbb{E}\|e_t^g\|^2 + \frac{2(\eta_{t+1}^g)^2}{\alpha_t}\sigma_g^2 + 16L_g^2 \alpha_t \mathbb{E}\|h_t^f\|^2 + \frac{32L_g^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2, \tag{46}$$

where we use the fact that $0 < 1 - \eta_t^g \leq 1$for all $t \in \{0, 1, ..., T-1\}$. Let us consider the coefficient of the first term on the right hand side of the above eq. (46). In specific, we have

$$\begin{aligned}\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} &\leq \frac{(1-\eta_{t+1}^g)}{\alpha_t}(1 + 32L_g^2\beta_t^2) - \frac{1}{\alpha_{t-1}} \\ &= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} + \frac{32L_g^2\beta_t^2}{\alpha_t} - c_{\eta_g}\alpha_t(1 + 32L_g^2\beta_t^2),\end{aligned}$$

which, combined with eq. (43) that $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f}$ and the definition of $\beta_t = c_\beta \alpha_t$, yields

$$\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f} + 32L_g^2 c_\beta^2 \alpha_t - c_{\eta_g}\alpha_t. \tag{47}$$

Recall $\bar{c}_{\eta_g}$ from eq. (33) that we choose, then we have

$$c_{\eta_g} = \frac{1}{3L_f} + 32L_g^2 c_\beta^2 + \frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\bar{c}_{\eta_g},$$

which, in conjunction with eq. (47), yields

$$\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq -\frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\bar{c}_{\eta_g}\alpha_t. \tag{48}$$

Substituting eq. (48) into eq. (46) yields

$$\frac{1}{\bar{c}_{\eta_g}}\left[\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}}\right] \leq -\frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\alpha_t \mathbb{E}\|e_t^g\|^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2 + \frac{\alpha_t}{16}\mathbb{E}\|h_t^f\|^2 + \frac{L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g + L_g)}\alpha_t \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2. \tag{49}$$

Step 5. Bound $\mathbb{E}\|e_t^R\|^2$ in eq. (32).

Next, from Lemma 10, we have

$$\begin{aligned}\frac{\mathbb{E}\|e_{t+1}^R\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^R\|^2}{\alpha_{t-1}} \leq &\left[\frac{(1-\eta_{t+1}^R)^2(1 + 48L_g^2\lambda_t^2)}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right]\mathbb{E}\|e_t^R\|^2 \\ &+ \left[\frac{4(\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2)}{\alpha_t}\right] + 48(1-\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2)\alpha_t \mathbb{E}\|h_t^f\|^2 \\ &+ 48(1-\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2)c_\beta^2 \alpha_t(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2) \\ &+ 48(1-\eta_{t+1}^R)^2 L_g^4 c_\lambda^2 \alpha_t \mathbb{E}\|v_t - v^*\|^2.\end{aligned} \tag{50}$$

For the first term of the right-hand side of eq. (50), we have

$$\begin{aligned}\frac{(1-\eta_{t+1}^R)^2(1 + 48L_g^2\lambda_t^2)}{\alpha_t} - \frac{1}{\alpha_{t-1}} &\leq \frac{1-\eta_{t+1}^R}{\alpha_t}(1 + 48L_g^2\lambda_t^2) - \frac{1}{\alpha_{t-1}} \\ &= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\eta_{t+1}^R}{\alpha_t} + \frac{1-\eta_{t+1}^R}{\alpha_t} \cdot 48L_g^2\lambda_t^2\end{aligned}$$

$$\overset{(a)}{=} \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - c_{\eta_R}\alpha_t + \left(\frac{1}{\alpha_t} - c_{\eta_R}\alpha_t\right) \cdot 48L_g^2 c_\lambda^2 \alpha_t^2$$
$$\overset{(b)}{\leq} \frac{\alpha_t}{3L_f} + 48L_g^2 c_\lambda^2 \alpha_t - c_{\eta_R}\alpha_t, \tag{51}$$

where (a) follows from the definition that $\eta_{t+1}^R = c_{\eta_R}\alpha_t^2$, and (b) follows from eq. (43) that $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f}$. Recall $\bar{c}_{\eta_R}$ from eq. (33) that

$$c_{\eta_R} = \frac{1}{3L_f} + 48L_g^2 c_\lambda^2 + \frac{16C_{g_{xy}}}{L_{\mu_g}^2}\bar{c}_{\eta_R},$$

which, in conjunction with eq. (51), yields

$$\frac{(1-\eta_{t+1}^R)^2(1+48L_g^2\lambda_t^2)}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq -\frac{16C_{g_{xy}}}{L_{\mu_g}^2}\bar{c}_{\eta_R}\alpha_t. \tag{52}$$

Incorporating eq. (52) into eq. (50), recalling from eq. (33) that $\sigma_R := \sqrt{\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2}$, and multiplying both sides of eq. (50) by $\frac{1}{\bar{c}_{\eta_R}}$, we have

$$\begin{aligned}\frac{1}{\bar{c}_{\eta_R}}\mathbb{E}\left[\frac{\|e_{t+1}^R\|^2}{\alpha_t} - \frac{\|e_{t+1}^R\|^2}{\alpha_{t-1}}\right] \leq & -\frac{16C_{g_{xy}}}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^R\|^2 + \frac{4(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2 + \frac{\alpha_t}{16}\mathbb{E}\|h_t^f\|^2 \\ & + \frac{L_{f_y}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2 + \frac{L_{f_y}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 \\ & + C_{g_{xy}}\alpha_t\mathbb{E}\|v_t - v_t^*\|^2. \end{aligned} \tag{53}$$

Step 6. Merging the results of Step 1-5 to prove eq. (32).

Finally, adding eq. (37), eq. (40) eq. (45), eq. (49), eq. (53) and the result of Lemma 6 with $\alpha_t \leq \frac{1}{3L_f}$ yields

$$\mathbb{E}[V_{t+1} - V_t] \leq -\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2 + \frac{4(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2. \tag{54}$$

Then, the proof is complete. □

C.7 Proof of Theorem 2

Proof. Summing the result of Lemma 12 for $t = 0$ to $T-1$, dividing by T on both sides and using the definitions that $\eta_{t+1}^f := c_{\eta_f}\alpha_t^2$, $\eta_{t+1}^g := c_{\eta_g}\alpha_t^2$, $\eta_{t+1}^R := c_{\eta_R}\alpha_t^2$, we have

$$\begin{aligned}\frac{\mathbb{E}[V_T - V_0]}{T} \leq & -\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \\ & + \frac{1}{T}\left[\frac{2(c_{\eta_f})^2}{\bar{c}_{\eta_f}}\sigma_f^2 + \frac{2(c_{\eta_g})^2}{\bar{c}_{\eta_g}}\sigma_g^2 + \frac{4(c_{\eta_R})^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\sum_{t=0}^{T-1}\alpha_t^3. \end{aligned} \tag{55}$$

Next based on the definition of α_t in eq. (33), we have

$$\sum_{t=0}^{T-1}\alpha_t^3 = \sum_{t=0}^{T-1}\frac{1}{w+t} \overset{(a)}{\leq} \sum_{t=0}^{T-1}\frac{1}{1+t} \leq \log(T+1) \tag{56}$$

where inequality (a) results from the fact that we choose $w \geq 1$. By plugging eq. (56) in eq. (55), we have

$$\frac{\mathbb{E}[V_T - V_0]}{T} \leq -\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \left[\frac{2c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2 + \frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2 + \frac{4c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{T}. \tag{57}$$

Rearrange the terms in eq. (57), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \leq \frac{\mathbb{E}[V_0 - l^*]}{T} + \left[\frac{2c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2 + \frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2 + \frac{4c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{T},$$

which, in conjunction with the fact that α_t is decreasing w.r.t. t and multiplying by $2/\alpha_T$ on both sides, yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \leq \frac{2\mathbb{E}[V_0 - l^*]}{\alpha_T T} + \left[\frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2 + \frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2 + \frac{8c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{\alpha_T T}. \tag{58}$$

Finally, based on the definition of the potential function, we have

$$\begin{aligned}
\mathbb{E}[V_0] :=& \mathbb{E}\Big[\Phi(x_0) + \frac{2L}{3\sqrt{2}L_y}\|y_0 - y^*(x_0)\|^2 + \frac{4C_B}{C_\lambda L_{\mu_g}}\|v_0 - v^*(x_0, y_0)\|^2 \\
&+ \frac{1}{\bar{c}_{\eta_f}}\frac{\|e_t^f\|^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_g}}\frac{\|e_t^g\|^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_R}}\frac{\|e_t^R\|^2}{\alpha_{t-1}}\Big] \\
\leq& \Phi(x_0) + \frac{2L}{3\sqrt{2}L_y}\|y_0 - y^*(x_0)\|^2 + \frac{4C_B}{C_\lambda L_{\mu_g}}\|v_0 - v^*(x_0, y_0)\|^2 \\
&+ \frac{1}{\bar{c}_{\eta_f}}\frac{\sigma_f^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_g}}\frac{\sigma_g^2}{\alpha_{t-1}} + \frac{1}{\bar{c}_{\eta_R}}\frac{\sigma_R^2}{\alpha_{t-1}}
\end{aligned} \tag{59}$$

where the inequality follows from Assumption 1, 2, 3, lemma 3 and the definitions of h_t^f, h_t^g and h_t^R in eq. (13), eq. (6), eq. (12). Then, substituting eq. (59) into eq. (58) yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|^2 \leq& \frac{2[\Phi(x_0) - \Phi^*]}{\alpha_t T} + \frac{4L}{3\sqrt{2}L_y}\frac{\|y_0 - y^*(x_0)\|^2}{\alpha_t T} + \frac{8C_B}{C_\lambda L_{\mu_y}}\frac{\|v_0 - v^*(x_0, y_0)\|^2}{\alpha_t T} \\
&+ \frac{2}{\alpha_{-1}\alpha_T T}\left(\frac{\sigma_f^2}{\bar{c}_{\eta_f}} + \frac{\sigma_g^2}{\bar{c}_{\eta_g}} + \frac{\sigma_R^2}{\bar{c}_{\eta_R}}\right) + \left[\frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2 + \frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2 + \frac{8c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{\alpha_T T},
\end{aligned}$$

which, combined with the definitions of $\alpha_T := \frac{1}{(\omega+T)^{1/3}}$ and $\alpha_{-1} = \alpha_0$, yields

$$\begin{aligned}
\mathbb{E}\|\nabla\Phi(x_a(T))\|^2 \leq \widetilde{\mathcal{O}}\Bigg(& \frac{\Phi(x_0) - \Phi^*}{T^{2/3}} + \frac{\|y_0 - y^*(x_0)\|^2}{T^{2/3}} + \frac{\|v_0 - v^*(x_0, y_0)\|^2}{T^{2/3}} \\
&+ \frac{\sigma_f^2}{T^{2/3}} + \frac{\sigma_g^2}{T^{2/3}} + \frac{\sigma_R^2}{T^{2/3}}\Bigg).
\end{aligned}$$

Finally, the proof is complete. □

D Proof of Theorem 1 (FdeHBO with first order approximation)

D.1 Descent in the function value

Lemma 13. *For non-convex and smooth $\widetilde{\Phi}(\cdot, \delta_\epsilon)$,with e_t^f defined as: $e_t^f := \widetilde{h}_t^f - \bar{\nabla} f(x_t, y_t, v_t)$, the consecutive iterates of Algorithm 2 satisfy:*

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\Phi}(x_{t+1}, \delta_\epsilon)\right] \leq \mathbb{E}\Big[&\widetilde{\Phi}(x_t, \delta_\epsilon) - \frac{\alpha_t}{2}\|\nabla\widetilde{\Phi}(x_t, \delta_\epsilon)\|^2 - \frac{\alpha_t}{2}(1 - \alpha_t L_f)\|\widetilde{h}_t^f\|^2 \\
&+ \alpha_t\|e_t^f\|^2 + 4\alpha_t\Big(L_{f_x}^2 + \frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\|y_t - y_t^*\|^2 + 2\alpha_t C_{g_{xy}}\|v_t - v_t^*\|^2\Big]
\end{aligned}$$

for all $t \in \{0, 1, ..., T-1\}$.

Proof. The proof follows the same steps as in Lemma 6. □

D.2 Descent in the iterates of the lower level function

Lemma 14. *Define $e_t^g := h_t^g - \nabla_y g(x_t, y_t)$. Then the iterates of solving the lower-level problem generated by Algorithm 1 satisfy*

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y_{t+1}^*\|^2 \\
&\leq (1+\gamma_t)(1+\delta_t)\left(1 - 2\beta_t \frac{\mu_g L g}{\mu_g + Lg}\right)\mathbb{E}\|y_t - y^*(x_t)\|^2 + \left(1 + \frac{1}{\gamma_t}\right) L_y^2 \alpha_t^2 \mathbb{E}\|\widetilde{h}_t^f\|^2 \\
&\quad - (1+\gamma_t)(1+\delta_t)\left(\frac{2\beta_t}{\mu_g + Lg} - \beta_t^2\right)\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 + (1+\gamma_t)(1+\frac{1}{\delta_t})\beta_t^2 \mathbb{E}\|e_t^g\|^2
\end{aligned}$$

for all $t \in \{0, ..., T-1\}$ with some $\gamma_t, \delta_t > 0$.

Proof. The proof follows the same steps as in Lemma C.2 in [34]. □

D.3 Descent in the gradient estimation error of the upper function

Lemma 15 (Restatement of Proposition 1)**.** *For any ξ, define $e_t^f := \widetilde{h}_t^f - \bar{\nabla} f(x_t, y_t, v_t)$ and $e_t^J := \widetilde{J}(x_t, y_t, v_t, \delta_\epsilon; \xi) - \nabla_{xy}^2 g(x_t, y_t; \xi) v_t$. Under Assumption 3, the iterates of the outer problem generated by Algorithm 1 satisfy*

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^f\|^2 \leq &\Big[(1-\eta_{t+1}^f)^2 + 4L_{g_{xy}} r_v^2 \delta_\epsilon\Big]\mathbb{E}\|e_t^f\|^2 + 4(\eta_{t+1}^f)^2 \sigma_f^2 + \big(4L_{g_{xy}} r_v^2 \delta_\epsilon + 16 L_{g_{xy}}^2 r_v^4 \delta_\epsilon^2\big) \\
&+ 6(1-\eta_{t+1}^f)^2\Big[L_F^2 \alpha_t^2 \mathbb{E}\|\widetilde{h}_t^f\|^2 + 2L_F^2 \beta_t^2\big(\mathbb{E}\|e_t^g\|^2 + \|\nabla_y g(x_t, y_t)\|^2\big) \\
&+ 2C_{g_{xy}} \lambda_t^2\big(\mathbb{E}\|e_t^R\|^2 + L_g^2 \mathbb{E}\|v_t - v_t^*\|^2\big)\Big]
\end{aligned}$$

for all $t \in \{0, ..., T-1\}$ with L_F in Lemma 2.

Proof. Based on the definition of $\bar{\nabla} f(x_t, y_t, v_t; \xi)$ in eq. (3), the definition of $\widetilde{\nabla} f(x_t, y_t, v_t; \xi)$ in Section 2.2 and the definition of e_t^J above, we have

$$\bar{\nabla} f(x_t, y_t, v_t; \xi) = \widetilde{\bar{\nabla}} f(x_t, y_t, v_t, \delta_\epsilon; \xi) + e_t^J \tag{60}$$

for any data sample ξ. From the definition of e_t^f, we have

$$\begin{aligned}
&\mathbb{E}\|e_{t+1}^f\|^2 \\
&= \mathbb{E}\|\widetilde{h}_{t+1}^f - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|\eta_{t+1}^f \widetilde{\bar{\nabla}} f(x_{t+1}, y_{t+1}, v_{t+1}, \delta_\epsilon; \xi_{t+1}) + (1-\eta_{t+1}^f)\big(\widetilde{h}_t^f + \widetilde{\bar{\nabla}} f(x_{t+1}, y_{t+1}, v_{t+1}, \delta_\epsilon; \xi_{t+1}) \\
&\qquad - \widetilde{\bar{\nabla}} f(x_t, y_t, v_t, \delta_\epsilon; \xi_{t+1})\big) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\|^2 \\
&\overset{(b)}{=} \mathbb{E}\|\eta_{t+1}^f \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) + (1-\eta_{t+1}^f)\big(\widetilde{h}_t^f + \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) \\
&\quad - \bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1})\big) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}) - \big(e_{t+1}^J - (1-\eta_{t+1}^f) e_t^J\big)\|^2 \\
&\overset{(c)}{=} \mathbb{E}\|(1-\eta_{t+1}^f) e_t^f + \eta_{t+1}^f\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}) \\
&\qquad + (1-\eta_{t+1}^f)\Big(\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\Big) - \big(e_{t+1}^J - (1-\eta_{t+1}^f) e_t^J\big)\|^2 \\
&\overset{(d)}{\leq} \Big[(1-\eta_{t+1}^f)^2 + 2L_{g_{xy}} r_v^2 \delta_\epsilon\Big]\mathbb{E}\|e_t^f\|^2 \\
&\quad + \mathbb{E}\|\eta_{t+1}^f\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\quad + (1-\eta_{t+1}^f)\Big(\big(\bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1}; \xi_{t+1}) - \bar{\nabla} f(x_{t+1}, y_{t+1}, v_{t+1})\big) \\
&\quad - \big(\bar{\nabla} f(x_t, y_t, v_t; \xi_{t+1}) - \bar{\nabla} f(x_t, y_t, v_t)\big)\Big) - \big(e_{t+1}^J - (1-\eta_{t+1}^f) e_t^J\big)\|^2 + 2L_{g_{xy}} r_v^2 \delta_\epsilon
\end{aligned}$$

$$
\begin{aligned}
&\overset{(e)}{\leq} \Big[(1-\eta_{t+1}^f)^2 + 2L_{g_{xy}} r_v^2 \delta_\epsilon\Big]\mathbb{E}\|e_t^f\|^2 + 2L_{g_{xy}} r_v^2\delta_\epsilon + 4(\eta_{t+1}^f)^2\sigma_f^2 + 4\mathbb{E}\|e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\|^2 \\
&\quad + 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\quad - \big(\bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t)\big)\|^2 \\
&\leq \Big[(1-\eta_{t+1}^f)^2 + 4L_{g_{xy}} r_v^2 \delta_\epsilon\Big]\mathbb{E}\|e_t^f\|^2 + 4L_{g_{xy}} r_v^2\delta_\epsilon + 4(\eta_{t+1}^f)^2\sigma_f^2 + 16L_{g_{xy}}^2 r_v^4\delta_\epsilon^2 \\
&\quad + 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\quad - \big(\bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t)\big)\|^2 \qquad (61)
\end{aligned}
$$

where (a) uses the definition of $\widetilde{h}_{t+1}^f$ in eq. (10), (b) uses eq. (60), (c) uses the definition that $e_t^f := \widetilde{h}_t^f - \nabla f(x_t,y_t,v_t)$, (d) follows because for $\Sigma_{t+1} = \sigma\{y_0,x_0,v_0,...,y_t,x_t,v_t,y_{t+1},x_{t+1},v_{t+1}\}$,

$$
\begin{aligned}
&\mathbb{E}\Big\langle (1-\eta_{t+1}^f)e_t^f, \big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t)\big)\Big) - \big(e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\big)\|^2 | \Sigma_{t+1}\Big\rangle \\
&= \mathbb{E}\Big\langle (1-\eta_{t+1}^f)e_t^f, \mathbb{E}\Big[\big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\qquad - \big(\bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t)\big)\Big) - \big(e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\big)\|^2\Big] | \Sigma_{t+1}\Big\rangle \\
&= \mathbb{E}\Big\langle (1-\eta_{t+1}^f)e_t^f, -\big(e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\big)\Big\rangle \\
&\leq \sqrt{\mathbb{E}\|e_t^f\|^2}\cdot\sqrt{\mathbb{E}\|e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\|^2} \\
&\leq \max\{1,\mathbb{E}\|e_t^f\|^2\}\cdot\sqrt{\mathbb{E}\|e_{t+1}^J - (1-\eta_{t+1}^f)e_t^J\|^2} \\
&\leq (1+\mathbb{E}\|e_t^f\|^2)\cdot\sqrt{2\mathbb{E}\|e_{t+1}^J\|^2 + 2\mathbb{E}\|e_t^J\|^2} \\
&\leq (1+\mathbb{E}\|e_t^f\|^2)\cdot\big(2L_{g_{xy}} r_v^2\delta_\epsilon\big) \\
&= \big(2L_{g_{xy}} r_v^2\delta_\epsilon\big)\mathbb{E}\|e_t^f\|^2 + 2L_{g_{xy}} r_v^2\delta_\epsilon,
\end{aligned}
$$

which follows from Lemma 5, and (e) follows from lemma 3.

Next, we bound the last term of eq. (61) as

$$
\begin{aligned}
&2(1-\eta_{t+1}^f)^2\mathbb{E}\|\big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1})\big) \\
&\qquad \big(\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t)\big)\|^2 \\
&\overset{(a)}{\leq} 2(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1})\|^2 \\
&\leq 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_{t+1},v_{t+1};\xi_{t+1})\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_t,y_{t+1},v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_{t+1};\xi_{t+1})\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\bar{\nabla} f(x_t,y_t,v_{t+1};\xi_{t+1}) - \bar{\nabla} f(x_t,y_t,v_t;\xi_{t+1})\|^2 \\
&\overset{(b)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|x_{t+1}-x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2\mathbb{E}\|\nabla_{xy}^2 g(x_t,y_t)(v_{t+1}-v_t)\|^2 \\
&\overset{(c)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|x_{t+1}-x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2 C_{g_{xy}}\mathbb{E}\|v_{t+1}-v_t\|^2 \\
&\overset{(d)}{\leq} 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|x_{t+1}-x_t\|^2 + 6(1-\eta_{t+1}^f)^2 L_F^2\mathbb{E}\|y_{t+1}-y_t\|^2 \\
&\quad + 6(1-\eta_{t+1}^f)^2 C_{g_{xy}}\mathbb{E}\|w_{t+1}-v_t\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(e)}{\leq} 6(1-\eta_{t+1}^f)^2\Big[L_F^2\big(\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+\beta_t^2\mathbb{E}\|h_t^g\|^2\big)+C_{g_{xy}}\lambda_t^2\mathbb{E}\|\widetilde{h}_t^R\|^2\Big]\\
&\overset{(f)}{\leq} 6(1-\eta_{t+1}^f)^2\Big[L_F^2\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+2L_F^2\beta_t^2\big(\mathbb{E}\|e_t^g\|^2+\|\nabla_y g(x_t,y_t)\|^2\big)\\
&\quad+2C_{g_{xy}}\lambda_t^2\big(\mathbb{E}\|e_t^R\|^2+\mathbb{E}\|\nabla R_v(x_t,y_t,v_t)\|^2\big)\Big]\\
&\overset{(g)}{\leq} 6(1-\eta_{t+1}^f)^2\Big[L_F^2\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+2L_F^2\beta_t^2\big(\mathbb{E}\|e_t^g\|^2+\|\nabla_y g(x_t,y_t)\|^2\big)\\
&\quad+2C_{g_{xy}}\lambda_t^2\big(\mathbb{E}\|e_t^R\|^2+\mathbb{E}\|v_t-v_t^*\|^2\big)\Big],
\end{aligned}
\tag{62}
$$

where (a) follows from the mean variance inequality: For a random variable Z we have $\mathbb{E}\|Z-\mathbb{E}[Z]\|^2\leq\mathbb{E}\|Z\|^2$ with Z defined as $Z:=\bar{\nabla}f(x_{t+1},y_{t+1},v_{t+1};\xi_{t+1})-\bar{\nabla}f(x_t,y_t,v_t;\xi_{t+1})$, (b) follows from Lemma 2 and eq. (5), (c) uses Assumption 2, (d) uses the nonexpansiveness of projection, (e) uses the definition of $\widetilde{h}_t^f$ in eq. (10), h_t^g in eq. (6) and $\widetilde{h}_t^R$ in eq. (7), (f) follows from the definition that $e_t^g:=g_t^R-\nabla g(x_t,y_t)$ and $e_t^R:=\widetilde{h}_t^R-\nabla_v R(x_t,y_t,v_t)$, (g) uses the result of eq. (21).

Finally, substituting eq. (62) in eq. (61), we finish the proof. □

D.4 Descent in the gradient estimation error of the inner function

Lemma 16. *Define $e_t^g:=h_t^g-\nabla_y g(x_t,y_t)$. Under Assumption 2 and 3, the iterates generated from Algorithm 1 satisfy*

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2\leq&\big((1-\eta_{t+1}^g)^2\mathbb{E}\|e_t^g\|^2+32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\big)\,\mathbb{E}\|e_t^g\|^2+2(\eta_{t+1}^g)^2\sigma_g^2\\
&+16(1-\eta_{t+1}^g)^2L_g^2\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2
\end{aligned}
$$

for all $t\in\{0,1,...,T-1\}$.

Proof. The proof follows the same steps as in Lemma 9. □

D.5 Descent in the gradient estimation error of the R function

Lemma 17 (Restatement of Proposition 2)**.** *For any ψ, define $e_t^R:=\widetilde{h}_t^R-\nabla_v R(x_t,y_t,v_t)$ and $e_t^H:=\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon;\psi)-\nabla_{yy}^2 g(x_t,y_t;\psi)v_t$. Under Assumption 1, 2, 3, the iterates generated by Algorithm 1 satisfy*

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^R\|^2\leq&\big[(1-\eta_{t+1}^R)^2(1+96L_g^2\lambda_t^2)+4L_{g_{yy}}r_v^2\delta_\epsilon\big]\mathbb{E}\|e_t^R\|^2+\big(4L_{g_{yy}}r_v^2\delta_\epsilon+8L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big)\\
&+96(1-\eta_{t+1}^R)^2\big(L_{g_{yy}}^2r_v^2+L_{f_y}^2\big)\Big[\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+2\beta_t^2(\mathbb{E}\|e_t^g\|^2+\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2)\Big]\\
&+96(1-\eta_{t+1}^R)^2L_g^2\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2+L_g^2\mathbb{E}\|v_t-v_t^*\|^2\right)+8(\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2r_v^2+\sigma_{f_y}^2),
\end{aligned}
$$

for all $t\in\{0,1,...,T-1\}$.

Proof. From the definition of $\nabla_v R(x_t,y_t,v_t;\psi_t)$ in eq. (5), the definition of $\widetilde{\nabla}_v R(x_t,y_t,v_t,\delta_\epsilon;\psi_t)$ in eq. (8) and the definition of e_t^H, we have

$$
\begin{aligned}
&\widetilde{\nabla}_v R(x_t,y_t,v_t,\delta_\epsilon;\psi)-\nabla_v R(x_t,y_t,v_t;\psi)\\
&\qquad=\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon;\psi)-\nabla_{yy}^2 g(x_t,y_t;\psi)v_t=e_t^H
\end{aligned}
\tag{63}
$$

for any data sample ψ. For the gradient estimation error of the R function, we have

$$
\begin{aligned}
&\mathbb{E}\|e_{t+1}^R\|^2\\
&=\mathbb{E}\|\widetilde{h}_{t+1}^R-\nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2\\
&\overset{(a)}{=}\mathbb{E}\|\widetilde{\nabla}_v R(x_{t+1},y_{t+1},v_{t+1},\delta_\epsilon;\psi_{t+1})+(1-\eta_{t+1}^R)\widetilde{h}_t^R-(1-\eta_{t+1}^R)\widetilde{\nabla}_v R(x_t,y_t,v_t,\delta_\epsilon;\psi_{t+1})\\
&\qquad-\nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(b)}{=} \mathbb{E}\|\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) + (1-\eta_{t+1}^R)\widetilde{h}_t^R - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) \\
&\qquad - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) + \big(e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\big)\|^2 \\
&\overset{(c)}{=} \mathbb{E}\|(1-\eta_{t+1}^R)e_t^R + (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) \\
&\qquad - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) + \big(e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\big)\|^2 \\
&\overset{(d)}{\leq} \big[(1-\eta_{t+1}^R)^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon\big]\mathbb{E}\|e_t^R\|^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon + \mathbb{E}\|(1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) \\
&\quad - (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) \\
&\quad + \big(e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\big)\|^2 \\
&\leq \big[(1-\eta_{t+1}^R)^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon\big]\mathbb{E}\|e_t^R\|^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon + 2\mathbb{E}\|(1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) \\
&- (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\quad + 2\mathbb{E}\|e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\|^2 \\
&\overset{(e)}{\leq} \big[(1-\eta_{t+1}^R)^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon\big]\mathbb{E}\|e_t^R\|^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon + 2\mathbb{E}\|(1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t) \\
&- (1-\eta_{t+1}^R)\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) + \nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1})\|^2 \\
&\quad + 8L_{g_{yy}}^2 r_v^2\delta_\epsilon^2, \qquad (64)
\end{aligned}
$$

where (a) follows from the definition of $\widetilde{h}_t^R$ in eq. (7), (b) follows from eq. (63), (c) uses the definition of $e_t^R := \widetilde{h}_t^R - \nabla_v R(x_t,y_t,v_t)$, (d) follows from the fact that

$$
\begin{aligned}
&\mathbb{E}\Big\langle (1-\eta_{t+1}^R)e_t^R, \big(\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla_v R(x_{t+1},y_{t+1},v_{t+1}) \\
&\qquad - (1-\eta_{t+1}^R)\big(\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) - \nabla_v R(x_t,y_t,v_t)\big) + \big(e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\big)\Big\rangle \\
&\quad = \mathbb{E}\Big\langle (1-\eta_{t+1}^R)e_t^R, \mathbb{E}\big[\big(\nabla_v R(x_{t+1},y_{t+1},v_{t+1};\psi_{t+1}) - \nabla R(x_{t+1},y_{t+1},v_{t+1})\big) \\
&\qquad - (1-\eta_{t+1}^R)\big(\nabla_v R(x_t,y_t,v_t;\psi_{t+1}) - \nabla_v R(x_t,y_t,v_t)\big) + \big(e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\big)|\Sigma_{t+1}\big]\Big\rangle \\
&\quad = \mathbb{E}\Big\langle (1-\eta_{t+1}^R)e_t^R, e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\Big\rangle \\
&\quad \leq \sqrt{\mathbb{E}\|e_t^R\|^2}\cdot\sqrt{\mathbb{E}\|e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\|^2} \\
&\quad \leq \max\{1,\mathbb{E}\|e_t^R\|^2\}\cdot\sqrt{\mathbb{E}\|e_{t+1}^H - (1-\eta_{t+1}^R)e_t^H\|^2} \\
&\quad \leq (1+\mathbb{E}\|e_t^R\|^2)\cdot\sqrt{2\mathbb{E}\|e_{t+1}^H\|^2 + 2\mathbb{E}\|e_t^H\|^2} \\
&\quad \leq (1+\mathbb{E}\|e_t^R\|^2)\cdot\big(2L_{g_{yy}}r_v^2\delta_\epsilon\big) \\
&\quad \leq \big(2L_{g_{yy}}r_v^2\delta_\epsilon\big)\mathbb{E}\|e_t^R\|^2 + 2L_{g_{yy}}r_v^2\delta_\epsilon,
\end{aligned}
$$

for $\Sigma_{t+1} = \sigma\{y_0,v_0,x_0,...,y_t,v_t,x_t,y_{t+1},v_{t+1},x_{t+1}\}$. Incorporating eq. (28) into eq. (64), we have

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^R\|^2 \leq\ & \big[(1-\eta_{t+1}^R)^2(1+96L_g^2\lambda_t^2) + 4L_{g_{yy}}r_v^2\delta_\epsilon\big]\mathbb{E}\|e_t^R\|^2 + 4L_{g_{yy}}r_v^2\delta_\epsilon + 8L_{g_{yy}}^2 r_v^4\delta_\epsilon^2 \\
&+ 96(1-\eta_{t+1}^R)^2\left(L_{g_{yy}}^2 r_v^2 + L_{f_y}^2\right)\Big[\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2 + 2\beta_t^2(\mathbb{E}\|e_t^g\|^2 + \mathbb{E}\|\nabla_y g(x_t,y_t)\|^2)\Big] \\
&+ 96(1-\eta_{t+1}^R)^2 L_g^2\lambda_t^2\left(\mathbb{E}\|e_t^R\|^2 + L_g^2\mathbb{E}\|v_t - v_t^*\|^2\right) + 8(\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2),
\end{aligned}
$$

which finishes the proof. □

D.6 Descent in iterates of the LS problem

Lemma 18. *Define $e_t^R := \widetilde{h}_t^R - \nabla_v R(x_t,y_t,v_t)$. Under the Assumption 1, 2, the iterates of the LS problem generated according to Algorithm 1 satisfy*

$$\mathbb{E}\|v_{t+1} - v_{t+1}^*\|^2$$

$$
\begin{aligned}
&\leq (1+\gamma_t')\,(1+\delta_t')\left[\left(1-2\lambda_t\frac{(L_g+L_g^3)\mu_g}{\mu_g+L_g}+\lambda_t^2L_g^2\right)\mathbb{E}\|v_t-v_t^*\|^2\right]\\
&\quad+(1+\gamma_t')\left(1+\frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2\\
&\quad+(1+\frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2}+\frac{2C_{f_y}^2L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+\beta_t^2\left(2\mathbb{E}\|e_t^g\|^2+2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\right].
\end{aligned}
$$

for all $t\in\{0,...,T-1\}$ *with some* $\gamma_t'>0$ *and* $\delta_t'>0$.

Proof. From eq. (65) we have that there exist $\gamma_t'\geq 0$, $\delta_t'\geq 0$ such that

$$
\begin{aligned}
\mathbb{E}\|v_{t+1}-v_{t+1}^*\|^2 &\leq (1+\gamma_t')(1+\delta_t')\mathbb{E}\|v_t-\lambda_t\nabla_v R(x_t,y_t,v_t)-v_t^*)\|^2\\
&\quad+(1+\gamma_t')(1+\frac{1}{\delta_t'})\lambda_t^2\mathbb{E}\|\widetilde{h}_t^R-\nabla_v R(x_t,y_t,v_t)\|^2+(1+\frac{1}{\gamma_t'})\mathbb{E}\|v_t^*-v_{t+1}^*\|^2. \qquad (65)
\end{aligned}
$$

Incorporating eq. (30) and eq. (31) into eq. (65), similarly to Lemma 18, we have

$$
\begin{aligned}
&\mathbb{E}\|v_{t+1}-v_{t+1}^*\|^2\\
&\quad\leq (1+\gamma_t')\,(1+\delta_t')\left[\left(1-2\lambda_t\frac{(L_g+L_g^3)\mu_g}{\mu_g+L_g}+\lambda_t^2L_g^2\right)\mathbb{E}\|v_t-v_t^*\|^2\right]\\
&\quad\quad+(1+\gamma_t')\left(1+\frac{1}{\delta_t'}\right)\lambda_t^2\mathbb{E}\|e_t^R\|^2\\
&\quad\quad+(1+\frac{1}{\gamma_t'})\left(\frac{2L_{f_y}^2}{\mu_g^2}+\frac{2C_{f_y}L_{g_{yy}}^2}{\mu_g^4}\right)\left[\alpha_t^2\mathbb{E}\|\widetilde{h}_t^f\|^2+\beta_t^2\left(2\mathbb{E}\|e_t^g\|^2+2\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\right)\right].
\end{aligned}
$$

which finishes the proof. □

D.7 Descent in the Potential Function

Define the potential function as

$$
\begin{aligned}
V_t :=&\Phi(x_t,\delta_\epsilon)+K_1\|y_t-y^*(x_t)\|^2+K_2\|v_t-v^*(x_t,y_t)\|^2\\
&+\frac{1}{\bar{c}_{\eta_f}}\frac{\|e_t^f\|^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_g}}\frac{\|e_t^g\|^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_R}}\frac{\|e_t^R\|^2}{\alpha_{t-1}}, \qquad (66)
\end{aligned}
$$

where the coefficients are given by

$$
\begin{aligned}
K_1&=\frac{8(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{c_\beta L_{\mu_g}},\quad K_2=\frac{4C_{g_{xy}}}{c_\lambda L_{\mu_g}};\\
\bar{c}_{\eta_f}&=\max\left\{96L_F^2,\ 12L_g^2c_\lambda^2,\ \frac{48L_{\mu_g}L_f^2c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}},\ \frac{3}{2}L_{\mu_g}^2c_\lambda^2\right\},\\
\bar{c}_{\eta_g}&=\max\left\{256L_g^2,\ \frac{128L_{\mu_g}L_g^2c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}\right\},\\
\bar{c}_{\eta_R}&=\max\left\{1536(L_{g_{yy}}^2r_v^2+L_{f_y}^2),\ \frac{96L_g^4c_\lambda^2}{C_{g_{xy}}},\ \frac{768L_{\mu_g}(L_{g_{yy}}^2r_v^2+L_{f_y}^2)c_\beta^2\max\{L_{\mu_g},\mu_g+L_g\}}{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}\right\}.
\end{aligned}
$$

Lemma 19. *Suppose Assumptions 1 ,2 and 3 are satisfied. Choose the parameters of Algorithm 1 as*

$$
\alpha_t:=\frac{1}{(w+t)^{1/3}},\ \beta_t:=c_\beta\alpha_t,\ \lambda_t:=c_\lambda\alpha_t;\ \eta_{t+1}^f:=c_{\eta_f}\alpha_t^2,\ \eta_{t+1}^g:=c_{\eta_g}\alpha_t^2,\ \eta_{t+1}^R:=c_{\eta_R}\alpha_t^2,
$$

where the constants are given by

$$
\begin{aligned}
&c_\beta \geq \sqrt{\frac{512L_y^2(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}},\\
&c_\lambda \geq \sqrt{\max\left\{\frac{1024C_{g_{xy}}}{L_{\mu_g}^2}\Big(\frac{L_{f_y}^2}{\mu_g^2}+\frac{C_{f_y}^2L_{g_{yy}}^2}{\mu_g^4}\Big), \frac{128(\mu_g+L_g)C_{g_{xy}}}{L_{\mu_g}}c_\beta^2, 128C_{g_{xy}}c_\beta^2\right\}};\\
&c_{\eta_f} = \frac{2}{3L_f}+2\bar{c}_{\eta_f},\ c_{\eta_g} = \frac{1}{3L_f}+32L_g^2c_\beta^2+\left[\frac{17(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\right]\bar{c}_{\eta_g},\\
&c_{\eta_R} = \frac{2}{3L_f}+192L_g^2c_\lambda^2+\left[\frac{32C_{g_{xy}}}{L_{\mu_g}^2}\right]\bar{c}_{\eta_R};\ \ \sigma_R := \sqrt{\sigma_{g_{yy}}^2r_v^2+\sigma_{f_y}^2},\\
&w \geq \left(\max\left\{c_\beta(\mu_g+L_g),\ \frac{c_\lambda(\mu_g+L_g)}{2\mu_gL_g}\right\}\right)^3-1,\\
&\delta_\epsilon \leq \min\left\{\frac{c_{\eta_f}}{8(L_{g_{xy}}r_v^2(w+T-1)^{2/3})}, \frac{c_{\eta_R}}{8(L_{g_{xy}}r_v^2(w+T-1)^{2/3})}\right\}.
\end{aligned}
\tag{67}
$$

for all $t \in \{0,1,...,T-1\}$. Then the iterates generated by Algorithm 1 satisfy

$$
\begin{aligned}
\mathbb{E}[V_{t+1}-V_t] \leq &-\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2+\frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2+\frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2+\frac{4(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2\\
&+\frac{1}{\alpha_t\bar{c}_{\eta_f}}\big(4L_{g_{xy}}r_v^2\delta_\epsilon+16L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+\frac{1}{\alpha_t\bar{c}_{\eta_R}}\big(4L_{g_{yy}}r_v^2\delta_\epsilon+8L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big),
\end{aligned}
$$

for all $t \in \{0,1,...,T-1\}$.

Proof. Based on the definition of V_t in eq. (66), it can be seen that V_t contains six parts. We next develop five important inequalities to prove Lemma 19.

Step 1. Bound $\mathbb{E}\|y_t-y^*(x_t)\|^2$ in eq. (66).

Same as Step 1 in Lemma 12, by Lemma 14, we have

$$
\begin{aligned}
&K_1\mathbb{E}\big[\|y_{t+1}-y^*(x_{t+1})\|^2-\|y_t-y^*(x_t)\|^2\big]\\
&\quad\leq -4\Big(L_{f_x}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}\Big)\alpha_t\mathbb{E}\|y_t-y^*(x_t)\|^2-\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\\
&\qquad+\frac{16(L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2+\frac{\alpha_t}{32}\mathbb{E}\|h_t^f\|^2.
\end{aligned}
\tag{68}
$$

Step 2. Bound $\mathbb{E}\|v_t-v^*(x_t)\|^2$ in eq. (66).

Same as Step 1 in Lemma 12, by Lemma 18, we have

$$
\begin{aligned}
&K_2\mathbb{E}\big[\|v_{t+1}-v_{t+1}^*\|^2-\|v_t-v_t^*\|^2\big]\\
&\quad\leq -2C_{g_{xy}}\alpha_t\mathbb{E}\|v_t-v_t^*\|^2+\frac{8C_{g_{xy}}^2}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^R\|^2+\frac{\alpha_t}{32}\mathbb{E}\|h_t^f\|^2\\
&\qquad+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2.
\end{aligned}
\tag{69}
$$

Step 3. Bound $\mathbb{E}\|e_t^f\|^2$ in eq. (66).

Next, we obtain from Lemma 15 that

$$\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}} \leq \left[\frac{(1-\eta_{t+1})^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}+\frac{4L_{g_{xy}}r_v^2\delta_\epsilon}{\alpha_t}\right]\mathbb{E}\|e_t^f\|^2+\frac{4(\eta_{t+1}^f)^2}{\alpha_t}\sigma_f^2 \\ +6L_F^2\alpha_t\mathbb{E}\|\widetilde{h}_t^f\|^2+\frac{12L_F^2\beta_t^2}{\alpha_t}\mathbb{E}\|e_t^g\|^2+\frac{12L_F^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 \\ +12C_{g_{xy}}\frac{\lambda_t^2}{\alpha_t}\Big(\mathbb{E}\|e_t^R\|^2+L_g^2\mathbb{E}\|v_t-v_t^*\|^2\Big)+\big(4L_{g_{xy}}r_v^2\delta_\epsilon+16L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big), \quad (70)$$

where the inequality follows from the fact that $0<1-\eta_t<1$ for all $t\in\{0,1,...,T-1\}$. Now considering the coefficient of the first term on the right hand side of the above eq. (70) and recalling the δ_ϵ in eq. (67), we have

$$\frac{(1-\eta_{t+1}^f)^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}+\frac{4L_{g_{xy}}r_v^2\delta_\epsilon}{\alpha_t}\leq\frac{1}{\alpha_t}-\frac{\eta_{t+1}^f}{2\alpha_t}-\frac{1}{\alpha_{t-1}}. \quad (71)$$

Using the definition of α_t in eq. (67), we have

$$\frac{1}{\alpha_t}-\frac{1}{\alpha_{t-1}}=(w+t)^{1/3}-(w+t-1)^{1/3}\overset{(a)}{\leq}\frac{1}{3(w+t-1)^{2/3}}\overset{(b)}{\leq}\frac{1}{3(w/2+t)^{2/3}} \\ =\frac{2^{2/3}}{3(w+2t)^{2/3}}\leq\frac{2^{2/3}}{3(w+t)^{2/3}}\overset{(c)}{\leq}\frac{2^{2/3}}{3}\alpha_t^2\overset{(d)}{\leq}\frac{\alpha_t}{3L_f}, \quad (72)$$

where (a) follows from $(x+y)^{1/3}-x^{1/3}\leq y/(3x^{2/3})$, (b) follows because we choose $w\geq 2$ hence $1\leq w/2$, (c) follows from the definition of α_t and (d) follows because we choose $\alpha_t\leq 1/3L_f$. Substituting eq. (72) into eq. (71) and using $\eta_{t+1}^f=c_{\eta_f}\alpha_t^2$, we have

$$\frac{(1-\eta_{t+1}^f)^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\leq\frac{\alpha_t}{3L_f}-\frac{c_{\eta_f}\alpha_t}{2}\leq-\bar{c}_{\eta_f}\alpha_t, \quad (73)$$

where the inequalities follow from $c_{\eta_f}=\frac{2}{3L_f}+2\bar{c}_{\eta_f}$ with $\bar{c}_{\eta_f}$ in eq. (67). Then substituting eq. (73) into eq. (70) yields

$$\frac{1}{\bar{c}_{\eta_f}}\mathbb{E}\left[\frac{\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\|e_t^f\|^2}{\alpha_{t-1}}\right] \\ \leq-\alpha_t\mathbb{E}\|e_t^f\|^2+\frac{2(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2+\frac{\alpha_t}{16}\mathbb{E}\|\widetilde{h}_t^f\|^2+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2 \\ +\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2+\frac{8C_{g_{xy}}}{L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^R\|^2+C_{g_{xy}}\alpha_t\mathbb{E}\|v_t-v_t^*\|^2. \quad (74)$$

Step 4. Bound $\mathbb{E}\|e_t^g\|^2$ in eq. (66).

Next, from Lemma 16, we have

$$\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t}-\frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}}\leq\left[\frac{(1-\eta_{t+1}^g)^2+32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\right]\mathbb{E}\|e_t^g\|^2+\frac{2(\eta_{t+1}^g)^2}{\alpha_t}\sigma_g^2 \\ +16L_g^2\alpha_t\mathbb{E}\|\widetilde{h}_t^f\|^2+\frac{32L_g^2\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2, \quad (75)$$

where we use the fact that $0<1-\eta_t^g\leq 1$for all $t\in\{0,1,...,T-1\}$. Let us consider the coefficient of the first term on the right hand side of the above eq. (75). In specific, we have

$$\frac{(1-\eta_{t+1}^g)^2+32(1-\eta_{t+1}^g)^2L_g^2\beta_t^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\leq\frac{(1-\eta_{t+1}^g)}{\alpha_t}(1+32L_g^2\beta_t^2)-\frac{1}{\alpha_{t-1}} \\ =\frac{1}{\alpha_t}-\frac{1}{\alpha_{t-1}}+\frac{32L_g^2\beta_t^2}{\alpha_t}-c_{\eta_g}\alpha_t(1+32L_g^2\beta_t^2),$$

which, combined with eq. (72) that $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f}$ and the definition of $\beta_t = c_\beta \alpha_t$, yields

$$\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f} + 32L_g^2 c_\beta^2 \alpha_t - c_{\eta_g}\alpha_t. \tag{76}$$

Recall from eq. (67) that we choose

$$c_{\eta_g} = \frac{1}{3L_f} + 32L_g^2 c_\beta^2 + \frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\bar{c}_{\eta_g},$$

which, in conjunction with eq. (76), yields

$$\frac{(1-\eta_{t+1}^g)^2 + 32(1-\eta_{t+1}^g)^2 L_g^2 \beta_t^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq -\frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\bar{c}_{\eta_g}\alpha_t. \tag{77}$$

Substituting eq. (77) into eq. (75) yields

$$\begin{aligned}\frac{1}{\bar{c}_{\eta_g}}\left[\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}}\right] \leq & -\frac{17(L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2})}{L_{\mu_g}^2}\alpha_t \mathbb{E}\|e_t^g\|^2 + \frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2 \\ & + \frac{\alpha_t}{16}\mathbb{E}\|h_t^f\|^2 + \frac{L_{f_y}^2 + \frac{C_{f_y} L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g + L_g)}\alpha_t \mathbb{E}\|\nabla_y g(x_t, y_t)\|^2.\end{aligned} \tag{78}$$

Step 5. Bound $\mathbb{E}\|e_t^R\|^2$ in eq. (66).

Next, from Lemma 17, we have

$$\begin{aligned}\frac{\mathbb{E}\|e_{t+1}^R\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^R\|^2}{\alpha_{t-1}} \leq & \left[\frac{(1-\eta_{t+1}^R)^2(1+96L_g^2\lambda_t^2)}{\alpha_t} - \frac{1}{\alpha_{t-1}} + 4L_{g_{yy}} r_v^2 \delta_\epsilon\right]\mathbb{E}\|e_t^R\|^2 \\ & + \left[\frac{8(\eta_{t+1}^R)^2(\sigma_{g_{yy}}^2 r_v^2 + \sigma_{f_y}^2)}{\alpha_t}\right] + 96(1-\eta_{t+1}^R)^2(L_{g_{yy}}^2 r_v^2 + L_{f_y}^2)\alpha_t \mathbb{E}\|\widetilde{h}_t^f\|^2 \\ & + 96(1-\eta_{t+1}^R)^2(L_{g_{yy}}^2 r_v^2 + L_{f_y}^2)c_\beta^2\alpha_t(2\mathbb{E}\|e_t^g\|^2 + 2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2) \\ & + 96(1-\eta_{t+1}^R)^2 L_g^4 c_\lambda^2 \alpha_t \mathbb{E}\|v_t - v^*\|^2 + \frac{1}{\alpha_t}\big(4L_{g_{yy}} r_v^2 \delta_\epsilon + 8L_{g_{yy}}^2 r_v^4 \delta_\epsilon^2\big).\end{aligned} \tag{79}$$

For the first term of right hand side of eq. (79), recalling the δ_ϵ in eq. (67), we have

$$\begin{aligned}& \frac{(1-\eta_{t+1}^R)^2(1+96L_g^2\lambda_t^2)}{\alpha_t} - \frac{1}{\alpha_{t-1}} + \frac{4L_{g_{yy}} r_v^2 \delta_\epsilon}{\alpha_t} \\ & \leq \frac{1-\eta_{t+1}^R}{\alpha_t}(1+96L_g^2\lambda_t^2) - \frac{1}{\alpha_{t-1}} + \frac{4L_{g_{yy}} r_v^2 \delta_\epsilon}{\alpha_t} \\ & = \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\eta_{t+1}^R}{2\alpha_t} + \frac{1-\eta_{t+1}^R}{\alpha_t}\cdot 96L_g^2\lambda_t^2 \\ & \overset{(a)}{=} \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{c_{\eta_R}\alpha_t}{2} + \left(\frac{1}{\alpha_t} - c_{\eta_R}\alpha_t\right)\cdot 96L_g^2 c_\lambda^2 \alpha_t^2 \\ & \overset{(b)}{\leq} \frac{\alpha_t}{3L_f} + 96L_g^2 c_\lambda^2 \alpha_t - \frac{c_{\eta_R}\alpha_t}{2},\end{aligned} \tag{80}$$

where (a) follows from the definition that $\eta_{t+1}^R = c_{\eta_R}\alpha_t^2$, and (b) follows from eq. (72) that $\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{3L_f}$. Recalling $\bar{c}_{\eta_R}$ from eq. (67) that

$$c_{\eta_R} = \frac{2}{3L_f} + 192L_g^2 c_\lambda^2 + \frac{32C_{g_{xy}}}{L_{\mu_g}^2}\bar{c}_{\eta_R}$$

which, in conjunction with eq. (80), yields

$$\frac{(1-\eta_{t+1}^R)^2(1+96L_g^2\lambda_t^2)}{\alpha_t}-\frac{1}{\alpha_{t-1}}+\frac{4L_{g_{yy}}r_v^2\delta_\epsilon}{\alpha_t}v\leq-\frac{16C_{g_{xy}}}{L_{\mu_g}^2}\bar{c}_{\eta_R}. \tag{81}$$

Incorporating eq. (81) into eq. (79), recalling from eq. (67) that $\sigma_R:=\sqrt{\sigma_{g_{yy}}^2r_v^2+\sigma_{f_y}^2}$, and multiplying both sides of eq. (79) by $\frac{1}{\bar{c}_{\eta_R}}$, we have

$$\begin{aligned}\frac{1}{\bar{c}_{\eta_R}}\mathbb{E}\left[\frac{\|e_{t+1}^R\|^2}{\alpha_t}-\frac{\|e_{t+1}^R\|^2}{\alpha_{t-1}}\right]\leq&-\frac{16C_{g_{xy}}}{L_{\mu_g}^2}\alpha_t\|e_t^R\|^2+\frac{8(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2+\frac{\alpha_t}{16}\mathbb{E}\|\widetilde{h}_t^f\|^2\\&+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}^2}\alpha_t\mathbb{E}\|e_t^g\|^2+\frac{L_{f_y}^2+\frac{C_{f_y}L_{g_{xy}}^2}{\mu_g^2}}{4L_{\mu_g}(\mu_g+L_g)}\alpha_t\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2\\&+C_{g_{xy}}\alpha_t\|v_t-v_t^*\|^2+\frac{1}{\alpha_t\bar{c}_{\eta_R}}\big(4L_{g_{yy}}r_v^2\delta_\epsilon+8L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big).\end{aligned} \tag{82}$$

Step 6. Merging the results of Step 1-5 to prove eq. (66).

Finally, adding eq. (68), eq. (69) eq. (74), eq. (78), eq. (82) and the result of Lemma 13 with $\alpha_t\leq\frac{1}{3L_f}$ yields

$$\begin{aligned}\mathbb{E}[V_{t+1}-V_t]\leq&-\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t,\delta_\epsilon)\|^2+\frac{4(\eta_{t+1}^f)^2}{\bar{c}_{\eta_f}\alpha_t}\sigma_f^2+\frac{2(\eta_{t+1}^g)^2}{\bar{c}_{\eta_g}\alpha_t}\sigma_g^2+\frac{8(\eta_{t+1}^R)^2}{\bar{c}_{\eta_R}\alpha_t}\sigma_R^2\\&+\frac{1}{\alpha_t\bar{c}_{\eta_f}}\big(4L_{g_{xy}}r_v^2\delta_\epsilon+16L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+\frac{1}{\alpha_t\bar{c}_{\eta_R}}\big(4L_{g_{yy}}r_v^2\delta_\epsilon+8L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big).\end{aligned} \tag{83}$$

Then, the proof is complete. □

D.8 Proof of Theorem 1

Proof. Summing up the result of Lemma 19 for $t=0$ to $T-1$, dividing by T on both sides and using the definitions that $\eta_{t+1}^f:=c_{\eta_f}\alpha_t^2$, $\eta_{t+1}^g:=c_{\eta_g}\alpha_t^2$, $\eta_{t+1}^R:=c_{\eta_R}\alpha_t^2$, we have

$$\begin{aligned}\frac{\mathbb{E}[V_T-V_0]}{T}\leq&-\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2+\frac{1}{T}\left[\frac{4(c_{\eta_f})^2}{\bar{c}_{\eta_f}}\sigma_f^2+\frac{2(c_{\eta_g})^2}{\bar{c}_{\eta_g}}\sigma_g^2+\frac{8(c_{\eta_R})^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\sum_{t=0}^{T-1}\alpha_t^3\\&+\sum_{t=0}^{T-1}\frac{1}{\alpha_t\bar{c}_{\eta_f}}\big(4L_{g_{xy}}r_v^2\delta_\epsilon+16L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+\sum_{t=0}^{T-1}\frac{1}{\alpha_t\bar{c}_{\eta_R}}\big(4L_{g_{yy}}r_v^2\delta_\epsilon+8L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big).\end{aligned} \tag{84}$$

Next based on the definition of α_t in eq. (67), we have

$$\begin{aligned}\sum_{t=0}^{T-1}\alpha_t^3=\sum_{t=0}^{T-1}\frac{1}{w+t}\overset{(a)}{\leq}\sum_{t=0}^{T-1}\frac{1}{1+t}\leq\log(T+1)\\\sum_{t=0}^{T-1}\frac{1}{\alpha_t}=\sum_{t=0}^{T-1}\frac{1}{(w+t)^{1/3}}\overset{(a)}{\leq}\sum_{t=0}^{T-1}\frac{1}{(1+t)^{1/3}}\leq\frac{3}{2}T^{2/3}.\end{aligned} \tag{85}$$

where inequality (a) results from the fact that we choose $w\geq1$.

By plugging eq. (85) in eq. (84), we have

$$\begin{aligned}\frac{\mathbb{E}[V_T-V_0]}{T}\leq&-\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2+\left[\frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2+\frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2+\frac{8c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{T}\\&+T^{2/3}\big(6L_{g_{xy}}r_v^2\delta_\epsilon+24L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+T^{2/3}\big(6L_{g_{yy}}r_v^2\delta_\epsilon+12L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big).\end{aligned} \tag{86}$$

Rearrange the terms in eq. (86), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \leq \frac{\mathbb{E}[V_0-\Phi^*]}{T}+\left[\frac{4c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2+\frac{2c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2+\frac{8c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{T} + T^{2/3}\big(6L_{g_{xy}}r_v^2\delta_\epsilon+24L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+T^{2/3}\big(6L_{g_{yy}}r_v^2\delta_\epsilon+12L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big),$$

which, in conjunction with the fact that α_t is decreasing w.r.t. t and multiplying by $2/\alpha_T$ on both sides, yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(x_t)\|^2 \leq \frac{2\mathbb{E}[V_0-\Phi^*]}{\alpha_T T}+\left[\frac{8c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2+\frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2+\frac{16c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{\alpha_T T} + \frac{T^{2/3}}{\alpha_t}\big(6L_{g_{xy}}r_v^2\delta_\epsilon+24L_{g_{xy}}^2r_v^4\delta_\epsilon^2\big)+\frac{T^{2/3}}{\alpha_t}\big(6L_{g_{yy}}r_v^2\delta_\epsilon+12L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big). \quad (87)$$

Finally, based on the definition of the potential function, we have

$$\begin{aligned}\mathbb{E}[V_0] :=& \mathbb{E}\Big[\Phi(x_0)+\frac{2L}{3\sqrt{2}L_y}\|y_0-y^*(x_0)\|^2+\frac{4C_B}{C_\lambda L_{\mu_g}}\|v_0-v^*(x_0,y_0)\|^2 \\ &+\frac{1}{\bar{c}_{\eta_f}}\frac{\|e_t^f\|^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_g}}\frac{\|e_t^g\|^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_R}}\frac{\|e_t^R\|^2}{\alpha_{t-1}}\Big] \\ \leq& \Phi(x_0)+\frac{2L}{3\sqrt{2}L_y}\|y_0-y^*(x_0)\|^2+\frac{4C_B}{C_\lambda L_{\mu_g}}\|v_0-v^*(x_0,y_0)\|^2 \\ &+\frac{1}{\bar{c}_{\eta_f}}\frac{\sigma_f^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_g}}\frac{\sigma_g^2}{\alpha_{t-1}}+\frac{1}{\bar{c}_{\eta_R}}\frac{\sigma_R^2}{\alpha_{t-1}}\end{aligned} \quad (88)$$

where the inequality follows from Assumption 1, 2, 3, lemma 3 and the definitions of $\widetilde{h}_t^f$, h_t^g and $\widetilde{h}_t^R$ in eq. (10), eq. (6), eq. (7). Then, substituting eq. (88) into eq. (87) yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|^2 \leq& \frac{2[\Phi(x_0)-\Phi^*]}{\alpha_t T}+\frac{4L}{3\sqrt{2}L_y}\frac{\|y_0-y^*(x_0)\|^2}{\alpha_t T}+\frac{8C_B}{C_\lambda L_{\mu_y}}\frac{\|v_0-v^*(x_0,y_0)\|^2}{\alpha_t T} \\ &+\frac{2}{\alpha_{-1}\alpha_T T}\left(\frac{\sigma_f^2}{\bar{c}_{\eta_f}}+\frac{\sigma_g^2}{\bar{c}_{\eta_g}}+\frac{\sigma_R^2}{\bar{c}_{\eta_R}}\right)+\left[\frac{8c_{\eta_f}^2}{\bar{c}_{\eta_f}}\sigma_f^2+\frac{4c_{\eta_g}^2}{\bar{c}_{\eta_g}}\sigma_g^2+\frac{16c_{\eta_R}^2}{\bar{c}_{\eta_R}}\sigma_R^2\right]\frac{\log(T+1)}{\alpha_T T} \\ &+\frac{T^{2/3}}{\alpha_t}\big(6L_{g_{xy}}r_v^2\delta_\epsilon+24L_{g_{xy}}^2r_v^4\delta_\epsilon^2+6L_{g_{yy}}r_v^2\delta_\epsilon+12L_{g_{yy}}^2r_v^4\delta_\epsilon^2\big).\end{aligned} \quad (89)$$

Recalling δ_ϵ in eq. (67), we choose our δ_ϵ as

$$\delta_\epsilon \leq \min\left\{\frac{c_{\eta_f}}{8(L_{g_{xy}}r_v^2(w+T-1)^{2/3})},\frac{c_{\eta_R}}{8(L_{g_{xy}}r_v^2(w+T-1)^{2/3})},\frac{(w+T)^{\frac{1}{3}}}{12T^{\frac{4}{3}}L_{g_{xy}}r_v^2},\frac{(w+T)^{\frac{1}{3}}}{12T^{\frac{4}{3}}L_{g_{yy}}r_v^2}\right\}. \quad (90)$$

Substituting eq. (90) and the definitions of $\alpha_T := \frac{1}{(\omega+T)^{1/3}}$ and $\alpha_{-1}=\alpha_0$ into eq. (89), yields

$$\mathbb{E}\|\nabla\Phi\big(x_a(T)\big)\|^2 \leq \widetilde{\mathcal{O}}\Bigg(\frac{\Phi(x_0)-\Phi^*}{T^{2/3}}+\frac{\|y_0-y^*(x_0)\|^2}{T^{2/3}}+\frac{\|v_0-v^*(x_0,y_0)\|^2}{T^{2/3}} +\frac{1}{T^{2/3}}+\frac{\sigma_f^2}{T^{2/3}}+\frac{\sigma_g^2}{T^{2/3}}+\frac{\sigma_R^2}{T^{2/3}}\Bigg).$$

Finally, the proof is complete. □