

Tuning-Free Bilevel Optimization: New Algorithms and Convergence Analysis

Yifan Yang¹, Hao Ban¹, Minhui Huang², Shiqian Ma³, and Kaiyi Ji¹

¹Department of Computer Science and Engineering, University at Buffalo

²AI Research for Monetization, Meta

³Department of Computational Applied Math and Operations Research, Rice University

Oct 3, 2024

Abstract

Bilevel optimization has recently attracted considerable attention due to its abundant applications in machine learning problems. However, existing methods rely on prior knowledge of problem parameters to determine stepsizes, resulting in significant effort in tuning stepsizes when these parameters are unknown. In this paper, we propose two novel tuning-free algorithms, D-TFBO and S-TFBO. D-TFBO employs a double-loop structure with stepsizes adaptively adjusted by the "inverse of cumulative gradient norms" strategy. S-TFBO features a simpler fully single-loop structure that updates three variables simultaneously with a theory-motivated joint design of adaptive stepsizes for all variables. We provide a comprehensive convergence analysis for both algorithms and show that D-TFBO and S-TFBO respectively require $\mathcal{O}(\frac{1}{\epsilon})$ and $\mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$ iterations to find an ϵ -accurate stationary point, (nearly) matching their well-tuned counterparts using the information of problem parameters. Experiments on various problems show that our methods achieve performance comparable to existing well-tuned approaches, while being more robust to the selection of initial stepsizes. To the best of our knowledge, our methods are the first to completely eliminate the need for stepsize tuning, while achieving theoretical guarantees.

1 Introduction

Bilevel optimization has gained considerable attention recently due to its widespread use in various machine learning applications, such as meta-learning [23, 7, 66], hyperparameter optimization [73, 20], reinforcement learning [40, 32], robotics [82], communication [37] and federated learning [78]. In this paper, we study a standard bilevel optimization problem that takes the following mathematical formulation:

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}} \Phi(x) &:= f(x, y^*(x)) \\ \text{s.t. } y^*(x) &= \arg \min_{y \in \mathbb{R}^{d_y}} g(x, y), \end{aligned} \tag{1}$$

¹Email: {yyang99, haoban, kaiyiji}@buffalo.edu.

²Email: mhhuang@meta.com.

³Email: sqma@rice.edu.

where f and g are jointly continuously differentiable outer (upper-level) and inner (lower-level) functions. In this paper, we focus on the nonconvex-strongly-convex setting, where the lower-level function g is strongly convex w.r.t. y and the outer function $\Phi(x)$ is possibly nonconvex.

Recent years have witnessed the rapid development of bilevel optimization algorithms, which can be categorized into approximate implicit differentiation (AID) [36, 14] based, iterative differentiation (ITD) [35, 27] based, and value-function based [41, 48] approaches. However, these methods often require substantial effort to tune a couple of hyperparameters like stepsizes, which typically depend on **unknown** problem parameters (such as Lipschitzness parameters, strong convexity parameters, and optimal function values). This emphasizes the importance of *adaptive and tuning-free* methods in bilevel optimization. *In this paper, an algorithm is considered tuning-free if it does not need to know the problem parameters in advance but can still achieve almost the same convergence rate guarantee as its well-tuned counterpart using this information.* Despite several recent efforts to reduce dependence on problem-specific parameters [19, 2], developing a fully tuning-free bilevel optimization algorithm remains an open challenge. For instance, [19] utilizes Polyak’s stepsizes to automate both inner and outer updates but still requires information such as gradient Lipschitzness parameters and optimal lower-level function values. Similarly, [1] introduces an "on-the-fly" accumulation strategy for (hyper)gradient norms, which removes the reliance on inner and outer gradient Lipschitzness parameters but still depends on the strong convexity parameter for the inner AdaNGD-type updates.

This paper aims to close this gap by introducing two novel fully tuning-free bilevel optimization algorithms named D-TFBO and S-TFBO (where D and S represent double- and single-loop approaches), along with a comprehensive convergence analysis demonstrating their competitive performance compared to existing well-tuned approaches (which tune their hyperparameters like stepsizes based on the problem parameters). Our key contributions are outlined below.

- Our algorithms are inspired by the "inverse of cumulative gradient norms" strategy introduced by [86, 84], adapting the stepsizes based on accumulated (hyper)gradient norms. D-TFBO utilizes two optimization sub-loops: one for solving the inner problem and another for addressing a linear system (LS), which approximates the Hessian-inverse-vector product of each hypergradient. Unlike previous approaches, D-TFBO introduces cold-start adaptive stepsizes that accumulate gradients exclusively within the sub-loops. This method establishes a tighter lower bound on stepsizes, improving gradient complexity. In contrast, S-TFBO adopts a single-loop structure, where all variables are updated simultaneously in each iteration. Rather than applying the "inverse of cumulative gradient norms" uniformly to all updates, our error analysis motivates a joint design of adaptive stepsizes for y , v , and x , which correspond to solving the inner problem, LS, and outer problem, respectively. For instance, the stepsize for v is coupled with that for y , while the stepsize for x depends on both y and v .
- Compared to the well-tuned AID methods in [35], our D-TFBO method achieves the same $\mathcal{O}(\frac{1}{T})$ convergence rate. Similarly, our S-TFBO method attains an $\tilde{\mathcal{O}}(\frac{1}{T})$ convergence rate, matching that of well-tuned counterparts, up to polylogarithmic factors. The complexity analysis shows that D-TFBO and S-TFBO require $\mathcal{O}(\frac{1}{\epsilon^2})$ and $\tilde{\mathcal{O}}(\frac{1}{\epsilon})$ gradient computations, respectively, to reach an ϵ -accurate stationary point. This comparison differs from the observation in well-tuned bilevel optimization, where double-loop approaches generally achieve lower gradient complexity than single-loop methods [35]. This is because the inner tuning-free solver requires $\mathcal{O}(\frac{1}{\epsilon})$ more iterations than well-tuned methods to achieve ϵ -level accuracy.
- The theoretical analysis is inspired by the two-stage framework in [86, 84], where the stages describe the

relationship between the stepsizes and certain constants that depend on the problem parameters. However, exploring this technical framework in bilevel problems is far more challenging because the stages for analyzing each stepsize interact with those for other stepsizes, resulting in intertwined multi-stage dynamics across different variables. For instance, the error analysis for the updates on v must account for the accumulated gradient norms from the updates on y . This motivates us to couple the stepsize for v with the adaptive stepsize for y to prevent the propagation of accumulated errors. In addition, our analysis requires establishing precise upper and lower bounds for all stepsizes to ensure convergence results that match those achieved under well-tuned stepsizes.

- We validate the effectiveness of our methods through experiments on regularization selection, data hyper-cleaning, and coresets selection for continual learning. The results show that our methods perform comparably to existing well-tuned methods. More importantly, our methods demonstrate greater robustness to different initial stepsizes, due to the tuning-free design.

2 Related Work

Bilevel Optimization. Bilevel optimization, initially introduced by [9], has been extensively studied for decades. Early works [29, 75] approached the bilevel problem from a constrained optimization perspective. More recently, gradient-based methods have gained significant attention for their efficiency and effectiveness. Among these, Approximate Implicit Differentiation (AID) methods [17, 46, 36, 14] leverage the implicit derivation of the hypergradient by approximating it through the solution of a linear system. In contrast, Iterative Differentiation (ITD) methods [54, 22] estimate the hypergradient using automatic differentiation, employing either forward or reverse mode. Recently, a range of stochastic bilevel methods have been developed and analyzed, using techniques such as Neumann series [12, 36], recursive momentum [87, 28], and variance reduction [87]. Another class of methods formulates the lower-level problem as a value-function-based constraint [41, 83], enabling the solution of bilevel problems without the need for second-order gradients. A more detailed discussion of related work can be found in the Appendix.

Adaptive and Tuning-free Algorithms. Adaptive gradient descent has achieved remarkable success and is widely studied and applied in modern machine learning. Early adaptive algorithms trace back to line search methods, such as backtracking [25], and Polyak’s stepsize [65], both of which have inspired numerous recent variants [4, 6, 71, 81, 31, 51, 63]. To reduce the computational cost of line search and avoid the reliance on an unknown optimal function value, the Barzilai-Borwein stepsize [5, 67, 15] was introduced, drawing inspiration from quasi-Newton schemes. Normalized gradient descent [13, 61, 60] preserves the direction of the gradient while ignoring its magnitude, removing the need for prior knowledge about the function. [18] and [58] pioneered AdaGrad, an adaptive gradient-based method, which proved efficient in solving online convex optimization problems. AdaGrad rapidly evolved for deep learning applications, giving rise to numerous methods, including popular variants like Adam [16, 69, 53, 85], RMSprop [79], and Adadelta [89]. Specifically, normalized versions of AdaGrad, such as AdaNGD $_k$ [42], AcceleGrad [43], and AdaGrad-Norm [84, 86], introduced adaptive stepsizes that require no problem-specific parameters, making them tuning-free approaches. Recent work by [55] further established lower bounds for minimizing the deterministic gradient l_1 -norm. Additional methods, such as Lipschitzness parameter approximation [56] and restart techniques [57], have also been explored. A more comprehensive discussion refers to [38].

Adaptive and Tuning-free Bilevel Algorithms. Instead of focusing on single-level problems, [34] extended Adam to bilevel optimization algorithms. [19] introduced adaptive stepsizes for bilevel problems, based

on Polyak’s stepsize and line search techniques. Most recently, [2] proposed a novel framework that applies adaptive normalized gradient descent to the strongly convex inner problem and AdaGrad-Norm to the nonconvex outer problem, allowing the algorithm to update adaptively with fewer problem-specific parameters.

3 Algorithm

3.1 Standard Bilevel Optimization

A key challenge in bilevel optimization is calculating the hypergradient $\nabla\Phi(x)$, which, according to the implicit function theorem, is given by:

$$\nabla\Phi(x) = \nabla_x f(x, y^*(x)) - \nabla_x \nabla_y g(x, y^*(x)) [\nabla_y \nabla_y g(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)),$$

when g is twice differentiable, $\nabla_y g$ is continuously differentiable and the Hessian $\nabla_y \nabla_y g(x, y^*(x))$ is invertible. In practice, $y^*(x)$ is not directly accessible, and one often use an iterative algorithm to obtain an estimate \hat{y} instead. Since computing the Hessian inverse is prohibitively expensive, a more efficient way is to approximate the Hessian-inverse-vector product in the above hypergradient $\nabla\Phi(x)$ by solving the following linear system:

$$\min_v R(x, \hat{y}, v) = \frac{1}{2} v^T \nabla_y \nabla_y g(x, \hat{y}) v - v^T \nabla_y f(x, \hat{y}). \quad (2)$$

Similarly, an iterative algorithm is usually deployed to obtain an approximate solution \hat{v} of the problem in eq. (2). Given the approximates \hat{y} and \hat{v} , the variable x is then updated with a hypergradient estimate given by

$$\bar{\nabla} f(x, \hat{y}, \hat{v}) = \nabla_x f(x, \hat{y}) - \nabla_x \nabla_y g(x, \hat{y}) \hat{v}. \quad (3)$$

Standard bilevel optimization approaches select the stepsizes for updating y , v , and x based on problem-specific parameters, such as Lipschitzness and strong convexity parameters [14, 36, 35]. However, these parameters are often difficult to obtain or approximate in practice, leading to significant tuning efforts. This challenge motivates the development of adaptive bilevel optimization algorithms that require less to no tuning.

3.2 Existing Adaptive Bilevel Optimization Methods

Among the existing adaptive bilevel methods, the most closely related to this work are [19] and [2]. [19] utilizes Polyak’s stepsizes and a line search to automate the stepsizes for both inner and outer updates. [2] applies AdaNGD [42] to solve the inner problem and updates x using the inverse of cumulative hypergradient norms, where the hypergradient norms are approximated via gradient mapping [61] with Fenchel coupling [59].

However, these methods are not entirely tuning-free. For instance, the initialization of Polyak’s stepsizes in [19] depends on Lipschitzness parameters, strong convexity parameters, and the optimal lower-level function values. While the line search approach in [19] bypasses the need for problem-specific parameters, it lacks theoretical convergence guarantees. Similarly, [1] requires the strong convexity parameter for the inner AdaNGD updates.

Algorithm 1 Double-loop Tuning-Free Bilevel Optimizer (D-TFBO)

1: **Input:** initialization $x_0, y_0, v_0, \alpha_0 > 0, \beta_0 > 0, \gamma_0 > 0$, total iteration rounds T , and $\epsilon_y = \epsilon_v = \frac{1}{T}$
2: **for** $t = 0, 1, 2, \dots, T - 1$ **do**
3: $p = 0, q = 0$, set $y_t^0 = y_{t-1}^0, v_t^0 = v_{t-1}^0$ if $t > 0$ and y_0, v_0 otherwise
4: **while** $\|\nabla_y g(x_t, y_t^p)\|^2 > \epsilon_y$ **do**
5: $\beta_{p+1}^2 = \beta_p^2 + \|\nabla_y g(x_t, y_t^p)\|^2, y_t^{p+1} = y_t^p - \frac{1}{\beta_{p+1}} \nabla_y g(x_t, y_t^p), p = p + 1$
6: **end while**
7: $P_t = p$
8: **while** $\|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2 > \epsilon_v$ **do**
9: $\gamma_{q+1}^2 = \gamma_q^2 + \|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2, v_t^{q+1} = v_t^q - \frac{1}{\gamma_{q+1}} \nabla_v R(x_t, y_t^{P_t}, v_t^q), q = q + 1$
10: **end while**
11: $Q_t = q$
12: $\alpha_{t+1}^2 = \alpha_t^2 + \|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2, x_{t+1} = x_t - \frac{1}{\alpha_{t+1}} \bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})$
13: **end for**

3.3 Double-Loop Tuning-Free Bilevel Optimization- D-TFBO

As shown in Algorithm 1, our D-TFBO method follows a double-loop structure, where two sub-loops of iterations are used to solve the lower-level and linear system problems. In the first sub-loop, we employ the idea of "inverse of cumulative gradient norm" to design the adaptive updates as

$$y_t^{p+1} \leftarrow y_t^p - \frac{1}{\beta_{p+1}} \nabla_y g(x_t, y_t^p), \quad \text{with } \beta_{p+1}^2 = \beta_p^2 + \|\nabla_y g(x_t, y_t^p)\|^2.$$

It can be seen from Algorithm 1 that our D-TFBO algorithm employs a stopping criterion based on the gradient norm: $\|\nabla_y g(x_t, y_t^p)\|^2 \leq \epsilon_y$, where ϵ_y (**defaulted to $1/T$ for convergence analysis**) is independent of problem parameters. The rationale behind this design is that if the stopping criterion is not met (i.e., $\|\nabla_y g(x_t, y_t^p)\|^2 > \epsilon_y$), the accumulation β_p of gradient norms continues to increase. This increase causes the stepsize $\frac{1}{\beta_p}$ to decrease to a value at which a descent in the optimality gap is guaranteed. A similar stopping criterion applies to the updates of v_t^q when solving the linear system.

Notably, both sub-loops utilize warm-start variable values but reset the stepsizes at each iteration (cold-start stepsizes). The warm-start variables ensure that the initial point is reasonably close to the optimal solution, while the cold-start scheme guarantees stepsizes to achieve stronger lower bounds. Finally, the update of x_t is based on the accumulation of hypergradient estimates $\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})$.

Remark 1 (Extension to a tunable version with problem-parameter-free tuning coefficients.). *Although Algorithm 1 is designed as a tuning-free method, a tunable version with the flexibility to preset hyperparameters can still achieve the same convergence rate and gradient complexity. The stepsizes for $\{x, y, v\}$ can be set as $\{\eta_x/\alpha_t, \eta_y/\beta_p, \eta_v/\gamma_q\}$ and the sub-loops stopping criteria can be set to $\{c_y/T, c_v/T\}$, where $\{\eta_x, \eta_y, \eta_v, c_y, c_v\}$ are configurable hyperparameters that are independent of the problem parameters such as strong-convexity and Lipschitzness parameters.*

3.4 Single-Loop Tuning-Free Bilevel Optimization- S-TFBO

The two sub-loops in D-TFBO may complicate the implementation, and increase the number of iterations to meet the stopping criterion. In this section, we propose a much simpler fully single-loop tuning-free bilevel optimization method named S-TFBO, as described in Algorithm 2.

Algorithm 2 Single-loop Tuning-Free Bilevel Optimizer (S-TFBO)

1: **Input:** initialization $x_0, y_0, v_0, \alpha_0 \geq 1, \beta_0 > 0, \gamma_0 > 0$, number of iteration rounds T
2: **for** $t = 0, 1, 2, \dots, T - 1$ **do**
3: $\beta_{t+1}^2 = \beta_t^2 + \|\nabla_y g(x_t, y_t)\|^2$
4: $\gamma_{t+1}^2 = \gamma_t^2 + \|\nabla_v R(x_t, y_t, v_t)\|^2$
5: $\varphi_{t+1} = \max\{\beta_{t+1}, \gamma_{t+1}\}$
6: $\alpha_{t+1}^2 = \alpha_t^2 + \|\bar{\nabla} f(x_t, y_t, v_t)\|^2$
7: $y_{t+1} = y_t - \frac{1}{\beta_{t+1}} \nabla_y g(x_t, y_t)$
8: $v_{t+1} = v_t - \frac{1}{\varphi_{t+1}} \nabla_v R(x_t, y_t, v_t)$
9: $x_{t+1} = x_t - \frac{1}{\alpha_{t+1}\varphi_{t+1}} \bar{\nabla} f(x_t, y_t, v_t)$
10: **end for**

The design of stepsizes in Algorithm 2 follows a similar idea in Algorithm 1. In each iteration t , we update $\alpha_t, \beta_t, \gamma_t$ as accumulations of gradient norms of $\bar{\nabla} f, \nabla_y g$, and $\nabla_v R$ from the previous $t - 1$ iterations. We then update variables y_t, v_t and x_t simultaneously with adaptive stepsizes $\left\{ \frac{1}{\beta_t}, \frac{1}{\max\{\beta_t, \gamma_t\}}, \frac{1}{\alpha_t \max\{\beta_t, \gamma_t\}} \right\}$. However, the stepsizes for v and x are not straightforward and require careful designs guided by our theoretical analysis, as elaborated below.

Design of stepsize for v_t . Instead of simply using $\frac{1}{\gamma_t}$, we introduce $\frac{1}{\varphi_t} := \frac{1}{\max\{\beta_t, \gamma_t\}}$ as the stepsize. This adjustment is necessary because $\nabla_v R(x_t, y_t, v_t)$ involves the approximation error $\|y_t - y^*(x_t)\|^2$. Since this error is proportional to $\|\nabla_y g(x_t, y_t)\|^2$, using $\frac{1}{\beta_t}$ helps control this error and prevents it from exploding after accumulation, as validated in our theoretical analysis later.

Design of stepsize for x_t . Similarly, we use $\frac{1}{\alpha_t \varphi_t}$ as the stepsize for updating x_t , where the coupled factor $\frac{1}{\varphi_t}$ is introduced to mitigate the approximation errors from the y_t and v_t updates, leading to a more stable convergence.

Remark 2 (Extension to a tunable version with problem-parameter-free tuning coefficients.). *Similarly to Remark 1, Algorithm 2 can extend to a tunable version with the same convergence rate and gradient complexity. The stepsizes for $\{x, y, v\}$ can be set as $\{\eta_x/\alpha_t\varphi_t, \eta_y/\beta_t, \eta_v/\varphi_t\}$, where $\{\eta_x, \eta_y, \eta_v\}$ are configurable hyperparameters that are independent of the problem parameters.*

4 Theoretical Analysis

4.1 Technical Challenges

Compared to existing single-level tuning-free approaches, fully tuning-free bilevel optimization poses unique challenges that have not been addressed well.

- Compared to single-level problems, bilevel problems involve interdependent variable updates, resulting in more complex and interconnected stepsize designs.
- The stages for analyzing each stepsize interact with those of other stepsizes, leading to intertwined multi-stage dynamics across various variables.
- The optimization error of each variable can accumulate (hyper)gradient norms from previous iterations due to the adaptive stepsize designs, complicating the error analysis.

In Section 4.2, we introduce the standard definitions and assumptions. Next, in Section 4.3 and 4.4, we provide a detailed convergence analysis, explaining how we address the above challenges.

4.2 Assumptions and Definitions

We make the following definitions and assumptions for outer- and inner-objective functions, as also adopted by [24, 12, 39].

Definition 1. A mapping f is L -Lipschitz continuous if $\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|$ for $\forall x_1, x_2$.

Since the outer objective function $\Phi(x)$ is non-convex, we aim to find an ϵ -accurate stationary point, as defined below.

Definition 2. An output \bar{x} of an algorithm is the ϵ -accurate stationary point of the objective function $\Phi(x)$ if $\|\nabla\Phi(\bar{x})\|^2 \leq \epsilon$, where $\epsilon \in (0, 1)$.

Assumption 1. Functions $f(x, y)$ and $g(x, y)$ are twice continuously differentiable and $g(x, y)$ is μ strongly convex w.r.t. y , for $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$.

The following assumption imposes the Lipschitz continuity on the outer and inner functions and their derivatives.

Assumption 2. Function $f(x, y)$ is $L_{f,0}$ -Lipschitz continuous; the gradients $\nabla f(x, y)$ and $\nabla g(x, y)$ are $L_{f,1}$ and $L_{g,1}$ -Lipschitz continuous, respectively; the second-order gradients $\nabla_x \nabla_y g(x, y)$ and $\nabla_y \nabla_y g(x, y)$ are $L_{g,2}$ -Lipschitz continuous.

Rather than directly using the Lipschitz continuity parameters as bounds on gradients-which can cause dimensional inconsistencies during logarithmic operations-we offer the following remark:

Remark 3. Assumption 2 indicates that there exist constants C_{f_x} , C_{f_y} , $C_{g_{xy}}$ and $C_{g_{yy}}$ such that $\|\nabla_x f(x, y)\| \leq C_{f_x}$, $\|\nabla_y f(x, y)\| \leq C_{f_y}$, $\|\nabla_x \nabla_y g(x, y)\| \leq C_{g_{xy}}$ and $\|\nabla_y \nabla_y g(x, y)\| \leq C_{g_{yy}}$.

Assumption 3. There exists $m \in \mathbb{R}$ such that $\inf_x \Phi(x) \geq m$.

Next, we present the main convergence theorems for Algorithm 1 and Algorithm 2, along with key propositions that provide insights into these theorems. **A proof sketch is provided in Appendix C.**

4.3 Convergence and Complexity Analysis for Algorithm 1

Firstly, we explain the two-stage framework used in our analysis.

Proposition 1. Suppose the iteration rounds to update $\{x, y, v\}$ are $\{T_1, T_2, T_3\}$ and $\{\alpha_t, \beta_t, \gamma_t\}$ are generated by Algorithm 1 or 2. For any $C_\alpha \geq \alpha_0$, $C_\beta \geq \beta_0$, $C_\gamma \geq \gamma_0$, we have

- (a) either $\alpha_t \leq C_\alpha$ for any $t \leq T_1$, or $\exists k_1 \leq T_1$ such that $\alpha_{k_1} \leq C_\alpha$, $\alpha_{k_1+1} > C_\alpha$;
- (b) either $\beta_t \leq C_\beta$ for any $t \leq T_2$, or $\exists k_2 \leq T_2$ such that $\beta_{k_2} \leq C_\beta$, $\beta_{k_2+1} > C_\beta$;
- (c) either $\gamma_t \leq C_\gamma$ for any $t \leq T_3$, or $\exists k_3 \leq T_3$ such that $\gamma_{k_3} \leq C_\gamma$, $\gamma_{k_3+1} > C_\gamma$.

The analysis for each stepsize is divided into two cases. Let us take (a) as an illustration example. Case 1: the accumulation α_t of gradient norms is bounded by a constant C_α before the end of the iteration. In this case, the average gradient norm square can be bound as $\frac{C_\alpha^2}{T_1}$, which decreases with T_1 . Case 2: the accumulation α_{T_1} exceeds C_α , and hence α_t experiences two stages: in stage 1, $\alpha_t \leq C_\alpha$, and in stage 2, $\alpha_t > C_\alpha$. The error analysis for stage 1 is similar to that of case 1. In stage 2, the stepsizes are small enough to show the gradient norm decreases via a descent lemma.

Proposition 2. *Recall that for t_{th} iteration, the sub-loops in Algorithm 1 aim to find $y_t^{P_t}$ and $v_t^{Q_t}$ such that $\|\nabla_y g(x_t, y_t^{P_t})\|^2 \leq \epsilon_y$ and $\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \leq \epsilon_v$. Under Assumptions 1, 2, we have*

$$\begin{cases} P_t \leq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1 + \epsilon_y/C_\beta^2)} + \frac{\beta_{\max}}{\mu} \log\left(\frac{L_{g,1}^2(\beta_{\max} - C_\beta)}{\epsilon_y}\right), \\ Q_t \leq \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1 + \epsilon_v/C_\gamma^2)} + \frac{\gamma_{\max}}{\mu} \log\left(\frac{C_{g,y}^2(\gamma_{\max} - C_\gamma)}{\epsilon_v}\right), \end{cases}$$

where $\{C_\beta, C_\gamma\}$, β_{\max} , γ_{\max} are denied in eq. (5), eq. (22), eq. (29) in the Appendix, respectively.

Proposition 2 provides upper bounds on P_t and Q_t , which correspond to the total numbers of iterations of the two sub-loops. This result is the same as that of the standard AdaGrad-Norm in the strongly convex setting [86]. For the sub-loop for y , in Case 1 above, the loop terminates within $\log(C_\beta^2/\beta_0^2)/\log(1 + \epsilon_y/C_\beta^2)$ steps; and in Case 2, it takes at most $\log(C_\beta^2/\beta_0^2)/\log(1 + \epsilon_y/C_\beta^2)$ steps for stage 1 and it takes at most $\frac{\beta_{\max}}{\mu} \log(L_{g,1}^2(\beta_{\max} - C_\beta)/\epsilon_y)$ steps for stage 2. For ϵ_y small enough, it can be seen that P_t takes an order of $1/\epsilon_y$, which is typically larger than those obtained with well-tuned stepsizes. Based on this proposition, we can derive the following convergence results.

Theorem 1. *Suppose Assumptions 1,2,3 are satisfied. By setting $\epsilon_y = 1/T$ and $\epsilon_v = 1/T$, the iterates generated by Algorithm 1 satisfy*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 \leq \frac{c_1(C_\alpha + 2c_1)}{T} = \mathcal{O}\left(\frac{1}{T}\right),$$

where C_α and c_1 are constants defined in eq. (5) and eq. (37), respectively.

Corollary 1. *Under the same setting Theorem 1, to achieve an ϵ -accurate stationary point, Algorithm 1 needs $T = \mathcal{O}(1/\epsilon)$, $\{P_t, Q_t\} = \mathcal{O}(1/\epsilon)$, and the gradient complexity (i.e., the number of gradient evaluations) is $\text{Gc}(\epsilon) = \mathcal{O}(1/\epsilon^2)$.*

Theorem 1 shows that the convergence rate of Algorithm 1 matches that of the standard double-loop bilevel algorithms [36, 35]. According to Proposition 2, the sub-loops for updating y and v require $\mathcal{O}(1/\epsilon_y)$ iterations to ensure an ϵ_y -level approximation accuracy, which is worse than the $\mathcal{O}(1)$ results achieved by well-tuned bilevel optimization methods. This is because more iterations are needed to ensure high accuracy in both sub-loops, due to the lack of information about the Lipschitzness parameters and strong convexity parameters. Consequently, the gradient complexity of our D-TFBO method is worse than those of well-tuned double-loop methods by an order of $1/\epsilon$.

4.4 Convergence and Complexity Analysis for Algorithm 2

Differently from D-TFBO that uses sub-loops to achieve high-accurate y and v iterates, the main challenge for analyzing S-TFBO lies in dealing with the accumulated approximations errors for updating all variables over iterations. In the following propositions, we will show how we upper-bound such cumulative approximation errors and lower-bound the adaptive stepsizes.

First, we present a descent result for the objective function $\Phi(\cdot)$.

Proposition 3. *Under Assumptions 1, 2, for Algorithm 2, suppose the total iteration number is T . No matter k_1 in Proposition 1 exists or not, we always have*

$$\begin{aligned} \Phi(x_{t+1}) - \Phi(x_t) \leq & -\frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \left(1 - \frac{L_\Phi}{\alpha_{t+1}\varphi_{t+1}}\right) \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ & + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1}\varphi_{t+1}} + \frac{\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1}\varphi_{t+1}}. \end{aligned}$$

If in addition, k_1 in Proposition 1 exists, then for $t \geq k_1$, we further have

$$\begin{aligned} \Phi(x_{t+1}) - \Phi(x_t) \leq & -\frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{4\alpha_{t+1}\varphi_{t+1}} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ & + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1}\varphi_{t+1}} + \frac{\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1}\varphi_{t+1}}, \end{aligned}$$

where $\bar{L} := \max \left\{ 2(C_{f_y}^2 L_{g,2}^2 / \mu^2 + L_{f,1}^2)^{\frac{1}{2}}, \sqrt{2}C_{g_{yy}} \right\}$.

It can be seen from Proposition 1 that we derive two distinct forms of descent results for the objective function based on the relationship between α_{t+1} and C_α (whose form is specified in eq. (41) in the appendix). Their key difference is that the second inequality is tighter for the case $t \geq k_1$ by eliminating a term of $\frac{L_\Phi}{2\alpha_{t+1}^2\varphi_{t+1}^2} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2$. Both upper bounds consist of two parts: (i) the approximation errors $\mathcal{O}(\|\nabla_y g(x_t, y_t)\|^2 + \|\nabla_v R(x_t, y_t, v_t)\|^2) / (\alpha_{t+1}\varphi_{t+1})$ induced by the updates on y and v ; (ii) the descent term $-\|\nabla\Phi(x_t)\|^2 / (\alpha_{t+1}\varphi_{t+1})$. It can be seen that there exists a trade-off: a smaller $\alpha_t\varphi_t$ leads to a more notable descent, but larger approximation errors. However, due to the lack of information about the problem parameters, the value of $\alpha_t\varphi_t$ remains unknown, making it infeasible to determine the optimal trade-off. Instead, we adjust this trade-off based on an overall bound on the descent and approximation errors, derived by telescoping all descent inequalities.

Next, we investigate the upper bounds on the summations of the positive error terms in Proposition 3.

Proposition 4. *Under Assumptions 1, 2, for any $0 \leq k_0 < t$, for the positive error terms in Proposition 3, we have the upper bounds in terms of logarithmic functions as*

$$\sum_{k=k_0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \leq a_2 \log(t+1) + b_2, \quad \sum_{k=k_0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \leq a_3 \log(t+1) + b_3,$$

where a_2, b_2, a_3, b_3 are defined in eq. (75) in the Appendix.

Proposition 5. *Under Assumptions 1, 2, 3, suppose the total iteration rounds is T . For any case in Proposition 1, we have the upper-bound of φ_t and α_t in Algorithm 2 as*

$$\varphi_t \leq a_1 \log(t) + b_1, \quad \alpha_t \leq C_\alpha + (a_4 \log(t) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)))\varphi_t,$$

where a_1, b_1 are defined in eq. (65) and a_4, b_4 are defined in eq. (79) in the Appendix.

Proposition 4 provides the upper bounds on the accumulated positive error terms in Proposition 3, and Proposition 5 shows that the cumulative gradient norms for all variables increase only logarithmically. By rearranging the terms and taking the average, we have the upper bound for the average squared hypergradient norm $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\|^2$, establishing the final convergence rate of Algorithm 2, as shown in the following theorem and corollary.

Theorem 2. *Suppose Assumptions 1,2,3 are satisfied. The iterates generated by Algorithm 2 satisfy*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\|^2 &\leq \frac{1}{2T} \left[\left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right)^2 (a_1 \log(T) + b_1)^2 \right. \\ &\quad \left. + C_\alpha \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) (a_1 \log(T) + b_1) \right] = \mathcal{O}\left(\frac{\log^4(T)}{T}\right), \end{aligned}$$

where $\{C_\alpha, a_1, b_1, a_4, b_4\} = \mathcal{O}(1)$ are defined in eq. (41), eq. (65), eq. (79) in the Appendix.

Corollary 2. *Under the same setting Theorem 2, to achieve an ϵ -accurate stationary point, Algorithm 2 needs $T = \mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$ and the gradient complexity is $\text{Gc}(\epsilon) = \mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$.*

Theorem 2 shows that the proposed Algorithm 2 achieves a convergence rate of $\mathcal{O}(\log^4(T)/T)$ and a gradient complexity of $\mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$, both of which nearly match the results in [35] of the standard well-tuned bilevel optimization methods up to polylogarithmic factors.

Remark 4. *Note that the difference of $\frac{1}{\epsilon}$ in gradient complexity between double-loop and single-loop methods has not been observed in previous works on well-tuned bilevel optimization. This difference stems from the design of the sub-loops. In previous double-loop works, carefully selected stepsizes are used to ensure that the iterates of each sub-loop converge linearly, up to an approximation error caused by the shift in x . However, due to the precise control of stepsizes, tuning-free approaches can only guarantee a sub-linear convergence for each sub-loop (as shown in Proposition 2).*

5 Experiments

In this section, we evaluate the effectiveness of our proposed algorithm on practical applications including regularization selection, data hyper-cleaning [22], and coresets selection for continual learning [30]. Our implementation is based on the benchmark provided in [14] and [30], respectively. *Please refer to Appendix B for more details about practical implementation, experiment configurations, and additional plots.*

5.1 Regularization Selection

The selection of regularization can be framed as a bilevel optimization problem, where the inner objective focuses on optimizing the model parameters θ on the training set $\mathcal{S}_T = \{(d_i^{train}, y_i^{train})\}_{1 \leq i \leq n}$, while the outer objective aims to determine the best regularization term λ on the validation set $\mathcal{S}_V = \{(d_j^{val}, y_j^{val})\}_{1 \leq j \leq m}$. Denote the model parameters by $\theta \in \mathbb{R}^p$ and regularization term by $\lambda \in \mathbb{R}^p$, then the outer and inner problems can be formulated as

$$f(\theta, \lambda) = \frac{1}{m} \sum_{j=1}^m l((d_j^{val}, y_j^{val}), \theta); \quad g(\theta, \lambda) = \frac{1}{n} \sum_{i=1}^n l((d_i^{train}, y_i^{train}), \theta) + \mathcal{R}(\theta, \lambda),$$

where the loss $l((d_i, y_i), \theta) = \log(1 + \exp(-y_i d_i^\top \theta))$, and $\mathcal{R}(\theta, \lambda) = \frac{1}{2} \sum_{k=1}^p \exp(\lambda_k) \theta_k^2$ represents the regularization, where each element θ_k is regularized with strength $\exp(\lambda_k)$. We compare our proposed algorithm with benchmark bilevel algorithms including AmIGO [3], BSA [24], FSLA [45], MRBO [87], SOBA [14], StocBiO [36], SUSTAIN [39], TTSA [33], VRBO [87] on the Covtype dataset. More details are provided in Appendix B.

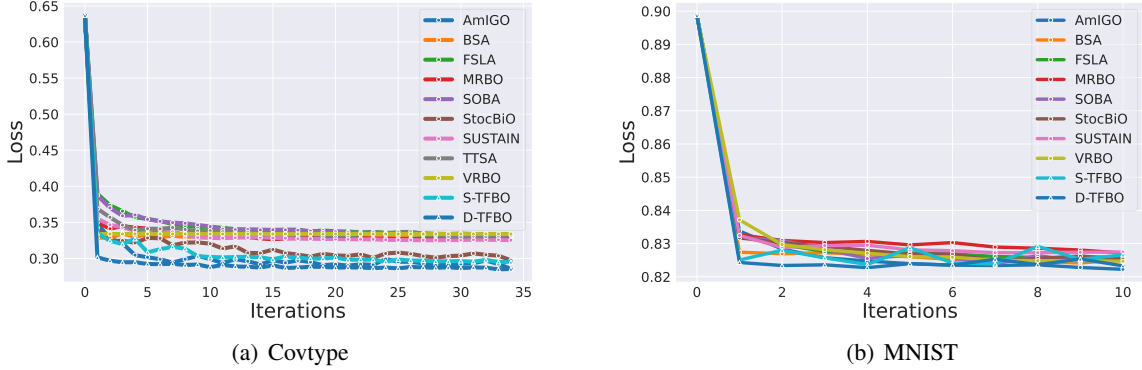


Figure 1: Comparison with other bilevel methods. (a) Regularization selection on Covtype dataset. (b) Data hyper-cleaning on MNIST dataset.

As shown in Figure 1(a), our D-TFBO achieves the fastest convergence rate, while S-TFBO converges slightly more slowly but remains comparable to other well-tuned methods.

5.2 Data Hyper-Cleaning

The training set $\mathcal{S}_T = \{(d_i^{train}, y_i^{train})\}_{1 \leq i \leq n}$ have been corrupted in this scenario, where the label of a data sample could be replaced by a random label with a certain probability p . It is important to note that we do not have prior knowledge about which data samples have been corrupted. The objective is to develop a model that can effectively fit the corrupted training set while performing well on the clean validation set $\mathcal{S}_V = \{(d_j^{val}, y_j^{val})\}_{1 \leq j \leq m}$. We conduct experiments on the MNIST dataset, where we aim to learn a set of weights λ , one for each training sample, in addition to the model parameters θ . Hence, the outer and inner problems are

$$f(\theta, \lambda) = \frac{1}{m} \sum_{j=1}^m l((d_j^{val}, y_j^{val}), \theta); \quad g(\theta, \lambda) = \frac{1}{n} \sum_{i=1}^n \sigma(\lambda_i) l((d_i^{train}, y_i^{train}), \theta) + C \|\theta\|^2,$$

where $\sigma(\cdot)$ is sigmoid function, C is a regularization constant, and loss function $l((d_i, y_i), \theta) = 1/(1 + \exp(-y_i d_i^\top \theta))$. Ideally, we would like the weights to be 0 for the corrupted sample and 1 for the clean sample. More details can be found in Appendix B. We compare the performance with other bilevel optimization methods including AmIGO [3], BSA [24], FSLA [45], MRBO [87], SOBA [14], StocBiO [36], SUSTAIN [39], VRBO [87]. The results presented in Figure 1(b) demonstrate that our algorithms achieve a convergence rate comparable to other baselines.

Table 1: Results on Split CIFAR100. The best and second best results are in bold and underlined.

| Method | Balanced | | Imbalanced | |
|----------------------|-------------------|--------------------|-------------------|--------------------|
| | A_T | FGT_T | A_T | FGT_T |
| k -means features | 57.82±0.69 | 0.070±0.003 | 45.44±0.76 | 0.037±0.002 |
| k -means embedding | 59.77±0.24 | 0.061±0.001 | 43.91±0.15 | 0.044±0.001 |
| Uniform Sampling | 58.99±0.54 | 0.074±0.004 | 44.73±0.11 | 0.033±0.007 |
| iCaRL | 60.74±0.09 | 0.044±0.026 | 44.25±2.04 | 0.042±0.019 |
| Grad Matching | 59.17±0.38 | 0.067±0.003 | 45.44±0.64 | 0.038±0.001 |
| GCR | 58.73±0.43 | 0.073±0.013 | 44.48±0.05 | 0.035±0.005 |
| Greedy Coreset | 59.39±0.16 | 0.066±0.017 | 43.80±0.01 | 0.039±0.007 |
| PBCS | 55.64±2.26 | 0.062±0.001 | 39.87±1.12 | 0.076±0.011 |
| BCSR | 61.60±0.14 | 0.051±0.015 | 47.30±0.57 | 0.022±0.005 |
| S-TFBO | 58.90±0.75 | 0.046±0.009 | 45.78±0.70 | 0.036±0.005 |
| D-TFBO | 59.54±0.45 | 0.041±0.005 | <u>46.68±0.72</u> | <u>0.029±0.002</u> |

5.3 Coreset Selection for Continual Learning

Coreset selection aims to improve training efficiency by selecting a subset of the most informative data samples, which can be used as an approximation of the entire dataset. Thus, the model that minimizes the loss on the coreset can also minimize the loss on the entire dataset. Following the design in [30], we apply the proposed algorithms to coreset selection for continual learning. The inner problem learns model parameters θ , and the outer problem determines the distribution λ ($0 \leq \lambda_{(i)} \leq 1$ and $\|\lambda\|_1 = 1$) over the entire dataset

$$f(\theta, \lambda) = \sum_{i=1}^n l_i(\theta) + C\mathcal{R}(\lambda); \quad g(\theta, \lambda) = \sum_{i=1}^n \lambda_{(i)} l_i(\theta),$$

where n is the sample size, C is a constant, $\lambda_{(i)}$ is the i -th entry. $\mathcal{R}(\lambda) = -\sum_{i=1}^K \mathbb{E}(\lambda + \delta z)_{[i]}$ denotes the smoothed top- K regularizer, where δ is a constant and $z \sim \mathcal{N}(0, 1)$, $\lambda_{[i]}$ is the i -th largest component. The regularizer encourages the distribution to have K non-zero entries, corresponding to the size of the selected coreset. Following [90], we use the Split CIFAR100 dataset and conduct experiments in the balanced and imbalanced scenarios. We compare the proposed algorithms with various methods, including k -means features [62], k -means embedding [72], Uniform Sampling, iCaRL [68], Grad Matching [11], GCR [80], Greedy Coreset [8], PBCS [90], and BCSR [30], with the last three being bilevel optimization-based methods. We evaluate the performance using the average accuracy and forgetting measure across all tasks after learning task T . The former is defined as $A_T = \frac{1}{T} \sum_{i=1}^T a_{T,i}$, where $a_{T,i}$ is the test accuracy of the i -th task after learning task T . The latter is defined as $FGT_T = \frac{1}{T} \sum_{i=1}^T [\max_{j \in \{1, \dots, T-1\}} (a_{j,i} - a_{T,i})]$. The results are shown in Table 1. Each experiment is repeated three times and the average is reported. It can be observed that our D-TFBO achieves the best FGT_T under the balanced setting and the second-best performance under the imbalanced setting.

Sensitivity analysis w.r.t. different initial learning rates. The tuning-free design provides another benefit. The proposed algorithms demonstrate more robustness compared to the [30]. We conduct a simple sensitivity analysis under the balanced setting, regarding the learning rates in the inner and outer loops. Specifically, we

Table 2: Experiment results of sensitivity analysis on Split CIFAR100. The initial values refer to the constant learning rates in BCSR or $\alpha_0, \beta_0, \gamma_0$ in S-TFBO and D-TFBO.

| Method | <i>initial</i> = 2 | <i>initial</i> = 4 | <i>initial</i> = 6 | <i>initial</i> = 8 | Relative Average Change |
|--------|--------------------|--------------------|--------------------|--------------------|-------------------------|
| BCSR | 59.42 | 56.25 | 58.75 | 57.55 | 5.8% |
| S-TFBO | 58.85 | 58.55 | 58.69 | 58.47 | 0.4% |
| D-TFBO | 59.71 | 59.62 | 59.11 | 59.08 | 0.3% |

set the initial learning rates in [30] and $\alpha_0, \beta_0, \gamma_0$ in S-TFBO and D-TFBO for the inner and outer loops to $\{2, 4, 6, 8\}$, where the original values are set to 5. We run one experiment for each learning rate. Further, we compare the changes in average accuracy A_T . We also compute the average and report the relative change compared to the results presented in Table 1. The code is available at <https://github.com/OptMN-Lab/tfbo>.

6 Conclusion

We introduce two fully tuning-free bilevel optimization algorithms, D-TFBO and S-TFBO. Both methods adaptively update stepsizes without requiring prior knowledge of problem parameters, while achieving convergence rates comparable to their well-tuned counterparts. The experimental results show that our tuning-free design performs comparably to existing well-tuned methods and is more robust to initial stepsizes. We anticipate that the proposed algorithms and the developed analysis can be extended to the stochastic setting, and the proposed algorithms may be applied to other applications such as meta-learning, few-shot learning, and fair machine learning.

References

- [1] K. Antonakopoulos, S. Sabach, L. Viano, M. Hong, and V. Cevher. Adaptive bilevel optimization. 2023.
- [2] K. Antonakopoulos, S. Sabach, L. Viano, M. Hong, and V. Cevher. Adaptive bilevel optimization. *ICLR revision*, 2024.
- [3] M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021.
- [4] L. Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.
- [5] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.
- [6] J. Y. Bello Cruz and T. T. Nghia. On the convergence of the forward–backward splitting method with linesearches. *Optimization Methods and Software*, 31(6):1209–1238, 2016.
- [7] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR)*, 2018.

- [8] Z. Borsos, M. Mutny, and A. Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33:14879–14890, 2020.
- [9] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.
- [10] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- [11] T. Campbell and T. Broderick. Automated scalable bayesian inference via hilbert coresets. *Journal of Machine Learning Research*, 20(15):1–38, 2019.
- [12] T. Chen, Y. Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.
- [13] J. Cortés. Finite-time convergent gradient flows with applications to network consensus. *Automatica*, 42(11):1993–2000, 2006.
- [14] M. Dagréou, P. Ablin, S. Vaïter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.
- [15] Y.-H. Dai and L.-Z. Liao. R-linear convergence of the barzilai and borwein gradient method. *IMA Journal of numerical analysis*, 22(1):1–10, 2002.
- [16] P. K. Diederik. Adam: A method for stochastic optimization. (*No Title*), 2014.
- [17] J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.
- [18] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- [19] C. Fan, G. Choné-Ducasse, M. Schmidt, and C. Thrampoulidis. Bisls/sps: Auto-tune step sizes for stable bi-level optimization. *Advances in Neural Information Processing Systems*, 36, 2024.
- [20] M. Feurer and F. Hutter. Hyperparameter optimization. In *Automated Machine Learning*, pages 3–33. Springer, Cham, 2019.
- [21] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.
- [22] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, pages 1165–1173, 2017.
- [23] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning (ICML)*, pages 1568–1577, 2018.

- [24] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- [25] A. A. Goldstein. Cauchy’s method of minimization. *Numerische Mathematik*, 4(1):146–150, 1962.
- [26] S. Gould, B. Fernando, A. Cherian, P. Anderson, R. S. Cruz, and E. Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.
- [27] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.
- [28] Z. Guo and T. Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.
- [29] P. Hansen, B. Jaumard, and G. Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1194–1217, 1992.
- [30] J. Hao, K. Ji, and M. Liu. Bilevel coreset selection in continual learning: A new formulation and algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.
- [31] E. Hazan and S. Kakade. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.
- [32] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.
- [33] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.
- [34] F. Huang and H. Huang. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.
- [35] K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. *Advances in Neural Information Processing Systems*, 35:3011–3023, 2022.
- [36] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.
- [37] K. Ji and L. Ying. Network utility maximization with general and unknown utility functions: A distributed, data-driven bilevel optimization approach. *Submitted*, 2022.
- [38] A. Khaled and C. Jin. Tuning-free stochastic optimization. *arXiv preprint arXiv:2402.07793*, 2024.
- [39] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.
- [40] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems (NeurIPS)*, pages 1008–1014, 2000.

- [41] J. Kwon, D. Kwon, S. Wright, and R. Nowak. A fully first-order method for stochastic bilevel optimization. *arXiv preprint arXiv:2301.10945*, 2023.
- [42] K. Levy. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30, 2017.
- [43] K. Y. Levy, A. Yurtsever, and V. Cevher. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 31, 2018.
- [44] J. Li, B. Gu, and H. Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.
- [45] J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.
- [46] R. Liao, Y. Xiong, E. Fetaya, L. Zhang, K. Yoon, X. Pitkow, R. Urtasun, and R. Zemel. Reviving and improving recurrent back-propagation. In *International Conference on Machine Learning*, pages 3082–3091. PMLR, 2018.
- [47] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.
- [48] R. Liu, X. Liu, X. Yuan, S. Zeng, and J. Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International Conference on Machine Learning (ICML)*, 2021.
- [49] R. Liu, Y. Liu, S. Zeng, and J. Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8662–8675, 2021.
- [50] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning (ICML)*, 2020.
- [51] N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.
- [52] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.
- [53] L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.
- [54] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

- [55] D. Maladkar, R. Jiang, and A. Mokhtari. Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. *arXiv preprint arXiv:2406.04592*, 2024.
- [56] Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.
- [57] N. Marumo and A. Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. *SIAM Journal on Optimization*, 34(2):2093–2120, 2024.
- [58] H. B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.
- [59] P. Mertikopoulos and W. H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.
- [60] R. Murray, B. Swenson, and S. Kar. Revisiting normalized gradient descent: Fast evasion of saddle points. *IEEE Transactions on Automatic Control*, 64(11):4818–4824, 2019.
- [61] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [62] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.
- [63] A. Orvieto, S. Lacoste-Julien, and N. Loizou. Dynamics of sgd with stochastic polyak stepsizes: Truly adaptive variants and convergence to exact solution. *Advances in Neural Information Processing Systems*, 35:26943–26954, 2022.
- [64] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pages 737–746. PMLR, 2016.
- [65] B. T. Polyak. Minimization of nonsmooth functionals. *Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki*, 9(3):509–521, 1969.
- [66] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 113–124, 2019.
- [67] M. Raydan. On the barzilai and borwein choice of steplength for the gradient method. *IMA Journal of Numerical Analysis*, 13(3):321–326, 1993.
- [68] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [69] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
- [70] S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

- [71] S. Salzo. The variable metric forward-backward splitting algorithm under mild differentiability assumptions. *SIAM Journal on Optimization*, 27(4):2153–2181, 2017.
- [72] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.
- [73] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1723–1732, 2019.
- [74] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.
- [75] C. Shi, J. Lu, and G. Zhang. An extended kuhn–tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005.
- [76] A. Sinha, P. Malo, and K. Deb. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.
- [77] D. Sow, K. Ji, Z. Guan, and Y. Liang. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.
- [78] D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. *arXiv preprint arXiv:2205.02215*, 2022.
- [79] T. Tieleman and G. Hinton. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 17, 2012.
- [80] R. Tiwari, K. Killamsetty, R. Iyer, and P. Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2022.
- [81] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019.
- [82] C. Wang, K. Ji, J. Geng, Z. Ren, T. Fu, F. Yang, Y. Guo, H. He, X. Chen, Z. Zhan, et al. Imperative learning: A self-supervised neural-symbolic learning framework for robot autonomy. *arXiv preprint arXiv:2406.16087*, 2024.
- [83] X. Wang, R. Pan, R. Pi, and T. Zhang. Effective bilevel optimization via minimax reformulation. *arXiv preprint arXiv:2305.13153*, 2023.
- [84] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.
- [85] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [86] Y. Xie, X. Wu, and R. Ward. Linear convergence of adaptive stochastic gradient descent. In *International conference on artificial intelligence and statistics*, pages 1475–1485. PMLR, 2020.

- [87] J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13670–13682, 2021.
- [88] Y. Yang, P. Xiao, and K. Ji. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [89] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- [90] X. Zhou, R. Pi, W. Zhang, Y. Lin, Z. Chen, and T. Zhang. Probabilistic bilevel coreset selection. In *International Conference on Machine Learning*, pages 27287–27302. PMLR, 2022.

Supplementary material

A Supplementary Related Work on Bilevel Optimization

Initially introduced by [9], bilevel optimization has been extensively studied for decades. Early works [29, 75, 26, 76] solved the bilevel problem from a constrained optimization perspective. More recently, gradient-based bilevel methods have gained significant attention for their efficiency and effectiveness in addressing machine learning problems. Among them, approaches based on Approximate Implicit Differentiation (AID) [17, 46, 64, 52, 27, 36, 3, 33] exploit the implicit derivation of the hypergradient, approximating it by solving a linear system.

On the other hand, approaches based on Iterative Differentiation (ITD) [54, 22, 21, 73, 27] estimate the hypergradient by employing automatic differentiation, utilizing either forward or reverse mode.

A series of stochastic bilevel approaches has been developed and analyzed recently, utilizing Neumann series [12, 36, 3], recursive momentum [87, 34, 28], and variance reduction [87, 14], etc. For the lower-level problem with multiple solutions, several approaches were proposed based on upper- and lower-level gradient aggregation [70, 50, 44], barrier types of regularization [48, 47], penalty-based formulations [74], primal-dual techniques [77], and dynamic system-based methods [49]. Another class of approaches formulated the lower-level problem as a value-function-based constraint [41, 83] to solve bilevel problems without second-order gradients.

B Specifications of Experiments

B.1 Practical Implementation

For regularization selection and data hyper-cleaning, we use the benchmark provided in [14]. For coresets selection, we use the codebase from [30]. We implement D-TFBO using “for loops” as an approximation, since the magnitude of $\|\nabla_v R(x, y, v)\|$ in Algorithm 1 varies across different experiments. Specifically, the number of loops for updating y and v in regularization selection and data hyper-cleaning are both set to 10, while the numbers of loops for updating y and v in coresets selection are 5 and 3, respectively.

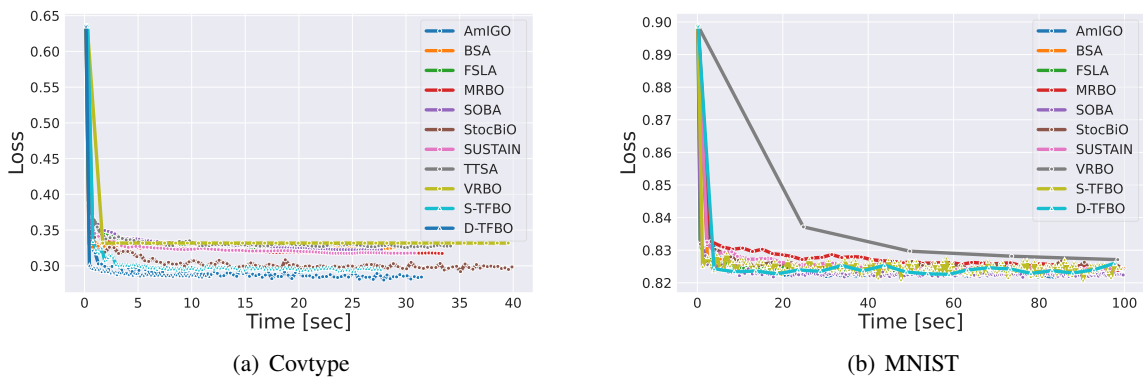


Figure 2: Comparison of running time with other bilevel optimization methods.

B.2 Configuration

We adopt the default configuration for regularization selection and data hyper-cleaning. The batch size is 64. The maximum iterations are 2048 and 512, respectively. The data corruption ratio in hyper-cleaning is 0.1. For coreset selection, we also use the default configuration except for the learning rates, due to the tuning-free design. The α_0 , β_0 , and γ_0 values are set to 5.

B.3 Additional Results

For regularization and data hyper-cleaning, we also present the loss curves regarding running time in Figure 2. Our methods exhibit a faster running time than other baselines on the Covtype dataset.

C Proof Sketch

The proofs of Propositions 1, 2, 3, 4 and 5 can be found in Lemma 4, 9, 11, 15,17, respectively. In this section, we present a high level proof sketch that outlines the convergence and gradient complexity analysis of Algorithm 1 and Algorithm 2, emphasizing the key challenges and our technical innovations.

Proof sketch of Algorithm 1:

Step 1: We first discuss the two-stage framework in our problem in Lemma 4 and we develop two forms of descent lemma of the objective function in Lemma 7 based on the two stages of α_t in Lemma 4.

Step 2: We developed upper bounds of α_t under the two stages in Lemma 4.

Step 3: We provide the maximum iteration numbers for the sub-loops approximating $y^*(x_t)$ and $v^*(x_t)$.

Step 4: Combining the results in Step 1 and Step 2, we telescope and take the average of the inequalities in the descent lemma of the objective function, then we obtain the convergence rate.

Step 5: Combining the maximum iteration numbers in Step 3 and convergence rate in Step 4, we obtain the gradient computation complexity to find ϵ -stationary point. Then the proof is complete.

Proof sketch of Algorithm 2:

Step 1: We first discuss the two-stage framework in our problem in Lemma 4 and we develop two forms of descent lemma of the objective function in Lemma 11 based on the two stages of α_t in Lemma 4.

Step 2: We develop a rough upper bound of two important components in the descent lemma in Lemma 11: $\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$ and $\sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}$, where k_2 and k_3 represents the second stage in Lemma 4.

Step 3: Following the results in Step 2 and the upper bound of v_t in Lemma 10, we obtain a two-way relationship between φ_{t+1} and $\sum_{k=0}^t \frac{\|\nabla f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$, which further indicates the logarithmic upper bounds of both terms in Lemma 15 and Lemma 16, respectively.

Step 4: Incorporating the results from Step 3 into the rough bounds from Step 2, we can also obtain the logarithmic upper bounds of $\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$ and $\sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}$ in Lemma 16.

Step 5: We rearrange the terms in Lemma 11 and incorporate in the results in Step 4, we obtain two forms of the upper bound of α_t in Lemma 17.

Step 6: Combining the results in Steps 3, 4, 5, we telescope and take the average of the inequalities in the descent lemma of the objective function, then we obtain the convergence rate.

Step 7: Without sub-loops, via the convergence rate in Step 6, we can directly obtain the gradient computation complexity to find ϵ -stationary point. Then the proof is complete.

D Proofs of Preliminary Lemmas

Lemma 1 ([84] Lemma 3.2). *For any non-negative a_1, \dots, a_T , and $a_1 \geq 1$, we have*

$$\sum_{l=1}^T \frac{a_l}{\sum_{i=1}^l a_i} \leq \log \left(\sum_{l=1}^T a_l \right) + 1. \quad (4)$$

Lemma 2. *Under Assumptions 1, 2, we have basic properties as follows:*

- (a) $\Phi(x)$ is L_Φ -smooth w.r.t x , where $L_\Phi := \left(L_{f,1} + \frac{L_{g,2}C_{fy}}{\mu} \right) \left(1 + \frac{C_{gxy}}{\mu} \right)^2$;
- (b) $y^*(x)$ is L_y -Lipschitz continuous w.r.t. x , where $L_y := \frac{C_{gxy}}{\mu}$;
- (c) the gradient estimator $\bar{\nabla} f(x, y, v)$ is $(L_{g,2}\|v\| + L_{f,1})$ -Lipschitz continuous w.r.t. (x, y) , and $L_{g,1}$ -Lipschitz continuous w.r.t. v ;
- (d) $\bar{\nabla} f(x, y, v)$ can be bounded as $\|\bar{\nabla} f(x, y, v)\| \leq C_{gxy}\|v\| + C_{fx}$.

Proof. The proof of (a) and (b) can refer to [24]. For (c), under Assumption 2, we have

$$\begin{aligned} \|\bar{\nabla} f(x_1, y_1, v) - \bar{\nabla} f(x_2, y_2, v)\| &\leq \|\nabla_x \nabla_y g(x_1, y_1) - \nabla_x \nabla_y g(x_2, y_2)\| \cdot \|v\| \\ &\quad + \|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \\ &\leq (L_{g,2}\|v\| + L_{f,1})(\|x_1 - x_2\| + \|y_1 - y_2\|) \\ \|\bar{\nabla} f(x, y, v_1) - \bar{\nabla} f(x, y, v_2)\| &\leq \|\nabla_x \nabla_y g(x, y)\| \cdot \|v_1 - v_2\| \leq L_{g,1}\|v_1 - v_2\|. \end{aligned}$$

By Assumption 2 and Remark 3, we can easily prove (d) as

$$\|\bar{\nabla} f(x, y, v)\| \leq \|\nabla_x \nabla_y g(x, y)\| \cdot \|v\| + \|\nabla_x f(x, y)\| \leq C_{gxy}\|v\| + C_{fx}.$$

Then the proof is complete. \square

Lemma 3. *Under Assumptions 1, 2, we have basic properties of linear system function R in eq. (2) as follows:*

- (a) $R(x, y, v)$ is μ -strongly convex and $C_{g_{yy}}$ -smooth w.r.t. v ;
- (b) $\nabla_v R(x, y, v)$ is $(L_{g,2}\|v\| + L_{f,1})$ -Lipschitz continuous w.r.t. (x, y) ;
- (c) $\nabla_v R(x, y, v)$ can be bounded as $\|\nabla_v R(x, y, v)\| \leq C_{g_{yy}}\|v\| + C_{fy}$;
- (d) $v^*(x)$ in eq. (2) can be bounded as $\|v^*(x)\| \leq \frac{C_{fy}}{\mu}$, and $\hat{v}^*(x, y) := \arg \min_v R(x, y, v)$ can also be bounded as $\|\hat{v}^*(x, y)\| \leq \frac{C_{fy}}{\mu}$;
- (e) $v^*(x)$ is L_v -Lipschitz continuous w.r.t. x and $\hat{v}^*(x, y)$ is \bar{L}_v -Lipschitz continuous w.r.t. y , where $L_v := \left(\frac{L_{f,1}}{\mu} + \frac{C_{fy}L_{g,2}}{\mu^2} \right) (1 + L_y)$ and $\bar{L}_v := \frac{L_{f,1}}{\mu} + \frac{C_{fy}L_{g,2}}{\mu^2}$.

Proof. First of all, since $\nabla_v \nabla_v R(x, y, v) = \nabla_y \nabla_y g(x, y)$, we know $\mu I \preceq \nabla_y \nabla_y g(x, y)$. Thus, according to Assumptions 1, 2, we have

$$\|\nabla_v \nabla_v R(x, y, v_1) - \nabla_v \nabla_v R(x, y, v_2)\| \leq \|\nabla_y \nabla_y g(x, y)\| \|v_1 - v_2\| \leq C_{g_{yy}} \|v_1 - v_2\|.$$

Then (a) is proved. Next, by using Lipschitz continuity in Assumption 2, we have

$$\begin{aligned} \|\nabla_v R(x_1, y_1, v) - \nabla_v R(x_2, y_2, v)\| &\leq \|\nabla_y \nabla_y g(x_1, y_1) - \nabla_y \nabla_y g(x_2, y_2)\| \cdot \|v\| \\ &\quad + \|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \\ &\leq (L_{g,2} \|v\| + L_{f,1})(\|x_1 - x_2\| + \|y_1 - y_2\|). \end{aligned}$$

Then (b) is proved. By Assumption 2, we can easily prove (c) as

$$\|\nabla_v R(x, y, v)\| \leq \|\nabla_y \nabla_y g(x, y)\| \cdot \|v\| + \|\nabla_y f(x, y)\| \leq C_{g_{yy}} \|v\| + C_{f_y}.$$

Next, for $\hat{v}^*(x, y)$, we have

$$\nabla_v R(x, y, \hat{v}^*(x, y)) = \nabla_y \nabla_y g(x, y) \hat{v}^*(x, y) - \nabla_y f(x, y) = 0,$$

which indicates that

$$\|\hat{v}^*(x, y)\| = \|\left[\nabla_y \nabla_y g(x, y)\right]^{-1} \nabla_y f(x, y)\| \leq \|\left[\nabla_y \nabla_y g(x, y)\right]^{-1}\| \cdot \|\nabla_y f(x, y)\| \leq \frac{C_{f_y}}{\mu}.$$

Since $v^*(x)$ is a special case as $v^*(x) = \hat{v}^*(x, y^*(x))$, (d) is proved. The proof of the first part of (e) can refer to Lemma 4 in [88]; for the second part, we have

$$\begin{aligned} &\|\hat{v}^*(x, y_1) - \hat{v}^*(x, y_2)\| \\ &= \|\left[\nabla_y \nabla_y g(x, y_1)\right]^{-1} \nabla_y f(x, y_1) - \left[\nabla_y \nabla_y g(x, y_2)\right]^{-1} \nabla_y f(x, y_2)\| \\ &\leq \|\left[\nabla_y \nabla_y g(x, y_1)\right]^{-1} (\nabla_y f(x, y_1) - \nabla_y f(x, y_2))\| \\ &\quad + \|\left(\left[\nabla_y \nabla_y g(x, y_1)\right]^{-1} - \left[\nabla_y \nabla_y g(x, y_2)\right]^{-1}\right) \nabla_y f(x, y_2)\| \\ &\leq \frac{L_{f,1}}{\mu} \|y_1 - y_2\| + C_{f_y} \|\left(\left[\nabla_y \nabla_y g(x, y_1)\right]^{-1} (\nabla_y \nabla_y g(x, y_2) - \nabla_y \nabla_y g(x, y_1)) \left[\nabla_y \nabla_y g(x, y_2)\right]^{-1}\right)\| \\ &\leq \left(\frac{L_{f,1}}{\mu} + \frac{C_{f_y} L_{g,2}}{\mu^2}\right) \|y_1 - y_2\|. \end{aligned}$$

Thus, the second part of (e) is proved and the proof of Lemma 3 is complete. \square

Lemma 4. Suppose the iteration rounds to update $\{x, y, v\}$ are $\{T_1, T_2, T_3\}$ and $\{\alpha_t, \beta_t, \gamma_t\}$ are generated by Algorithm 1 or 2. For any $C_\alpha \geq \alpha_0$, $C_\beta \geq \beta_0$, $C_\gamma \geq \gamma_0$, we have

- (a) either $\alpha_t \leq C_\alpha$ for any $t \leq T_1$, or $\exists k_1 \leq T_1$ such that $\alpha_{k_1} \leq C_\alpha$, $\alpha_{k_1+1} > C_\alpha$;
- (b) either $\beta_t \leq C_\beta$ for any $t \leq T_2$, or $\exists k_2 \leq T_2$ such that $\beta_{k_2} \leq C_\beta$, $\beta_{k_2+1} > C_\beta$;
- (c) either $\gamma_t \leq C_\gamma$ for any $t \leq T_3$, or $\exists k_3 \leq T_3$ such that $\gamma_{k_3} \leq C_\gamma$, $\gamma_{k_3+1} > C_\gamma$.

Proof. The proof resembles the Lemma 4.1 in [84]. Here we only prove part (a), and the other two are similar. Note that if $\alpha_{T_1} > C_\alpha$, then there must exist $k_1 \leq T_1$ such that $\alpha_{k_1} \leq C_\alpha$, $\alpha_{k_1+1} > C_\alpha$, because $C_\alpha \geq \alpha_0$ and the sequence $\{\alpha_k\}$ is monotonically increasing. Otherwise, we have $\alpha_t \leq \alpha_{T_1} \leq C_\alpha$ for any $t \leq T_1$. This completes the proof of part (a). \square

E Proof of Theorem 1

We define some notation for convenience before proving Theorem 1.

E.1 Notation

Here, we define the following constants as thresholds for parameters $\beta_p, \gamma_q, \alpha_t$ in Algorithm 1 as

$$C_\alpha := \max \{2L_\Phi, \alpha_0\}, \quad C_\beta := \max \{L_{g,1}, \beta_0\}, \quad C_\gamma := \max \{C_{g_{yy}}, \gamma_0\}. \quad (5)$$

E.2 Proofs of Preliminary Lemmas

Lemma 5. *Under Assumptions 1, 2, for any $t \geq 0$ in Algorithm 1, we have*

$$\|y_t^{P_t} - y^*(x_t)\|^2 \leq \frac{\epsilon_y}{\mu^2}, \quad \|v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t})\|^2 \leq \frac{\epsilon_v}{\mu^2},$$

where ϵ_y and ϵ_v are sub-loop stopping criteria in Algorithm 1.

Proof. For the k_{th} iteration, according to the stop criteria of the sub-loops, we have

$$\|\nabla_y g(x_t, y_t^{P_t})\|^2 \leq \epsilon_y, \quad \|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \leq \epsilon_v.$$

By using Assumptions 1,2, we have

$$\begin{aligned} \|y_t^{P_t} - y^*(x_t)\|^2 &\leq \frac{1}{\mu^2} \|\nabla_y g(x_t, y_t^{P_t}) - \nabla_y g(x_t, y^*(x_t))\|^2 \leq \frac{\epsilon_y}{\mu^2}, \\ \|v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t})\|^2 &\leq \frac{1}{\mu^2} \|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t}) - \nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2 \leq \frac{\epsilon_v}{\mu^2}, \end{aligned}$$

since $\|\nabla_y g(x_t, y^*(x_t))\|^2 = 0$ and $\|\nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2 = 0$. Thus, the proof is complete. \square

Lemma 6. *Under Assumptions 1, 2, for any $t \geq 0$ in Algorithm 1, we have $\|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \leq C_f^2$, where $C_f := \left(\frac{2C_{g_{xy}}^2 \epsilon_v}{\mu^2} + \frac{4C_{g_{xy}}^2 C_{f_y}^2}{\mu^2} + 4C_{f_y}^2 \right)^{\frac{1}{2}}$.*

Proof. For the k_{th} iteration, we have

$$\begin{aligned} &\|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &\leq 2\|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t}) - \bar{\nabla} f(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2 + 2\|\bar{\nabla} f(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2 \\ &= 2\|\nabla_x \nabla_y g(x_t, y_t^{P_t})(v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t}))\|^2 + 2\|\nabla_x \nabla_y g(x_t, y_t^{P_t})\hat{v}^*(x_t, y_t^{P_t}) - \nabla_y f(x_t, y_t^{P_t})\|^2 \\ &\leq 2\|\nabla_x \nabla_y g(x_t, y_t^{P_t})\|^2 \cdot \|v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + 2\|\nabla_x \nabla_y g(x_t, y_t^{P_t})\hat{v}^*(x_t, y_t^{P_t}) - \nabla_y f(x_t, y_t^{P_t})\|^2 \\ &\stackrel{(a)}{\leq} \frac{2C_{g_{xy}}^2 \epsilon_v}{\mu^2} + \frac{4C_{g_{xy}}^2 C_{f_y}^2}{\mu^2} + 4C_{f_y}^2, \end{aligned}$$

where (a) uses Assumption 1, Remark 3, Lemma 3 and Lemma 5. Then, the proof is complete. \square

E.3 Descent in Objective Function

Lemma 7. *Under Assumptions 1, 2, for Algorithm 1, suppose the total iteration number is T . No matter k_1 in Lemma 4 exists or not, we always have*

$$\Phi(x_{t+1}) \leq \Phi(x_t) - \frac{1}{2\alpha_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}} \left(1 - \frac{L_\Phi}{2\alpha_{t+1}}\right) \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{\epsilon'}{2\alpha_{t+1}}. \quad (6)$$

If in addition, k_1 in Lemma 4 exists, then for $t \geq k_1$, we further have

$$\Phi(x_{t+1}) \leq \Phi(x_t) - \frac{1}{2\alpha_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{4\alpha_{t+1}} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{\epsilon'}{2\alpha_{t+1}}, \quad (7)$$

where $\epsilon' := \frac{\bar{L}^2}{\mu^2}(\epsilon_y + \epsilon_v)$ and $\bar{L} := \max\{2\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2 + C_{g_{yy}}^2 \bar{L}_v^2\right)^{\frac{1}{2}}, \sqrt{2}C_{g_{yy}}\}$.

Proof. From Lemma 2, we have $\Phi(x)$ is L_Φ -smooth. So we can apply the descent lemma to Φ as

$$\begin{aligned} \Phi(x_{t+1}) &\leq \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t \rangle + \frac{L_\Phi}{2} \|x_{t+1} - x_t\|^2 \\ &= \Phi(x_t) - \frac{1}{\alpha_{t+1}} \langle \nabla\Phi(x_t), \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t}) \rangle + \frac{L_\Phi}{2\alpha_{t+1}^2} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &= \Phi(x_t) - \frac{1}{2\alpha_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &\quad + \frac{1}{2\alpha_{t+1}} \|\nabla\Phi(x_t) - \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{L_\Phi}{2\alpha_{t+1}^2} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2, \end{aligned} \quad (8)$$

where the approximation error

$$\begin{aligned} &\|\nabla\Phi(x_t) - \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &= \|\bar{\nabla}f(x_t, y^*(x_t), v^*(x_t)) - \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &\leq 2\|\bar{\nabla}f(x_t, y^*(x_t), v^*(x_t)) - \bar{\nabla}f(x_t, y_t^{P_t}, v^*(x_t))\|^2 \\ &\quad + 2\|\bar{\nabla}f(x_t, y_t^{P_t}, v^*(x_t)) - \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \\ &\leq 4\|\nabla_y \nabla_y g(x_t, y^*(x_t))v^*(x_t) - \nabla_y \nabla_y g(x_t, y_t^{P_t})v^*(x_t)\|^2 \\ &\quad + 4\|\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_t^{P_t})\|^2 + 2\|\nabla_y \nabla_y g(x_t, y_t^{P_t})(v^*(x_t) - v_t^{Q_t})\|^2 \\ &\stackrel{(a)}{\leq} 4\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2\right) \|y_t^{P_t} - y^*(x_t)\|^2 + 2C_{g_{yy}}^2 \|v_t^{Q_t} - v^*(x_t)\|^2 \\ &\leq 4\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2\right) \|y_t^{P_t} - y^*(x_t)\|^2 + 4C_{g_{yy}}^2 \|v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + 4C_{g_{yy}}^2 \|\hat{v}^*(x_t, y_t^{P_t}) - v^*(x_t)\|^2 \\ &\stackrel{(b)}{\leq} 4\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2 + C_{g_{yy}}^2 \bar{L}_v^2\right) \|y_t^{P_t} - y^*(x_t)\|^2 + 4C_{g_{yy}}^2 \|v_t^{Q_t} - \hat{v}^*(x_t, y_t^{P_t})\|^2 \\ &\leq \bar{L}^2 (\|y_t^{P_t} - y^*(x_t)\|^2 + \|v_t^{Q_t} - \hat{v}^*(x_t, y^*(x_t))\|^2), \end{aligned} \quad (9)$$

where (a) uses Assumption 2, Remark 3 and Lemma 3; (b) uses $v^*(x_t) = \hat{v}^*(x_t, y^*(x_t))$ and Lemma 3. By using Lemma 5, we have

$$\|\nabla\Phi(x_t) - \bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \leq \frac{\bar{L}^2}{\mu^2}(\epsilon_y + \epsilon_v) =: \epsilon'. \quad (10)$$

By plugging eq. (10) into eq. (8), we obtain (6).

Now if in addition, k_1 in Lemma 4 exists, then for $t \geq k_1$, we have $\alpha_{t+1} > C_\alpha \geq 2L_\Phi$. From (6) we can immediately obtain (7). Thus, the proof is complete. \square

E.4 The bound of α_t

Lemma 8. *Under Assumptions 1, 2, 3, suppose the number of total iteration rounds in Algorithm 1 is T . If there exists $k_1 \leq T$ described in Lemma 4, we have*

$$\begin{cases} \alpha_t \leq C_\alpha, & t \leq k_1; \\ \alpha_t \leq C_\alpha + 2c_0 + \frac{2t\epsilon'}{\alpha_0}, & t \geq k_1, \end{cases}$$

where we define

$$c_0 := 2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2}. \quad (11)$$

When such k_1 does not exist, we have $\alpha_t \leq C_\alpha$ for any $t \leq T$.

Proof. According to Lemma 4, the proof can be split into the following three cases.

Case 1: if $\alpha_T \leq C_\alpha$, for any $t < T$, we have the upper bound of α_{t+1} as $\alpha_{t+1} \leq C_\alpha$.

Case 2: if $\alpha_T > C_\alpha$, there exists $k_1 \leq T$ described in Lemma 4. Then we have the upper bound of α_{t+1} as $\alpha_{t+1} \leq C_\alpha$ for any $t < k_1$.

Case 3: in the remaining proof, we only consider and explore the case $k_1 \leq t \leq T$ when $\alpha_T > C_\alpha$.

From Lemma 7, for $k \geq k_1$, we have

$$\Phi(x_{k+1}) \leq \Phi(x_k) - \frac{1}{2\alpha_{k+1}} \|\nabla \Phi(x_k)\|^2 - \frac{1}{4\alpha_{k+1}} \|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2 + \frac{\epsilon'}{2\alpha_{k+1}},$$

which indicates that

$$\frac{\|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2}{\alpha_{k+1}} \leq 4(\Phi(x_k) - \Phi(x_{k+1})) + \frac{2\epsilon'}{\alpha_{k+1}}.$$

By taking summation over $k = k_1, \dots, t$, we have

$$\begin{aligned} \sum_{k=k_1}^t \frac{\|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2}{\alpha_{k+1}} &\leq 4 \sum_{k=k_1}^t (\Phi(x_k) - \Phi(x_{k+1})) + \sum_{k=k_1}^t \frac{2\epsilon'}{\alpha_{k+1}} \\ &= 4(\Phi(x_{k_1}) - \Phi(x_{t+1})) + \sum_{k=k_1}^t \frac{2\epsilon'}{\alpha_{k+1}}. \end{aligned} \quad (12)$$

For $\Phi(x_{k_1})$, by telescoping (6), we get

$$\Phi(x_{k_1}) \leq \Phi(x_0) + \sum_{k=0}^{k_1-1} \frac{L_\Phi}{4\alpha_{k+1}^2} \|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2 + \sum_{k=0}^{k_1-1} \frac{\epsilon'}{2\alpha_{k+1}}. \quad (13)$$

Plugging eq. (13) into eq. (12), we obtain

$$\begin{aligned}
\sum_{k=k_1}^t \frac{\|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2}{\alpha_{k+1}} &\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \sum_{k=0}^{k_1-1} \frac{L_\Phi}{\alpha_{k+1}^2} \|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2 + \sum_{k=0}^t \frac{2\epsilon'}{\alpha_{k+1}} \\
&\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi \sum_{k=0}^{k_1-1} \|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2}{\alpha_0^2} + \sum_{k=0}^t \frac{2\epsilon'}{\alpha_{k+1}} \\
&\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi \alpha_{k_1}^2}{\alpha_0^2} + \frac{2(t+1)\epsilon'}{\alpha_0} \\
&\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{\alpha_0^2} + \frac{2(t+1)\epsilon'}{\alpha_0}. \tag{14}
\end{aligned}$$

Inspired by [84] and using telescoping, we have

$$\begin{aligned}
\alpha_{t+1} &= \alpha_t + \frac{\|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2}{\alpha_{t+1} + \alpha_t} \\
&\leq \alpha_t + \frac{\|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2}{\alpha_{t+1}} \\
&\leq \alpha_{k_1} + \sum_{k=k_1}^t \frac{\|\bar{\nabla} f(x_k, y_k^{P_k}, v_k^{Q_k})\|^2}{\alpha_{k+1}} \\
&\leq C_\alpha + 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{\alpha_0^2} + \frac{2(t+1)\epsilon'}{\alpha_0}.
\end{aligned}$$

Thus, the proof is complete. \square

E.5 Convergence Analysis of Sub-loops

Lemma 9. Recall that for the t th iteration, the sub-loops in Algorithm 1 aim to find $y_t^{P_t}$ and $v_t^{Q_t}$ such that $\|\nabla_y g(x_t, y_t^{P_t})\|^2 \leq \epsilon_y$ and $\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 \leq \epsilon_v$. Here we prove that

$$P_t \leq P' := \frac{\log(C_\beta^2/\beta_0^2)}{\log(1 + \epsilon_y/C_\beta^2)} + \frac{\beta_{\max}}{\mu} \log\left(\frac{L_{g,1}^2(\beta_{\max} - C_\beta)}{\epsilon_y}\right), \tag{15a}$$

$$Q_t \leq Q' := \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1 + \epsilon_v/C_\gamma^2)} + \frac{\gamma_{\max}}{\mu} \log\left(\frac{C_{g_{yy}}^2(\gamma_{\max} - C_\gamma)}{\epsilon_v}\right), \tag{15b}$$

where $\beta_{\max} := C_\beta + L_{g,1}\left(\frac{2\epsilon_y}{\mu^2} + \frac{2C_{g_{xy}}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1\right)$ and $\gamma_{\max} := C_\gamma + C_{g_{yy}}\left(\frac{2\epsilon_y}{\mu^2} + \frac{8C_{fy}^2}{\mu^2} + 2\log(C_\gamma/\gamma_0) + 1\right)$.

Proof. The proof is split into the following two parts.

Part I: maximum number for convergence of $g(x_t, y_t^{P_t})$.

Inspired by [86], we split the analysis into the following two cases.

Case 1: k_2 does not exist before we find P_t . This indicates $\beta_{P_t} < C_\beta$. Referring to Lemma 2 in [86], we have $P_t < \frac{\log(C_\beta^2/\beta_0^2)}{\log(1 + \epsilon_y/C_\beta^2)}$ and therefore the desired upper bound for P_t holds. This can be proved as follows. If

$P_t \geq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1+\epsilon_y/C_\beta^2)}$, we have the following result.

$$\begin{aligned}
\beta_{P_t}^2 &= \beta_{P_t-1}^2 + \|\nabla_y g(x_t, y_t^{P_t-1})\|^2 \\
&= \beta_{P_t-1}^2 \left(1 + \frac{\|\nabla_y g(x_t, y_t^{P_t-1})\|^2}{\beta_{P_t-1}^2}\right) \\
&\geq \beta_0^2 \prod_{p=0}^{P_t-1} \left(1 + \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_p^2}\right) \\
&\geq \beta_0^2 \left(1 + \frac{\epsilon_y}{C_\beta^2}\right)^{P_t} \geq C_\beta^2.
\end{aligned} \tag{16}$$

This contradicts $\beta_{P_t} < C_\beta$.

Case 2: k_2 exists and $P_t \geq k_2$. Here we have $\beta_{k_2} \leq C_\beta$ and $\beta_{k_2+1} > C_\beta$.

Firstly, we prove $k_2 \leq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1+\epsilon_y/C_\beta^2)}$. Similar to **Case 1**, if $k_2 > \frac{\log(C_\beta^2/\beta_0^2)}{\log(1+\epsilon_y/C_\beta^2)}$, following eq. (16) by replacing P_t with k_2 , we have

$$\beta_{k_2}^2 \geq \beta_0^2 \left(1 + \frac{\epsilon_y}{C_\beta^2}\right)^{k_2} > C_\beta^2,$$

which contradicts $\beta_{k_2} \leq C_\beta$.

Secondly, referring to Lemma 3 in [86], we have the bound of $\|y_t^{k_2} - y^*(x_t)\|^2$ as

$$\begin{aligned}
&\|y_t^{k_2} - y^*(x_t)\|^2 \\
&= \left\| y_t^{k_2-1} - \frac{\nabla_y g(x_t, y_t^{k_2-1})}{\beta_{k_2}} - y^*(x_t) \right\|^2 \\
&= \|y_t^{k_2-1} - y^*(x_t)\|^2 + \left\| \frac{\nabla_y g(x_t, y_t^{k_2-1})}{\beta_{k_2}} \right\|^2 - 2 \left\langle y_t^{k_2-1} - y^*(x_t), \frac{\nabla_y g(x_t, y_t^{k_2-1})}{\beta_{k_2}} \right\rangle \\
&\stackrel{(a)}{\leq} \|y_t^{k_2-1} - y^*(x_t)\|^2 + \left\| \frac{\nabla_y g(x_t, y_t^{k_2-1})}{\beta_{k_2}} \right\|^2 - \frac{2}{\beta_{k_2} L_{g,1}} \|\nabla_y g(x_t, y_t^{k_2-1}) - \nabla_y g(x_t, y^*(x_t))\|^2 \\
&\leq \|y_t^{k_2-1} - y^*(x_t)\|^2 + \frac{\|\nabla_y g(x_t, y_t^{k_2-1})\|^2}{\beta_{k_2}^2} \\
&\leq \|y_t^0 - y^*(x_t)\|^2 + \sum_{p=0}^{k_2-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_{p+1}^2} \\
&\stackrel{(b)}{\leq} \|y_{t-1}^{P_t-1} - y^*(x_t)\|^2 + \sum_{p=0}^{k_2-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2 / \beta_0^2}{\sum_{k=0}^p \|\nabla_y g(x_t, y_t^k)\|^2 / \beta_0^2} \\
&\stackrel{(c)}{\leq} 2\|y_{t-1}^{P_t-1} - y^*(x_{t-1})\|^2 + 2\|y^*(x_{t-1}) - y^*(x_t)\|^2 + \log \left(\sum_{p=0}^{k_2-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_0^2} \right) + 1 \\
&\stackrel{(d)}{\leq} \frac{2\epsilon_y}{\mu^2} + \frac{2C_{g_{xy}}^2 \|\bar{\nabla} f(x_{t-1}, y_{t-1}^{P_t-1}, v_{t-1}^{Q_{t-1}})\|^2}{\mu^2 \alpha_t^2} + \log \left(\sum_{p=0}^{k_2-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_0^2} \right) + 1
\end{aligned}$$

$$\stackrel{(e)}{\leq} \frac{2\epsilon_y}{\mu^2} + \frac{2C_{gxy}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1, \quad (17)$$

where (a) uses Assumptions 1,2; (b) refers to the warm start of y_t^0 ; (c) uses Lemma 1; (d) uses Lemmas 2 and 5; (e) follows from Lemma 6 and $\beta_{k_2} \leq C_\beta$.

Last, following [86], for all $P > k_2$, we have the bound of $\|y_t^P - y^*(x_t)\|^2$ as

$$\begin{aligned} \|y_t^P - y^*(x_t)\|^2 &= \|y_t^{P-1} - y^*(x_t)\|^2 + \frac{\|\nabla_y g(x_t, y_t^{P-1})\|^2}{\beta_P^2} - \frac{2\langle y_t^{P-1} - y^*(x_t), \nabla_y g(x_t, y_t^{P-1}) \rangle}{\beta_P} \\ &\leq \|y_t^{P-1} - y^*(x_t)\|^2 - \frac{1}{\beta_P} \left(2 - \frac{L_{g,1}}{\beta_P}\right) \langle y_t^{P-1} - y^*(x_t), \nabla_y g(x_t, y_t^{P-1}) \rangle \\ &\stackrel{(a)}{\leq} \|y_t^{P-1} - y^*(x_t)\|^2 - \frac{1}{\beta_P} \langle y_t^{P-1} - y^*(x_t), \nabla_y g(x_t, y_t^{P-1}) \rangle \\ &\stackrel{(b)}{\leq} \left(1 - \frac{\mu}{\beta_P}\right) \|y_t^{P-1} - y^*(x_t)\|^2 \\ &\stackrel{(c)}{\leq} e^{-\mu(P-k_2)/\beta_P} \|y_t^{k_2} - y^*(x_t)\|^2 \\ &\stackrel{(d)}{\leq} e^{-\mu(P-k_2)/\beta_P} \left(\frac{2\epsilon_y}{\mu^2} + \frac{2C_{gxy}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1 \right), \end{aligned} \quad (18)$$

where (a) uses $\beta_P \geq C_\beta \geq L_{g,1}$; (b) uses Assumption 1; (c) follows from $\beta_P \geq C_\beta \geq L_{g,1} \geq \mu$ and $1 - m \leq e^{-m}$ for $0 < m < 1$; (d) refers to eq. (17). Inspired by Lemma 4 in [86], we have the upper-bound of β_P as

$$\beta_P = \beta_{P-1} + \frac{\|\nabla_y g(x_t, y_t^{P-1})\|^2}{\beta_P + \beta_{P-1}} \leq \beta_{k_2} + \sum_{p=k_2}^{P-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_{p+1}}. \quad (19)$$

To further bound the last term of the right-hand side of eq. (19), using Assumption 2, we have the following result:

$$\begin{aligned} &\|y_t^P - y^*(x_t)\|^2 \\ &= \|y_t^{P-1} - y^*(x_t)\|^2 + \frac{\|\nabla_y g(x_t, y_t^{P-1})\|^2}{\beta_P^2} - \frac{2\langle y_t^{P-1} - y^*(x_t), \nabla_y g(x_t, y_t^{P-1}) \rangle}{\beta_P} \\ &\stackrel{(a)}{\leq} \|y_t^{P-1} - y^*(x_t)\|^2 + \frac{\|\nabla_y g(x_t, y_t^{P-1})\|^2}{\beta_P^2} - \frac{2\|\nabla_y g(x_t, y_t^{P-1}) - \nabla_y g(x_t, y^*(x_t))\|^2}{\beta_P L_{g,1}} \\ &\stackrel{(b)}{\leq} \|y_t^{P-1} - y^*(x_t)\|^2 - \frac{\|\nabla_y g(x_t, y_t^{P-1})\|^2}{\beta_P L_{g,1}} \\ &\leq \|y_t^{k_2} - y^*(x_t)\|^2 - \sum_{p=k_2}^{P-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_{p+1} L_{g,1}}, \end{aligned} \quad (20)$$

where (a) uses Assumptions 1,2; (b) refers to $\beta_P \geq C_\beta \geq L_{g,1}$. By rearranging eq. (20) and using eq. (17), we have

$$\sum_{p=k_2}^{P-1} \frac{\|\nabla_y g(x_t, y_t^p)\|^2}{\beta_{p+1}} \leq L_{g,1} (\|y_t^{k_2} - y^*(x_t)\|^2 - \|y_t^P - y^*(x_t)\|^2)$$

$$\begin{aligned}
&\leq L_{g,1} \|y_t^{k_2} - y^*(x_t)\|^2 \\
&\leq L_{g,1} \left(\frac{2\epsilon_y}{\mu^2} + \frac{2C_{gxy}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1 \right). \tag{21}
\end{aligned}$$

Plugging eq. (21) into eq. (19), we obtain the upper-bound of β_P as

$$\beta_P \leq C_\beta + L_{g,1} \left(\frac{2\epsilon_y}{\mu^2} + \frac{2C_{gxy}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1 \right) =: \beta_{\max}. \tag{22}$$

Then, by plugging eq. (22) into eq. (18), we have the upper bound of $\|y_t^P - y^*(x_t)\|^2$ as

$$\|y_t^P - y^*(x_t)\|^2 \leq e^{-\mu(P-k_2)/\beta_{\max}} \left(\frac{2\epsilon_y}{\mu^2} + \frac{2C_{gxy}^2 C_f^2}{\mu^2 \alpha_0^2} + 2\log(C_\beta/\beta_0) + 1 \right). \tag{23}$$

Recall we have the upper bound $k_2 \leq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1+\epsilon_y/C_\beta^2)}$. Note that P' defined in (15a) satisfies

$$P' \geq k_2 + \frac{\beta_{\max}}{\mu} \log(L_{g,1}^2(\beta_{\max} - C_\beta)/\epsilon_y).$$

By replacing P with P' in eq. (23), we have

$$\|\nabla_y g(x_t, y_t^{P'})\|^2 \leq L_{g,1}^2 \|y_t^{P'} - y^*(x_t)\|^2 \leq e^{-\mu(P'-k_2)/\beta_{\max}} L_{g,1}^2 (\beta_{\max} - C_\beta) \leq \epsilon_y.$$

Therefore, $P_t \leq P'$ and this completes the proof of (15a).

Part II: maximum number for convergence of $R(x_t, y_t^{P_t}, v_t^{Q_t})$.

Similarly to **Part I**, we split the analysis into the following two cases.

Case 1: k_3 does not exist before we find Q_t . This indicates $\gamma_{Q_t} < C_\gamma$. Then we have $Q_t < \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1+\epsilon_v/C_\gamma^2)}$.

Otherwise, if $Q_t \geq \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1+\epsilon_v/C_\gamma^2)}$, we have the following result.

$$\begin{aligned}
\gamma_{Q_t}^2 &= \gamma_{Q_t-1}^2 + \|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t-1})\|^2 \\
&= \gamma_{Q_t-1}^2 \left(1 + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t-1})\|^2}{\gamma_{Q_t-1}^2} \right) \\
&\geq \gamma_0^2 \prod_{q=0}^{Q_t-1} \left(1 + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q_t-1})\|^2}{\gamma_{Q_t-1}^2} \right) \\
&\geq \gamma_0^2 \left(1 + \frac{\epsilon_v}{C_\gamma^2} \right)^{Q_t} \geq C_\gamma^2.
\end{aligned}$$

This contradicts $\gamma_{Q_t} < C_\gamma$.

Case 2: k_3 exists and $Q_t \geq k_3$. Here we have $\gamma_{k_3} \leq C_\gamma$ and $\gamma_{k_3+1} > C_\gamma$.

Firstly, we have $k_3 \leq \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1+\epsilon_v/C_\gamma^2)}$. Similar to **Case 1**, if $k_3 > \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1+\epsilon_v/C_\gamma^2)}$, following eq. (16), by replacing Q_t with k_3 , we have

$$\gamma_{k_3}^2 \geq \gamma_0^2 \left(1 + \frac{\epsilon_v}{C_\gamma^2} \right)^{k_3} > C_\gamma^2,$$

which contradicts $\gamma_{k_3} \leq C_\gamma$.

Secondly, referring to Lemma 3 in [86], we have the bound of $\|v_t^{k_3} - v^*(x_t)\|^2$ as following:

$$\begin{aligned}
\|v_t^{k_3} - \hat{v}^*(x_t, y_t^{P_t})\|^2 &= \left\| v_t^{k_3-1} - \frac{\nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1})}{\gamma_{k_3}} - \hat{v}^*(x_t, y_t^{P_t}) \right\|^2 \\
&= \|v_t^{k_3-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \left\| \frac{\nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1})}{\gamma_{k_3}} \right\|^2 \\
&\quad - \frac{2}{\gamma_{k_3}} \langle v_t^{k_3-1} - \hat{v}^*(x_t, y_t^{P_t}), \nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1}) \rangle \\
&\stackrel{(a)}{\leq} \|v_t^{k_3-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \left\| \frac{\nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1})}{\gamma_{k_3}} \right\|^2 \\
&\quad - \frac{2}{\gamma_{k_3} C_{g_{yy}}} \|\nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1}) - \nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2 \\
&\leq \|v_t^{k_3-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \left\| \frac{\nabla_v R(x_t, y_t^{P_t}, v_t^{k_3-1})}{\gamma_{k_3}} \right\|^2 \\
&\leq \|v_t^0 - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \sum_{q=0}^{k_3-1} \left\| \frac{\nabla_v R(x_t, y_t^{P_t}, v_t^q)}{\gamma_{k_3}} \right\|^2 \\
&\stackrel{(b)}{\leq} \|v_t^0 - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \sum_{q=0}^{k_3-1} \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2 / \gamma_0^2}{\sum_{k=0}^q \|\nabla_v R(x_t, y_t^{P_t}, v_t^k)\|^2 / \gamma_0^2} \\
&\stackrel{(c)}{\leq} 2\|v_{t-1}^{P_t-1} - \hat{v}^*(x_{t-1}, y_{t-1}^{P_t-1})\|^2 + 2\|\hat{v}^*(x_{t-1}, y_{t-1}^{P_t-1}) - \hat{v}^*(x_t, y_t^{P_t})\|^2 \\
&\quad + \log \left(\sum_{q=0}^{k_3-1} \|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2 / \gamma_0^2 \right) + 1 \\
&\leq 2\|v_{t-1}^{P_t-1} - \hat{v}^*(x_{t-1}, y_{t-1}^{P_t-1})\|^2 + 4\|\hat{v}^*(x_{t-1}, y_{t-1}^{P_t-1})\|^2 + 4\|\hat{v}^*(x_t, y_t^{P_t})\|^2 \\
&\quad + \log \left(\sum_{q=0}^{k_3-1} \|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2 / \gamma_0^2 \right) + 1 \\
&\stackrel{(d)}{\leq} \frac{2\epsilon_y}{\mu^2} + \frac{8C_{f_y}^2}{\mu^2} + 2\log(C_\gamma/\gamma_0) + 1, \tag{24}
\end{aligned}$$

where (a) uses Lemma 3 and $\nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t})) = 0$; (b) refers to the warm start of v_t^0 ; (c) uses Lemma 1; (d) follows from Lemma 3,5 and $\gamma_{k_3} \leq C_\gamma$.

Last, similar to **Part I**, for all $Q > k_3$, we explore the bound of $\|v_t^Q - v^*(x_t)\|^2$ as

$$\begin{aligned}
&\|v_t^Q - \hat{v}^*(x_t, y_t^{P_t})\|^2 \\
&= \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1})\|^2}{\gamma_Q^2} \\
&\quad - \frac{2\langle v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t}), \nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1}) \rangle}{\gamma_Q} \\
&\stackrel{(a)}{\leq} \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 - \frac{1}{\gamma_Q} \left(2 - \frac{C_{g_{yy}}}{\gamma_Q} \right) \langle v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t}), \nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1}) \rangle
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\leq} \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 - \frac{1}{\gamma_Q} \langle v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t}), \nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1}) \rangle \\
&\stackrel{(c)}{\leq} \left(1 - \frac{\mu}{\gamma_Q}\right) \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 \\
&\stackrel{(d)}{\leq} e^{-\mu(Q-k_3)/\gamma_Q} \|v_t^{k_3} - \hat{v}^*(x_t, y_t^{P_t})\|^2 \\
&\stackrel{(e)}{\leq} e^{-\mu(Q-k_3)/\gamma_Q} \left(\frac{2\epsilon_y}{\mu^2} + \frac{8C_{f_y}^2}{\mu^2} + 2 \log(C_\gamma/\gamma_0) + 1 \right), \tag{25}
\end{aligned}$$

where (a) uses Lemma 3; (b) follows from $\gamma_Q > C_\gamma \geq C_{g_{yy}}$; (c) uses $\nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t})) = 0$ and Lemma 3; (d) follows from $\gamma_Q \geq C_\gamma \geq C_{g_{yy}} \geq \mu$ and $1 - m \leq e^{-m}$ for $0 < m < 1$; (e) uses eq. (24). Similar to eq. (19), we have the upper-bound of γ_Q as

$$\gamma_Q = \gamma_{Q-1} + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1})\|^2}{\gamma_Q + \gamma_{Q-1}} \leq \gamma_{k_3} + \sum_{q=k_3}^{Q-1} \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2}{\gamma_{q+1}}. \tag{26}$$

To further bound the last term on the right-hand side of eq. (26), we can have the following result:

$$\begin{aligned}
\|v_t^Q - \hat{v}^*(x_t, y_t^{P_t})\|^2 &= \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1})\|^2}{\gamma_Q^2} \\
&\quad - \frac{2\langle v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t}), \nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1}) \rangle}{\gamma_Q} \\
&\stackrel{(a)}{\leq} \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 + \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1})\|^2}{\gamma_Q^2} \\
&\quad - \frac{2\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1}) - \nabla_v R(x_t, y_t^{P_t}, \hat{v}^*(x_t, y_t^{P_t}))\|^2}{\gamma_Q C_{g_{yy}}} \\
&\stackrel{(b)}{\leq} \|v_t^{Q-1} - \hat{v}^*(x_t, y_t^{P_t})\|^2 - \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q-1})\|^2}{\gamma_Q C_{g_{yy}}} \\
&\leq \|v_t^{k_3} - \hat{v}^*(x_t, y_t^{P_t})\|^2 - \sum_{q=k_3}^{Q-1} \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2}{\gamma_{q+1} C_{g_{yy}}}, \tag{27}
\end{aligned}$$

where (a) uses Lemma 3; (b) refers to $\gamma_Q \geq C_\gamma \geq C_{g_{yy}}$. By rearranging eq. (27) and using eq. (24), we have

$$\begin{aligned}
\sum_{q=k_3}^{Q-1} \frac{\|\nabla_v R(x_t, y_t^{P_t}, v_t^q)\|^2}{\gamma_{q+1}} &\leq C_{g_{yy}} (\|v_t^{k_3} - \hat{v}^*(x_t, y_t^{P_t})\|^2 - \|v_t^Q - \hat{v}^*(x_t, y_t^{P_t})\|^2) \\
&\leq C_{g_{yy}} \left(\frac{2\epsilon_y}{\mu^2} + \frac{8C_{f_y}^2}{\mu^2} + 2 \log(C_\gamma/\gamma_0) + 1 \right). \tag{28}
\end{aligned}$$

Plugging eq. (23) into eq. (20), we obtain the upper-bound of γ_Q as

$$\gamma_Q \leq C_\gamma + C_{g_{yy}} \left(\frac{2\epsilon_y}{\mu^2} + \frac{8C_{f_y}^2}{\mu^2} + 2 \log(C_\gamma/\gamma_0) + 1 \right) =: \gamma_{\max}. \tag{29}$$

Then, we have the upper bound of $\|v_t^Q - \hat{v}^*(x_t, y_t^{P_t})\|^2$ as

$$\|v_t^Q - \hat{v}^*(x_t, y_t^{P_t})\|^2 \leq e^{-\mu(Q_t - k_3)/\gamma_{\max}} \left(\frac{2\epsilon_y}{\mu^2} + \frac{8C_{f_y}^2}{\mu^2} + 2\log(C_\gamma/\gamma_0) + 1 \right). \quad (30)$$

Recall we have the upper bound $k_3 \leq \frac{\log(C_\gamma^2/\gamma_0^2)}{\log(1+\epsilon_v/C_\gamma^2)}$. Note that Q' defined in (15b) satisfies

$$Q' \geq k_3 + \frac{\gamma_{\max}}{\mu} \log(C_{g_{yy}}^2(\gamma_{\max} - C_\gamma)/\epsilon_v).$$

By replacing Q with Q' in eq. (30), we have

$$\|\nabla_v R(x_t, y_t^{P_t}, v_t^{Q'})\|^2 \leq C_{g_{yy}}^2 \|v_t^{Q'} - \hat{v}^*(x_t, y_t^{P_t})\|^2 \leq e^{-\mu(Q' - k_3)/\gamma_{\max}} (\gamma_{\max} - C_\gamma) \leq \epsilon_v.$$

Therefore, $Q_t \leq Q'$ and this completes the proof of (15b). Thus, the proof is complete. \square

E.6 Proof of Theorem 1

Here we suppose the total iteration round is T . According to Lemma 4, the proof can be split into the following two cases.

Case 1: k_1 does not exist. Based on Lemma 4, we have $\alpha_T \leq C_\alpha$. Then by Lemma 7 we have

$$\frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} \leq 2(\Phi(x_t) - \Phi(x_{t+1})) + \frac{L_\Phi}{2\alpha_{t+1}^2} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{\epsilon'}{\alpha_{t+1}},$$

where ϵ' is defined in Lemma 7. By taking the average, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} &\leq \frac{2}{T} (\Phi(x_0) - \Phi(x_T)) + \frac{L_\Phi}{2\alpha_0^2} \frac{1}{T} \sum_{t=0}^{T-1} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{1}{T} \sum_{t=0}^{T-1} \frac{\epsilon'}{\alpha_{t+1}} \\ &\leq \frac{1}{T} \left(2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2} \right) + \frac{\epsilon'}{\alpha_0} = \frac{c_0}{T} + \frac{\epsilon'}{\alpha_0}, \end{aligned} \quad (31)$$

where c_0 is defined by eq. (11) in Lemma 8.

Case 2: k_1 exists. For $t < k_1$, according to Lemma 7, we still have

$$\frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} \leq 2(\Phi(x_t) - \Phi(x_{t+1})) + \frac{L_\Phi}{2\alpha_{t+1}^2} \|\bar{\nabla}f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{\epsilon'}{\alpha_{t+1}}. \quad (32)$$

For $t \geq k_1$, we have $\alpha_t \geq C_\alpha$. Using Lemma 7, we have

$$\frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} \leq 2(\Phi(x_t) - \Phi(x_{t+1})) + \frac{\epsilon'}{\alpha_{t+1}}. \quad (33)$$

By merging eq. (32) and eq. (33), and taking an average from $t = 0, \dots, T-1$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} = \frac{1}{T} \sum_{t=0}^{k_1-1} \frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}} + \frac{1}{T} \sum_{t=k_1}^{T-1} \frac{\|\nabla\Phi(x_t)\|^2}{\alpha_{t+1}}$$

$$\begin{aligned}
&\leq \frac{2}{T} (\Phi(x_0) - \Phi(x_T)) + \frac{L_\Phi}{2\alpha_0^2} \frac{1}{T} \sum_{t=0}^{k_1-1} \|\bar{\nabla} f(x_t, y_t^{P_t}, v_t^{Q_t})\|^2 + \frac{1}{T} \sum_{t=0}^{T-1} \frac{\epsilon'}{\alpha_{t+1}} \\
&\leq \frac{1}{T} \left(2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2} \right) + \frac{\epsilon'}{\alpha_0} = \frac{c_0}{T} + \frac{\epsilon'}{\alpha_0}, \tag{34}
\end{aligned}$$

where c_0 is defined in Lemma 8. This result is the same as eq. (31). Thus, for both **Case 1** and **Case 2**, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_T} \leq \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1}} \leq \frac{1}{T} \left(2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2} \right) + \frac{\epsilon'}{\alpha_0},$$

which indicates that

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 &\leq \left[\frac{1}{T} \left(2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2} \right) + \frac{\epsilon'}{\alpha_0} \right] \alpha_T \\
&\stackrel{(a)}{\leq} \frac{1}{T} \left[\left(2(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{2\alpha_0^2} \right) + \frac{T\epsilon'}{\alpha_0} \right] \\
&\quad \times \left[C_\alpha + 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{\alpha_0^2} + \frac{2T\epsilon'}{\alpha_0} \right], \tag{35}
\end{aligned}$$

where (a) uses Lemma 8. To achieve the $\mathcal{O}(1/T)$ convergence rate, we need $\epsilon' = \mathcal{O}(1/T)$ in eq. (35). This can be guaranteed by taking $\epsilon_y = 1/T$ and $\epsilon_v = 1/T$, which implies (see Lemma 7)

$$\epsilon' = \frac{1}{T} \left[\left(\frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right) + 1 \right) L_{g,1}^2 \bar{L}^2 + \frac{2\bar{L}^2}{\mu^2} \right]. \tag{36}$$

For symbol convenience, here we define

$$c_1 := c_0 + \frac{1}{\alpha_0} \left[\left(\frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right) + 1 \right) L_{g,1}^2 \bar{L}^2 + \frac{2\bar{L}^2}{\mu^2} \right], \tag{37}$$

where c_0 is defined in eq. (11). Thus, we can obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 \leq \frac{c_1(C_\alpha + 2c_1)}{T} = \mathcal{O}\left(\frac{1}{T}\right).$$

Thus, Theorem 1 is proved.

E.7 Complexity Analysis of Algorithm 1 (Proof of Corollary 1)

Recall in Theorem 1, we take $\epsilon_y = 1/T$, $\epsilon_v = 1/T$, and we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 \leq \frac{c_1(C_\alpha + 2c_1)}{T}.$$

To achieve ϵ -accurate stationary point, we need

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\|^2 \leq \frac{c_1(C_\alpha + 2c_1)}{T} \leq \epsilon \quad \text{i.e.,} \quad T = \mathcal{O}(1/\epsilon). \quad (38)$$

Recall in Lemma 9, we have

$$\begin{aligned} P_t &\leq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1 + \epsilon_y/C_\beta^2)} + \frac{\beta_{\max}}{\mu} \log\left(\frac{L_{g,1}^2(\beta_{\max} - C_\beta)}{\epsilon_y}\right) \\ &\leq \frac{\log(C_\beta^2/\beta_0^2)}{\log(1 + 1/C_\beta^2 T)} + \frac{\beta_{\max}}{\mu} \log\left(\frac{TL_{g,1}^2(\beta_{\max} - C_\beta)}{1}\right) = \mathcal{O}\left(\frac{1}{\log(1 + \epsilon)} + \log\left(\frac{1}{\epsilon}\right)\right). \end{aligned}$$

When ϵ is sufficiently small, we have

$$P_t = \mathcal{O}\left(\frac{1}{\log(1 + \epsilon)} + \log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left(\frac{1}{\epsilon} + \log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}(1/\epsilon). \quad (39)$$

Similarly, we have

$$Q_t = \mathcal{O}\left(\frac{1}{\log(1 + \epsilon)} + \log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left(\frac{1}{\epsilon} + \log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}(1/\epsilon). \quad (40)$$

We denote $\text{Gc}(\epsilon)$ as the gradient complexity, then we have

$$\text{Gc}(\epsilon) = T \cdot \max_t \{P_t + Q_t\} = \mathcal{O}(1/\epsilon^2).$$

Therefore Corollary 1 is proved.

F Proof of Theorem 2

We define some notation for convenience before proving Theorem 2.

F.1 Notation

Below, we define several preset constants for notational convenience at their first use. We first define some Lipschitzness parameters for $\Phi(x)$ as

$$\begin{aligned} L_\Phi &:= \left(L_{f,1} + \frac{L_{g,2}C_{f_y}}{\mu} \right) \left(1 + \frac{C_{g_{xy}}}{\mu} \right)^2 \\ \bar{L} &:= \max \left\{ 2 \left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2 \right)^{\frac{1}{2}}, \sqrt{2}C_{g_{yy}} \right\}. \end{aligned}$$

Next, we define the following constants as thresholds for parameters $\beta_k, \gamma_k, \alpha_k$ as

$$\begin{aligned} C_\alpha &:= \max \left\{ \frac{2L_\Phi}{\varphi_0}, \alpha_0 \right\}, \\ C_\beta &:= \max \left\{ \mu + L_{g,1}, \frac{2\mu L_{g,1}}{\mu + L_{g,1}}, \beta_0, 64a_0^2, 1 \right\}, \\ C_\gamma &:= \max \left\{ 2(\mu + C_{g_{yy}}), \frac{\mu C_{g_{yy}}}{\mu + C_{g_{yy}}}, \gamma_0, 64a_0^2, 1, C_{g_{yy}} \right\}, \\ C_\varphi &:= C_\beta + C_\gamma, \end{aligned} \tag{41}$$

where the constant α_0 is defined as

$$\begin{aligned} a_0 &:= \left(\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right) \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \\ &\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} \gamma_0}. \end{aligned}$$

F.2 A rough bound of v_k

Lemma 10. *Under Assumptions 1, 2, for any $t \geq 0$ in Algorithm 2, we have $\|v_t\| \leq \frac{\sqrt{2}}{\mu} \varphi_{t+1} + \frac{\sqrt{2}C_{f_y}}{\mu} \sqrt{t}$.*

Proof. By strong convexity of g in Assumption 1, we have

$$\begin{aligned} \sum_{k=1}^t \mu^2 \|v_k\|^2 &\leq \sum_{k=1}^t \|\nabla_y \nabla_y g(x_k, y_k) v_k\|^2 \\ &\leq \sum_{k=1}^t 2 \|\nabla_y \nabla_y g(x_k, y_k) v_k - \nabla_y f(x_k, y_k)\|^2 + \sum_{k=1}^t 2 \|\nabla_y f(x_k, y_k)\|^2 \\ &= \sum_{k=1}^t 2 \|\nabla_v R(x_k, y_k, v_k)\|^2 + \sum_{k=1}^t 2 \|\nabla_y f(x_k, y_k)\|^2 \\ &\leq 2\gamma_{t+1}^2 + 2tC_{f_y}^2, \end{aligned}$$

which indicates that for any $t \geq 0$, $\|v_t\|$ can be bounded as

$$\|v_t\| \leq \frac{(2\gamma_{t+1}^2 + 2tC_{f_y}^2)^{\frac{1}{2}}}{\mu} \leq \frac{(2\varphi_{t+1}^2 + 2tC_{f_y}^2)^{\frac{1}{2}}}{\mu} \leq \frac{\sqrt{2}(\varphi_{t+1} + \sqrt{t}C_{f_y})}{\mu}. \quad (42)$$

Then the proof is complete. \square

F.3 Descent in Objective Function

Lemma 11. *Under Assumptions 1, 2, for Algorithm 2, suppose the total iteration number is T . No matter k_1 in Lemma 4 exists or not, we always have*

$$\begin{aligned} \Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \left(1 - \frac{L_\Phi}{\alpha_{t+1}\varphi_{t+1}}\right) \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &\quad + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1}\varphi_{t+1}} + \frac{\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1}\varphi_{t+1}}. \end{aligned} \quad (43)$$

If in addition, k_1 in Lemma 4 exists, then for $t \geq k_1$, we further have

$$\begin{aligned} \Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{4\alpha_{t+1}\varphi_{t+1}} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &\quad + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1}\varphi_{t+1}} + \frac{\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1}\varphi_{t+1}}, \end{aligned} \quad (44)$$

where $\bar{L} := \max\left\{2\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2\right)^{\frac{1}{2}}, \sqrt{2}C_{g_{yy}}\right\}$.

Proof. From Lemma 2, we have $\Phi(x)$ is L_Φ -smooth. So we can apply the descent lemma to Φ as

$$\begin{aligned} \Phi(x_{t+1}) &\leq \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t \rangle + \frac{L_\Phi}{2} \|x_{t+1} - x_t\|^2 \\ &= \Phi(x_t) - \frac{1}{\alpha_{t+1}\varphi_{t+1}} \langle \nabla\Phi(x_t), \bar{\nabla}f(x_t, y_t, v_t) \rangle + \frac{L_\Phi}{2\alpha_{t+1}^2\varphi_{t+1}^2} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &= \Phi(x_t) - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &\quad + \frac{1}{2\alpha_{t+1}\varphi_{t+1}} \|\nabla\Phi(x_t) - \bar{\nabla}f(x_t, y_t, v_t)\|^2 + \frac{L_\Phi}{2\alpha_{t+1}^2\varphi_{t+1}^2} \|\bar{\nabla}f(x_t, y_t, v_t)\|^2, \end{aligned} \quad (45)$$

and the approximation error

$$\begin{aligned} &\|\nabla\Phi(x_t) - \bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &= \|\bar{\nabla}f(x_t, y^*(x_t), v^*(x_t)) - \bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &\leq 2\|\bar{\nabla}f(x_t, y^*(x_t), v^*(x_t)) - \bar{\nabla}f(x_t, y_t, v^*(x_t))\|^2 + 2\|\bar{\nabla}f(x_t, y_t, v^*(x_t)) - \bar{\nabla}f(x_t, y_t, v_t)\|^2 \\ &\leq 4\|\nabla_y \nabla_y g(x_t, y^*(x_t))v^*(x_t) - \nabla_y \nabla_y g(x_t, y_t)v^*(x_t)\|^2 \\ &\quad + 4\|\nabla_y f(x_t, y^*(x_t)) - \nabla_y f(x_t, y_t)\|^2 + 2\|\nabla_y \nabla_y g(x_t, y_t)(v^*(x_t) - v_t)\|^2 \end{aligned}$$

$$\begin{aligned}
&\leq 4\left(\frac{C_{f_y}^2 L_{g,2}^2}{\mu^2} + L_{f,1}^2\right)\|y_t - y^*(x_t)\|^2 + 2C_{g_{yy}}^2\|v_t - v^*(x_t)\|^2 \\
&\leq \bar{L}^2(\|y_t - y^*(x_t)\|^2 + \|v_t - v^*(x_t)\|^2),
\end{aligned} \tag{46}$$

where the third inequality used results from Lemma 3. By plugging eq. (46) into eq. (45), we have

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{1}{2\alpha_{t+1}\varphi_{t+1}}\|\nabla\Phi(x_t)\|^2 - \frac{1}{2\alpha_{t+1}\varphi_{t+1}}\left(1 - \frac{L_\Phi}{\alpha_{t+1}\varphi_{t+1}}\right)\|\bar{\nabla}f(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{\bar{L}^2}{2\alpha_{t+1}\varphi_{t+1}}(\|y_t - y^*(x_t)\|^2 + \|v_t - v^*(x_t)\|^2).
\end{aligned} \tag{47}$$

Note that $g(x, y)$ is μ -strongly convex in y and $R(x, y, v)$ is μ -strongly convex in v . So here we can bound the approximation gaps $\|y_t - y^*(x_t)\|^2 + \|v_t - v^*(x_t)\|^2$ by $\|\nabla_y g(x_t, y_t)\|^2$ and $\|\nabla_v R(x_t, y_t, v_t)\|^2$ as

$$\begin{aligned}
&\|y_t - y^*(x_t)\|^2 + \|v_t - v^*(x_t)\|^2 \\
&\stackrel{(a)}{\leq} \frac{1}{\mu^2}\|\nabla_y g(x_t, y_t) - \nabla_y g(x_t, y^*(x_t))\|^2 + \frac{1}{\mu^2}\|\nabla_v R(x_t, y_t, v_t) - \nabla_v R(x_t, y_t, v^*(x_t))\|^2 \\
&\stackrel{(b)}{\leq} \frac{1}{\mu^2}\|\nabla_y g(x_t, y_t)\|^2 + \frac{2}{\mu^2}\|\nabla_v R(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{2}{\mu^2}\|\nabla_v R(x_t, y_t, v^*(x_t)) - \nabla_v R(x_t, y^*(x_t), v^*(x_t))\|^2 \\
&\stackrel{(c)}{\leq} \frac{1}{\mu^2}\|\nabla_y g(x_t, y_t)\|^2 + \frac{2}{\mu^2}\|\nabla_v R(x_t, y_t, v_t)\|^2 + \frac{2}{\mu^2}\left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\|y_t - y^*(x_t)\|^2 \\
&\stackrel{(d)}{\leq} \left[\frac{1}{\mu^2} + \frac{2}{\mu^4}\left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2\right]\|\nabla_y g(x_t, y_t)\|^2 + \frac{2}{\mu^2}\|\nabla_v R(x_t, y_t, v_t)\|^2,
\end{aligned} \tag{48}$$

where (a) and (d) use the strong convexity; (b) and (d) result from $\nabla_y g(x, y^*(x)) = 0$ and $\nabla_v R(x, y^*(x), v^*(x)) = 0$; (c) uses Lemma 3. By plugging eq. (48) into eq. (47), we obtain eq. (43).

Now if in addition, k_1 in Lemma 4 exists, then for $t \geq k_1$, we have $\alpha_{t+1} > C_\alpha \geq 2L_\Phi/\varphi_0$. From (43) we can immediately obtain (44). Thus, the proof is complete. \square

Note that to further explore the bounds of the right-hand side of eq. (43) and eq. (44) in the above lemma, we next show the (summed) bounds of $\frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}}$ and $\frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}}$.

Lemma 12. *Under Assumptions 1, 2, for Algorithm 2, suppose the total iteration rounds is T . If k_2 in Lemma 4 exists within T iterations, for all integer $t \in [k_2, T]$, we have*

$$\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \leq \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}}.$$

Proof. For $k_2 \leq t < T$, we have $\beta_{k_2} \leq C_\beta$ and $\beta_{t+1} > C_\beta$. For any positive scalar $\bar{\lambda}_{t+1}$, using Young's inequality, we have

$$\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 + \bar{\lambda}_{t+1})\|y_{t+1} - y^*(x_t)\|^2 + \left(1 + \frac{1}{\bar{\lambda}_{t+1}}\right)\|y^*(x_t) - y^*(x_{t+1})\|^2. \tag{49}$$

For the first term on the right hand side of eq. (49), we have

$$\begin{aligned}
& \|y_{t+1} - y^*(x_t)\|^2 \\
&= \left\| y_t - \frac{1}{\beta_{t+1}} \nabla_y g(x_t, y_t) - y^*(x_t) \right\|^2 \\
&= \|y_t - y^*(x_t)\|^2 + \frac{1}{\beta_{t+1}^2} \|\nabla_y g(x_t, y_t)\|^2 - \frac{2}{\beta_{t+1}} \langle y_t - y^*(x_t), \nabla_y g(x_t, y_t) \rangle \\
&\stackrel{(a)}{\leq} \left(1 - \frac{2\mu L_{g,1}}{\beta_{t+1}(\mu + L_{g,1})} \right) \|y_t - y^*(x_t)\|^2 + \frac{1}{\beta_{t+1}} \left(\frac{1}{\beta_{t+1}} - \frac{2}{\mu + L_{g,1}} \right) \|\nabla_y g(x_t, y_t)\|^2 \\
&\stackrel{(b)}{\leq} \left(1 - \frac{2\mu L_{g,1}}{\beta_{t+1}(\mu + L_{g,1})} \right) \|y_t - y^*(x_t)\|^2 - \frac{1}{\beta_{t+1}(\mu + L_{g,1})} \|\nabla_y g(x_t, y_t)\|^2, \tag{50}
\end{aligned}$$

where (a) uses Lemma 3.11 in [10]; (b) follows from $\beta_{t+1} \geq C_\beta \geq \mu + L_{g,1}$. By plugging eq. (50) into eq. (49), we have

$$\begin{aligned}
& \|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\leq (1 + \bar{\lambda}_{t+1}) \left(1 - \frac{2\mu L_{g,1}}{\beta_{t+1}(\mu + L_{g,1})} \right) \|y_t - y^*(x_t)\|^2 - (1 + \bar{\lambda}_{t+1}) \frac{1}{\beta_{t+1}(\mu + L_{g,1})} \|\nabla_y g(x_t, y_t)\|^2 \\
&\quad + \left(1 + \frac{1}{\bar{\lambda}_{t+1}} \right) \|y^*(x_t) - y^*(x_{t+1})\|^2. \tag{51}
\end{aligned}$$

By rearranging the terms in eq. (51), we have

$$\begin{aligned}
& (1 + \bar{\lambda}_{t+1}) \frac{1}{\beta_{t+1}(\mu + L_{g,1})} \|\nabla_y g(x_t, y_t)\|^2 \\
&\leq (1 + \bar{\lambda}_{t+1}) \left(1 - \frac{2\mu L_{g,1}}{\beta_{t+1}(\mu + L_{g,1})} \right) \|y_t - y^*(x_t)\|^2 - \|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\quad + \left(1 + \frac{1}{\bar{\lambda}_{t+1}} \right) \|y^*(x_t) - y^*(x_{t+1})\|^2.
\end{aligned}$$

We take $\bar{\lambda}_{t+1} := \frac{2\mu L_{g,1}}{\beta_{t+1}(\mu + L_{g,1})}$. Since $\beta_{t+1} > C_\beta \geq \frac{2\mu L_{g,1}}{\mu + L_{g,1}}$ in eq. (41), we have $\bar{\lambda}_{t+1} \leq 1$. Then we have

$$\begin{aligned}
\frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}} &\leq (1 + \bar{\lambda}_{t+1}) \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}} \\
&\leq (\mu + L_{g,1}) (\|y_t - y^*(x_t)\|^2 - \|y_{t+1} - y^*(x_{t+1})\|^2) \\
&\quad + \frac{2(\mu + L_{g,1})}{\bar{\lambda}_{t+1}} \|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&= (\mu + L_{g,1}) (\|y_t - y^*(x_t)\|^2 - \|y_{t+1} - y^*(x_{t+1})\|^2) \\
&\quad + \frac{(\mu + L_{g,1})^2 \beta_{t+1}}{\mu L_{g,1}} \|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\stackrel{(a)}{\leq} (\mu + L_{g,1}) (\|y_t - y^*(x_t)\|^2 - \|y_{t+1} - y^*(x_{t+1})\|^2) \\
&\quad + \frac{(\mu + L_{g,1})^2 L_y^2 \beta_{t+1}}{\mu L_{g,1}} \|x_t - x_{t+1}\|^2,
\end{aligned}$$

where (a) uses Lemma 2. Summing the above inequality over $k = k_2, \dots, t$, we have

$$\begin{aligned}
& \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
& \leq \sum_{k=k_2-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
& \leq (\mu + L_{g,1}) \|y_{k_2-1} - y^*(x_{k_2-1})\|^2 + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2-1}^t \beta_{k+1} \|x_k - x_{k+1}\|^2 \\
& \stackrel{(a)}{\leq} \frac{\mu + L_{g,1}}{\mu^2} \|\nabla_y g(x_{k_2-1}, y_{k_2-1}) - \nabla_y g(x_{k_2-1}, y^*(x_{k_2-1}))\|^2 \\
& \quad + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2-1}^t \frac{\beta_{k+1}}{\alpha_{k+1}^2 \varphi_{k+1}^2} \|\bar{\nabla} f(x_k, y_k, v_k)\|^2 \\
& \leq \frac{\mu + L_{g,1}}{\mu^2} \|\nabla_y g(x_{k_2-1}, y_{k_2-1})\|^2 + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \stackrel{(b)}{\leq} \frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \stackrel{(c)}{\leq} \frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}}, \tag{52}
\end{aligned}$$

where (a) uses Assumption 1; (b) results from $\|\nabla_y g(x_{k_2-1}, y_{k_2-1})\|^2 \leq \beta_{k_2}^2 \leq C_\beta^2$; (c) denotes $\varphi_0 = \max\{\beta_0, \gamma_0\}$. Then, the proof is complete. \square

Lemma 13. *Under Assumptions 1, 2, for Algorithm 2, suppose the total iteration rounds is T . If k_3 in Lemma 4 exists within T iterations, for all integer $t \in [k_3, T)$, we have*

$$\begin{aligned}
& \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
& \leq \frac{4(\mu + C_{g_{yy}}) C_\beta^2}{\mu^4} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}}) C_\gamma}{\mu^2} \\
& \quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1}) L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& \quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}.
\end{aligned}$$

Proof. For $k_3 \leq t < T$, we have $\gamma_{t+1} > C_\gamma$. For any positive scalar $\hat{\lambda}_{t+1}$, using Young's inequality, we have

$$\|v_{t+1} - v^*(x_{t+1})\|^2 \leq (1 + \hat{\lambda}_{t+1})\|v_{t+1} - v^*(x_t)\|^2 + \left(1 + \frac{1}{\hat{\lambda}_{t+1}}\right)\|v^*(x_t) - v^*(x_{t+1})\|^2. \quad (53)$$

For the first term on the right hand side of eq. (53), we have

$$\begin{aligned} & \|v_{t+1} - v^*(x_t)\|^2 \\ &= \left\| v_t - \frac{1}{\varphi_{t+1}} \nabla_v R(x_t, y_t, v_t) - v^*(x_t) \right\|^2 \\ &= \|v_t - v^*(x_t)\|^2 + \frac{1}{\varphi_{t+1}^2} \|\nabla_v R(x_t, y_t, v_t)\|^2 - \frac{2}{\varphi_{t+1}} \langle v_t - v^*(x_t), \nabla_v R(x_t, y_t, v_t) \rangle. \end{aligned} \quad (54)$$

For the last term of the right-hand side of eq. (54), we have

$$\begin{aligned} & -\langle v_t - v^*(x_t), \nabla_v R(x_t, y_t, v_t) \rangle \\ &= -\langle v_t - v^*(x_t), \nabla_v R(x_t, y_t, v_t) - \nabla_v R(x_t, y_t, v^*(x_t)) \rangle \\ &\quad - \langle v_t - v^*(x_t), \nabla_v R(x_t, y_t, v^*(x_t)) - \nabla_v R(x_t, y^*(x_t), v^*(x_t)) \rangle \\ &\stackrel{(a)}{\leq} -\frac{1}{\mu + C_{g_{yy}}} \|\nabla_v R(x_t, y_t, v_t) - \nabla_v R(x_t, y_t, v^*(x_t))\|^2 - \frac{\mu C_{g_{yy}}}{\mu + C_{g_{yy}}} \|v_t - v^*(x_t)\|^2 \\ &\quad + \frac{\mu + C_{g_{yy}}}{2\mu C_{g_{yy}}} \|\nabla_v R(x_t, y_t, v^*(x_t)) - \nabla_v R(x_t, y^*(x_t), v^*(x_t))\|^2 \\ &\quad + \frac{\mu C_{g_{yy}}}{2(\mu + C_{g_{yy}})} \|v_t - v^*(x_t)\|^2 \\ &\stackrel{(b)}{\leq} -\frac{1}{2(\mu + C_{g_{yy}})} \|\nabla_v R(x_t, y_t, v_t)\|^2 + \frac{1}{\mu + C_{g_{yy}}} \|\nabla_v R(x_t, y_t, v^*(x_t))\|^2 \\ &\quad + \frac{\mu + C_{g_{yy}}}{2\mu C_{g_{yy}}} \|\nabla_v R(x_t, y_t, v^*(x_t)) - \nabla_v R(x_t, y^*(x_t), v^*(x_t))\|^2 \\ &\quad - \frac{\mu C_{g_{yy}}}{2(\mu + C_{g_{yy}})} \|v_t - v^*(x_t)\|^2 \\ &\stackrel{(c)}{=} -\frac{1}{2(\mu + C_{g_{yy}})} \|\nabla_v R(x_t, y_t, v_t)\|^2 - \frac{\mu C_{g_{yy}}}{2(\mu + C_{g_{yy}})} \|v_t - v^*(x_t)\|^2 \\ &\quad + \left(\frac{1}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{2\mu C_{g_{yy}}} \right) \|\nabla_v R(x_t, y_t, v^*(x_t)) - \nabla_v R(x_t, y^*(x_t), v^*(x_t))\|^2 \\ &\stackrel{(d)}{\leq} -\frac{1}{2(\mu + C_{g_{yy}})} \|\nabla_v R(x_t, y_t, v_t)\|^2 - \frac{\mu C_{g_{yy}}}{2(\mu + C_{g_{yy}})} \|v_t - v^*(x_t)\|^2 \\ &\quad + \left(\frac{1}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{2\mu C_{g_{yy}}} \right) (L_{g,2} \|v^*(x_t)\| + L_{f,1})^2 \|y_t - y^*(x_t)\|^2 \\ &\stackrel{(e)}{\leq} -\frac{1}{2(\mu + C_{g_{yy}})} \|\nabla_v R(x_t, y_t, v_t)\|^2 - \frac{\mu C_{g_{yy}}}{2(\mu + C_{g_{yy}})} \|v_t - v^*(x_t)\|^2 \\ &\quad + \left(\frac{1}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{2\mu C_{g_{yy}}} \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \|y_t - y^*(x_t)\|^2, \end{aligned} \quad (55)$$

where (a) follows from Lemma 3.11 in [10]; (b) uses $-||a - b||^2 \leq -\frac{1}{2}||a||^2 + ||b||^2$ since $||a - b + b||^2 \leq 2||a - b||^2 + 2||b||^2$; (c) uses $\nabla_v R(x_t, y^*(x_t), v^*(x_t)) = 0$; (d) and (e) follow from Lemma 3. Plugging eq. (55) into eq. (54), we have

$$\begin{aligned}
& \|v_{t+1} - v^*(x_t)\|^2 \\
& \leq \left(1 - \frac{\mu C_{g_{yy}}}{(\mu + C_{g_{yy}})\varphi_{t+1}}\right) \|v_t - v^*(x_t)\|^2 + \frac{1}{\varphi_{t+1}} \left(\frac{1}{\varphi_{t+1}} - \frac{1}{\mu + C_{g_{yy}}}\right) \|\nabla_v R(x_t, y_t, v_t)\|^2 \\
& \quad + \left(\frac{2}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{\mu C_{g_{yy}}}\right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2 \frac{1}{\varphi_{t+1}} \|y_t - y^*(x_t)\|^2 \\
& \stackrel{(a)}{\leq} \left(1 - \frac{\mu C_{g_{yy}}}{(\mu + C_{g_{yy}})\varphi_{t+1}}\right) \|v_t - v^*(x_t)\|^2 - \frac{1}{2(\mu + C_{g_{yy}})\varphi_{t+1}} \|\nabla_v R(x_t, y_t, v_t)\|^2 \\
& \quad + \left(\frac{2}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{\mu C_{g_{yy}}}\right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2 \frac{1}{\varphi_{t+1}} \|y_t - y^*(x_t)\|^2, \tag{56}
\end{aligned}$$

where (a) follows from $\varphi_{t+1} \geq \gamma_{t+1} \geq C_\gamma \geq 2(\mu + C_{g_{yy}})$. Combining eq. (56) with eq. (53), we have

$$\begin{aligned}
& \|v_{t+1} - v^*(x_{t+1})\|^2 \\
& \leq (1 + \hat{\lambda}_{t+1}) \left(1 - \frac{\mu C_{g_{yy}}}{(\mu + C_{g_{yy}})\varphi_{t+1}}\right) \|v_t - v^*(x_t)\|^2 \\
& \quad - (1 + \hat{\lambda}_{t+1}) \frac{1}{2(\mu + C_{g_{yy}})\varphi_{t+1}} \|\nabla_v R(x_t, y_t, v_t)\|^2 \\
& \quad + (1 + \hat{\lambda}_{t+1}) \left(\frac{2}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{\mu C_{g_{yy}}}\right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2 \frac{1}{\varphi_{t+1}} \|y_t - y^*(x_t)\|^2 \\
& \quad + \left(1 + \frac{1}{\hat{\lambda}_{t+1}}\right) \|v^*(x_t) - v^*(x_{t+1})\|^2. \tag{57}
\end{aligned}$$

By rearranging the terms in eq. (57), we have

$$\begin{aligned}
& (1 + \hat{\lambda}_{t+1}) \frac{1}{2(\mu + C_{g_{yy}})\varphi_{t+1}} \|\nabla_v R(x_t, y_t, v_t)\|^2 \\
& \leq (1 + \hat{\lambda}_{t+1}) \left(1 - \frac{\mu C_{g_{yy}}}{(\mu + C_{g_{yy}})\varphi_{t+1}}\right) \|v_t - v^*(x_t)\|^2 - \|v_{t+1} - v^*(x_{t+1})\|^2 \\
& \quad + (1 + \hat{\lambda}_{t+1}) \left(\frac{2}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{\mu C_{g_{yy}}}\right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1}\right)^2 \frac{1}{\varphi_{t+1}} \|y_t - y^*(x_t)\|^2 \\
& \quad + \left(1 + \frac{1}{\hat{\lambda}_{t+1}}\right) \|v^*(x_t) - v^*(x_{t+1})\|^2. \tag{58}
\end{aligned}$$

We now take $\hat{\lambda}_{t+1} := \frac{\mu C_{g_{yy}}}{(\mu + C_{g_{yy}})\varphi_{t+1}}$. Since $\varphi_{t+1} \geq \gamma_{t+1} \geq C_\gamma \geq \frac{\mu C_{g_{yy}}}{\mu + C_{g_{yy}}}$ in eq. (41), we have $\hat{\lambda}_{t+1} \leq 1$. Then we get

$$\begin{aligned}
\frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} & < (1 + \hat{\lambda}_{t+1}) \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} \\
& \stackrel{(a)}{\leq} 2(\mu + C_{g_{yy}}) (\|v_t - v^*(x_t)\|^2 - \|v_{t+1} - v^*(x_{t+1})\|^2)
\end{aligned}$$

$$\begin{aligned}
& + 4(\mu + C_{g_{yy}}) \left(\frac{2}{\mu + C_{g_{yy}}} + \frac{\mu + C_{g_{yy}}}{\mu C_{g_{yy}}} \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{\|y_t - y^*(x_t)\|^2}{\varphi_{t+1}} \\
& + 2(\mu + C_{g_{yy}}) \left(1 + \frac{(\mu + C_{g_{yy}}) \varphi_{t+1}}{\mu C_{g_{yy}}} \right) L_v^2 \|x_t - x_{t+1}\|^2 \\
& \stackrel{(b)}{\leq} 2(\mu + C_{g_{yy}}) (\|v_t - v^*(x_t)\|^2 - \|v_{t+1} - v^*(x_{t+1})\|^2) \\
& + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{\|y_t - y^*(x_t)\|^2}{\varphi_{t+1}} \\
& + \frac{4(\mu + C_{g_{yy}})^2 L_v^2 \varphi_{t+1}}{\mu C_{g_{yy}}} \|x_t - x_{t+1}\|^2, \tag{59}
\end{aligned}$$

where (a) multiplies both sides of eq. (58) by $2(\mu + C_{g_{xy}})$ and uses $\hat{\lambda}_{t+1} \leq 1$; (b) uses $\varphi_{t+1} \geq \gamma_{t+1} \geq C_\gamma \geq \frac{\mu C_{g_{yy}}}{\mu + C_{g_{yy}}}$. Take summation of eq. (59) and we have

$$\begin{aligned}
& \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
& \leq \sum_{k=k_3-1}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
& \leq 2(\mu + C_{g_{yy}}) \|v_{k_3-1} - v^*(x_{k_3-1})\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \varphi_{k+1} \|x_k - x_{k+1}\|^2 \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_3-1}^t \frac{\|y_k - y^*(x_k)\|^2}{\varphi_{k+1}} \\
& \leq 2(\mu + C_{g_{yy}}) \|v_{k_3-1} - v^*(x_{k_3-1})\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_3-1}^t \frac{\|y_k - y^*(x_k)\|^2}{\varphi_{k+1}} \\
& \leq 2(\mu + C_{g_{yy}}) \|v_{k_3-1} - v^*(x_{k_3-1})\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_3-1}^t \frac{\|y_k - y^*(x_k)\|^2}{\beta_{k+1}} \\
& \stackrel{(a)}{\leq} 2(\mu + C_{g_{yy}}) \|v_{k_3-1} - v^*(x_{k_3-1})\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k) - \nabla_y g(x_k, y^*(x_k))\|^2}{\beta_{k+1}} \\
& \stackrel{(b)}{\leq} 2(\mu + C_{g_{yy}}) \|v_{k_3-1} - v^*(x_{k_3-1})\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
& \quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{\leq} \frac{2(\mu + C_{g_{yy}})}{\mu^2} \left\| \nabla_v R(x_{k_3-1}, y_{k_3-1}, v_{k_3-1}) - \nabla_v R(x_{k_3-1}, y_{k_3-1}, v^*(x_{k_3-1})) \right\|^2 \\
&\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
&\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\stackrel{(d)}{\leq} \frac{4(\mu + C_{g_{yy}})}{\mu^2} \left\| \nabla_v R(x_{k_3-1}, y^*(x_{k_3-1}), v^*(x_{k_3-1})) - \nabla_v R(x_{k_3-1}, y_{k_3-1}, v^*(x_{k_3-1})) \right\|^2 \\
&\quad + \frac{4(\mu + C_{g_{yy}})}{\mu^2} \left\| \nabla_v R(x_{k_3-1}, y_{k_3-1}, v_{k_3-1}) - \nabla_v R(x_{k_3-1}, y^*(x_{k_3-1}), v^*(x_{k_3-1})) \right\|^2 \\
&\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
&\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\stackrel{(e)}{\leq} \frac{4(\mu + C_{g_{yy}})}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \|y_{k_3-1} - y^*(x_{k_3-1})\|^2 \\
&\quad + \frac{4(\mu + C_{g_{yy}})}{\mu^2} \left\| \nabla_v R(x_{k_3-1}, y_{k_3-1}, v_{k_3-1}) \right\|^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}}} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\
&\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}, \tag{60}
\end{aligned}$$

where (a) uses Assumption 1; (b) results from $\nabla_y g(x, y^*(x)) = 0$; (c) uses the strong convexity in Lemma 3; (d) uses $\nabla_v R(x, y^*(x), v^*(x)) = 0$; (e) follows from Lemma 3.

Our next step is bounding $\|y_{k_3-1} - y^*(x_{k_3-1})\|^2$ on the right hand side of eq. (60) in two cases. **The first case is $\beta_{k_3} \leq C_\beta$.** In this case, by using strong convexity of g and the definition of β_{k_3} , we can easily have

$$\begin{aligned}
\|y_{k_3-1} - y^*(x_{k_3-1})\|^2 &\leq \frac{1}{\mu^2} \left\| \nabla_y g(x_{k_3-1}, y_{k_3-1}) - \nabla_y g(x_{k_3-1}, y^*(x_{k_3-1})) \right\|^2 \\
&= \frac{1}{\mu^2} \left\| \nabla_y g(x_{k_3-1}, y_{k_3-1}) \right\|^2 \leq \frac{\beta_{k_3}^2}{\mu^2} \leq \frac{C_\beta^2}{\mu^2}. \tag{61}
\end{aligned}$$

The second case is $\beta_{k_3} > C_\beta$. This indicates that k_2 exists and $k_3 > k_2$ based on Lemma 4. By plugging $\bar{\lambda}_{k_3-1} := \frac{2\mu L_{g,1}}{\beta_{k_3-1}(\mu + L_{g,1})}$ into eq. (51), and noting $\bar{\lambda}_{k_3-1} \leq 1$, we have

$$\begin{aligned}
\|y_{k_3-1} - y^*(x_{k_3-1})\|^2 &\leq \|y_{k_3-2} - y^*(x_{k_3-2})\|^2 + \frac{(\mu + L_{g,1})\beta_{k_3-1}}{\mu L_{g,1}} \|y^*(x_{k_3-2}) - y^*(x_{k_3-1})\|^2 \\
&\stackrel{(a)}{\leq} \|y_{k_3-2} - y^*(x_{k_3-2})\|^2 + \frac{(\mu + L_{g,1})L_y^2 \beta_{k_3-1}}{\mu L_{g,1}} \|x_{k_3-2} - x_{k_3-1}\|^2 \\
&= \|y_{k_3-2} - y^*(x_{k_3-2})\|^2 + \frac{(\mu + L_{g,1})L_y^2 \beta_{k_3-1}}{\mu L_{g,1}} \frac{\|\bar{\nabla} f(x_{k_3-2}, y_{k_3-2}, v_{k_3-2})\|^2}{\alpha_{k_3-1}^2 \varphi_{k_3-1}^2} \\
&\leq \|y_{k_3-2} - y^*(x_{k_3-2})\|^2 + \frac{(\mu + L_{g,1})L_y^2}{\mu L_{g,1}} \frac{\|\bar{\nabla} f(x_{k_3-2}, y_{k_3-2}, v_{k_3-2})\|^2}{\alpha_{k_3-1}^2 \varphi_{k_3-1}}
\end{aligned}$$

$$\begin{aligned}
&\leq \|y_{k_3-2} - y^*(x_{k_3-2})\|^2 + \frac{(\mu + L_{g,1})L_y^2}{\mu L_{g,1}\varphi_0} \frac{\|\bar{\nabla}f(x_{k_3-2}, y_{k_3-2}, v_{k_3-2})\|^2}{\alpha_{k_3-1}^2} \\
&\leq \|y_{k_2-1} - y^*(x_{k_2-1})\|^2 + \frac{(\mu + L_{g,1})L_y^2}{\mu L_{g,1}\varphi_0} \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
&\stackrel{(b)}{\leq} \frac{C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})L_y^2}{\mu L_{g,1}\varphi_0} \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}, \tag{62}
\end{aligned}$$

where (a) uses Lemma 2; (b) uses eq. (61) by replacing k_3 by k_2 since $\beta_{k_2} \leq C_\beta$ (see Lemma 4). By combining eq. (61) and eq. (62), we obtain a **general** upper bound of $\|y_{k_3-1} - y^*(x_{k_3-1})\|^2$ as

$$\|y_{k_3-1} - y^*(x_{k_3-1})\|^2 \leq \frac{C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})L_y^2}{\mu L_{g,1}\varphi_0} \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}, \tag{63}$$

where we define $\sum_{k=m}^n l_k = 0$ for any $m > n$ and non-negative sequence $\{l_k\}$. By plugging eq. (63) into eq. (60) and using $\|\nabla_v R(x_{k_3-1}, y_{k_3-1}, v_{k_3-1})\|^2 \leq \gamma_{k_3}^2 \leq C_\gamma^2$, we have

$$\begin{aligned}
&\sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\leq \frac{4(\mu + C_{g_{yy}})C_\beta^2}{\mu^4} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})C_\gamma^2}{\mu^2} \\
&\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1}\varphi_0} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
&\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
&\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}.
\end{aligned}$$

Then, the proof is complete. \square

Supported by Lemma 12 and Lemma 13, we derive upper bounds of β_t and φ_t .

Lemma 14. *Suppose the total iteration rounds of Algorithm 2 is T . Under Assumptions 1, 2, if k_2 in Lemma 4 exists within T iterations, we have*

$$\begin{cases} \beta_{t+1} \leq C_\beta, & t < k_2; \\ \beta_{t+1} \leq \left(C_\beta + \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}\varphi_0} \right) + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}C_\beta} \sum_{k=k_2}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}, & t \geq k_2. \end{cases}$$

When such k_2 does not exist, $\beta_{t+1} \leq C_\beta$ holds for any $t < T$.

Proof. According to Lemma 4, the proof can be split into the following three cases.

Case 1: k_2 does not exist: In this case, based on Lemma 4, we have $\beta_T \leq C_\beta$, and hence $\beta_{t+1} \leq C_\beta$ for any $t < T$ because β_t is non-decreasing with t .

Case 2: k_2 exists and $t < k_2$: In this case, based on Lemma 4, we have $\beta_{t+1} \leq C_\beta$.

Case 3: k_2 exists and $t \geq k_2$: Inspired by [84] and using telescoping, we have

$$\begin{aligned}
\beta_{t+1} &= \beta_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1} + \beta_t} \\
&\leq \beta_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}} \\
&\leq \beta_{k_2} + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\stackrel{(a)}{\leq} \left(C_\beta + \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right) + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \sum_{k=k_2}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}, \tag{64}
\end{aligned}$$

where (a) uses lemma 12. Thus, the proof is complete. \square

Lemma 15. Under Assumptions 1, 2, suppose the total iteration rounds of Algorithm 2 is T . If at least one of k_2 and k_3 in Lemma 4 exists, we denote $k_{\min} := \min\{k_2, k_3\}$. Then we have the upper bound of φ_t as

$$\begin{cases} \varphi_t \leq C_\varphi, & t \leq k_{\min}; \\ \varphi_t \leq a_1 \log(t) + b_1, & t > k_{\min}, \end{cases}$$

where a_1, b_1 are defined as

$$a_1 := 6a_0, \quad b_1 := 4a_0 \log \left(1 + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 4a_0 \log(C_{g_{xy}} \bar{a}) + 4a_0 + 2b_0, \tag{65}$$

in which we define constants

$$\begin{aligned}
\bar{a} &:= \frac{\sqrt{2}}{\mu}, \quad \bar{b} := \frac{\sqrt{2} C_{f_y}}{\mu}, \\
a_0 &:= \left(\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right) \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \\
&\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1}) L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} \gamma_0}, \\
b_0 &:= C_\beta + C_\gamma + \frac{4(\mu + C_{g_{yy}}) C_\beta^2}{\mu^4} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}}) C_\gamma^2}{\mu^2} \\
&\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \left(\frac{C_\beta^2}{\beta_0} - \beta_0 \right) \\
&\quad + \left[\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right] \left(\frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right). \tag{66}
\end{aligned}$$

When such k_2 and k_3 do not exist, we have $\varphi_t \leq C_\varphi$ for all $t \leq T$.

Proof. To begin with, we first show the following result as the first two lines of eq. (64): since β_t and γ_t are positive and increasing monotonically with t , we can easily have

$$0 \leq \min\{\beta_{t+1}^2, \gamma_{t+1}^2\} - \min\{\beta_t^2, \gamma_t^2\}$$

$$\begin{aligned}
&= (\beta_{t+1}^2 + \gamma_{t+1}^2 - \max\{\beta_{t+1}^2, \gamma_{t+1}^2\}) - (\beta_t^2 + \gamma_t^2 - \max\{\beta_t^2, \gamma_t^2\}) \\
&\stackrel{(a)}{=} (\beta_{t+1}^2 + \gamma_{t+1}^2) - (\beta_t^2 + \gamma_t^2) - (\varphi_{t+1}^2 - \varphi_t^2),
\end{aligned}$$

where (a) uses the definition $\varphi_t := \max\{\beta_t, \gamma_t\}$. Similar to eq. (64), we have

$$\varphi_{t+1}^2 - \varphi_t^2 \leq (\beta_{t+1}^2 - \beta_t^2) + (\gamma_{t+1}^2 - \gamma_t^2) = \|\nabla_y g(x_t, y_t)\|^2 + \|\nabla_v R(x_t, y_t, v_t)\|^2,$$

which indicates that

$$\begin{aligned}
\varphi_{t+1} &\leq \varphi_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\varphi_{t+1} + \varphi_t} + \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1} + \varphi_t} \\
&\leq \varphi_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1} + \beta_t} + \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} \\
&\leq \varphi_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}} + \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}}.
\end{aligned} \tag{67}$$

Note that, to simplify the proof, we define $\sum_{k=m}^n l_k = 0$ for any $m > n$ and non-negative sequence $\{l_k\}$. According to the definitions of k_2 and k_3 in Lemma 4, the proof can be split into the following four cases.

Case 1: neither k_2 nor k_3 exists: for any $t \in (0, T)$, we can easily have $\varphi_t = \max\{\beta_t, \gamma_t\} \leq \max\{C_\beta, C_\gamma\} \leq C_\varphi$.

Case 2: k_2 exists but k_3 does not: by using the third line of eq. (64), for any $t \in (0, T)$, we have

$$\varphi_{t+1} \leq \beta_{t+1} + \gamma_{t+1} \leq C_\beta + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + C_\gamma, \tag{68}$$

where we take $\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} = 0$ for any $t < k_2$.

Case 3: k_3 exists but k_2 does not: from the second line of eq. (67), for any $t \in (0, T)$, we have

$$\begin{aligned}
\varphi_{t+1} &\stackrel{(a)}{\leq} \varphi_t + \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1} + \beta_t} + \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} \\
&\leq \varphi_{k_3} + \sum_{k=k_3}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1} + \beta_k} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\leq \beta_{k_3} + \gamma_{k_3} + \sum_{k=k_3}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1} + \beta_k} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\stackrel{(b)}{=} \beta_{t+1} + \gamma_{k_3} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\leq \beta_{t+1} + C_\gamma + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\leq C_\beta + C_\gamma + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}},
\end{aligned} \tag{69}$$

where (a) uses the second line of eq. (67); and we take $\sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} = 0$ for any $t < k_3$; (b) uses the first line of eq. (64).

Case 4: both k_2 and k_3 exist: from the third line of eq. (69), for any $t \in (0, T)$, we have

$$\begin{aligned}
\varphi_{t+1} &\leq \beta_{k_3} + \gamma_{k_3} + \sum_{k=k_3}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\stackrel{(a)}{\leq} \beta_{k_2} + \sum_{k=k_2}^{k_3-1} \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + C_\gamma + \sum_{k=k_3}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&= C_\beta + C_\gamma + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}, \tag{70}
\end{aligned}$$

where (a) uses the third line of eq. (64); and we take $\sum_{k=k_2}^{k_3-1} \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} = 0$ when $k_2 \geq k_3$, and $\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} = 0$ for any $t < k_2$ and $\sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\gamma_{k+1}} = 0$ for any $t < k_3$. It is easy to see that the upper bound of φ_{t+1} in eq. (70) is the largest among all cases. Thus, in the remaining proof, we only explore the upper bound of φ_t in **Case 4**.

To further explore the bound of φ_t , we need to use some auxiliary results and bounds. So we split them into three parts as follows.

Part I: an auxiliary bound of $\sum \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$.

To further explore **Case 4**, we begin with a common term $\sum_{k=k_0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$ for any $k_0 \leq t$. Recall in Lemma 10, we have

$$\|v_k\| \leq \frac{\sqrt{2}}{\mu} \varphi_{k+1} + \frac{\sqrt{2} C_{f_y}}{\mu} \sqrt{k} =: \bar{a} \varphi_{k+1} + \bar{b} \sqrt{k},$$

where \bar{a} and \bar{b} refer to eq. (66). According to Lemma 1, since $\alpha_0 \geq 1$, for any integer $t > 0$, we have

$$\begin{aligned}
\sum_{k=k_0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} &\leq \sum_{k=0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
&\leq \log \left(\sum_{k=0}^t \|\bar{\nabla} f(x_k, y_k, v_k)\|^2 + \alpha_0^2 \right) + 1 \\
&\stackrel{(a)}{\leq} \log \left(\sum_{k=0}^t (C_{g_{xy}} \bar{a} \varphi_{k+1} + C_{g_{xy}} \bar{b} \sqrt{k} + C_{f_x})^2 + \alpha_0^2 \right) + 1 \\
&\leq \log \left(\left(\sum_{k=0}^t C_{g_{xy}} \bar{a} \varphi_{k+1} + C_{g_{xy}} \bar{b} \sqrt{k} + C_{f_x} + \alpha_0 \right)^2 \right) + 1 \\
&= 2 \log \left(\sum_{k=0}^t C_{g_{xy}} \bar{a} \varphi_{k+1} + C_{g_{xy}} \bar{b} \sqrt{k} + C_{f_x} + \alpha_0 \right) + 1 \\
&\leq 2 \log \left((t+1) (C_{g_{xy}} \bar{a} \varphi_{t+1} + C_{g_{xy}} \bar{b} \sqrt{t} + C_{f_x} + \alpha_0) \right) + 1
\end{aligned}$$

$$\begin{aligned}
&= 2\log(t+1) + 2\log(C_{g_{xy}}\bar{a}\varphi_{t+1} + C_{g_{xy}}\bar{b}\sqrt{t} + C_{f_x} + \alpha_0) + 1 \\
&\leq 2\log(t+1) + 2\log((C_{g_{xy}}\bar{a}\varphi_{t+1} + C_{g_{xy}}\bar{b} + C_{f_x} + \alpha_0)\sqrt{t}) + 1 \\
&\leq 3\log(t+1) + 2\log(C_{g_{xy}}\bar{a}\varphi_{t+1} + C_{g_{xy}}\bar{b} + C_{f_x} + \alpha_0) + 1, \tag{71}
\end{aligned}$$

where (a) follows from Remark 3 and Lemma 10. Therefore, we obtain the upper bound of $\sum_{k=k_0}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$ for any $k_0 \leq t$ in eq. (71). **Part I** is completed.

Part II: a more general bound of $\sum \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$.

In Lemma 12, we show the bound of $\sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$ when k_2 exists. In **Part II**, we further provide a rough bound of $\sum_{k=\tilde{k}}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$ for any potential $\tilde{k} \leq T$. Firstly, if $\tilde{k} \geq k_2$, it is easy to have

$$\sum_{k=\tilde{k}}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \leq \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}};$$

secondly, if $\tilde{k} < k_2$, we have

$$\begin{aligned}
\sum_{k=\tilde{k}}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} &\leq \sum_{k=\tilde{k}}^{k_2-1} \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\leq \frac{\sum_{k=\tilde{k}}^{k_2-1} \|\nabla_y g(x_k, y_k)\|^2}{\beta_0} + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\leq \frac{\beta_{k_2}^2 - \beta_{\tilde{k}}^2}{\beta_0} + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\leq \frac{C_\beta^2 - \beta_0^2}{\beta_0} + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&= \frac{C_\beta^2}{\beta_0} - \beta_0 + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}.
\end{aligned}$$

Combining these two situations, since $C_\beta \geq \beta_0$, for any $\tilde{k} \leq t$, we have

$$\begin{aligned}
\sum_{k=\tilde{k}}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} &\leq \frac{C_\beta^2}{\beta_0} - \beta_0 + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\stackrel{(a)}{\leq} \frac{C_\beta^2}{\beta_0} - \beta_0 + \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \\
&\quad + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}}, \tag{72}
\end{aligned}$$

where (a) uses Lemma 12. Thus, **Part II** is completed.

Part III: the bound of φ_t in Case 4.

Here, we explore the upper bound of φ_t in **Case 4**. Recalling eq. (70), we have

$$\varphi_{t+1} \leq C_\beta + C_\gamma + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} = C_\beta + C_\gamma = C_\varphi,$$

for $t \leq k_{\min} := \min\{k_2, k_3\}$. For $t > k_{\min}$, we have

$$\begin{aligned} \varphi_{t+1} &\leq C_\beta + C_\gamma + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\ &\stackrel{(a)}{\leq} C_\beta + C_\gamma + \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\ &\quad + \frac{4(\mu + C_{g_{yy}})C_\beta^2}{\mu^4} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})C_\gamma^2}{\mu^2} \\ &\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1}\varphi_0} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\ &\stackrel{(b)}{\leq} \left[\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right] \sum_{k=k_2}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\ &\quad + C_\beta + C_\gamma + \frac{4(\mu + C_{g_{yy}})C_\beta^2}{\mu^4} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})C_\gamma^2}{\mu^2} \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \left(\frac{C_\beta^2}{\beta_0} - \beta_0 \right) \\ &\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1}\varphi_0} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\stackrel{(c)}{\leq} \left(\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right) \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1}} \sum_{k=k_2}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2 \varphi_{k+1}} \\ &\quad + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1}\varphi_0} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\quad + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\quad + C_\beta + C_\gamma + \frac{4(\mu + C_{g_{yy}})C_\beta^2}{\mu^4} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})C_\gamma^2}{\mu^2} \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \left(\frac{C_\beta^2}{\beta_0} - \beta_0 \right) \end{aligned}$$

$$\begin{aligned}
& + \left[\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right] \left(\frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right) \\
\leq & \left[\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right] \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \\
& + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1}) L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& + \frac{4(\mu + C_{g_{yy}})^2 L_y^2}{\mu C_{g_{yy}} \gamma_0} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& + C_\beta + C_\gamma + \frac{4(\mu + C_{g_{yy}}) C_\beta^2}{\mu^4} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}}) C_\gamma^2}{\mu^2} \\
& + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \left(\frac{C_\beta^2}{\beta_0} - \beta_0 \right) \\
& + \left[\left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} + 1 \right] \left(\frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right) \\
\stackrel{(d)}{=} & a_0 \sum_{k=\min\{k_2-1, k_3-1\}}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} + b_0 \\
\leq & a_0 \sum_{k=\min\{k_2, k_3\}}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} + a_0 + b_0 \\
\stackrel{(e)}{\leq} & a_0 \left[3 \log(t+1) + 2 \log \left(\varphi_{t+1} + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 2 \log(C_{g_{xy}} \bar{a}) + 1 \right] + a_0 + b_0, \tag{73}
\end{aligned}$$

where (a) uses Lemma 13; (b) uses the first line in eq. (72) by replacing \tilde{k} with $k_3 - 1$; (c) results from eq. (52); (d) refers to eq. (66); (e) uses eq. (71). Since $\min\{k_2, k_3\} \leq T$, we have $\varphi_{t+1} \geq \min\{C_\beta, C_\gamma\} \geq \max\{64a_0^2, 1\}$, which indicate that

(i) if $8a_0 \leq 1$, we have

$$4a_0 \log(\varphi_{t+1}) \leq \frac{\log(\varphi_{t+1})}{2} \leq \frac{\varphi_{t+1}}{2} \leq \varphi_{t+1};$$

(ii) if $8a_0 > 1$, we have

$$\varphi_{t+1} - 4a_0 \log(\varphi_{t+1}) = \varphi_{t+1} - 8a_0 \log(\sqrt{\varphi_{t+1}}) \geq 8a_0 (\sqrt{\varphi_{t+1}} - \log(\sqrt{\varphi_{t+1}})) \geq 0.$$

Combining (i) and (ii), we have $4a_0 \log(\varphi_{t+1}) \leq \varphi_{t+1}$. Then we obtain

$$\begin{aligned}
\varphi_{t+1} & \leq a_0 \left[3 \log(t+1) + 2 \log \left(\varphi_{t+1} + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 2 \log(C_{g_{xy}} \bar{a}) + 1 \right] + a_0 + b_0 \\
& \leq a_0 \left[3 \log(t+1) + 2 \log(\varphi_{t+1}) + 2 \log \left(1 + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 2 \log(C_{g_{xy}} \bar{a}) + 1 \right] + a_0 + b_0 \\
& \leq \frac{1}{2} \varphi_{t+1} + a_0 \left[3 \log(t+1) + 2 \log \left(1 + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 2 \log(C_{g_{xy}} \bar{a}) + 1 \right] + a_0 + b_0,
\end{aligned}$$

which indicates that

$$\varphi_{t+1} \leq 6a_0 \log(t+1) + 4a_0 \log \left(1 + \frac{C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0}{C_{g_{xy}} \bar{a}} \right) + 4a_0 \log(C_{g_{xy}} \bar{a}) + 4a_0 + 2b_0$$

$$\stackrel{(a)}{=} a_1 \log(t+1) + b_1, \quad (74)$$

where (a) refers to eq. (65). Thus, **Part III** is completed and the proof of this lemma is completed. \square

Lemma 16. *Under Assumptions 1, 2, for any integer $k_0 \in [0, t)$, we have the upper bounds in terms of logarithmic functions as*

$$\begin{aligned} \sum_{k=k_0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} &\leq 5 \log(t+1) + c_2, \\ \sum_{k=k_0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} &\leq a_2 \log(t+1) + b_2, \\ \sum_{k=k_0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} &\leq a_3 \log(t+1) + b_3, \end{aligned}$$

where referring to eq. (65), eq. (66), c_2, a_2, b_2, a_3, b_3 are defined as

$$\begin{aligned} c_2 &:= 2 \log(C_{g_{xy}} \bar{a} a_1 + C_{g_{xy}} \bar{a} b_1 + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0) + 1, \\ a_2 &:= \frac{5(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta}, \quad b_2 := \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} c_2 + \left(\frac{C_\beta^2}{\beta_0} - \beta_0 + \frac{(\mu + L_{g,1}) C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right), \\ a_3 &:= \frac{20(\mu + C_{g_{yy}})(\mu + L_{g,1}) L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{20(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{a_2}{\mu^2}, \\ b_3 &:= \frac{C_\gamma^2}{\gamma_0} - \gamma_0 + \frac{4(\mu + C_{g_{yy}}) C_\beta^2}{\mu^4} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}}) C_\gamma^2}{\mu^2} \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1}) L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \right) c_2 \\ &\quad + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{b_2}{\mu^2}. \end{aligned} \quad (75)$$

Proof. Based on the logarithmic-function form bound in Lemma 15, we can further have the logarithmic-function form bounds of the components in Lemma 11 as the following 3 parts.

Part I: the bound of $\sum \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$ in terms of logarithmic function.

Firstly, we bound $\sum_{k=k_0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2}$ for arbitrary $k_0 < t$. Back to eq. (71), by plugging in eq. (74), we have

$$\begin{aligned} &\sum_{k=k_0}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\ &\leq 3 \log(t+1) + 2 \log(C_{g_{xy}} \bar{a} \varphi_{t+1} + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0) + 1 \\ &\stackrel{(a)}{\leq} 3 \log(t+1) + 2 \log(C_{g_{xy}} \bar{a} a_1 \log(t+1) + C_{g_{xy}} \bar{a} b_1 + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0) + 1 \\ &\leq 3 \log(t+1) + 2 \log(C_{g_{xy}} \bar{a} a_1 (t+1) + C_{g_{xy}} \bar{a} b_1 + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0) + 1 \end{aligned}$$

$$\begin{aligned}
&\leq 3 \log(t+1) + 2 \log((C_{g_{xy}} \bar{a} a_1 + C_{g_{xy}} \bar{a} b_1 + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0)(t+1)) + 1 \\
&\leq 5 \log(t+1) + 2 \log(C_{g_{xy}} \bar{a} a_1 + C_{g_{xy}} \bar{a} b_1 + C_{g_{xy}} \bar{b} + C_{f_x} + \alpha_0) + 1 \\
&\stackrel{(b)}{=} 5 \log(t+1) + c_2,
\end{aligned} \tag{76}$$

where (a) results from eq. (74); (b) refers to eq. (75).

Part II: the bound of $\sum \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$ in terms of logarithmic function.

Secondly, we bound $\sum_{k=k_0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}}$. We split this part into two cases using Lemma 4.

Case 1: If $\beta_{t+1} \leq C_\beta$, we have

$$\sum_{k=k_0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \leq \frac{\sum_{k=k_0}^t \|\nabla_y g(x_k, y_k)\|^2}{\beta_0} \leq \frac{\beta_{t+1}^2 - \beta_{k_0}^2}{\beta_0} \leq \frac{C_\beta^2 - \beta_0^2}{\beta_0} = \frac{C_\beta^2}{\beta_0} - \beta_0 \leq b_2.$$

Case 2: If $\beta_{t+1} > C_\beta$, we have $k_2 \leq t$, where k_2 refers to Lemma 4. Then we can use eq. (72), which indicates

$$\begin{aligned}
&\sum_{k=k_0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\leq \left(\frac{C_\beta^2}{\beta_0} - \beta_0 + \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right) + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \sum_{k=k_2}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
&\leq \frac{5(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} \log(t+1) + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} C_\beta} c_2 + \left(\frac{C_\beta^2}{\beta_0} - \beta_0 + \frac{(\mu + L_{g,1})C_\beta^2}{\mu^2} + \frac{(\mu + L_{g,1})^2 L_y^2}{\mu L_{g,1} \varphi_0} \right) \\
&\stackrel{(a)}{=} a_2 \log(t+1) + b_2,
\end{aligned} \tag{77}$$

where the second inequality uses (76), and (a) refers to eq. (75). Since the upper bound of **Case 2** is larger, we take eq. (77) as our final result.

Part III: the bound of $\sum \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}$ in terms of logarithmic function.

Last, we bound $\sum_{k=k_0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}$. We split this part into two cases using Lemma 4.

Case 1: If $\gamma_{t+1} \leq C_\gamma$, we have

$$\sum_{k=k_0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \leq \frac{\sum_{k=k_0}^t \|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_0} \leq \frac{C_\gamma^2 - \gamma_0^2}{\gamma_0} \leq \frac{C_\gamma^2}{\gamma_0} - \gamma_0 \leq b_3.$$

Case 2: If $\gamma_{t+1} > C_\gamma$, we have $k_3 \leq t$, where k_3 refers to Lemma 4.

$$\begin{aligned}
&\sum_{k=k_0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\stackrel{(a)}{\leq} \sum_{k=k_0}^{k_3-1} \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} + \sum_{k=k_3}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} \\
&\stackrel{(b)}{\leq} \frac{C_\gamma^2}{\gamma_0} - \gamma_0 + \frac{4(\mu + C_{g_{yy}})C_\beta^2}{\mu^4} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})C_\gamma^2}{\mu^2}
\end{aligned}$$

$$\begin{aligned}
& + \frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \sum_{k=k_2-1}^{k_3-2} \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \sum_{k=k_3-1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}^2} \\
& + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} \sum_{k=k_3-1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
& \stackrel{(c)}{\leq} \frac{C_\gamma^2}{\gamma_0} - \gamma_0 + \frac{4(\mu + C_{g_{yy}}) C_\beta^2}{\mu^4} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}}) C_\gamma^2}{\mu^2} \\
& + \left(\frac{4(\mu + C_{g_{yy}})(\mu + L_{g,1})L_y^2}{\mu^3 L_{g,1} \varphi_0} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 + \frac{4(\mu + C_{g_{yy}})^2 L_v^2}{\mu C_{g_{yy}} C_\gamma} \right) (5 \log(t+1) + c_2) \\
& + \left(\frac{4(\mu + C_{g_{yy}})^2}{\mu C_{g_{yy}}} + 8 \right) \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \frac{1}{\mu^2} (a_2 \log(t+1) + b_2) \\
& \stackrel{(d)}{=} a_3 \log(t+1) + b_3, \tag{78}
\end{aligned}$$

where (a) allows $\sum_{k=k_0}^{k_3-1} \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} = 0$ when $k_0 \geq k_3$; (b) uses $C_\gamma \geq \gamma_0$ and Lemma 13; (c) follows from eq. (76) and eq. (77); (d) refers to eq. (75). Since the upper bound of **Case 2** is larger, we take eq. (78) as our final result.

Thus, the proof is complete. \square

Next, we show the upper bound of α_t .

Lemma 17 (The upper bound of α_t). *Under Assumptions 1, 2, 3, suppose the number of total iteration rounds in Algorithm 2 is T . If there exists $k_1 \leq T$ described in Lemma 4, we have*

$$\begin{cases} \alpha_t \leq C_\alpha, & t \leq k_1; \\ \alpha_t \leq C_\alpha + \left(a_4 \log(t) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) \varphi_t, & t \geq k_1, \end{cases}$$

where a_4, b_4 are defined as

$$\begin{aligned}
a_4 & := \frac{2\bar{L}^2 a_2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] + \frac{4\bar{L}^2 a_3}{\mu^2 C_\alpha} \\
b_4 & := \frac{2\bar{L}^2 b_2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] + \frac{4\bar{L}^2 b_3}{\mu^2 C_\alpha} + \frac{2L_\Phi C_\alpha^2}{\varphi_0^2 \alpha_0^2}, \tag{79}
\end{aligned}$$

and the upper bound of $\varphi_t := \max\{\beta_t, \gamma_t\}$ refers to Lemma 15. When such k_1 does not exist, we have $\alpha_t \leq C_\alpha$ for any $t \leq T$.

Proof. According to Lemma 4, the proof can be split into the following three cases.

Case 1: if $\alpha_T \leq C_\alpha$, for any $t < T$, we have the upper bound of α_{t+1} as $\alpha_{t+1} \leq C_\alpha$.

Case 2: if $\alpha_T > C_\alpha$, there exists $k_1 \leq T$ described in Lemma 4. Then we have the upper bound of α_{t+1} as $\alpha_{t+1} \leq C_\alpha$ for any $t < k_1$.

Case 3: in the remaining proof, we only consider and explore the case $k_1 \leq t \leq T$ when $\alpha_T > C_\alpha$.

From Lemma 11, for $k \geq k_1$, we have

$$\begin{aligned} \Phi(x_{k+1}) &\leq \Phi(x_k) - \frac{1}{2\alpha_{k+1}\varphi_{k+1}} \|\nabla\Phi(x_k)\|^2 - \frac{1}{4\alpha_{k+1}\varphi_{k+1}} \|\bar{\nabla}f(x_k, y_k, v_k)\|^2 \\ &\quad + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{\|\nabla_y g(x_k, y_k)\|^2}{\alpha_{k+1}\varphi_{k+1}} + \frac{\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\alpha_{k+1}\varphi_{k+1}}, \end{aligned}$$

which indicates that

$$\begin{aligned} \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}\varphi_{k+1}} &\leq 4(\Phi(x_k) - \Phi(x_{k+1})) \\ &\quad + \frac{2\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{\|\nabla_y g(x_k, y_k)\|^2}{\alpha_{k+1}\varphi_{k+1}} + \frac{4\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\alpha_{k+1}\varphi_{k+1}}. \end{aligned}$$

By taking summation, we have

$$\begin{aligned} &\sum_{k=k_1}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}\varphi_{k+1}} \\ &\leq 4(\Phi(x_{k_1}) - \inf_x \Phi(x)) + \frac{2\bar{L}^2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{k=k_1}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\varphi_{k+1}} \\ &\quad + \frac{4\bar{L}^2}{\mu^2 C_\alpha} \sum_{k=k_1}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}}. \end{aligned} \tag{80}$$

For $\Phi(x_{k_1})$, by telescoping eq. (43) in Lemma 11, we get

$$\begin{aligned} \Phi(x_{k_1}) &\leq \Phi(x_0) + \frac{L_\Phi}{2} \sum_{k=0}^{k_1-1} \frac{\|\bar{\nabla}f(x_t, y_t, v_t)\|^2}{\alpha_{t+1}^2 \varphi_{t+1}^2} \\ &\quad + \frac{\bar{L}^2}{2\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{k=0}^{k_1-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1}\varphi_{t+1}} \\ &\quad + \frac{\bar{L}^2}{\mu^2} \sum_{k=0}^{k_1-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1}\varphi_{t+1}}. \end{aligned} \tag{81}$$

By plugging eq. (81) into eq. (80), we have

$$\begin{aligned} &\sum_{k=k_1}^t \frac{\|\bar{\nabla}f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}\varphi_{k+1}} \\ &\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{2\bar{L}^2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2}C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{k=0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\varphi_{k+1}} \\ &\quad + \frac{4\bar{L}^2}{\mu^2 C_\alpha} \sum_{k=0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} + \frac{2L_\Phi}{\varphi_0^2} \frac{C_\alpha^2}{\alpha_0^2} \end{aligned}$$

$$\begin{aligned}
&\leq 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{2\bar{L}^2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{k=0}^t \frac{\|\nabla_y g(x_k, y_k)\|^2}{\beta_{k+1}} \\
&\quad + \frac{4\bar{L}^2}{\mu^2 C_\alpha} \sum_{k=0}^t \frac{\|\nabla_v R(x_k, y_k, v_k)\|^2}{\varphi_{k+1}} + \frac{2L_\Phi C_\alpha^2}{\varphi_0^2 \alpha_0^2} \\
&\stackrel{(a)}{\leq} 4(\Phi(x_0) - \inf_x \Phi(x)) + \frac{2\bar{L}^2}{\mu^2 C_\alpha} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] (a_2 \log(t+1) + b_2) \\
&\quad + \frac{4\bar{L}^2}{\mu^2 C_\alpha} (a_3 \log(t+1) + b_3) + \frac{2L_\Phi C_\alpha^2}{\varphi_0^2 \alpha_0^2} \\
&\stackrel{(b)}{=} a_4 \log(t+1) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)), \tag{82}
\end{aligned}$$

where (a) plugs in eq. (77) and eq. (78); (b) refers to eq. (79). This immediately implies

$$\sum_{k=k_1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}} \leq \left(a_4 \log(t+1) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) \varphi_{t+1}. \tag{83}$$

Similarly, we can have the upper bound of α_{t+1} as

$$\begin{aligned}
\alpha_{t+1} &\leq \alpha_{k_1} + \sum_{k=k_1}^t \frac{\|\bar{\nabla} f(x_k, y_k, v_k)\|^2}{\alpha_{k+1}} \\
&\leq C_\alpha + \left(a_4 \log(t+1) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) \varphi_{t+1}. \tag{84}
\end{aligned}$$

Then the upper bound of α_{t+1} is proved. \square

F.4 Proof of Theorem 2

Here we still assume the total iteration rounds of Algorithm 2 is T . According to Lemma 4, the proof can be split into the following two cases.

Case 1: If $\alpha_T \leq C_\alpha$, then by Lemma 11 and Lemma 17, we have

$$\begin{aligned}
\frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} &\leq 2(\Phi(x_t) - \Phi(x_{t+1})) + \frac{L_\Phi}{\alpha_{t+1}^2 \varphi_{t+1}^2} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} + \frac{2\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}},
\end{aligned}$$

By taking the average, we have

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} &\leq \frac{2}{T} (\Phi(x_0) - \Phi(x_T)) + \frac{L_\Phi}{\alpha_0^2 \varphi_0^2} \frac{1}{T} \sum_{t=0}^{T-1} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
&\quad + \frac{2\bar{L}^2}{\mu^2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{T} (\Phi(x_0) - \Phi(x_T)) + \frac{L_\Phi C_\alpha^2}{T \alpha_0^2 \varphi_0^2} \\
&\quad + \frac{\bar{L}^2}{\mu^2 \alpha_0 T} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{t=0}^{T-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\beta_{t+1}} \\
&\quad + \frac{2\bar{L}^2}{\mu^2 \alpha_0 T} \sum_{t=0}^{T-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} \\
&\stackrel{(a)}{\leq} \frac{2}{T} (\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{T \alpha_0^2 \varphi_0^2} \\
&\quad + \frac{\bar{L}^2}{\mu^2 \alpha_0 T} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] (a_2 \log(T) + b_2) \\
&\quad + \frac{2\bar{L}^2}{\mu^2 \alpha_0 T} (a_3 \log(T) + b_3) \\
&= \frac{1}{2T} \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right), \tag{85}
\end{aligned}$$

where (a) uses Lemma 16 with $k_0 = 0$.

Case 2: If $\alpha_T > C_\alpha$, by Lemma 4, there exists $k_1 \leq T_0$ such that $\alpha_{k_1} \leq C_\alpha$, $\alpha_{k_1+1} > C_\alpha$.

Then for $t < k_1$ when $\alpha_T > C_\alpha$, from Lemma 11, we have

$$\begin{aligned}
\frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} &\leq 2(\Phi(x_t) - \Phi(x_{t+1})) + \frac{L_\Phi}{\alpha_{t+1}^2 \varphi_{t+1}^2} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} + \frac{2\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}}.
\end{aligned}$$

For $t \geq k_1$ when $\alpha_T > C_\alpha$, from Lemma 11, we have

$$\begin{aligned}
\frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} &\leq 2(\Phi(x_t) - \Phi(x_{t+1})) \\
&\quad + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} + \frac{2\bar{L}^2}{\mu^2} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}}.
\end{aligned}$$

By taking the average, we can merge $t < k_1$ and $t \geq k_1$ as

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} &= \frac{1}{T} \sum_{t=0}^{k_1-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} + \frac{1}{T} \sum_{t=k_1}^{T-1} \frac{\|\nabla \Phi(x_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
&\leq \frac{2}{T} (\Phi(x_0) - \Phi(x_{k_1})) + \frac{L_\Phi}{\alpha_0^2 \varphi_0^2} \frac{1}{T} \sum_{t=0}^{k_1-1} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
&\quad + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{1}{T} \sum_{t=0}^{k_1-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
&\quad + \frac{2\bar{L}^2}{\mu^2} \frac{1}{T} \sum_{t=0}^{k_1-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}}
\end{aligned}$$

$$\begin{aligned}
& + \frac{2}{T} (\Phi(x_{k_1}) - \Phi(x_T)) \\
& + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{1}{T} \sum_{t=k_1}^{T-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
& + \frac{2\bar{L}^2}{\mu^2} \frac{1}{T} \sum_{t=k_1}^{T-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
& \leq \frac{2}{T} (\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi}{\alpha_0^2 \varphi_0^2} \frac{1}{T} \sum_{t=0}^{k_1-1} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
& + \frac{\bar{L}^2}{\mu^2} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
& + \frac{2\bar{L}^2}{\mu^2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\alpha_{t+1} \varphi_{t+1}} \\
& \leq \frac{2}{T} (\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi}{\alpha_0^2 \varphi_0^2} \frac{1}{T} \sum_{t=0}^{k_1-1} \|\bar{\nabla} f(x_t, y_t, v_t)\|^2 \\
& + \frac{\bar{L}^2}{\mu^2 \alpha_0 T} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] \sum_{t=0}^{T-1} \frac{\|\nabla_y g(x_t, y_t)\|^2}{\varphi_{t+1}} \\
& + \frac{2\bar{L}^2}{\mu^2 \alpha_0 T} \sum_{t=0}^{T-1} \frac{\|\nabla_v R(x_t, y_t, v_t)\|^2}{\varphi_{t+1}} \\
& \stackrel{(a)}{\leq} \frac{2}{T} (\Phi(x_0) - \inf_x \Phi(x)) + \frac{L_\Phi C_\alpha^2}{T \alpha_0^2 \varphi_0^2} \\
& + \frac{\bar{L}^2}{\mu^2 \alpha_0 T} \left[1 + \frac{2}{\mu^2} \left(\frac{L_{g,2} C_{f_y}}{\mu} + L_{f,1} \right)^2 \right] (a_2 \log(T) + b_2) \\
& + \frac{2\bar{L}^2}{\mu^2 \alpha_0 T} (a_3 \log(T) + b_3) \\
& = \frac{1}{2T} \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right), \tag{86}
\end{aligned}$$

where (a) uses Lemma 16 by plugging in $k_0 = 0$.

Note that **Case 1** and **Case 2** indicate the same result. Thus, we have

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 & \leq \frac{1}{2T} \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) \alpha_T \varphi_T \\
& \stackrel{(a)}{\leq} \frac{1}{2T} \left[\left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right)^2 \varphi_T^2 \right. \\
& \quad \left. + C_\alpha \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) \varphi_T \right] \\
& \stackrel{(b)}{\leq} \frac{1}{2T} \left[\left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right)^2 (a_1 \log(T) + b_1)^2 \right]
\end{aligned}$$

$$\begin{aligned}
& + C_\alpha \left(a_4 \log(T) + b_4 + 4(\Phi(x_0) - \inf_x \Phi(x)) \right) (a_1 \log(T) + b_1) \Big] \\
& = \mathcal{O} \left(\frac{\log^4(T)}{T} \right).
\end{aligned}$$

where (a) follows from Lemma 17; (b) results from Lemma 15. Thus, the proof is finished.

F.5 Complexity Analysis of Algorithm 2 (Proof of Corollary 2)

Recall in Theorem 2, we know that there exist a constant M such that

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \Phi(x_t)\|^2 \leq \frac{M \log^4(T)}{T}.$$

When we set the iteration number $T = \frac{MN}{\epsilon} \log^4(\frac{M}{\epsilon})$ and assume the constant $N = 12^4$, we have

$$\begin{aligned}
\frac{M \log^4(T)}{T} &= \frac{M \log^4(\frac{MN}{\epsilon} \log^4(\frac{M}{\epsilon}))}{\frac{MN}{\epsilon} \log^4(\frac{M}{\epsilon})} \\
&\leq \frac{[\log(N) + \log(\frac{M}{\epsilon}) + 4 \log(\log(\frac{M}{\epsilon}))]^4}{N \log^4(\frac{M}{\epsilon})} \cdot \epsilon \\
&\stackrel{(a)}{\leq} \left(\frac{\log(N) + 2 \log(\frac{M}{\epsilon})}{N^{\frac{1}{4}} \log(\frac{M}{\epsilon})} \right)^4 \cdot \epsilon \stackrel{(b)}{\leq} \epsilon,
\end{aligned}$$

where (a) follows from the inequality $\log(\log(\frac{M}{\epsilon})) \leq \frac{1}{4} \log(\frac{M}{\epsilon})$ for sufficiently small ϵ ; (b) holds because $\log(N) + 2 \log(\frac{M}{\epsilon}) \leq N^{\frac{1}{4}} \log(\frac{M}{\epsilon})$ for $N = 12^4$ and ϵ is sufficiently small. Thus, to achieve ϵ -accurate stationary point, we require $T = \frac{MN}{\epsilon} \log^4(\frac{M}{\epsilon}) = \mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$, and the gradient complexity is given by $\text{Gc}(\epsilon) = \Omega(T) = \mathcal{O}(\frac{1}{\epsilon} \log^4(\frac{1}{\epsilon}))$.