Journal of Machine Learning Research 26 (2025) 1-61 Submitted 2/22; Revised 1/25; Published 1/25

Efficiently Escaping Saddle Points in Bilevel Optimization

Minhui Huang* MHHUANG@UCDAVIS.EDU
Department of Electrical and Computer Engineering
University of California,
Davis, CA 95616, USA

Xuxing Chen* XUXCHEN@UCDAVIS.EDU
Department of Mathematics
University of California,
Davis, CA 95616, USA

Kaiyi Ji KAIYIJI@BUFFALO.EDU
Department of Computer Science and Engineering
University at Buffalo,
Buffalo, NY 14260, USA

Shiqian Ma SQMA@RICE.EDU
Department of Computational Applied Mathematics and Operations Research
Rice University,
Houston, TX 77005, USA

Lifeng Lai LFLAI@UCDAVIS.EDU
Department of Electrical and Computer Engineering
University of California,
Davis, CA 95616, USA

Editor: Silvia Villa

Abstract

Bilevel optimization is one of the fundamental problems in machine learning and optimization. Recent theoretical developments in bilevel optimization focus on finding the first-order stationary points for nonconvex-strongly-convex cases. In this paper, we analyze algorithms that can escape saddle points in nonconvex-strongly-convex bilevel optimization. Specifically, we show that the perturbed approximate implicit differentiation (AID) with a warm start strategy finds an ϵ-approximate local minimum of bilevel optimization in $\tilde{O}(\epsilon^{-2})$ iterations with high probability. Moreover, we propose an inexact NEgative-curvature-Originated-from-Noise Algorithm (iNEON), an algorithm that can escape saddle point and find local minimum of stochastic bilevel optimization. As a by-product, we provide the first nonasymptotic analysis of perturbed multi-step gradient descent ascent (GDmax) algorithm that converges to local minimax point for minimax problems.

Keywords: Bilevel optimization, minimax problem, local minimax point, saddle point, inexact NEON

1. Introduction

Bilevel optimization has become a powerful tool in various machine learning fields including reinforcement learning (Hong et al., 2020), hyperparameter optimization (Franceschi et al., 2018; Feurer and Hutter, 2019), meta learning (Franceschi et al., 2018; Ji et al., 2020) and signal processing (Kunapuli et al., 2008). A general formulation of bilevel optimization problem can be written as

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \Phi(x) := f(x, y^*(x)), \\ \text{s.t.} \quad & y^*(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n} g(x, y). \end{aligned} \tag{1.1}$$

In this paper, we focus on the nonconvex-strongly-convex case where the lower level function $g(x, y)$ is smooth and strongly convex with respect to y and the overall objective function $\Phi(x)$ is smooth but possibly nonconvex. One crucial but challenging task in the bilevel optimization is the computation of the hypergradient $\nabla\Phi(x)$, which, via chain rule, can be written as

$$\nabla\Phi(x) = \nabla_x f(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \cdot \nabla_y f(x, y^*(x)), \tag{1.2}$$

where $\frac{\partial y^*(x)}{\partial x} \in \mathbb{R}^{d\times n}$. Note that the differentiability of $y^*(x)$ is a direct result of the Implicit Function Theorem, as mentioned in Lemma 2.1 of Ghadimi and Wang (2018). By taking derivative with respect to x on the optimality condition: $\nabla_y g(x, y) = 0$, we have the relation

$$\nabla^2_{xy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \nabla^2_{yy} g(x, y^*(x)) = 0, \tag{1.3}$$

which implies

$$\frac{\partial y^*(x)}{\partial x} = -\nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g(x, y^*(x))^{-1}. \tag{1.4}$$

Substituting (1.4) to (1.2), we get

$$\nabla\Phi(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g(x, y^*(x))^{-1} \nabla_y f(x, y^*(x)). \tag{1.5}$$

Note that the above hypergradient $\nabla\Phi(x)$ involves computationally intractable components such as the exact solution $y^*(x)$ and the Hessian inverse $\nabla^2_{yy} g(x, y^*(x))^{-1}$. To address such difficulties, various computing approaches have been proposed, which include popular Approximate Implicit Differentiation (AID) (Domke, 2012; Pedregosa, 2016; Gould et al., 2016; Ghadimi and Wang, 2018; Grazzi et al., 2020; Ji et al., 2021) and Iterative Differentiation (ITD) (Domke, 2012; Maclaurin et al., 2015; Shaban et al., 2019; Grazzi et al., 2020; Ji et al., 2021). Among them, Ghadimi and Wang (2018) and Ji et al. (2021) further analyze the computational complexities of these two types of approaches in finding a stationary point. Besides these nested-loop approaches, Hong et al. (2020) and Chen et al. (2022) propose single-loop algorithms with convergence analysis to stationary points.

However, it still remains unknown how to provably find a **local minimum** for bilevel optimization. This type of study is important as it has been widely shown that saddle points (which are also stationary points) can seriously undermine the quality of solutions (Choromanska et al., 2015; Dauphin et al., 2014). To address this issue, this paper focuses on escaping saddle points for bilevel optimization. We are interested in finding an approximate local minimum for $\Phi(x)$ defined as follows.

Definition 1 (ϵ-local minimum) *We say x is an ϵ-local minimum for bilevel optimization* (1.1) *if*

$$\|\nabla\Phi(x)\| \leq \epsilon, \quad \lambda_{\min}\left(\nabla^2\Phi(x)\right) \geq -\sqrt{\rho_\phi \epsilon}, \tag{1.6}$$

where $\lambda_{\min}(Z)$ denotes the minimum eigenvalue of a matrix Z and ρ_ϕ is the Lipschitz constant of $\nabla^2\Phi(x)$, i.e.,

$$\left\|\nabla^2\Phi(x) - \nabla^2\Phi(x')\right\| \leq \rho_\phi\|x - x'\|, \quad \forall x, x' \in \mathbb{R}^d. \tag{1.7}$$

For completeness, we also define stationary points and saddle points as follows.

Definition 2 *We say x is an ϵ-stationary point of bilevel optimization* (1.1) *if $\|\nabla\Phi(x)\| \leq \epsilon$. We say x is a saddle point of bilevel optimization* (1.1) *if $\nabla\Phi(x) = 0$, and x is not a local maximum point or local minimum point.*

Motivated by the recent demand in solving online or large-scale bilevel optimization problems, we also generalize our technique to the following stochastic bilevel optimization:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \Phi(x) = f(x, y^*(x)) = \mathbb{E}_\xi\left[F(x, y^*(x); \xi)\right] \\ &\text{s.t. } y^*(x) = \operatorname*{argmin}_{y\in\mathbb{R}^s} g(x, y) = \mathbb{E}_\zeta\left[G(x, y; \zeta)\right], \end{aligned} \tag{1.8}$$

where $f(x, y)$ and $g(x, y)$ take the expectation form with respect to the random variables ξ and ζ. There is a line of work studying stochastic bilevel algorithms that converge to the stationary point (Ghadimi and Wang, 2018; Ji et al., 2021; Hong et al., 2020). Comparing with these results, we are interested in providing new stochastic algorithms that provably converge to the local minimum.

1.1 Our Contributions

In this paper, we derive a framework of adding perturbation to gradient sequence for bilevel optimization and design various new bilevel algorithms that provably escape saddle points and find local minimum. Our approach is mostly inspired by existing works for nonconvex minimization and minimax problems (Jin et al., 2021; Xu et al., 2018; Allen-Zhu and Li, 2018). Our main contributions are summarized below.

(i) For deterministic bilevel optimization, we propose the perturbed AID with warm start strategy. We prove that the proposed algorithm achieves ϵ-local minimum of $\Phi(x)$ in at most $\tilde{\mathcal{O}}(\epsilon^{-2})$ iterations. Here the notation $\tilde{\mathcal{O}}(\cdot)$ hides logorithmic terms and absolute constants.

(ii) For the minimax problem, which is a special case of bilevel optimization, we prove that the strict local minimum of $\Phi(x)$ is equivalent to strict local minimax point (Jin et al., 2020) and propose the perturbed GDmax algorithm with a nonasymptotic convergence rate to local minimax point. To the best of our knowledge, this is the first nonasympototic analysis for gradient algorithms escaping saddle point in minimax problem.

(iii) For stochastic bilevel optimization, we propose inexact NEgative-curvature-Originated-from-Noise Algorithm (iNEON), a deterministic algorithm that extracts negative curvature descent direction with high probability. Combining iNEON with stocBiO (Ji et al., 2021), we obtain a stochastic first-order algorithm with a gradient complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$. To the best of our knowledge, our algorithms – perturbed AID and stocBiO+iNEON – are the first ones that provably converge to local minimum of bilevel optimization.

1.2 Related Work

Escaping Saddle Point. Most existing works for finding local minimum focus on classical optimization problems (i.e., minimization problems) and derive the complexity for reaching an ϵ-local minimum. Nesterov and Polyak (2006) and Curtis et al. (2017) proposed second-order methods for obtaining an ϵ-local minimum. To avoid Hessian computation required in Nesterov and Polyak (2006) and Curtis et al. (2017), Carmon et al. (2018) and Agarwal et al. (2017) proposed to use Hessian-vector product and achieved convergence rate of $O(\epsilon^{-7/4})$. Recently, the complexity results of pure first-order methods for obtaining local minimum have been studied (see, e.g., Ge et al. (2015); Daneshmand et al. (2018); Jin et al. (2021); Fang et al. (2019)). Lee et al. (2016) provided asymptotic results showing that gradient descent (GD) method converges to a local minimizer almost surely. Jin et al. (2017, 2021) proved that the perturbed GD can converge to a local minimizer in a number of iterations that depends poly-logarithmically on the dimension, reaching a nonasymptotic iteration complexity of $\tilde{O}(\epsilon^{-2}\log(d)^4)$ for nonconvex minimization. For stochastic optimization problems, Jin et al. (2021), Jin et al. (2018), and Fang et al. (2019) provided nonasymptotic rate for finding local minimizers. How to escape saddle points for constrained problems and nonsmooth problems are also studied in the literature. In particular, Lu et al. (2020a), Criscitiello and Boumal (2019), and Sun et al. (2019) studied escaping saddle points for constrained optimization. Davis and Drusvyatskiy (2021), Davis et al. (2021), and Huang (2021) studied escaping saddle points for nonsmooth problems. All these algorithms are for solving the minimization problems, and to the best of our knowledge, how to escape saddle points in bilevel optimization has not been addressed in the literature.

Minimax Optimization. Motivated by its applications in adversarial learning (Goodfellow et al., 2015; Sinha et al., 2018), training Generative Adversarial Nets (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017) and optimal transport (Lin et al., 2020a; Huang et al., 2021a,b,c), the convergence theory of nonconvex minimax problems has been extensively studied in the literature. Specifically, Nouiehed et al. (2019) and Jin et al. (2020) studied the complexity of multistep gradient descent ascent (GDmax). Lin et al. (2020b) and Lu et al. (2020b) provided the first convergence analysis for the single loop gradient descent ascent (GDA) algorithm. More recently, Luo et al. (2020) applied the stochastic variance reduction technique to the nonconvex-strongly-concave case and achieved the best known stochastic gradient complexity. Zhang et al. (2020) proposed smoothed GDA, which stabilizes GDA algorithm and helps achieve a better complexity for the nonconvex-concave case. However, all the previous works targeted finding stationary point of $\Phi(x)$. Very recently, Chen et al. (2021b) and Luo and Chen (2021) proposed cubic regularized GDA, a

second-order algorithm that provably converges to a local minimum. Fiez et al. (2021) provided asymptotic results showing that GDA converges to local minimax point almost surely. To the best of our knowledge, the convergence rate of first-order methods for obtaining a local minimax point has been missing in the literature.

Bilevel Optimization. The bilevel optimization has a long history and dates back to Bracken and McGill (1973). Recently, bilevel programming has been successfully applied to meta-learning (Snell et al., 2017; Rajeswaran et al., 2019; Franceschi et al., 2018; Ji et al., 2022) and hyperparameter optimization (Pedregosa, 2016; Franceschi et al., 2018; Shaban et al., 2019; Sow et al., 2021). Theoretically, Ghadimi and Wang (2018) provided the first convergence rate for the AID approach. Ji et al. (2021) further improved their complexity dependence on the condition number and analyzed the convergence of the ITD approach. Both AID and ITD have an iteration complexity of $\mathcal{O}(\epsilon^{-2})$. Ji and Liang (2021) provided lower bounds for a class of AID and ITD-based bilevel algorithms. For stochastic bilevel problems, Ghadimi and Wang (2018) and Ji et al. (2021) proposed Bilevel Stochastic Approximation (BSA) and stochastic bilevel optimizer (stocBiO) methods respectively, which are both double-loop algorithms inspired by AID. Hong et al. (2020) proposed a two-timescale framework for bilevel optimization (TTSA), a provable single-loop algorithm that updates two variables in an alternating way with a convergence rate of $\mathcal{O}(\epsilon^{-5})$. Chen et al. (2021a) proposed ALternating Stochastic gradient dEscenT (ALSET), a simple SGD type approach, and improved the convergence rate to $\mathcal{O}(\epsilon^{-4})$. Very recently, Khanduri et al. (2021), Yang et al. (2021), Chen et al. (2022), and Guo and Yang (2021) studied stochastic algorithms with variance reduction and momentum techniques, and provided the cutting-edge first-order oracle complexity, which is $\mathcal{O}(\epsilon^{-3})$. All these previous analyses have focused on finding stationary points and algorithm for finding a local minimum is still missing. It is unclear if the obtained stationary points of the existing bilevel optimization algorithms are local minimum or saddle points. In other words, finding an ϵ-local minimum, i.e., escaping saddle points in bilevel optimization, receives little attention. Our work proposes a new algorithm with guaranteed convergence to ϵ-local minimum of bilevel optimization problem, which is much more challenging than most of the prior works because converging to ϵ-local minima is much stronger than merely converging to stationary points. Furthermore, we note that existing works on escaping saddle points only consider single-level optimization problems. It is more challenging in bilevel problem. The reason is that the hypergradient of the bilevel problem, i.e., $\nabla\Phi(x)$ in (1.5), involves the first-order information of the upper-level problem and the second-order information of the lower-level problem. This brings more challenges to the design of our algorithm as well as the convergence analysis as we need to handle the convergence error of the lower-level problem as well as the hypergradient estimation error.

Notation. Let $\hat{\Phi}(x), \widehat{\nabla}\Phi(x), \widehat{\nabla}^2\Phi(x)$ be the inexact function value, gradient and Hessian, respectively. Denote $G(f,\epsilon), JV(f,\epsilon), HV(f,\epsilon)$ as the complexity of gradient evaluations, Jacobian-vector product evaluations, and Hessian-vector product evaluations of function f, respectively. In particular, for matrix-vector product oracles, say Hessian-vector products, $HV(f,\epsilon)$ represents the total number of (deterministic or stochastic) $(\nabla^2 f)\cdot v$ computation in our algorithm. Typically computing a Hessian-vector product is as cheap as computing a gradient (Pearlmutter, 1994). Let κ be the condition number of the lower-level

problem. We use notation $\mathcal{O}(\cdot)$ to hide only absolute constants which do not depend on any problem parameters and $\tilde{\mathcal{O}}(\cdot)$ to further hide additional log factors.

2. Escape Saddle Points in General Bilevel Optimization

In this section, we propose novel algorithms for general bilevel optimization (1.1) that are guaranteed to converge to local minimum. We consider one of the popular approaches AID to estimate the hypergradient $\nabla\Phi(x)$. The AID approach is a nested-loop algorithm, which first update the lower-level variable y with D steps of gradient descent, and then construct an estimate of the upper-level hypergradient. To efficiently approximate the Hessian inverse in the hypergradient (1.5), AID solves the linear system:

$$\nabla^2_{yy} g(x_k, y_k^D) v = \nabla_y f(x_k, y_k^D) \tag{2.1}$$

using N steps of the conjugate gradient (CG) method. The resulting vector v_k^N is used as an approximation to the solution of (2.1): $\nabla^2_{yy} g(x_k, y_k^D)^{-1} \nabla_y f(x_k, y_k^D)$. The hypergradient is then constructed as

$$\widehat{\nabla}\Phi(x_k) = \nabla_x f(x_k, y_k^D) - \nabla^2_{xy} g(x_k, y_k^D) v_k^N. \tag{2.2}$$

However, current AID-based approach can only guarantee the convergence to the first-order stationary point. In Algorithm 1, we propose perturbed AID (i.e., Algorithm 1 with option **AID** in step 9) for solving bilevel optimization (1.1) with convergence guarantee to second-order stationarity. In the proposed algorithms, we update variable x with the hypergradient $\widehat{\nabla}\Phi(x)$ estimated by AID. When the norm of $\widehat{\nabla}\Phi(x)$ is small, we add random noise sampled from a uniform ball and keep running AID for at least $\mathscr{T}$ steps (see steps 10-13 of Algorithm 1). If the current point is a saddle point of $\nabla\Phi(x)$, we show that with high probability the function value $\Phi(x)$ has sufficient decrease after $\mathscr{T}$ steps so it can escape the current saddle point.

2.1 Convergence Analysis

We first state assumptions needed for our analysis.

Assumption 3 *Assume the upper level function $f(x, y)$ and the lower level function $g(x, y)$ satisfy the following assumptions:*

(i) Function $g(x, y)$ is three times differentiable and μ-strongly convex with respect to y for any fixed x.

(ii) Function $f(x, y)$ is twice differentiable and $f(x, y)$ is M-Lipschitz continuous with respect to x and y.

(iii) Gradients $\nabla f(x, y)$ and $\nabla g(x, y)$ are ℓ-Lipschitz continuous with respect to x and y.

(iv) Jacobian matrices $\nabla^2_{xx} f(x, y)$, $\nabla^2_{xy} f(x, y)$, $\nabla^2_{yy} f(x, y)$, $\nabla^2_{xy} g(x, y)$ and $\nabla^2_{yy} g(x, y)$ are ρ-Lipschitz continuous with respect to x and y.

(v) Third-order derivatives $\nabla^3_{xyx} g(x, y)$, $\nabla^3_{yxy} g(x, y)$ and $\nabla^3_{yyy} g(x, y)$ are ν-Lipschitz continuous with respect to x and y.

Algorithm 1 Perturbed Algorithms for Minimax and Bilevel Optimization Problems

Input: Iteration Numbers K, D, N, Step Sizes τ, η, Accuracy ϵ, Radius r, Perturbation Time $\mathscr{T}$.
Initialization: x_0, y_0, v_0.
Set $k_{perturb} = 0$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Set $y_k^0 = y_{k-1}^D$ if $k > 0$, otherwise y_0
  for $t = 0, 1, 2, \ldots, D$ **do**
    $y_k^t = y_k^{t-1} - \tau \cdot \nabla_y g(x_k, y_k^{t-1})$
  end for
  Estimating Hypergradient:
  Option 1 (for minimax problem) GDmax: compute $\widehat{\nabla}\Phi(x_k) = \nabla_x f(x_k, y_k^D)$
  Option 2 (for bilevel optimization) AID:
    1) set $v_k^0 = v_{k-1}^N$ if $k > 0$ and v_0 otherwise
    2) solve v_k^N from $\nabla_{yy}^2 g(x_k, y_k^D)v = \nabla_y f(x_k, y_k^D)$ via N steps of CG starting from v_k^0
    3) get Jacobian-vector product $\nabla_{xy}^2 g(x_k, y_k^D)v_k^N$ via automatic differentiation
    4) $\widehat{\nabla}\Phi(x_k) = \nabla_x f(x_k, y_k^D) - \nabla_{xy}^2 g(x_k, y_k^D)v_k^N$
  if $\|\widehat{\nabla}\Phi(x_k)\| \le \frac{4}{5}\epsilon$ and $k - k_{perturb} > \mathscr{T}$ **then**
    $x_k = x_k - \eta \cdot u, \quad (u \sim \text{Uniform}(\mathbb{B}(r)))$
    $k_{perturb} = k$
  end if
  $x_{k+1} = x_k - \eta \cdot \widehat{\nabla}\Phi(x_k)$
end for
Output: x_K.

Remark 4 *Compared with assumptions in recent bilevel optimization literature (Ghadimi and Wang, 2018; Ji et al., 2021), we further assume the Lipschitz continuity of the Hessian of $f(x, y)$ and the third-order derivative of $g(x, y)$. These assumptions are required to prove the Hessian Lipschitz continuity of $\Phi(x)$, which is a common condition required in the literature of escaping saddle points (Jin et al., 2021).*

Remark 5 *It should be noted that although we assume the third-order partial derivatives of $g(x, y)$ to be Lipschitz continuous, this assumption is for the theoretical analysis only, we do not compute any third-order derivatives in our algorithms.*

One of the key elements in our proof technique is to show that under Assumption 3, function $\Phi(x)$ is Hessian Lipschitz continuous, as shown in the following lemma.

Lemma 6 *Suppose Assumption 3 holds, then $\Phi(x)$ is ρ_ϕ-Hessian Lipschitz continuous, i.e.,* (1.7) *holds, where $\rho_\phi = \mathcal{O}(\kappa^5)$ and is defined in* (B.17).

For the AID approach, the main results are in the following theorem.

Theorem 7 (Convergence of Perturbed AID) *Suppose that Assumption 3 holds. Set parameters of Algorithm 1 as*

$$\tau = \frac{1}{\ell}, \quad \eta = \frac{1}{L_\phi}, \quad r = \frac{\epsilon}{400\iota^3}, \quad \mathscr{T} = \frac{L_\phi}{\sqrt{\rho_\phi \epsilon}} \cdot \iota,$$

and

$$D = \mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)\right), \quad N = \mathcal{O}\left(\sqrt{\kappa} \log\left(\epsilon^{-1}\right)\right). \tag{2.3}$$

With probability at least $1-\delta$*, the iteration number of the perturbed AID algorithm for visiting an* ϵ*-local minimum of* $\Phi(x)$ *is*

$$K = \tilde{\mathcal{O}}\left(\kappa^3 \epsilon^{-2}\right). \tag{2.4}$$

Here $\delta \in (0,1)$*,* L_ϕ *is the Lipschitz constant of* $\nabla\Phi$*,* ρ_ϕ *is the Lipschitz constant of* $\nabla^2\Phi$ *(see Lemma 36), and* ι *is a constant satisfying*

$$\iota > 1, \text{ and } \delta \geq \frac{L_\phi \sqrt{d}}{\sqrt{\rho_\phi \epsilon}} \cdot \iota^2 2^{8-\iota}. \tag{2.5}$$

Corollary 8 *The gradient complexities of the perturbed AID algorithm for finding an* ϵ*-local minimum of* $\Phi(x)$ *are*

$$G(f,\epsilon) = \tilde{\mathcal{O}}(\kappa^3\epsilon^{-2}), \quad G(g,\epsilon) = \tilde{\mathcal{O}}(\kappa^4\epsilon^{-2}).$$

The Jacobian- and Hessian-vector product complexities are

$$JV(g,\epsilon) = \tilde{\mathcal{O}}(\kappa^3\epsilon^{-2}), \quad HV(g,\epsilon) = \tilde{\mathcal{O}}(\kappa^{3.5}\epsilon^{-2}).$$

Remark 9 *Though the complexities of the perturbed AID method are worse than the results in Ji et al. (2021) by a* log *factor, it should be noted that the algorithms converge to different points. Specifically, our perturbed AID method converges to a local minimum of* $\Phi(x)$*, whereas the algorithms in Ji et al. (2021) are only guaranteed to converge to first-order stationarity.*

2.2 Proof sketch

We briefly describe the main elements in proving the above theorems. The main ideas follow (Jin et al., 2021). However, in contrast to the problem studied in Jin et al. (2021), we do not have access to the exact hypergradient of $\Phi(x)$ in bilevel optimization problems. Therefore, we need to deal with the error introduced by this approximation. We first provide the inexact descent lemma.

Lemma 10 (Inexact Descent Lemma) *Suppose Assumption 3 holds and set* $\eta = 1/L_\phi$*, then the inexact gradient sequence* $\{x_k\}$ *satisfies:*

$$\Phi(x_{k+1}) - \Phi(x_k) \leq -\frac{\eta}{4}\left\|\widehat{\nabla}\Phi(x_k)\right\|^2 + \eta\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2. \tag{2.6}$$

Secondly, the following lemma shows that with high probability, adding random noise sampled uniformly from a ball helps escape saddle points of $\Phi(x)$.

Lemma 11 (Escaping Saddle Points) *Assume Assumption 3 holds. Assume* $\tilde{x}$ *satisfies* $\|\nabla\Phi(\tilde{x})\| \leq \epsilon$*, and* $\lambda_{\min}(\nabla^2\Phi(\tilde{x})) \leq -\sqrt{\rho_\phi\epsilon}$*, where* $\rho_\phi = \mathcal{O}(\kappa^5)$ *and is defined in* (B.17)*.*

Let $x_0 = \tilde{x} + \eta u$ ($u \sim$ *Uniform*$(B_0(r))$). *With parameters given in* (3.2), *as long as the following inequality holds in each iteration,*

$$\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\| \leq \min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\} \cdot \epsilon, \tag{2.7}$$

with probability at least $1-\delta$, *it holds that*

$$\Phi(x_{\mathscr{T}}) - \Phi(\tilde{x}) \leq -\mathscr{F}/2, \tag{2.8}$$

where $x_{\mathscr{T}}$ *is the* $\mathscr{T}^{th}$ *gradient descent iterate starting from* x_0, ι *satisfies* (3.3) *and*

$$\mathscr{F} = \frac{1}{100\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}}. \tag{2.9}$$

Finally, by Lemma 10, Lemma 11 and with a proper choice of parameters D and N, we can bound the total iteration number of Algorithm 1 by

$$K = \frac{(\Phi(x_0) - \Phi*)\mathscr{T}}{\mathscr{F}} + \frac{L_\phi(\Phi(x_0) - \Phi*)}{\epsilon^2}.$$

Here the perturbation time $\mathscr{T}$ will be specified later in the proof of our main theorem. In practice we set it as a hyperparameter to tune.

3. Escape Saddle Points for Minimax Problem

In this section, we consider the following nonconvex-strongly-concave minimax problem:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n} \quad f(x,y), \tag{3.1}$$

where $f(x,y)$ is nonconvex with respect to x and μ-strongly concave with respect to y. We note that the problem aims at minimizing $\max_{y\in\mathbb{R}^n} f(x,y)$ for each x. Thus, by defining the function $\Phi(x) = \max_y f(x,y)$, (3.1) reduces to a smooth nonconvex minimization problem $\min_{x\in\mathbb{R}^d}\Phi(x)$. Note that this is also a special case of the bilevel optimization problem by setting $g(x,y) = -f(x,y)$ in (1.1), which leads to the following problem:

$$\min_x f(x, y^*(x)), \text{ s.t. } y^*(x) = \operatorname*{argmin}_y g(x,y) := -f(x,y),$$

where $y^*(x)$ is uniquely defined for any x, since $f(x,y)$ is strongly concave with respect to y. The minimax problem (3.1) seeks the Nash equilibrium of $f(x,y)$. However, when considering nonconvex minimax problem, the global Nash equilibrium does not exist in general. Instead, one is more interested in finding the local Nash equilibrium (Daskalakis and Panageas, 2018; Mazumdar et al., 2019) and the local minimax point (Jin et al., 2020). Therefore, the following question arises naturally:

- *What is the relationship between the local minimum of* $\Phi(x)$ *and the local optimality of the minimax problem* (3.1)*?*

The answer to this question has been so far still ambiguous. We first discuss the relationship between the local minimum of $\Phi(x)$ and the local Nash equilibrium of (3.1). The Nash equilibrium and its local alternative are defined as below.

Definition 12 *(Mazumdar et al., 2019)[Local Nash Equilibrium] We say $(x^\star, y^\star)$ is a **Nash equilibrium** of function f, if for any (x, y):*

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star).$$

*Point $(x^\star, y^\star)$ is a **local Nash equilibrium** of f if there exists $\delta > 0$ such that for any (x, y) satisfying $\|x - x^\star\| \leq \delta$ and $\|y - y^\star\| \leq \delta$ we have:*

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star).$$

Remark 13 *It is worth noting that, Nash equilibrium is sometimes called 'saddle point' in minimax optimization literature (Lin et al., 2020c). We highlight here that, throughout this paper, our notion of saddle points follows Definition 2. In other words, we first view Problem* (3.1) *as a bilevel problem, in which we have $\Phi(x)$, and then define the saddle points of Φ by Definition 2. The readers should not confuse with these two completely different notions of saddle points.*

The local Nash equilibrium can be characterized in terms of first-order and second-order conditions. Specifically, when assuming f is twice-differentiable, any stationary point (i.e., (x, y) such that $\nabla f(x, y) = 0$) is a **strict local Nash equilibrium** if and only if

$$\nabla^2_{yy} f(x, y) \prec 0, \text{ and } \nabla^2_{xx} f(x, y) \succ 0.$$

We have the following proposition, showing that the local minimum of $\Phi(x)$ is indeed superior to its saddle point regarding whether it is a local Nash equilibrium or not.

Proposition 14 *For any smooth nonconvex-strongly-concave function $f(x, y)$, define $\Phi(x) = \max_{y \in \mathbb{R}^n} f(x, y)$. Then we have*

(i) A saddle point of $\Phi(x)$ cannot be a strict local Nash equilibrium of $f(x, y)$.

(ii) A strict local Nash equilibrium of $f(x, y)$ must be a local minimum of $\Phi(x)$.

Moreover, Jin et al. (2020) introduced the concept of local minimax point, which is a weakened notion of the local Nash equilibrium. Compared with the local Nash equilibrium, the local minimax point alleviates the non-existence issue[1] and is the first proper mathematical definition of local optimality for the two-player sequential games.

Definition 15 *(Jin et al., 2020)[Strict Local Minimax Point] For any twice differentiable function $f(x, y)$, a point (x, y) is a strict local minimax point if it satisfies $\nabla f(x, y) = 0$, $\nabla^2_{yy} f(x, y) \prec 0$ and*

$$\nabla^2_{xx} f(x, y) - \nabla^2_{xy} f(x, y) \nabla^2_{yy} f(x, y)^{-1} \nabla^2_{xy} f(x, y) \succ 0.$$

[1] In the Proposition 6 of (Jin et al., 2020), the authors constructed a two dimensional function showing that the Nash equilibria may not exist, and this is known as the "non-existence issue".

The following proposition shows the equivalence between a strict local minimax point and a strict local minimum of $\Phi(x)$.

Proposition 16 *For nonconvex-strongly-concave minimax problem* (3.1)*, suppose* $\Phi(x)$ *has a strict local minimum, then a strict local minimax point of* (3.1) *always exists and is equivalent to a strict local minimum of* $\Phi(x)$*.*

Most existing convergence theory for minimax problems focuses on finding ϵ-stationary point. Recently, Chen et al. (2021b) and Luo and Chen (2021) proposed second-order algorithms for minimizing $\Phi(x)$, which is guaranteed to converge to local minimum. Second-order methods enjoy faster convergence rate than gradient methods, but require solving nonconvex subproblems in each iteration. Moreover, second-order methods are difficult to be implemented in large-scale problems due to the heavy computation of Hessian matrices. Fiez et al. (2021) proved that GDA asymptotically converges to strict local minimax point almost surely. However, no convergence rate for finding a local minimax point was given in Fiez et al. (2021).

The above facts motivate us to propose the perturbed GDmax Algorithm (Algorithm 1 with option **GDmax** in step 9), a first-order nested-loop algorithm that provably escapes saddle points in minimax problems. In the inner loop, the perturbed GDmax runs D steps of gradient ascent for solving the y-subproblem inexactly. With the warm start strategy, we set the initial point in k-th iteration y_k^0 to be the output of the inner loop in the previous iteration y_{k-1}^D. In the outer loop, we estimate the hypergradient $\nabla\Phi(x)$ by

$$\widehat{\nabla}\Phi(x) = \nabla_x f(x_k, y_k^D),$$

and update x by one step of inexact gradient descent. When the first-order stationary condition is satisfied (step 10 in Algorithm 1), we add a random noise vector sampled uniformly from a ball with radius of r and centered at the current iterate.

Now we analyze the convergence of the perturbed GDmax algorithm. We first list our assumptions.

Assumption 17 *For the minimax problem* (3.1)*,* $f(x, y)$ *satisfies the following assumptions:*

(i) $f(x, y)$ *is twice differentiable,* μ*-strongly concave with respect to* y *and non-convex with respect to* x*.*

(ii) Denote $z = (x, y)$*.* $f(z)$ *is* ℓ*-smooth, i.e., for any* z, z'*, it holds:*

$$\|\nabla f(z) - \nabla f(z')\| \le \ell\|z - z'\|.$$

(iii) The Hessian and Jacobian matrices $\nabla_{xx}^2 f(x, y)$*,* $\nabla_{xy}^2 f(x, y)$*, and* $\nabla_{yy}^2 f(x, y)$ *are* ρ*-Lipschitz continuous.*

(iv) Function $\Phi(x)$ *is bounded below and has compact sub-level sets.*

Remark 18 *Compared with assumptions for general bilevel optimization problem (Assumption 3 in Section 2), we do not require any third-order derivative information for the minimax problem.*

Our perturbed GDmax algorithm is the first pure gradient algorithm with a nonasymptotic convergence rate for finding a local minimax point. The main results for the perturbed GDmax algorithm are given in the following theorem.

Theorem 19 (Convergence of Perturbed GDmax) *Suppose $f(x,y)$ satisfies Assumption 17. Set parameters as*

$$\tau = \frac{1}{\ell}, \quad \eta = \frac{1}{L_\phi}, \quad r = \frac{\epsilon}{400\iota^3}, \quad \mathscr{T} = \frac{L_\phi}{\sqrt{\rho_\phi \epsilon}} \cdot \iota \tag{3.2}$$

*and $D = \mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)\right)$, with probability at least $1-\delta$, the perturbed GDmax Algorithm (i.e., Algorithm 1 with option **GDmax** in step 9) obtains an ϵ-local minimum of $\Phi(x) = \max_y f(x,y)$ in*

$$K = \tilde{\mathcal{O}}\left(\frac{L_\phi(\Phi(x_0) - \Phi(x^*))}{\epsilon^2}\right) = \tilde{\mathcal{O}}(\kappa\epsilon^{-2})$$

iterations. Here $\delta \in (0,1)$, L_ϕ is the Lipschitz constant of $\nabla\Phi$, ρ_ϕ is the Lipschitz constant of $\nabla^2\Phi$ (see Lemma 36), and ι is a constant satisfying

$$\iota > 1, \text{ and } \delta \geq \frac{L_\phi\sqrt{d}}{\sqrt{\rho_\phi\epsilon}} \cdot \iota^2 2^{8-\iota}. \tag{3.3}$$

Remark 20 *Throughout this paper, we also assume*

$$L_\phi/\sqrt{\rho_\phi\epsilon} \geq 1. \tag{3.4}$$

We make this assumption because if (3.4) does not hold, finding the ϵ-local minimum is straightforward, see Jin et al. (2021).

Remark 21 *Note that in practice, we may choose ι sufficiently large so that δ in (3.3) can be small, which leads to the fact that Theorem 17 holds with probability at least $1-\delta$.*

Corollary 22 *The complexity of the gradient evaluations of the perturbed GDmax algorithm for finding an ϵ-local minimum of $\Phi(x) = \max_y f(x,y)$ is*

$$G(f,\epsilon) = D \cdot K = \tilde{\mathcal{O}}\left(\kappa^2\epsilon^{-2}\right).$$

Remark 23 *Compared with results of second-order methods escaping saddle points for minimax problem (Chen et al., 2021b; Luo and Chen, 2021), our perturbed GDmax algorithm is purely first order, which means we do not need to compute Hessian-vector product. Moreover, algorithms in Chen et al. (2021b) and Luo and Chen (2021) require solving a nonconvex cubic sub-problem and multiple linear systems in each iteration. All these expensive computations are avoided in our perturbed GDmax algorithm, which makes it practical in real applications.*

Remark 24 *Compared with asymptotic results in Fiez et al. (2021), we provide nonasymptotic convergence rate for finding a local minimax point for minimax problems.*

Remark 25 *The dependence on the conditional number κ for the perturbed GDmax algorithm is $\tilde{\mathcal{O}}(\kappa^2)$, which matches the order in Lin et al. (2020b) and is better than the results in Chen et al. (2021b).*

4. Inexact NEON and Stochastic Bilevel Algorithms

In this section, we consider escaping saddle points for stochastic bilevel optimization problem (1.8). Inspired by recent work (Xu et al., 2018) and (Allen-Zhu and Li, 2018), we propose inexact NEON (iNEON) that helps escape saddle points in stochastic bilevel optimization (1.8).

4.1 Inexact NEON

Recently, Xu et al. (2018) and Allen-Zhu and Li (2018) proposed NEgative curvature Originated from Noise (NEON) and NEON2, two pure first-order methods that extract negative curvature descent direction. NEON turns almost all stationary-point finding algorithms into local-minimum finding algorithms. The work of Xu et al. (2018) was inspired by the connection between perturbed gradient descent method (Jin et al., 2017) and the power method, while the idea of Allen-Zhu and Li (2018) is based on the result of Oja's algorithm (Allen-Zhu and Li, 2017). Compared with classical optimization problems, bilevel optimization no longer has access to the exact gradient, which motivates us to propose the inexact NEON (iNEON). The proposed iNEON update is

$$u_{k+1} = u_k - \eta(\widehat{\nabla}\Phi(\tilde{x}+u_k) - \widehat{\nabla}\Phi(\tilde{x})), \tag{4.1}$$

where $\widehat{\nabla}\Phi(\tilde{x}+u_k)$ and $\widehat{\nabla}\Phi(\tilde{x})$ are the hypergradient estimates. Our iNEON algorithm is described in Algorithm 2. Intuitively, the iNEON can be viewed as an approximate power method. More specifically, note that (4.1) is equivalent to

$$\begin{aligned} u_{k+1} =& u_k - \eta(\nabla\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x})) + \delta_k \\ \approx& (I - \eta\nabla^2\Phi(\tilde{x}))u_k + \delta_k, \end{aligned} \tag{4.2}$$

where $\delta_k = \eta(\widehat{\nabla}\Phi(\tilde{x}+u_k) - \widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(\tilde{x}+u_k) + \nabla\Phi(\tilde{x}))$ is the gradient estimation error and in the last step we use the approximation: $\nabla\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x}) \approx \nabla^2\Phi(\tilde{x})u_k$ as long as $\|u_k\|$ is small. Therefore, (4.1) is equivalent to applying approximate power method to the matrix $I - \eta\nabla^2\Phi(\tilde{x})$ starting with initial vector u_0.

We next show that iNEON can extract negative gradient descent direction with high probability.

Lemma 26 *Suppose Assumption 3 holds. Choose parameters*

$$\tau = \frac{1}{\ell}, \quad \eta = \frac{1}{L_\phi}, \quad r = \frac{\epsilon}{400\iota^3}, \quad \mathscr{T} = \frac{L_\phi}{\sqrt{\rho_\phi\epsilon}} \cdot \frac{\iota}{4}, \tag{4.3}$$

$$\mathscr{F} = \frac{1}{25\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}}, \tag{4.4}$$

and $D = \mathcal{O}(\kappa\log(\epsilon^{-1}))$ in Algorithm 2. Let $\tilde{x}$ satisfy $\|\nabla\Phi(\tilde{x})\| \le \epsilon$, and $\lambda_{\min}(\nabla^2\Phi(\tilde{x})) \le -\sqrt{\rho_\phi\epsilon}$, where $\rho_\phi = \mathcal{O}(\kappa^5)$ and is defined in (B.17)*. Denote u_{out} as the output of Algorithm 2. If $u_{out} \neq 0$, we have with probability at least $1-\delta$ that*

$$\frac{u_{out}^\top \nabla^2\Phi(\tilde{x})u_{out}}{\|u_{out}\|^2} \le -\frac{1}{40\iota}\sqrt{\rho_\phi\epsilon}, \tag{4.5}$$

Algorithm 2 Inexact NEgative-curvature-Originated-from-Noise Algorithm (iNEON)

Input: Iteration Numbers $\mathscr{T}, D$, Step Sizes τ, η, Accuracy ϵ, Radius r, Potential Saddle Point $\tilde{x}$, Initial Point y_0
Select $u_0 \sim Uniform(\mathbb{B}(\eta r))$
Set $\tilde{y}^0 = y_0$
for $t = 0, 1, 2, \ldots, D$ **do**
  $\tilde{y}^t = \tilde{y}^{t-1} - \tau \cdot \nabla_y g(\tilde{x}, \tilde{y}^{t-1})$
end for
Compute $\widehat{\nabla}\Phi(\tilde{x})$ using AID and $\tilde{y}^D$
for $k = 0, 1, 2, \ldots, \mathscr{T}$ **do**
  Set $y_k^0 = y_{k-1}^D$ if $k > 0$, otherwise y_0
  for $t = 0, 1, 2, \ldots, D$ **do**
    $y_k^t = y_k^{t-1} - \tau \cdot \nabla_y g(\tilde{x} + u_k, y_k^{t-1})$
  end for
  Compute $\widehat{\nabla}\Phi(\tilde{x} + u_k)$ using AID and y_k^D
  $u_{k+1} = u_k - \eta(\widehat{\nabla}\Phi(\tilde{x} + u_k) - \widehat{\nabla}\Phi(\tilde{x}))$
  if $\hat{\Phi}(\tilde{x} + u_{k+1}) - \hat{\Phi}(\tilde{x}) - \widehat{\nabla}\Phi(\tilde{x})^\top u_{k+1} \leq -\frac{11519}{12800}\mathscr{F}$ **then**
    return $u_{out} = u_{k+1}/\|u_{k+1}\|$
  end if
end for
return 0

where ι is a constant satisfying

$$\iota > 1, \text{ and } \delta > \frac{\ell\sqrt{d}}{\sqrt{\rho\epsilon}} \cdot \iota^2 2^{8-\iota/4}. \tag{4.6}$$

If $u_{out} = 0$, then we conclude that $\lambda_{\min}(\nabla^2\Phi(x)) \geq -\sqrt{\rho_\phi \epsilon}$ with high probability $1 - O(\delta)$.

Remark 27 *Compared with results in Xu et al. (2018), we provide a simplified proof that can handle the gradient estimation error.*

So far, we treat iNEON as a deterministic algorithm that extracts the descent direction for a deterministic objective function $\Phi(x)$. We will show how to apply iNEON to stochastic bilevel algorithms in the next section.

4.2 StocBiO Escapes Saddle Point

In this section, we apply iNEON to a popular algorithm for stochastic bilevel optimization, StocBiO (Ji et al., 2021). StocBiO is a double-loop batch stochastic algorithm, which has similar structure as AID. In its inner loop, it runs D steps of stochastic gradient descent (SGD) for an estimated solution y_k^D. Let $\Pi_{Q+1}^{Q}(\cdot) = I$. In the outer loop, StocBiO samples data batches $\mathcal{D}_F, \mathcal{D}_H = \{\mathcal{B}_j, j = 1, ..., Q\}$ and $\mathcal{D}_G$ and constructs v_Q as an approximate

Algorithm 3 StocBiO with iNEON

1: **Input:** K, D, Q, batch size S, stepsizes α and β, initializations x_0 and y_0.
2: **for** $k = 1, \ldots, K-1$ **do**
3: Set $y_k^0 = y_{k-1}^D$ if $k > 0$, otherwise y_0
4: **for** $t = 1, ..., D$ **do**
5: Draw a sample batch $\mathcal{S}_{t-1}$ with batch size S
6: Update $y_k^t = y_k^{t-1} - \alpha \nabla_y G(x_k, y_k^{t-1}; \mathcal{S}_{t-1})$
7: **end for**
8: Draw sample batches $\mathcal{D}_F, \mathcal{D}_H$ and $\mathcal{D}_G$
9: Compute $\widehat{\nabla}\Phi(x_k)$ by (4.7) - (4.8)
10: Update $x_{k+1} = x_k - \beta \widehat{\nabla}\Phi(x_k)$
11: $k \leftarrow k+1$;
12: Compute $\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)$ via AID
13: **if** $\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)\| \le \frac{4}{5}\epsilon$ **then**
14: $u = iNEON(x_k, \mathscr{T}, r, f_{\mathcal{D}_F}, g_{\mathcal{D}_G})$
15: **if** $u = 0$ **then**
16: Return x_k;
17: **else**
18: Select Rademacher variable $\bar{\xi} \in \{1, -1\}$
19: $x_{k+1} = x_k - \frac{\bar{\xi}}{80}\sqrt{\frac{\epsilon}{\rho_\phi}}u$
20: $k \leftarrow k+1$;
21: **end if**
22: **end if**
23: **end for**

solution of the linear system (2.1) as follows:

$$\begin{aligned} v_0 =& \nabla_y F(x_k, y_k^D; \mathcal{D}_F), \\ v_Q =& \eta \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} (I - \eta \nabla_{yy}^2 G(x_k, y_k^D; \mathcal{B}_j)) v_0. \end{aligned} \tag{4.7}$$

The stochastic hypergradient can be constructed as

$$\widehat{\nabla}\Phi(x_k) = \nabla_x F(x_k, y_k^D; \mathcal{D}_F) - \nabla_{xy}^2 G(x_k, y_k^D; \mathcal{D}_G) v_Q. \tag{4.8}$$

When the norm of the batch gradient is small (see step 12 of Algorithm 3), we fix sample batches $\mathcal{D}_F, \mathcal{D}_G$ and call iNEON. Denote $f_{\mathcal{D}_F}(x,y) = \frac{1}{D_f}\sum_{i=1}^{D_f} F(x,y;\xi_i)$, $g_{\mathcal{D}_G}(x,y) = \frac{1}{D_g}\sum_{i=1}^{D_g} G(x,y;\zeta_i)$ and $\Phi_{\mathcal{D}}(x) = f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))$. iNEON finds the descent direction for $\Phi_{\mathcal{D}}(x)$ at saddle points with high probability. We list the assumptions for Algorithm 3 as following.

Assumption 28 *For the stochastic case, Assumption 3 holds for $F(x,y;\xi)$ and $G(x,y;\zeta)$ for any given ξ and ζ.*

Assumption 29 *The variance of gradient $\nabla G(x, y; \zeta)$ is bounded:*

$$\mathbb{E}_\zeta \|\nabla G(x, y; \zeta) - \nabla g(x, y)\|^2 \leq \sigma^2.$$

The following theorem provides our main results on stochastic bilevel optimization.

Theorem 30 *Suppose Assumptions 28 and 29 hold. Set parameters as* (4.3) *and* (4.4)*, and let $\alpha = \frac{2}{\ell+\mu}$, $\beta = \frac{1}{4L_\phi}$, $D = \mathcal{O}\left(\kappa \log(\epsilon^{-1})\right), Q = \mathcal{O}\left(\kappa \log(\epsilon^{-1})\right), |\mathcal{B}_{Q+1-j}| = BQ(1-\eta\mu)^{j-1}$, for $j = 1, ..., Q$, where $B = \mathcal{O}\left(\kappa^2 \cdot \epsilon^{-2}\right)$, and $S = \mathcal{O}\left(\kappa^5 \cdot \epsilon^{-2}\right), D_f = \tilde{\mathcal{O}}\left(\kappa^2 \cdot \epsilon^{-2}\right), D_g = \tilde{\mathcal{O}}\left(\kappa^6 \cdot \epsilon^{-2}\right)$ in Algorithm 3. With high probability, the total iteration number of the Algorithm 3 for visiting an ϵ-local minimum of* (1.8) *can bounded by*

$$K = \tilde{\mathcal{O}}\left(\frac{L_\phi(\Phi(x_0) - \Phi(x^*))}{\epsilon^2}\right) = \tilde{\mathcal{O}}(\kappa^3 \epsilon^{-2}),$$

where L_ϕ is the Lipschitz constant of $\nabla\Phi(x)$, which is defined in Lemma 34.

Corollary 31 *For Algorithm 3, the complexities of gradient evaluations for finding an ϵ-local minimum of* (1.8) *are*

$$G(f, \epsilon) = \mathcal{O}(\kappa^5 \epsilon^{-4}), \quad G(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^{10} \epsilon^{-4}),$$

and the Jacobian- and Hessian-vector product complexities are

$$JV(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^9 \epsilon^{-4}), \quad HV(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^{9.5} \epsilon^{-4}).$$

Remark 32 *Compared with the StocBiO complexity results in Ji et al. (2021), our results have a worse dependence on the condition number κ. This is because we set a larger sample size D_g in order to obtain a high probability result.*

5. Numerical Experiments

5.1 Deterministic case

In this section we present the experimental results to demonstrate the efficiency of our algorithm. We reformulate the problem in Du et al. (2017) as a bilevel optimization problem (1.1) and then compare our Algorithm 1 with AID-BiO in Ji et al. (2021). More precisely, we consider the following bilevel optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \Phi(x) := f_1(x, y^*(x)), \\ \text{s.t.} \quad & y^*(x) = \operatorname*{argmin}_{y \in \mathbb{R}} f_2(x, y). \end{aligned} \tag{5.1}$$

Motivated by Du et al. (2017), we construct the following functions. For the upper level function we have

$$f_1(x, y) = \begin{cases} f_{i,1}(x, y) & x_1, ..., x_{i-1} \in [2\tau, 6\tau],\ x_i \in [0, \tau],\ x_{i+1}, ..., x_d \in [0, \tau],\ 1 \leq i \leq d-1 \\ f_{i,2}(x, y) & x_1, ..., x_{i-1} \in [2\tau, 6\tau],\ x_i \in [\tau, 2\tau],\ x_{i+1}, ..., x_d \in [0, \tau],\ 1 \leq i \leq d-1 \\ f_{d,1}(x, y) & x_1, ..., x_{d-1} \in [2\tau, 6\tau],\ x_d \in [0, \tau] \\ f_{d,2}(x, y) & x_1, ..., x_{d-1} \in [2\tau, 6\tau],\ x_d \in [\tau, 2\tau] \\ f_{d+1,1}(x, y) & x_1, ..., x_d \in [2\tau, 6\tau], \end{cases} \tag{5.2}$$

where

$$f_{i,1}(x,y) = \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 - \gamma x_i^2 + \sum_{j=i+1}^{d} L x_j^2 - (i-1)\nu,\ 1 \le i \le d-1, \tag{5.3}$$

$$f_{i,2}(x,y) = \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 + y + \sum_{j=i+2}^{d} L x_j^2 - (i-1)\nu,\ 1 \le i \le d-1, \tag{5.4}$$

$$f_{d,1}(x,y) = \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 - \gamma x_d^2 - (d-1)\nu, \tag{5.5}$$

$$f_{d,2}(x,y) = \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 + y - (d-1)\nu, \tag{5.6}$$

$$f_{d+1,1}(x,y) = \sum_{j=1}^{d} L(x_j - 4\tau)^2 - d\nu. \tag{5.7}$$

The lower level function is defined as

$$f_2(x,y) = \frac{y^2}{2} - g(x)y, \tag{5.8}$$

where

$$g(x) = \begin{cases} h_1(x_i) + h_2(x_i)x_{i+1}^2 & x_1,...,x_{i-1} \in [2\tau, 6\tau],\ x_i \in [\tau, 2\tau],\ x_{i+1},...,x_d \in [0,\tau], \\ & 1 \le i \le d-1 \\ h_1(x_d) & x_1,...,x_{d-1} \in [2\tau, 6\tau],\ x_d \in [\tau, 2\tau] \\ 0 & \text{elsewhere} \end{cases} \tag{5.9}$$

$$h_1(c) = -\gamma c^2 + \frac{(-14L + 10\gamma)(c-\tau)^3}{3\tau} + \frac{(5L - 3\gamma)(c-\tau)^4}{2\tau^2} \tag{5.10}$$

$$h_2(c) = -\gamma - \frac{10(L+\gamma)(c-2\tau)^3}{\tau^3} - \frac{15(L+\gamma)(c-2\tau)^4}{\tau^4} - \frac{6(L+\gamma)(c-2\tau)^5}{\tau^5} \tag{5.11}$$

The constants satisfy

$$L > 0,\ \gamma > 0,\ \tau = e,\ \nu = -h_1(2\tau) + 4L\tau^2.$$

Note that from (5.2) we know the function $\Phi(x)$ is only defined on the following domain (see also Eq. (5) in Du et al. (2017)):

$$D_0 = \bigcup_{i=1}^{d+1} \left\{ x \in \mathbb{R}^d : 6\tau \ge x_1,...,x_{i-1} \ge 2\tau, 2\tau \ge x_i \ge 0, \tau \ge x_{i+1},...,x_d \ge 0 \right\}. \tag{5.12}$$

By Lemma A.3 in Du et al. (2017) we know there are d saddle points in D_0:

$$(0,...,0)^\top, (4\tau,0,...,0)^\top, ..., (4\tau,...,4\tau,0)^\top.$$

Moreover, the only local optimum is $(4\tau, ..., 4\tau)^\top$. One can follow Steps 2 and 3 in Section A.1 of Du et al. (2017) to extend the domain to $\mathbb{R}^d$. For simplicity we omit the extension here. We refer the interested readers to Section 4 and Appendix of Du et al. (2017) for details of the motivation for constructing these functions. In our experiments, we choose the total number of iterations to be 1000 and all stepsizes to be 0.05 in both Algorithm 1 and AID-BiO. Following Du et al. (2017), we conduct the comparison using different choices of problem parameters. In Figure 1 we plot the learning curves of $\Phi(x) - \min \Phi(x)$ vs. Iteration number. Our algorithm is denoted as "PBO", and AID-BiO is denoted as "BO" in Figure 1. Note that each learning curve is nearly a step function which consists of vertical and horizontal line segments. The horizontal segment indicates that the function value does not change and thus we may deduce that the iterates are stuck at a saddle point. Each vertical segment indicates that a perturbation successfully helps the iterate escape the saddle point. We observe that under different parameter choices our Algorithm 1 escapes saddle points more efficiently than standard bilevel optimization algorithm.

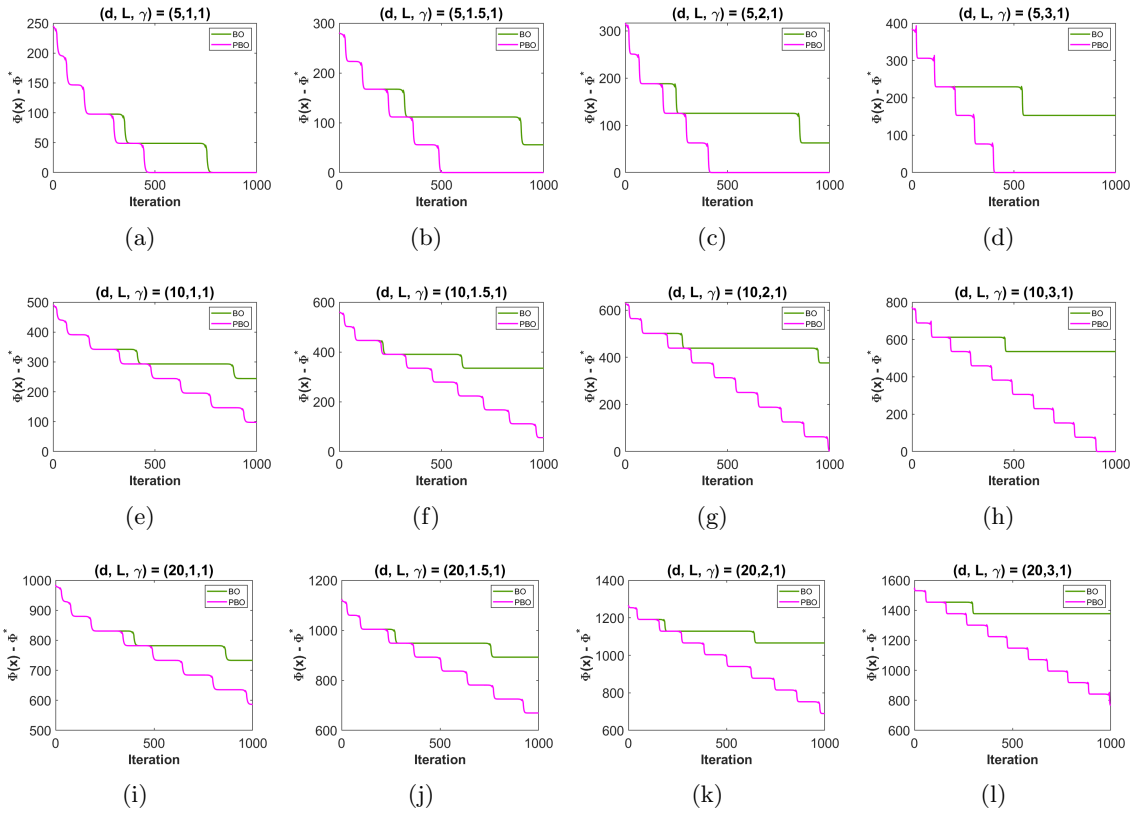


Figure 1: Comparison between Algorithm 1 and AID-BiO in Ji et al. (2021). PBO and BO represent Algorithm 1 and AID-BiO respectively. d is the dimension of the upper level function in (1.1). L and γ are problem parameters used to generate the functions in both levels.

5.2 Stochastic case

Next we investigate the performance of our Algorithm 3 under the stochastic setting. Consider symmetric matrix sensing problem (Ge et al., 2017) as follows

$$\min_{U\in\mathbb{R}^{d\times r}} \frac{1}{2m}\sum_{i=1}^{m}(\langle A_i, UU^\top\rangle - b_i)^2 \tag{5.13}$$

where $\langle A, B\rangle = \text{Trace}(AB^\top)$ for any matrices A, B and $0 < r < d$. The matrices $\{A_i : i = 1, ..., m\}$ are sensing matrices and $b_i = \langle A_i, M^*\rangle$ denotes the observation that corresponds to A_i, where $M^* = U^*(U^*)^\top$ with $U^* \in \mathbb{R}^{d\times r}$ as the ground-truth matrix. The goal of (5.13) is to recover the low-rank matrix U^* based on (A_i, b_i) pairs. Note that (5.13) can be reformulated as follows.

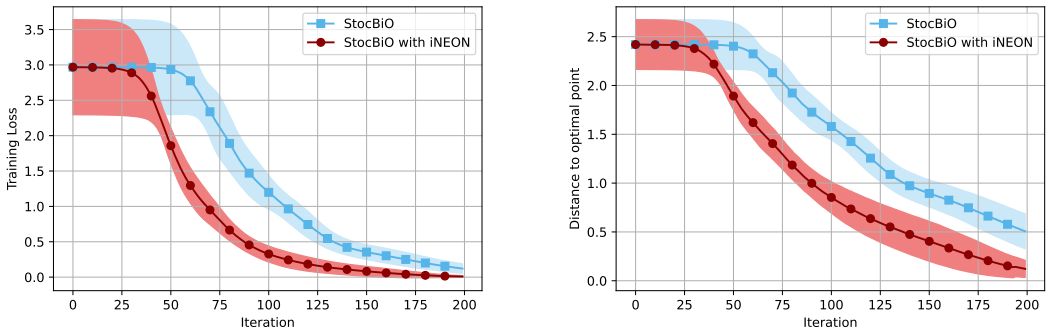
$$\begin{aligned}\min_{U\in\mathbb{R}^{d\times r}} \quad & \frac{1}{2m}\sum_{i=1}^{m}(\langle A_i, Y^*(U)\rangle - b_i)^2 \\ \text{s.t.} \quad & Y^*(U) = \operatorname*{argmin}_{Y\in\mathbb{R}^{d\times d}} \Big(\langle Y, Y\rangle - 2\langle Y, UU^\top\rangle\Big).\end{aligned} \tag{5.14}$$

By reformulating the original problem (5.13) as a bi-level one in (5.14), the objective functions in both levels are now convex in Y. We compare StocBiO (Ji et al., 2021) and our Algorithm 3 for solving (5.14). Note that StocBiO is essentially our Algorithm 3 without lines 12-22. We first generate U^* in which every entry is sampled from the normal distribution $\mathcal{N}(0, \frac{1}{d})$, and then generate m matrices $\{A_i : i = 1, ..., m\}$ where each entry is sampled from standard normal distribution $\mathcal{N}(0, 1)$. For both algorithms, we set $d = 50, r = 5, m = 2000, K = 200, D = 5, \alpha = 0.1, \beta = 0.03$ and the batch-size to be 200. For Algorithm 3, we set $\mathscr{T} = 5, r = 0.5, \epsilon = 0.03, \rho_\phi = 0.001, \tau = 0.03, \eta = 0.1, \mathscr{F} = 0.001$. Each entry of the initial U is uniformly sampled from $[0, 0.001]$ and the initial Y is set to a zero matrix. Denote by (U_k, Y_k) the matrices obtained at the k-th iteration (in StocBiO or our Algorithm 3). In Figure 2(a) we plot the training loss $\frac{1}{2m}\sum_{i=1}^{m}(\langle A_i, Y_k\rangle - b_i)^2$, and in Figure 2(b) we plot the distance between iterate and the ground-truth $\sqrt{\langle U_k - U^*, U_k - U^*\rangle}$ in each iteration. From the figures we can observe that both algorithms get stuck near a saddle point in the first few iterations, and then escape it. Our Algorithm 3 escapes the saddle point faster than StocBiO, which further validates our theory under the stochastic setting.

6. Conclusion

In this paper, we have proposed the perturbed AID algorithm that provably converges to an ϵ-local minimum in bilevel optimization. As a byproduct, we have provided the first nonasymptotic convergence rate for minimax problem converging to local minimax point with first-order method. Moreover, we have proposed inexact NEON that can extract negative gradient descent direction at saddle points. By combining the inexact NEON with StocBiO, we have proposed the first algorithm that converges to local minimum for stochastic bilevel optimization.

Acknowledgments



(a) (b)

Figure 2: Comparison between Algorithm 3 and StocBiO in Ji et al. (2021).

We are grateful to two anonymous reviewers for their insightful comments and suggestions that have helped improve the presentation of this paper. This work has been supported in part by NSF grants DMS-2243650, CCF-2308597, CCF-1934568, ECCS-2000415, CCF-2112504, CCF-2232907, CCF-2311275, and ECCS-2326591, ONR grant N00014-24-1-2705, UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program, and a startup fund at Rice University.

Appendix A. Preliminaries

Under Assumption 3, we have the following proposition. The proof can be found in Chen et al. (2021b)[Lemma 1].

Proposition 33 *Suppose Assumption 3 holds. We have the following bounds hold*

$$\begin{aligned}
&\|\nabla^2_{yy} g(x,y)^{-1}\| \leq \frac{1}{\mu},\\
&\max\left\{\|\nabla_x f(x,y)\|, \|\nabla_y f(x,y)\|, \|\nabla_y g(x,y)\|\right\} \leq M,\\
&\max\left\{\|\nabla^2_{xx} f(x,y)\|, \|\nabla^2_{yy} f(x,y)\|, \|\nabla^2_{xy} f(x,y)\|, \|\nabla^2_{xy} g(x,y)\|, \|\nabla^2_{yy} g(x,y)\|\right\} \leq \ell,\\
&\max\left\{\|\nabla^3_{xxy} g(x,y)\|, \|\nabla^3_{yxy} g(x,y)\|, \|\nabla^3_{yyy} g(x,y)\|\right\} \leq \rho.
\end{aligned}$$

Under Assumption 3, the gradient of $\Phi(x)$ is Lipschitz continuous.

Lemma 34 *Ghadimi and Wang (2018)[Lemma 2.2] Suppose Assumption 3 holds for $f(x,y)$ and $g(x,y)$, Then $\Phi(x)$ is L_ϕ smooth and the following inequality holds for any $x, x' \in \mathbb{R}^n$:*

$$\left\|\nabla\Phi(x) - \nabla\Phi(x')\right\| \leq L_\phi \|x - x'\|, \tag{A.1}$$

where

$$L_\phi = \ell + \frac{2\ell^2 + \rho M^2}{\mu} + \frac{\ell^3 + 2\rho\ell M}{\mu^2} + \frac{\rho\ell^2 M}{\mu^3}. \tag{A.2}$$

We postpone the proof of Theorem 19 to Section C and focus on the general bilevel optimization first.

Appendix B. Proofs of Results in Section 2

B.1 Proof of Lemma 6

Proof. For general bilevel optimization problem (1.1), the Hessian of function $\Phi(x)$ can be computed as

$$\begin{aligned}
\nabla^2\Phi(x) =& \nabla^2_{xx} f(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \cdot \nabla^2_{yx} f(x, y^*(x)) + \frac{\partial^2 y^*(x)}{\partial^2 x} \cdot \nabla_y f(x, y^*(x))\\
&+ \frac{\partial y^*(x)}{\partial x} \cdot \nabla^2_{xy} f(x, y^*(x))^\top + \frac{\partial y^*(x)}{\partial x} \cdot \nabla^2_{yy} f(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top,
\end{aligned} \tag{B.1}$$

which is obtained by taking derivative on (1.2). By further taking the derivative with respect to x on (1.3), we obtain:

$$\begin{aligned}
&\nabla^3_{xxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) + \frac{\partial^2 y^*(x)}{\partial^2 x} \cdot \nabla^2_{yy} g(x, y^*(x))\\
&+ \frac{\partial y^*(x)}{\partial x} \nabla^3_{xyy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top = 0.
\end{aligned} \tag{B.2}$$

By (B.1), we have

$$
\begin{aligned}
&\|\nabla^2\Phi(x) - \nabla^2\Phi(x')\| \\
=&\Bigg\|\nabla^2_{xx} f(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx} f(x, y^*(x)) + \frac{\partial^2 y^*(x)}{\partial^2 x}\cdot\nabla_y f(x, y^*(x)) \\
&+ \frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{xy} f(x, y^*(x))^\top + \frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yy} f(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top \\
&- \Bigg(\nabla^2_{xx} f(x', y^*(x')) + \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yx} f(x', y^*(x')) + \frac{\partial^2 y^*(x')}{\partial^2 x'}\cdot\nabla_y f(x', y^*(x')) \\
&+ \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{xy} f(x', y^*(x'))^\top + \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yy} f(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\Bigg)\Bigg\| \\
\le&\underbrace{\left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f(x', y^*(x'))\right\|}_{(I)} \\
&+ 2\underbrace{\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx} f(x, y^*(x)) - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yx} f(x', y^*(x'))\right\|}_{(II)} \\
&+ \underbrace{\left\|\frac{\partial^2 y^*(x)}{\partial^2 x}\cdot\nabla_y f(x, y^*(x)) - \frac{\partial^2 y^*(x')}{\partial^2 x'}\cdot\nabla_y f(x', y^*(x'))\right\|}_{(III)} \\
&+ \underbrace{\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yy} f(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yy} f(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\right\|}_{(IV)},
\end{aligned}
\tag{B.3}
$$

where the inequality is due to the triangle inequality and the fact that $\nabla^2_{yx} f(x, y^*(x)) = \nabla^2_{xy} f(x, y^*(x))^\top$ for any smooth functions $f(x, y)$. We then bound the terms $(I) - (IV)$. By Ghadimi and Wang (2018)[Lemma 2.2], we know that $y^*(x)$ is $\frac{\ell}{\mu}$-Lipschitz continuous. Therefore, we can bound the first term as

$$
\begin{aligned}
(I) =& \left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f(x', y^*(x'))\right\| \\
\le& \left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f(x', y^*(x))\right\| + \left\|\nabla^2_{xx} f(x', y^*(x)) - \nabla^2_{xx} f(x', y^*(x'))\right\| \\
\le& \rho\left(\|x - x'\| + \|y^*(x) - y^*(x')\|\right) \\
\le& \rho\left(1 + \frac{\ell}{\mu}\right)\|x - x'\|,
\end{aligned}
\tag{B.4}
$$

where the second inequality applies Assumption 3 and the last inequality is obtained by the Lipschitz continuity of $y^*(x)$. For the second term (II), we have the following bound:

$$\begin{aligned}
(II) =& 2\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx}f(x,y^*(x)) - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yx}f(x',y^*(x'))\right\| \\
\leq& 2\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx}f(x,y^*(x)) - \frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx}f(x',y^*(x'))\right\| \\
&+ 2\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx}f(x',y^*(x')) - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yx}f(x',y^*(x'))\right\| \\
\leq& 2\left\|\frac{\partial y^*(x)}{\partial x}\right\|\left\|\nabla^2_{yx}f(x,y^*(x)) - \nabla^2_{yx}f(x',y^*(x'))\right\| \\
&+ 2\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\|\left\|\nabla^2_{yx}f(x',y^*(x'))\right\| \\
\leq& \frac{2\ell}{\mu}\left\|\nabla^2_{yx}f(x,y^*(x)) - \nabla^2_{yx}f(x',y^*(x'))\right\| + 2\ell\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\| \\
\leq& \frac{2\ell}{\mu}\rho\left(1+\frac{\ell}{\mu}\right)\|x-x'\| + 2\ell\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\|,
\end{aligned} \tag{B.5}$$

where the second inequality is by the Cauchy-Schwarz inequality, and the last two inequalities are obtained by Assumption 3. Moreover, we can bound $\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\|$ as

$$\begin{aligned}
&\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\| \\
\leq& \left\|\nabla^2_{xy}g(x,y^*(x))\cdot\nabla^2_{yy}g(x,y^*(x))^{-1} - \nabla^2_{xy}g(x',y^*(x'))\cdot\nabla^2_{yy}g(x',y^*(x'))^{-1}\right\| \\
\leq& \left\|\nabla^2_{xy}g(x,y^*(x))\cdot\nabla^2_{yy}g(x,y^*(x))^{-1} - \nabla^2_{xy}g(x,y^*(x))\cdot\nabla^2_{yy}g(x',y^*(x'))^{-1}\right\| \\
&+ \left\|\nabla_{xy}g(x,y^*(x))\cdot\nabla^2_{yy}g(x',y^*(x'))^{-1} - \nabla^2_{xy}g(x',y^*(x'))\cdot\nabla^2_{yy}g(x',y^*(x'))^{-1}\right\| \\
\leq& \ell\left\|\nabla^2_{yy}g(x,y^*(x))^{-1} - \nabla^2_{yy}g(x',y^*(x'))^{-1}\right\| + \frac{1}{\mu}\left\|\nabla^2_{xy}g(x,y^*(x)) - \nabla^2_{xy}g(x',y^*(x'))\right\| \\
\leq& \frac{\ell}{\mu^2}\left\|\nabla^2_{yy}g(x,y^*(x)) - \nabla^2_{yy}g(x',y^*(x'))\right\| + \frac{1}{\mu}\left\|\nabla^2_{xy}g(x,y^*(x)) - \nabla^2_{xy}g(x',y^*(x'))\right\| \\
\leq& \left(\frac{\ell}{\mu^2}+\frac{1}{\mu}\right)\rho\left(\|x-x'\| + \|y^*(x)-y^*(x')\|\right) \\
\leq& \frac{\rho}{\mu}\left(1+\frac{\ell}{\mu}\right)^2\|x-x'\|,
\end{aligned} \tag{B.6}$$

where the first inequality is by equation (1.3), the third inequality applies Proposition 33 and the forth inequality follows the fact that $\|X^{-1}-Y^{-1}\| \leq \|X^{-1}\|\|X-Y\|\|Y^{-1}\|$ and Proposition 33. Therefore, we have the following bound for the second term:

$$(II) \leq \left(\frac{2\ell\rho}{\mu}\left(1+\frac{\ell}{\mu}\right) + \frac{2\ell\rho}{\mu}\left(1+\frac{\ell}{\mu}\right)^2\right)\|x-x'\|. \tag{B.7}$$

The third term (III) can be bounded by:

$$\begin{aligned}(III) =& \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \cdot \nabla_y f(x, y^*(x)) - \frac{\partial^2 y^*(x')}{\partial^2 x'} \cdot \nabla_y f(x', y^*(x')) \right\| \\ \leq& \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \right\| \left\| \nabla_y f(x, y^*(x)) - \nabla_y f(x', y^*(x')) \right\| \\ &+ \left\| \nabla_y f(x', y^*(x')) \right\| \cdot \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'} \right\|.\end{aligned} \tag{B.8}$$

Therefore, we need to give upper bounds for both $\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \right\|$ and $\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'} \right\|$. Note that equation (B.2) yields

$$\begin{aligned}\frac{\partial^2 y^*(x)}{\partial^2 x} =& - \left[\nabla^3_{xxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x} \nabla^3_{xyy} g(x, y^*(x)) \right. \\ &\left. + \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top \right] \nabla^2_{yy} g(x, y^*(x))^{-1},\end{aligned} \tag{B.9}$$

which, combined with Proposition 33, leads to

$$\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \right\| \leq \frac{1}{\mu} \left[\rho + 2 \frac{\ell}{\mu} \cdot \rho + \left(\frac{\ell}{\mu} \right)^2 \cdot \rho \right] = \frac{\rho}{\mu} \left(1 + \frac{\ell}{\mu} \right)^2. \tag{B.10}$$

Furthermore, $\left\|\frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'}\right\|$ can be upper bounded by

$$
\begin{aligned}
&\left\|\frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'}\right\| \\
\leq &\Bigg\| \Bigg[\nabla^3_{xxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{yxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{xyy} g(x, y^*(x)) \\
&+ \frac{\partial y^*(x)}{\partial x}\cdot\nabla^3_{yyy} g(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top\Bigg]\nabla^2_{yy} g(x, y^*(x))^{-1} \\
&- \Bigg[\nabla^3_{xxy} g(x', y^*(x')) + \frac{\partial y^*(x')}{\partial x'}\nabla^3_{yxy} g(x', y^*(x')) + \frac{\partial y^*(x')}{\partial x'}\nabla_x\nabla_y\nabla_y g(x', y^*(x')) \\
&+ \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^3_{yyy} g(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\Bigg]\nabla^2_{yy} g(x', y^*(x'))^{-1}\Bigg\| \\
\leq &\Bigg\| \Bigg[\nabla^3_{xxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{yxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{xyy} g(x, y^*(x)) \\
&+ \frac{\partial y^*(x)}{\partial x}\cdot\nabla^3_{yyy} g(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top\Bigg] - \Bigg[\nabla^3_{xxy} g(x', y^*(x')) + \frac{\partial y^*(x')}{\partial x'}\nabla^3_{yxy} g(x', y^*(x')) \\
&+ \frac{\partial y^*(x')}{\partial x'}\nabla^3_{xyy} g(x', y^*(x')) + \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^3_{yyy} g(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\Bigg]\Bigg\|\left\|\nabla^2_{yy} g(x', y^*(x'))^{-1}\right\| \\
&+ \Bigg\| \Bigg[\nabla^3_{xxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{yxy} g(x, y^*(x)) + \frac{\partial y^*(x)}{\partial x}\nabla^3_{xyy} g(x, y^*(x)) \\
&+ \frac{\partial y^*(x)}{\partial x}\cdot\nabla^3_{yyy} g(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top\Bigg]\Bigg\|\left\|\nabla^2_{yy} g(x, y^*(x))^{-1} - \nabla^2_{yy} g(x', y^*(x'))^{-1}\right\| \\
\leq &\Bigg[\frac{1}{\mu}\|\nabla^3_{xxy} g(x, y^*(x)) - \nabla^3_{xxy} g(x', y^*(x'))\| \\
&+ \frac{2}{\mu}\left\|\frac{\partial y^*(x)}{\partial x}\nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*(x')}{\partial x'}\nabla^3_{yxy} g(x', y^*(x'))\right\| \\
&+ \frac{1}{\mu}\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^3_{yyy} g(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^3_{yyy} g(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\right\|\Bigg] \\
&+ \rho\left(1 + \frac{\ell}{\mu}\right)^2\left\|\nabla^2_{yy} g(x, y^*(x))^{-1} - \nabla^2_{yy} g(x', y^*(x'))^{-1}\right\| \\
\leq &\Bigg[\frac{\nu}{\mu}\left(1 + \frac{\ell}{\mu}\right)\|x - x'\| + \frac{2}{\mu}\left\|\frac{\partial y^*(x)}{\partial x}\nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*(x')}{\partial x'}\nabla^3_{yxy} g(x', y^*(x'))\right\| \\
&+ \frac{1}{\mu}\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^3_{yyy} g(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^3_{yyy} g(x', y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\right\|\Bigg] \\
&+ \left(\frac{\rho}{\mu}\right)^2\left(1 + \frac{\ell}{\mu}\right)^3\|x - x'\|, \qquad \text{(B.11)}
\end{aligned}
$$

where in the third inequality, we used Proposition 33 and the last inequality is due to (B.10), Assumption 3, and $\|X^{-1} - Y^{-1}\| \leq \|X^{-1}\|\|X - Y\|\|Y^{-1}\|$. To bound the term

$$\left\| \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*(x')}{\partial x} \nabla^3_{yxy} g(x', y^*(x')) \right\|,$$

we follow the computation of (II) and get

$$\begin{aligned} &\left\| \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) \right\| \\ \leq &\left(\frac{\ell\nu}{\mu} \left(1 + \frac{\ell}{\mu}\right) + \frac{2\rho^2}{\mu} \left(1 + \frac{\ell}{\mu}\right)^2 \right) \|x - x'\|. \end{aligned} \tag{B.12}$$

Moreover, we have

$$\begin{aligned} &\left\| \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x', y^*(x')) \cdot \frac{\partial y^*(x')}{\partial x'}^\top \right\| \\ \leq &\left\| \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top \right\| \\ &+ \left\| \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x', y^*(x')) \cdot \frac{\partial y^*(x)}{\partial x}^\top \right\| \\ &+ \left\| \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x', y^*(x')) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'} \cdot \nabla^3_{yyy} g(x', y^*(x')) \cdot \frac{\partial y^*(x')}{\partial x'}^\top \right\| \\ \leq &2 \cdot \frac{\rho\ell}{\mu} \left\| \frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'} \right\| + \left(\frac{\ell}{\mu}\right)^2 \|\nabla^3_{yyy} g(x, y^*(x)) - \nabla^3_{yyy} g(x', y^*(x'))\| \\ \leq &\left(\frac{2\ell\rho^2}{\mu^2} \left(1 + \frac{\ell}{\mu}\right)^2 + \nu \left(\frac{\ell}{\mu}\right)^2 \left(1 + \frac{\ell}{\mu}\right) \right) \|x - x'\|, \end{aligned} \tag{B.13}$$

where the second inequality is due to $\left\| \frac{\partial y^*(x)}{\partial x} \right\| \leq \frac{\ell}{\mu}$ and Proposition 33 and the last inequality is obtained by (B.6) and Assumption 3. Combining (B.11) - (B.13) leads to

$$\begin{aligned} &\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'} \right\| \\ \leq &\left[\left(\frac{\nu}{\mu} + \frac{2\ell\nu}{\mu^2} + \frac{\ell^2\nu}{\mu^3} \right) \left(1 + \frac{\ell}{\mu}\right) + \left(\frac{4\rho^2}{\mu^2} + \frac{2\ell\rho^2}{\mu^3} \right) \left(1 + \frac{\ell}{\mu}\right)^2 + \frac{\rho^2}{\mu^2} \left(1 + \frac{\ell}{\mu}\right)^3 \right] \|x - x'\|. \end{aligned} \tag{B.14}$$

Combining (B.8), (B.10), and (B.14) yields

$$\begin{aligned}(III) \leq & \frac{\rho\ell}{\mu}\left(1+\frac{\ell}{\mu}\right)^3\|x-x'\| + M\left\|\frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*(x')}{\partial^2 x'}\right\| \\ \leq & \left[M\left(\frac{\nu}{\mu}+\frac{2\ell\nu}{\mu^2}+\frac{\ell^2\nu}{\mu^3}\right)\left(1+\frac{\ell}{\mu}\right) + M\left(\frac{4\rho^2}{\mu^2}+\frac{2\ell\rho^2}{\mu^3}\right)\left(1+\frac{\ell}{\mu}\right)^2\right. \\ & \left. + \left(\frac{M\rho^2}{\mu^2}+\frac{\rho\ell}{\mu}\right)\left(1+\frac{\ell}{\mu}\right)^3\right]\|x-x'\|.\end{aligned} \tag{B.15}$$

Finally, similar to the computation in (B.13), we have

$$\begin{aligned}(IV) = & \left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yy}f(x,y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*(x')}{\partial x'}\cdot\nabla^2_{yy}f(x',y^*(x'))\cdot\frac{\partial y^*(x')}{\partial x'}^\top\right\| \\ \leq & 2\cdot\frac{\ell^2}{\mu}\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\| + \left(\frac{\ell}{\mu}\right)^2\left\|\nabla^2_{yy}f(x,y^*(x)) - \nabla^2_{yy}f(x',y^*(x'))\right\| \\ \leq & \left(\frac{2\ell^2\rho}{\mu^2}\left(1+\frac{\ell}{\mu}\right)^2 + \rho\left(\frac{\ell}{\mu}\right)^2\left(1+\frac{\ell}{\mu}\right)\right)\|x-x'\|.\end{aligned} \tag{B.16}$$

The bounds of $(I)-(IV)$ together lead to

$$\begin{aligned}\rho_\phi = & \left[\left(\rho + \frac{2\ell\rho+M\nu}{\mu} + \frac{2M\ell\nu+\rho\ell^2}{\mu^2} + \frac{M\ell^2\nu}{\mu^3}\right)\left(1+\frac{\ell}{\mu}\right)\right. \\ & \left. + \left(\frac{2\ell\rho}{\mu} + \frac{4M\rho^2+2\ell^2\rho}{\mu^2} + \frac{2M\ell\rho^2}{\mu^3}\right)\left(1+\frac{\ell}{\mu}\right)^2 + \left(\frac{M\rho^2}{\mu^2}+\frac{\rho\ell}{\mu}\right)\left(1+\frac{\ell}{\mu}\right)^3\right],\end{aligned} \tag{B.17}$$

which completes the proof. □

In our proof of the main theorem, it is crucial to bound the estimation error for the hypergradient. For the AID method, we have the following lemma.

Lemma 35 *Suppose Assumption 3 holds. For the AID approach (i.e., the **AID** option in Algorithm 1) with parameters $D=\mathcal{O}(\kappa), N=\mathcal{O}(\sqrt{\kappa})$, we have:*

$$\left\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\right\| \leq \left(\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}} + 2\ell\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\right)\Gamma_1, \tag{B.18}$$

where

$$\Gamma_1 = \widehat{\Delta} + 2\eta\left(\kappa^2+2\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\left(M+\frac{\ell M}{\mu}\right), \tag{B.19}$$

$$\widehat{\Delta} = \|y^0-y^*(x_0)\| + \|v^0-v_0^*\|, \tag{B.20}$$

and $\widehat{v}_k = \nabla^2_{yy}g(x_k,y_k^D)^{-1}\nabla_y f(x_k,y_k^D)$.

Proof. First note that the convergence rates of CG for the quadratic programming (see, e.g., eq. (17) in Grazzi et al. (2020)) and GD for strongly convex optimization (see, e.g., Theorem 2.1.14 in Nesterov (2004)) yield

$$\|v_k^N - \widehat{v}_k\| \leq 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - \widehat{v}_k\|, \tag{B.21}$$

$$\|y_k^D - y^*(x_k)\| \leq \Big(1 - \frac{\mu}{\ell}\Big)^{\frac{D}{2}} \|y_k^0 - y^*(x_k)\|. \tag{B.22}$$

Denote $v_k^* = \nabla_{yy}^2 g(x_k, y^*(x_k))^{-1} \nabla_y f(x_k, y^*(x_k))$. Note that $\nabla\Phi(x_k)$ is defined in (1.5), i.e.,

$$\nabla\Phi(x_k) = \nabla_x f(x_k, y^*(x_k)) - \nabla_{xy}^2 g(x_k, y^*(x_k)) v_k^*,$$

and $\widehat{\nabla}\Phi(x)$ is defined in step 9 of Algorithm 1, i.e.,

$$\widehat{\nabla}\Phi(x_k) = \nabla_x f(x_k, y_k^D) - \nabla_{xy}^2 g(x_k, y_k^D) v_k^N.$$

We then have the following inequality holds

$$\begin{aligned}
&\Big\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\Big\| \\
\leq&\|\nabla_x f(x_k, y^*(x_k)) - \nabla_x f(x_k, y_k^D)\| + \|\nabla_{xy}^2 g(x_k, y_k^D)\|\|v_k^* - v_k^N\| \qquad \text{(B.23)}\\
&+ \|\nabla_{xy}^2 g(x_k, y^*(x_k)) - \nabla_{xy}^2 g(x_k, y_k^D)\|\|v_k^*\| \\
\leq&\ell\|y^*(x_k) - y_k^D\| + \ell\|v_k^* - v_k^N\| + \rho\|v_k^*\|\|y_k^D - y^*(x_k)\| \\
\leq&\left(\ell + \frac{\rho M}{\mu}\right)\|y^*(x_k) - y_k^D\| + \ell\|v_k^* - v_k^N\|. \qquad \text{(B.24)}
\end{aligned}$$

Here the last inequality follows from $\|v_k^*\| \leq \|(\nabla_{yy}^2 g(x_k, y^*(x_k)))^{-1}\|\|\nabla_y f(x_k, y^*(x_k))\| \leq \frac{M}{\mu}$ in which we use Proposition 33. Next we give an upper bound of $\|v_k^* - v_k^N\|$:

$$\begin{aligned}
&\|v_k^* - v_k^N\| \leq \|v_k^* - \widehat{v}_k\| + \|v_k^N - \widehat{v}_k\| \\
\leq&\|v_k^* - \widehat{v}_k\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - \widehat{v}_k\| \\
\leq&\left(1 + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\right)\|v_k^* - \widehat{v}_k\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - v_k^*\| \\
=&\left(1 + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\right)\|\nabla_{yy}^2 g(x_k, y_k^D)^{-1}\nabla_y f(x_k, y_k^D) - \nabla_{yy}^2 g(x_k, y^*(x_k))^{-1}\nabla_y f(x_k, y^*(x_k))\| \\
&+ 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - v_k^*\| \\
\leq&\left(1 + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\right)\Big(\frac{\ell}{\mu} + \frac{\rho M}{\mu^2}\Big)\|y_k^D - y^*(x_k)\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - v_k^*\| \\
\leq&\big(1 + 2\sqrt{\kappa}\big)\Big(\frac{\ell}{\mu} + \frac{\rho M}{\mu^2}\Big)\|y_k^D - y^*(x_k)\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \|v_k^0 - v_k^*\|, \qquad \text{(B.25)}
\end{aligned}$$

where the second inequality is due to (B.21) and in the second to last inequality we have used Assumption 3, Proposition 33, and $\|X^{-1}-Y^{-1}\| \leq \|X^{-1}\|\cdot\|X-Y\|\cdot\|Y^{-1}\|$. Plugging (B.22) and (B.25) into (B.24) leads to

$$\begin{aligned}\left\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\right\| \leq &\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\left\|y_k^0-y^*(x_k)\right\| \\ &+2\ell\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_k^0-v_k^*\|.\end{aligned} \tag{B.26}$$

Secondly, by the warm-start strategy $y_k^0 = y_{k-1}^D, v_k^0 = v_{k-1}^N$ in the inner loop, we have

$$\begin{aligned}\left\|y_k^0-y^*(x_k)\right\| &\leq \left\|y_{k-1}^D-y^*(x_{k-1})\right\| + \left\|y^*(x_{k-1})-y^*(x_k)\right\| \\ &\leq \left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\left\|y_{k-1}^0-y^*(x_{k-1})\right\| + \kappa\left\|x_{k-1}-x_k\right\| \\ &\leq \left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\left\|y_{k-1}^0-y^*(x_{k-1})\right\| + \kappa\eta\left\|\widehat{\nabla}\Phi(x_{k-1})\right\|,\end{aligned} \tag{B.27}$$

where the second inequality is due to (B.22) and the fact that $y^*(x)$ is κ-Lipschitz continuous Ghadimi and Wang (2018)[Lemma 2.2], the third inequality follows the update of x_k. The next step is to bound $\left\|v_k^0-v_k^*\right\|$. Before that, we prepare the following inequality:

$$\begin{aligned}&\left\|v_{k-1}^*-v_k^*\right\| \\ =&\left\|\nabla_{yy}^2 g(x_{k-1},y^*(x_{k-1}))^{-1}\nabla_y f(x_{k-1},y^*(x_{k-1})) - \nabla_{yy}^2 g(x_k,y^*(x_k))^{-1}\nabla_y f(x_k,y^*(x_k))\right\| \\ \leq& M\left\|\nabla_{yy}^2 g(x_{k-1},y^*(x_{k-1}))^{-1} - \nabla_{yy}^2 g(x_k,y^*(x_k))^{-1}\right\| \\ &+\frac{1}{\mu}\left\|\nabla_y f(x_{k-1},y^*(x_{k-1})) - \nabla_y f(x_k,y^*(x_k))\right\| \\ \leq&\left(\kappa^2+\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\|x_k-x_{k-1}\|,\end{aligned} \tag{B.28}$$

where in the first inequality we used Proposition 33 and the second inequality follows Assumption 3. We then have the bound of $\left\|v_k^0-v_k^*\right\|$ as follows

$$\begin{aligned}&\left\|v_k^0-v_k^*\right\| = \left\|v_{k-1}^N-v_k^*\right\| \leq \left\|v_{k-1}^N-v_{k-1}^*\right\| + \left\|v_{k-1}^*-v_k^*\right\| \\ \leq&\left(1+2\sqrt{\kappa}\right)\Big(\frac{\ell}{\mu}+\frac{\rho M}{\mu^2}\Big)\|y_{k-1}^D-y^*(x_{k-1})\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_{k-1}^0-v_{k-1}^*\| + \left\|v_{k-1}^*-v_k^*\right\| \\ \leq&\left(1+2\sqrt{\kappa}\right)\left(\frac{\ell}{\mu}+\frac{\rho M}{\mu^2}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\|y_{k-1}^0-y^*(x_{k-1})\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_{k-1}^0-v_{k-1}^*\| \\ &+\left(\kappa^2+\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\|x_k-x_{k-1}\| \\ \leq&\left(1+2\sqrt{\kappa}\right)\left(\frac{\ell}{\mu}+\frac{\rho M}{\mu^2}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\|y_{k-1}^0-y^*(x_{k-1})\| + 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_{k-1}^0-v_{k-1}^*\| \\ &+\eta\left(\kappa^2+\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\left\|\widehat{\nabla}\Phi(x_{k-1})\right\|,\end{aligned} \tag{B.29}$$

where the second inequality is by (B.25), the third inequality is obtained by (B.22) and (B.28). Summing (B.27) and (B.29), we obtain

$$\begin{aligned}
&\left\|y_k^0 - y^*(x_k)\right\| + \left\|v_k^0 - v_k^*\right\| \\
\leq &\left(\left(1+2\sqrt{\kappa}\right)\left(\frac{\ell}{\mu}+\frac{\rho M}{\mu^2}\right)+1\right)\cdot\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}}\left\|y_{k-1}^0 - y^*(x_{k-1})\right\| \\
&+ 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\|v_{k-1}^0 - v_{k-1}^*\| + \left(\kappa^2+2\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\eta\left\|\widehat{\nabla}\Phi(x_{k-1})\right\|.
\end{aligned} \tag{B.30}$$

Set the parameters as

$$\begin{aligned}
D \geq &2\log\frac{1}{2\left(\left(1+2\sqrt{\kappa}\right)\left(\frac{\ell}{\mu}+\frac{\rho M}{\mu^2}\right)+1\right)}/\log(1-\kappa^{-1}) = \mathcal{O}(\kappa), \\
N \geq &\log\frac{1}{4\sqrt{\kappa}}/\log\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big) = \mathcal{O}(\sqrt{\kappa}),
\end{aligned} \tag{B.31}$$

such that we can have

$$\begin{aligned}
&\left\|y_k^0 - y^*(x_k)\right\| + \left\|v_k^0 - v_k^*\right\| \\
\leq &\frac{1}{2}\left(\|y_{k-1}^0 - y^*(x_{k-1})\| + \|v_{k-1}^0 - v_{k-1}^*\|\right) + \left(\kappa^2+2\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\eta\left\|\widehat{\nabla}\Phi(x_{k-1})\right\| \\
\leq &\left(\frac{1}{2}\right)^k\widehat{\Delta} + \left(\kappa^2+2\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\eta\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\widehat{\nabla}\Phi(x_j)\right\| \\
\leq &\widehat{\Delta} + 2\eta\left(\kappa^2+2\kappa+\frac{\rho M(1+\kappa)}{\mu^2}\right)\left(M+\frac{\ell M}{\mu}\right) = \Gamma_1,
\end{aligned} \tag{B.32}$$

where the last inequality follows $\left\|\widehat{\nabla}\Phi(x)\right\| \leq \left(M+\frac{\ell M}{\mu}\right)$ by Proposition 33. Combining (B.32) with (B.26), we have

$$\begin{aligned}
&\left\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\right\| \\
\leq &\left(\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}} + 2\ell\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N\right)\Gamma_1,
\end{aligned} \tag{B.33}$$

where the inequality follows from the inequality $ab+cd \leq (a+c)(b+d)$ for any positive a, b, c, d. This completes the proof. □

B.2 Proof of Lemma 10

Proof. By Lemma 34, $\Phi(x)$ is L_ϕ-smooth, which yields

$$\begin{aligned}
\Phi(x_{k+1}) \leq & \Phi(x_k) + \langle \nabla\Phi(x_k), x_{k+1} - x_k\rangle + \frac{L_\Phi}{2}\|x_{k+1} - x_k\|_2^2 \\
\leq & \Phi(x_k) + \langle \widehat{\nabla}\Phi(x_k), x_{k+1} - x_k\rangle + \langle \nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k), x_{k+1} - x_k\rangle \\
& + \frac{L_\Phi}{2}\|x_{k+1} - x_k\|_2^2 \\
\leq & \Phi(x_k) + \langle \widehat{\nabla}\Phi(x_k), x_{k+1} - x_k\rangle + \frac{1}{4\eta}\|x_{k+1} - x_k\|^2 + \eta\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2 \\
& + \frac{L_\Phi}{2}\|x_{k+1} - x_k\|_2^2 \\
\leq & \Phi(x_k) - \frac{\eta}{4}\|\widehat{\nabla}\Phi(x_k)\|_2^2 + \eta\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\|^2,
\end{aligned}$$

where the third inequality is obtained by Young's inequality and the last inequality uses $\eta = \frac{1}{L_\phi}$. □

B.3 Proof of Lemma 11

Proof. The proof of Lemma 11 closely follows Jin et al. (2021)[Lemma 22]. We first define two sequences $\{x_k\}$, $\{x'_k\}$ that are generated by Algorithm 1 with initial points x_0 and x'_0, respectively. That is,

$$x_{k+1} = x_k - \eta\widehat{\nabla}\Phi(x_k), \qquad x'_{k+1} = x'_k - \eta\widehat{\nabla}\Phi(x'_k).$$

We require the two initial points to satisfy the following conditions:

- **Condition (i)**: $\max\{\|x_0 - \tilde{x}\|, \|x'_0 - \tilde{x}\|\} \leq \eta r$;
- **Condition (ii)**: $x_0 - x'_0 = \eta r_0 \mathbf{e}_1$, where $\mathbf{e}_1$ is the minimum eigenvector of $\nabla^2\Phi(\tilde{x})$ with $\|\mathbf{e}_1\| = 1$ and $r_0 > \omega := 2^{2-\iota}L_\phi\mathscr{S}$, and

$$\mathscr{S} = \frac{1}{4\iota}\sqrt{\frac{\epsilon}{\rho_\phi}}. \tag{B.34}$$

Note that the parameters are given in (3.2). We show that for these two sequences, the following inequality must hold:

$$\min\{\Phi(x_{\mathscr{T}}) - \Phi(x_0), \Phi(x'_{\mathscr{T}}) - \Phi(x'_0)\} \leq -\mathscr{F}, \tag{B.35}$$

where $\mathscr{F}$ is defined in (2.9). We now prove (B.35) by contradiction. Assume the contrary of (B.35) holds, i.e.:

$$\min\{\Phi(x_{\mathscr{T}}) - \Phi(x_0), \Phi(x'_{\mathscr{T}}) - \Phi(x'_0)\} > -\mathscr{F}. \tag{B.36}$$

First, by the update of x_k, we have for any $\tau \leq k \leq \mathscr{T}$:

$$\begin{aligned}
\|x_\tau - x_0\| &\leq \sum_{t=1}^{k} \|x_t - x_{t-1}\| \leq \left[k \sum_{t=1}^{k} \|x_t - x_{t-1}\|^2\right]^{\frac{1}{2}} \\
&= \left[\eta^2 k \sum_{t=1}^{k} \left\|\widehat{\nabla}\Phi(x_{t-1})\right\|^2\right]^{\frac{1}{2}} \\
&\leq \left[\eta^2 \mathscr{T} \sum_{t=1}^{\mathscr{T}} \left\|\widehat{\nabla}\Phi(x_{t-1})\right\|^2\right]^{\frac{1}{2}} \\
&\leq \sqrt{4\eta\mathscr{T}\left(\Phi(x_0) - \Phi(x_{\mathscr{T}}) + \eta \sum_{t=1}^{\mathscr{T}} \|\nabla\Phi(x_t) - \widehat{\nabla}\Phi(x_t)\|^2\right)}, \qquad \text{(B.37)}
\end{aligned}$$

where the last inequality is obtained by Lemma 10. We have for any $k \leq \mathscr{T}$:

$$\begin{aligned}
\max\{\|x_k - \tilde{x}\|, \|x'_k - \tilde{x}\|\} &\leq \max\{\|x_k - x_0\|, \|x'_k - x'_0\|\} + \max\{\|x_0 - \tilde{x}\|, \|x'_0 - \tilde{x}\|\} \\
&\leq \sqrt{4\eta\mathscr{T}\mathscr{F} + 4\eta^2\mathscr{T}^2 \cdot \frac{17}{6400\iota^4} \cdot \epsilon^2} + \eta r \\
&\leq \frac{9}{40\iota} \cdot \sqrt{\frac{\epsilon}{\rho_\phi}} + \frac{1}{400\iota^3} \cdot \frac{\epsilon}{L_\phi} \\
&\leq \frac{9}{40\iota} \cdot \sqrt{\frac{\epsilon}{\rho_\phi}} + \frac{1}{400\iota} \cdot \sqrt{\frac{\epsilon}{\rho_\phi}} \\
&\leq \mathscr{S}, \qquad \text{(B.38)}
\end{aligned}$$

where the second inequality uses (B.37), (B.36), (2.7), and **Condition (i)**, the third inequality is due to (3.2) and (2.9), and the fourth inequality is due to (3.3) and (3.4). On the other hand, we can write the update equation for the difference $\hat{x}_k := x_k - x'_k$ as:

$$\begin{aligned}
\hat{x}_{k+1} &= \hat{x}_k - \eta\left[\widehat{\nabla}\Phi(x_k) - \widehat{\nabla}\Phi(x'_k)\right] \\
&= \hat{x}_k - \eta\left[\nabla\Phi(x_k) - \nabla\Phi(x'_k)\right] - \eta\left[\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k) + \nabla\Phi(x'_k) - \widehat{\nabla}\Phi(x'_k)\right] \\
&= (I - \eta\mathcal{H})\hat{x}_k - \eta\left(\Delta_{1,k}\hat{x}_k + \Delta_{2,k}\right) \\
&= \underbrace{(I - \eta\mathcal{H})^{k+1}\hat{x}_0}_{p(k+1)} \underbrace{- \eta\sum_{t=0}^{k}(I - \eta\mathcal{H})^{k-t}\left(\Delta_{1,t}\hat{x}_t + \Delta_{2,t}\right)}_{q(k+1)}, \qquad \text{(B.39)}
\end{aligned}$$

where we denote

$$\begin{aligned}
\mathcal{H} &= \nabla^2\Phi(\tilde{x}), \\
\Delta_{1,k} &= \int_0^1 \left[\nabla^2\Phi(x'_k + \theta(x_k - x'_k)) - \mathcal{H}\right] d\theta, \\
\Delta_{2,k} &= \left[\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k) + \nabla\Phi(x'_k) - \widehat{\nabla}\Phi(x'_k)\right].
\end{aligned}$$

For the two parts $q(k), p(k)$, we show that $p(k)$ is the dominant term by proving

$$\|q(k)\| \le \|p(k)\|/2, \quad \forall k \in [\mathscr{T}]. \tag{B.40}$$

We now prove (B.40) by induction. First, (B.40) holds trivially when $k = 0$ becase $\|q(0)\| = 0$. Denote $\lambda_{\min}(\nabla^2\Phi(\tilde{x})) = -\gamma$, which implies $\gamma \ge \sqrt{\rho_\phi \epsilon}$. Assume (B.40) holds for any $t \le k$. Since $\hat{x}_0 = \eta r_0 \mathbf{e}_1$, we have for any $t \le k$:

$$\|\hat{x}_t\| \le \|p(t)\| + \|q(t)\| \le \frac{3}{2}\|p(t)\| = \frac{3}{2}\|(\mathbf{I} - \eta\mathcal{H})^t \hat{x}_0\| = \frac{3}{2}(1+\eta\gamma)^t \eta r_0. \tag{B.41}$$

Therefore, at step $k+1$ we have

$$\begin{aligned}
\|q(k+1)\| =& \left\| \eta \sum_{t=0}^{k} (I - \eta\mathcal{H})^{k-t} \left(\Delta_{1,t}\hat{x}_t + \Delta_{2,t}\right) \right\| \\
\le& \left\| \eta \sum_{t=0}^{k} (I - \eta\mathcal{H})^{k-t} \Delta_{1,t}\hat{x}_t \right\| + \left\| \eta \sum_{t=0}^{k} (I - \eta\mathcal{H})^{k-t} \Delta_{2,t} \right\| \\
\le& \eta\rho_\phi \mathscr{S} \sum_{t=0}^{k} \left\| (I - \eta\mathcal{H})^{k-t} \right\| \|\hat{x}_t\| + \eta \sum_{t=0}^{k} \left\| (I - \eta\mathcal{H})^{k-t} \right\| \|\Delta_{2,t}\| \\
\le& \frac{3}{2}\eta\rho_\phi \mathscr{S} \sum_{t=0}^{k} (1+\eta\gamma)^k \eta r_0 + 2\eta \sum_{t=0}^{k} (1+\eta\gamma)^k \|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\| \\
\le& \frac{3}{2}\eta\rho_\phi \mathscr{S}\mathscr{T} (1+\eta\gamma)^k \eta r_0 + 2\eta\mathscr{T} (1+\eta\gamma)^k \|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\| \\
\le& 2\eta\rho_\Phi \mathscr{S}\mathscr{T} (1+\eta\gamma)^k \eta r_0 \\
\le& 2\eta\rho_\Phi \mathscr{S}\mathscr{T} \|p(k+1)\|.
\end{aligned} \tag{B.42}$$

Here the second inequality is by $\|\Delta_{1,k}\| \le \rho_\phi \max\{\|x_k - \tilde{x}\|, \|x'_k - \tilde{x}\|\} \le \rho_\phi \mathscr{S}$ which uses (B.38). The third inequality is due to (B.41) and the fact that $I - \eta\mathcal{H} \succeq 0$ which is because $\eta = 1/L_\phi$ and $\lambda_{\max}(\mathcal{H}) \le L_\phi$. The fifth inequality applies (2.7), i.e., $\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\| \le \frac{\epsilon}{16\iota^2 2^\iota} \le \frac{1}{4}\rho_\phi \mathscr{S}\eta r_0$. By noting $2\eta\rho_\phi \mathscr{S}\mathscr{T} = 1/2$, we complete the proof of (B.40). Finally, (B.40) implies

$$\begin{aligned}
\max\{\|x_{\mathscr{T}} - \tilde{x}\|, \|x'_{\mathscr{T}} - \tilde{x}\|\} \ge& \frac{1}{2}\|\hat{x}_{\mathscr{T}}\| \ge \frac{1}{2}[\|p(\mathscr{T})\| - \|q(\mathscr{T})\|] \ge \frac{1}{4}\|p(\mathscr{T})\| \\
=& \frac{(1+\eta\gamma)^{\mathscr{T}}\eta r_0}{4} \ge \frac{(1+\eta\sqrt{\rho_\phi\epsilon})^{\mathscr{T}}\eta r_0}{4} \ge 2^{\iota-2}\eta r_0 > \mathscr{S},
\end{aligned}$$

where the second to last inequality uses the fact $(1+x)^{1/x} \ge 2$ for any $x \in (0, 1]$. This contradicts with (B.38), which finishes the proof of (B.35). We then characterize the probability, which follows the ideas in Jin et al. (2021). Recall $x_0 \sim \text{Uniform}(B_{\tilde{x}}(\eta r))$. We refer to $B_{\tilde{x}}(\eta r)$ the *perturbation ball*, and define the *stuck region* within the perturbation ball to be the set of points starting from which GD requires more than $\mathscr{T}$ steps to escape:

$$\mathcal{X} := \{x \in B_{\tilde{x}}(\eta r) \mid \{x_t\} \text{ is GD sequence with } x_0 = x, \text{and } \Phi(x_{\mathscr{T}}) - \Phi(x_0) > -\mathscr{F}\}.$$

Although the shape of the stuck region can be very complicated, we know that the width of $\mathcal{X}$ along the $\mathbf{e}_1$ direction is at most $\eta\omega$. That is, $\mathrm{Vol}(\mathcal{X}) \leq \mathrm{Vol}(\mathbb{B}_0^{d-1}(\eta r))\eta\omega$. Therefore,

$$\begin{aligned}\Pr(\mathbf{x}_0 \in \mathcal{X}) =& \frac{\mathrm{Vol}(\mathcal{X})}{\mathrm{Vol}(\mathbb{B}_{\tilde{\mathbf{x}}}^d(\eta r))} \leq \frac{\eta\omega \times \mathrm{Vol}(\mathbb{B}_0^{d-1}(\eta r))}{\mathrm{Vol}(\mathbb{B}_0^d(\eta r))} \\ =& \frac{\omega}{r\sqrt{\pi}} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{\omega}{r} \cdot \sqrt{\frac{d}{\pi}} \leq \frac{\ell\sqrt{d}}{\sqrt{\rho\epsilon}} \cdot \iota^2 2^{8-\iota}.\end{aligned}$$

On the event $\{x_0 \notin \mathcal{X}\}$, due to our parameter choice in (3.2), (3.3), (2.9) and (B.34), we have:

$$\Phi(x_{\mathscr{T}}) - \Phi(\tilde{x}) = [\Phi(x_{\mathscr{T}}) - \Phi(x_0)] + [\Phi(x_0) - \Phi(\tilde{x})] \leq -\mathscr{F} + \epsilon\eta r + \frac{L_\phi \eta^2 r^2}{2} \leq -\mathscr{F}/2,$$

where the first inequality uses the L_ϕ-smoothness of $\Phi(\cdot)$. This finishes the proof. □

B.4 Proof of Theorem 7

Proof. For the AID method, we characterize the iteration complexity for D and N so that (2.7) holds. By Lemma 35, we require D and N to satisfy

$$\begin{aligned}&\Gamma_1\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)\left(1-\frac{\mu}{\ell}\right)^{\frac{D}{2}} + 2\ell\sqrt{\kappa}\Gamma_1\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^N \\ \leq& \min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\} \cdot \epsilon.\end{aligned} \tag{B.43}$$

It is easy to verify that (B.43) holds when

$$\begin{aligned}D =& 2\log\left(\frac{2\Gamma_1\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)}{\min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\}\epsilon}\right) / \log\left(\frac{1}{1-\kappa^{-1}}\right) = \mathcal{O}\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right), \\ N =& \log\left(\frac{4\ell\sqrt{\kappa}\Gamma_1}{\min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\}\epsilon}\right) / \log\left(\frac{1+\sqrt{\kappa}}{1-\sqrt{\kappa}}\right) = \mathcal{O}\left(\sqrt{\kappa}\log\left(\frac{1}{\epsilon}\right)\right).\end{aligned} \tag{B.44}$$

Moreover, it is easy to verify that the right hand side of (B.43) is smaller than $\epsilon/5$. Therefore, by choosing D and N as in (B.44), we know that

$$\|\nabla\Phi(x_k) - \widehat{\nabla}\Phi(x_k)\| \leq \frac{\epsilon}{5}, \quad \forall k. \tag{B.45}$$

There are two possible cases to consider.

- **Case 1:** $\left\|\widehat{\nabla}\Phi(x_k)\right\| > \frac{4}{5}\epsilon$ and $k - k_{perturb} > \mathscr{T}$. In this case, combining (B.45) and Lemma 10 leads to

  $$\Phi(x_{k+1}) \leq \Phi(x_k) - \frac{4}{25L_\phi}\epsilon^2 + \frac{1}{25L_\phi}\epsilon^2 = \Phi(x_k) - \frac{3}{25L_\phi}\epsilon^2.$$

  Therefore, the total iteration number of **Case 1** can be bounded by

  $$\frac{25L_\phi(\Phi(x_0) - \Phi^*)}{3\epsilon^2}. \tag{B.46}$$

- **Case 2:** $k - k_{perturb} \leq \mathscr{T}$. This case means that we are within $\mathscr{T}$ iterations of the last perturbation step, i.e., the step 10 in Algorithm 1. Suppose the last perturbation step happened at the $\bar{k}$-th iteration. Therefore, from the step 10 in Algorithm 1 we know that $\|\widehat{\nabla}\Phi(x_{\bar{k}})\| \leq \frac{4}{5}\epsilon$. This together with (B.45) implies $\|\nabla\Phi(x_{\bar{k}})\| \leq \epsilon$. Now there are two cases to further consider. **Case 2(i).** If $\lambda_{\min}(\nabla^2\Phi(x_{\bar{k}})) \leq -\sqrt{\rho_\phi\epsilon}$, then according to Lemma 11 we know that with probability at least $1-\delta$ it holds that

$$\Phi(x_{\bar{k}+\mathscr{T}}) - \Phi(x_{\bar{k}}) \leq -\mathscr{F}/2.$$

  So the total iteration number in this case is bounded by

$$\frac{(\Phi(x_0) - \Phi^*)\mathscr{T}}{\mathscr{F}/2}. \tag{B.47}$$

  Case 2(ii). If $\lambda_{\min}(\nabla^2\Phi(x_{\bar{k}})) > -\sqrt{\rho_\phi\epsilon}$, then we have already found an ϵ-local minimum of $\Phi(x)$.

Therefore, combining (B.46) and (B.47) we know that the total iteration number before we visit an ϵ-local minimum can be bounded by

$$K = \frac{(\Phi(x_0) - \Phi^*)\mathscr{T}}{\mathscr{F}/2} + \frac{25L_\phi(\Phi(x_0) - \Phi^*)}{3\epsilon^2} = \tilde{\mathcal{O}}\left(\kappa^3\epsilon^{-2}\right).$$

This completes the proof. □

Appendix C. Proofs of Results in Section 3

C.1 Proof of Proposition 14

Proof. By Danskin's theorem, the gradient of $\Phi(x)$ is $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$. Therefore the Hessian of $\Phi(x)$ is given by

$$\nabla^2\Phi(x) = \nabla^2_{xx} f(x, y^*(x)) + \nabla^2_{xy} f(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}. \tag{C.1}$$

Note that the optimality condition for the max-player is $\nabla_y f(x, y^*(x)) = 0$, which leads to

$$\nabla^2_{yx} f(x, y^*(x)) + \nabla^2_{yy} f(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x} = 0. \tag{C.2}$$

Combining (C.1) and (C.2) yields

$$\nabla^2\Phi(x) = \nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xy} f(x, y^*(x))\nabla^2_{yy} f(x, y^*(x))^{-1}\nabla^2_{yx} f(x, y^*(x)). \tag{C.3}$$

Since $f(x, y)$ is μ-strongly concave with respect to y, the second term on the right hand side of (C.3), i.e., $-\nabla^2_{xy} f(x, y^*(x))\nabla^2_{yy} f(x, y^*(x))^{-1}\nabla^2_{yx} f(x, y^*(x))$, is always positive definite. Therefore, we have the following conclusions.

- A saddle point of $\Phi(x)$ satisfies $\lambda_{\min}(\nabla^2\Phi(x)) < 0$, which together with (C.3), implies $\lambda_{\min}(\nabla^2_{xx} f(x, y^*(x))) < 0$. Therefore, it cannot be a strict local Nash equilibrium.
- A strict local Nash equilibrium of $f(x, y)$ satisfies $\lambda_{\min}(\nabla^2_{xx} f(x, y^*(x))) > 0$, which yields $\lambda_{\min}(\nabla^2\Phi(x)) > 0$. So it must be a local minimum of $\Phi(x)$.

□

C.2 Proof of Proposition 16

Proof. A local minimum of $\Phi(x)$ satisfies

$$\nabla\Phi(x) = 0, \quad \nabla^2\Phi(x) \succ 0. \tag{C.4}$$

According to (C.3), the inequality in (C.4) is equivalent to

$$\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xy} f(x, y^*(x)) \nabla^2_{yy} f(x, y^*(x))^{-1} \nabla^2_{yx} f(x, y^*(x)) \succ 0. \tag{C.5}$$

Moreover, for nonconvex-strongly-concave problems, it holds that $\nabla^2_{yy} f(x, y) \prec 0$. Therefore, we only need to show that $\nabla\Phi(x) = 0$ is equivalent to $\nabla f(x, y) = 0$. Notice that for a pair (x, y) satisfying $\nabla_x f(x, y) = 0, \nabla_y f(x, y) = 0$, we have $y = y^*(x)$ from the strongly convexity and $\nabla_x f(x, y^*(x)) = \nabla\Phi(x) = 0$. Further more, when $\nabla\Phi(x) = 0$, we can always choose $y = y^*(x)$ so that $\nabla_x f(x, y) = 0, \nabla_y f(x, y) = 0$. Therefore, these two conditions are equivalent to each other. When function $\Phi(x)$ has a strict local minimum, the local minimax point is guaranteed to exist. □

C.3 Proof of Theorem 19

To prove Theorem 19, we need the following lemmas. The first lemma shows that under Assumption 17, the function $\Phi(x)$ is smooth and Hessian-Lipschitz continuous.

Lemma 36 *Chen et al. (2021b)[Proposition 1] Suppose $f(x, y)$ satisfies Assumption 17, we have*

- *$\Phi(x)$ is L_ϕ-smooth, where $L_\phi = \ell(1 + \kappa)$.*
- *$\Phi(x)$ is ρ_ϕ-Hessian Lipschitz continuous, i.e., (1.7) holds, where $\rho_\phi = \rho(1 + \kappa)^3$.*

The second lemma gives an upper bound for the gradient estimation error with the warm start strategy.

Lemma 37 *Suppose Assumption 17 holds. For the GDmax algorithm (i.e., the **GDmax** option in Algorithm 1) with parameters $D = \mathcal{O}(\kappa)$, we have,*

$$\left\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\right\| \leq \ell\left(\widehat{\Delta} + 2\eta\kappa\left(M + \frac{\ell M}{\mu}\right)\right)(1 - \kappa^{-1})^{\frac{D}{2}}, \tag{C.6}$$

where $\widehat{\Delta} = \|y^0 - y^*(x_0)\|$.

Proof. The gradient estimation error for minimax problem can be bounded by

$$\begin{aligned}\left\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\right\| &= \left\|\nabla_x f(x_k, y_k^D) - \nabla_x f(x_k, y^*(x_k))\right\| \leq \ell\|y_k^D - y^*(x_k)\| \\ &\leq \ell(1 - \kappa^{-1})^{\frac{D}{2}}\|y_k^0 - y^*(x_k)\|,\end{aligned} \tag{C.7}$$

where the last inequality follows (B.22). By the warm start strategy $y_k^0 = y_{k-1}^D$, we have

$$\begin{aligned}\|y_k^0 - y^*(x_k)\| &\leq \|y_{k-1}^D - y^*(x_{k-1})\| + \|y^*(x_{k-1}) - y^*(x_k)\| \\ &\leq (1 - \kappa^{-1})^{\frac{D}{2}}\|y_{k-1}^0 - y^*(x_{k-1})\| + \kappa\|x_k - x_{k-1}\| \\ &\leq (1 - \kappa^{-1})^{\frac{D}{2}}\|y_{k-1}^0 - y^*(x_{k-1})\| + \eta\kappa\|\widehat{\nabla}\Phi(x_{k-1})\|.\end{aligned} \tag{C.8}$$

By setting

$$D > 2\log 2/\log\left(\frac{1}{1-\kappa^{-1}}\right) = \mathcal{O}(\kappa), \tag{C.9}$$

we have

$$\begin{aligned}\|y_k^0 - y^*(x_k)\| \leq & \|y_{k-1}^D - y^*(x_{k-1})\| + \|y^*(x_{k-1}) - y^*(x_k)\| \\ \leq & \frac{1}{2}\|y_{k-1}^0 - y^*(x_{k-1})\| + \eta\kappa\|\widehat{\nabla}\Phi(x_{k-1})\| \\ \leq & \left(\frac{1}{2}\right)^k \|y^0 - y^*(x_0)\| + \eta\kappa\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\|\widehat{\nabla}\Phi(x_{k-1})\| \\ \leq & \widehat{\Delta} + 2\eta\kappa\left(M + \frac{\ell M}{\mu}\right),\end{aligned} \tag{C.10}$$

where the last inequality uses $\|\widehat{\nabla}\Phi(x_{k-1})\| \leq \left(M + \frac{\ell M}{\mu}\right)$ for any k. Combining (C.10) and (C.7) yields

$$\left\|\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\right\| \leq \ell\left(\widehat{\Delta} + 2\eta\kappa\left(M + \frac{\ell M}{\mu}\right)\right)(1-\kappa^{-1})^{\frac{D}{2}}, \tag{C.11}$$

which completes the proof. □

We now give the proof of Theorem 19.

Proof. By Lemma 37, we require D to satisfy

$$\ell\left(\widehat{\Delta} + 2\eta\kappa\left(M + \frac{\ell M}{\mu}\right)\right)(1-\kappa^{-1})^{\frac{D}{2}} \leq \min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\}\cdot\epsilon. \tag{C.12}$$

It is easy to verify that

$$D = 2\log\left(\frac{\ell\left(\widehat{\Delta} + 2\eta\kappa\left(M + \frac{\ell M}{\mu}\right)\right)}{\min\left\{\frac{\sqrt{17}}{80\iota^2}, \frac{1}{16\iota^2 2^\iota}\right\}\epsilon}\right)/\log\left(\frac{1}{1-\kappa^{-1}}\right) = \mathcal{O}\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right), \tag{C.13}$$

satisfies (C.12). The rest of the proof is the same as the proof of Theorem 7 in Section B.4. □

Appendix D. Proofs of Results in Section 4

D.1 Proof of Lemma 26

Proof. Define the following function:

$$\hat{\Phi}_x(u) = \Phi(x+u) - \Phi(x) - \nabla\Phi(x)^\top u. \tag{D.1}$$

We first characterize the required estimation error, which is used in the later proof. Specifically, we choose the inner iteration number D (step 9 of Algorithm 2) and N (used in steps

6 and 12 of Algorithm 2) such that for any $k \leq \mathscr{T}$, the following inequalities hold:

$$\begin{aligned} &\|\widehat{\nabla}\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x}+u_k)\| \leq \min\left\{\frac{\sqrt{17}}{40\iota^2}, \frac{1}{16\iota^2 2^{\iota/4}}, \frac{9^{1/3}-2}{8\iota}, \frac{1}{750\iota^2}\right\}\epsilon, \\ &M\left\|y_k^D(\tilde{x}+u_k) - y^*(\tilde{x}+u_k))\right\| \leq \frac{1}{750\iota^3 L_\phi}\epsilon^2. \end{aligned} \tag{D.2}$$

Note that both inequalities in (D.2) also imply the following inequality, which corresponds to the case $u_k = 0$:

$$\begin{aligned} &\|\widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(\tilde{x})\| \leq \min\left\{\frac{\sqrt{17}}{40\iota^2}, \frac{1}{16\iota^2 2^{\iota/4}}, \frac{9^{1/3}-2}{8\iota}, \frac{1}{750\iota^2}\right\}\epsilon, \\ &M\left\|y_k^D(\tilde{x}) - y^*(\tilde{x}))\right\| \leq \frac{1}{750\iota^3 L_\phi}\epsilon^2. \end{aligned} \tag{D.3}$$

Since the lower level problem is strongly convex, combining with Lemma 35, we can set D and N in Algorithm 2 as

$$\begin{aligned} D =& \max\left\{2\log\left(\frac{2\left(1+\frac{\ell}{\mu}\left(1+2\sqrt{\kappa}\right)\right)\left(\ell+\frac{\rho M}{\mu}\right)\Gamma_1}{\min\left\{\frac{\sqrt{17}}{40\iota^2}, \frac{1}{16\iota^2 2^{\iota/4}}, \frac{9^{1/3}-2}{8\iota}, \frac{1}{750\iota^2}\right\}\epsilon}\right), 2\log\left(\frac{750 M L_\phi \iota^3 \Gamma_1}{\epsilon^2}\right)\right\} / \log\left(\frac{1}{1-\kappa^{-1}}\right) \\ =& \mathcal{O}\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right), \\ N =& \log\left(\frac{4\ell\sqrt{\kappa}\Gamma_1}{\min\left\{\frac{\sqrt{17}}{40\iota^2}, \frac{1}{16\iota^2 2^{\iota/4}}, \frac{9^{1/3}-2}{8\iota}, \frac{1}{750\iota^2}\right\}\epsilon}\right) / \log\left(\frac{1+\sqrt{\kappa}}{1-\sqrt{\kappa}}\right) = \mathcal{O}\left(\sqrt{\kappa}\log\left(\frac{1}{\epsilon}\right)\right), \end{aligned} \tag{D.4}$$

such that (D.2) holds in each iteration. Next, we show that with probability at least $1 - \frac{L_\phi\sqrt{d}}{\sqrt{\rho_\phi\epsilon}} \cdot \iota^2 2^{8-\iota/4}$, the following inequality holds:

$$\hat{\Phi}_{\tilde{x}}(u_{\mathscr{T}}) - \hat{\Phi}_{\tilde{x}}(u_0) \leq -\mathscr{F}, \tag{D.5}$$

where $\hat{\Phi}_x(u)$ is defined in (D.1). Note that it is easy to verify that $\hat{\Phi}_x(u)$ is L_ϕ-smooth, and it yields

$$\begin{aligned}
\hat{\Phi}_{\tilde{x}}(u_{k+1}) \leq & \hat{\Phi}_{\tilde{x}}(u_k) + \langle \nabla \hat{\Phi}_{\tilde{x}}(u_k), u_{k+1} - u_k \rangle + \frac{L_\phi}{2} \|u_{k+1} - u_k\|_2^2 \\
= & \hat{\Phi}_{\tilde{x}}(u_k) + \langle \nabla \Phi(\tilde{x} + u_k) - \nabla \Phi(\tilde{x}), u_{k+1} - u_k \rangle + \frac{L_\phi}{2} \|u_{k+1} - u_k\|_2^2 \\
= & \hat{\Phi}_{\tilde{x}}(u_k) + \langle \widehat{\nabla} \Phi(\tilde{x} + u_k) - \widehat{\nabla} \Phi(\tilde{x}), u_{k+1} - u_k \rangle + \langle \nabla \Phi(\tilde{x} + u_k) - \widehat{\nabla} \Phi(\tilde{x} + u_k), u_{k+1} - u_k \rangle \\
& + \langle \widehat{\nabla} \Phi(\tilde{x}) - \nabla \Phi(\tilde{x}), u_{k+1} - u_k \rangle + \frac{L_\phi}{2} \|u_{k+1} - u_k\|_2^2 \\
\leq & \hat{\Phi}_{\tilde{x}}(u_k) - \frac{1}{\eta} \|u_{k+1} - u_k\|^2 + \frac{1}{8\eta} \|u_{k+1} - u_k\|^2 + 2\eta \|\nabla \Phi(\tilde{x} + u_k) - \widehat{\nabla} \Phi(\tilde{x} + u_k)\|^2 \\
& + \frac{1}{8\eta} \|u_{k+1} - u_k\|^2 + 2\eta \|\widehat{\nabla} \Phi(\tilde{x}) - \nabla \Phi(\tilde{x})\|^2 + \frac{1}{2\eta} \|u_{k+1} - u_k\|_2^2 \\
= & \hat{\Phi}_{\tilde{x}}(u_k) - \frac{1}{4\eta} \|u_{k+1} - u_k\|^2 + 2\eta \|\nabla \Phi(\tilde{x} + u_k) - \widehat{\nabla} \Phi(\tilde{x} + u_k)\|^2 \\
& + 2\eta \|\widehat{\nabla} \Phi(\tilde{x}) - \nabla \Phi(\tilde{x})\|^2,
\end{aligned} \tag{D.6}$$

where the first equality is from (D.1), the second inequality is from (4.1) and Young's inequality. We then follow the same ideas as in the proof of Lemma 11. We design two coupling sequences $\{u_t\}$, $\{w_t\}$ generated by iNEON (Algorithm 2) with initial points u_0 and w_0, respectively. We require the two sequences to satisfy: **Condition (i).** $\max\{\|u_0\|, \|w_0\|\} \leq \eta r$; and **Condition (ii).** $u_0 - w_0 = \eta r_0 \mathbf{e}_1$, where $\mathbf{e}_1$ is the minimum eigenvector of $\nabla^2 \Phi(\tilde{x})$ with $\|\mathbf{e}_1\| = 1$ and $r_0 > \omega := 2^{2-\iota/4} L_\phi \mathscr{S}$. The rest is to prove

$$\min\{\hat{\Phi}_x(u_{\mathscr{T}}) - \hat{\Phi}_x(u_0), \hat{\Phi}_x(w_{\mathscr{T}}) - \hat{\Phi}_x(w_0)\} \leq -\mathscr{F}. \tag{D.7}$$

We prove (D.7) by contradiction. Assume the contrary holds:

$$\min\{\hat{\Phi}_{\tilde{x}}(u_{\mathscr{T}}) - \hat{\Phi}_{\tilde{x}}(u_0), \hat{\Phi}_{\tilde{x}}(w_{\mathscr{T}}) - \hat{\Phi}_{\tilde{x}}(w_0)\} > -\mathscr{F}. \tag{D.8}$$

First, by the update of u_k (i.e., (4.1) or step 13 of Algorithm 2), we have for any $\tau \leq k$:

$$\begin{aligned}
& \|u_\tau - u_0\| \\
& \leq \sum_{t=1}^{k} \|u_t - u_{t-1}\| \leq \left[k \sum_{t=1}^{k} \|u_t - u_{t-1}\|^2 \right]^{\frac{1}{2}} \leq \left[\mathscr{T} \sum_{t=1}^{\mathscr{T}} \|u_t - u_{t-1}\|^2 \right]^{\frac{1}{2}} \\
& \leq \sqrt{4\eta \mathscr{T} \left(\hat{\Phi}_{\tilde{x}}(u_0) - \hat{\Phi}_{\tilde{x}}(u_{\mathscr{T}}) + 2\eta \left(\sum_{t=1}^{\mathscr{T}} \|\nabla \Phi(\tilde{x} + u_t) - \widehat{\nabla} \Phi(\tilde{x} + u_t)\|^2 + \|\widehat{\nabla} \Phi(\tilde{x}) - \nabla \Phi(\tilde{x})\|^2 \right) \right)} \\
& \leq \sqrt{4\eta \mathscr{T} \mathscr{F} + 8\eta^2 \mathscr{T}^2 \cdot \frac{17}{800\iota^4} \cdot \epsilon^2} + \eta r \leq \mathscr{S},
\end{aligned} \tag{D.9}$$

where the fourth inequality is obtained by (D.6), the fifth inequality follows from (D.8), (D.2) and (D.3), and the last inequality is due to the parameter choice in (4.3), and $\iota \geq 1$, $L_\phi / \sqrt{\rho_\phi \epsilon} \geq 1$. Therefore, from (D.9) we have for any $k \leq \mathscr{T}$:

$$\max\{\|u_k\|, \|w_k\|\} \leq \max\{\|u_k - u_0\|, \|w_k - w_0\|\} + \max\{\|u_0\|, \|w_0\|\} \leq \mathscr{S}. \tag{D.10}$$

On the other hand, we can write the update equation for the difference $v_k := u_k - w_k$ as:

$$\begin{aligned} v_{k+1} =& v_k - \eta \left[\widehat{\nabla}\Phi(\tilde{x}+u_k) - \widehat{\nabla}\Phi(\tilde{x}+w_k)\right] \\ =& v_k - \eta\left[\nabla\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x}+w_k)\right] \\ & - \eta\left[\widehat{\nabla}\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x}+u_k) + \nabla\Phi(\tilde{x}+w_k) - \widehat{\nabla}\Phi(\tilde{x}+w_k)\right] \\ =& \underbrace{(I-\eta\mathcal{H})^{k+1}v_0}_{p(k+1)} - \underbrace{\eta\sum_{t=0}^{k}(I-\eta\mathcal{H})^{k-t}\left(\Delta_{1,t}v_t + \Delta_{2,t}\right)}_{q(k+1)}, \end{aligned} \tag{D.11}$$

where we denote

$$\mathcal{H} = \nabla^2\Phi(\tilde{x}), \tag{D.12}$$

$$\Delta_{1,k} = \int_0^1 \left[\nabla^2\Phi(\tilde{x}+w_k+\theta(u_k-w_k)) - \mathcal{H}\right] d\theta, \tag{D.13}$$

$$\Delta_{2,k} = \left[\widehat{\nabla}\Phi(\tilde{x}+u_k) - \nabla\Phi(\tilde{x}+u_k) + \nabla\Phi(\tilde{x}+w_k) - \widehat{\nabla}\Phi(\tilde{x}+w_k)\right]. \tag{D.14}$$

We then prove $\|q(k)\| \le \|p(k)\|/2, \forall k \in [\mathscr{T}]$. We prove it by induction. It is easy to check that it holds at $k=0$. Assume it holds for any $t \le k$. Denote $\lambda_{\min}(\nabla^2\Phi(\tilde{x})) = -\gamma$. Since v_0 lies in the direction of the minimum eigenvector of $\nabla^2\Phi(\tilde{x}_0)$, we have for any $t \le k$:

$$\|v_t\| \le \|p(t)\| + \|q(t)\| \le \frac{3}{2}\|p(t)\| \le \frac{3}{2}(1+\eta\gamma)^t\eta r_0. \tag{D.15}$$

At step $k+1$, similar to (B.42), we have

$$\|q(k+1)\| \le 2\eta\rho_\Phi \mathscr{S}\mathscr{T}\|p(k+1)\|, \tag{D.16}$$

where we used (D.2). Combining (D.16) with the choice of parameters in (4.3) finishes the proof of $\|q(k)\| \le \|p(k)\|/2$. Therefore, we have

$$\begin{aligned} \max\{\|u_{\mathscr{T}}\|, \|w_{\mathscr{T}}\|\} \ge& \frac{1}{2}\|v(\mathscr{T})\| \ge \frac{1}{2}[\|p(\mathscr{T})\| - \|q(\mathscr{T})\|] \ge \frac{1}{4}[\|p(\mathscr{T})\| \\ =& \frac{(1+\eta\gamma)^{\mathscr{T}}\eta r_0}{4} \ge 2^{\iota/4-2}\eta r_0 > \mathscr{S}, \end{aligned}$$

where the second to last inequality uses the fact $(1+x)^{1/x} \ge 2$ for any $x \in (0,1]$. This contradicts with (D.10), which finishes the proof of (D.7). To characterize the probability, we define the *stuck region*:

$$\mathcal{X} := \{u \in \mathbb{B}_0(\eta r) \mid \{u_t\} \text{ is the iNEON sequence with } u_0 = u, \text{and } \hat{\Phi}_{\tilde{x}}(u_{\mathscr{T}}) - \hat{\Phi}_{\tilde{x}}(u_0) > -\mathscr{F}\}.$$

Although the shape of the stuck region can be very complicated, we know that the width of $\mathcal{X}$ along the $\mathbf{e}_1$ direction is at most $\eta\omega$. That is, $\mathrm{Vol}(\mathcal{X}) \le \mathrm{Vol}(\mathbb{B}_0^{d-1}(\eta r))\eta\omega$. Therefore:

$$\Pr(u_0 \in \mathcal{X}) \le \frac{\eta\omega \times \mathrm{Vol}(\mathbb{B}_0^{d-1}(\eta r))}{\mathrm{Vol}(\mathbb{B}_0^d(\eta r))} = \frac{\omega}{r\sqrt{\pi}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \le \frac{\omega}{r}\cdot\sqrt{\frac{d}{\pi}} \le \frac{\ell\sqrt{d}}{\sqrt{\rho\epsilon}}\cdot\iota^2 2^{8-\iota/4}.$$

On the event $\{u_0 \notin \mathcal{X}\}$, due to the choice of the parameters in (4.3), we have with probability at least $1-\delta$, where $\delta > \frac{\ell\sqrt{d}}{\sqrt{\rho\epsilon}} \cdot \iota^2 2^{8-\iota/4}$, that

$$\hat{\Phi}_{\tilde{x}}(u_{\mathscr{T}}) - \hat{\Phi}_{\tilde{x}}(u_0) < -\mathscr{F}.$$

Therefore, there exists some $k' \leq \mathscr{T}$ such that $\hat{\Phi}_{\tilde{x}}(u_{k'}) - \hat{\Phi}_{\tilde{x}}(u_0) < -\mathscr{F}$ and $\|u_\tau\| \leq \mathscr{S}, \forall \tau < k'$. In other words, k' is the first iteration that satisfies

$$\hat{\Phi}_{\tilde{x}}(u_{k'}) - \hat{\Phi}_{\tilde{x}}(u_0) < -\mathscr{F}. \tag{D.17}$$

By the update $u_{k'} = u_{k'-1} - \eta(\widehat{\nabla}\Phi(\tilde{x}+u_{k'-1}) - \widehat{\nabla}\Phi(\tilde{x}))$, we can bound the norm of $u_{k'}$:

$$\begin{aligned}
\|u_{k'}\| \leq& \|u_{k'-1}\| + \eta\|\nabla\Phi(\tilde{x}+u_{k'-1}) - \nabla\Phi(\tilde{x})\| + \eta\|\widehat{\nabla}\Phi(\tilde{x}+u_{k'-1}) - \nabla\Phi(\tilde{x}+u_{k'-1})\| \\
&+ \eta\|\widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(\tilde{x})\| \\
\leq& \|u_{k'-1}\| + \eta L_\phi\|u_{k'-1}\| + 2\eta\|\widehat{\nabla}\Phi(\tilde{x}+u_{k'-1}) - \nabla\Phi(\tilde{x}+u_{k'-1})\| \\
=& 2\mathscr{S} + 2\eta\|\widehat{\nabla}\Phi(\tilde{x}+u_{k'-1}) - \nabla\Phi(\tilde{x}+u_{k'-1})\| \\
\leq& 9^{1/3}\mathscr{S},
\end{aligned} \tag{D.18}$$

where the second inequality is by the smoothness of $\Phi(x)$ given in Lemma 36, the equality is due to the parameter choice in (4.3), and the last inequality is obtained by (D.2). Moreover, by (D.17) and the smoothness of $\Phi(x)$, we have

$$\begin{aligned}
\Phi(\tilde{x}+u_{k'}) - \Phi(\tilde{x}) - \nabla\Phi(\tilde{x})^\top u_{k'} \leq& \Phi(\tilde{x}+u_0) - \Phi(\tilde{x}) - \nabla\Phi(\tilde{x})^\top u_0 - \mathscr{F} \leq \frac{L_\phi}{2}\|u_0\|^2 - \mathscr{F} \\
\leq& \frac{L_\phi}{2} \cdot \frac{\epsilon^2}{160000\iota^6 L_\phi^2} - \frac{1}{25\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} \leq \left(\frac{1}{320000} - \frac{1}{25}\right)\frac{1}{\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}},
\end{aligned} \tag{D.19}$$

where the third inequality is due to the parameter choice in (4.3), and the last inequality applies $\iota > 1$ and $L_\phi/\sqrt{\rho_\phi\epsilon} > 1$. From (D.2) we can get

$$\begin{aligned}
&\left\|\hat{\Phi}(\tilde{x}+u_{k'}) - \hat{\Phi}(\tilde{x}) - \widehat{\nabla}\Phi(\tilde{x})^\top u_{k'} - \left(\Phi(\tilde{x}+u_{k'}) - \Phi(\tilde{x}) - \nabla\Phi(\tilde{x})^\top u_{k'}\right)\right\| \\
\leq& \left\|\hat{\Phi}(\tilde{x}+u_{k'}) - \Phi(\tilde{x}+u_{k'})\right\| + \left\|\hat{\Phi}(\tilde{x}) - \Phi(\tilde{x})\right\| + \left\|\widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(x)\right\|\|u_{k'}\| \\
\leq& \left\|f(\tilde{x}+u_{k'}, y_k^D(\tilde{x}+u_{k'})) - f(\tilde{x}+u_{k'}, y^*(\tilde{x}+u_{k'}))\right\| + \left\|f(\tilde{x}, y_k^D(\tilde{x})) - f(\tilde{x}, y^*(\tilde{x}))\right\| \\
&+ \left\|\widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(x)\right\|\|u_{k'}\| \\
\leq& M\left\|y_k^D(\tilde{x}+u_{k'}) - y^*(\tilde{x}+u_{k'}))\right\| + M\left\|y_k^D(\tilde{x}) - y^*(\tilde{x})\right\| + 3\mathscr{S}\left\|\widehat{\nabla}\Phi(\tilde{x}) - \nabla\Phi(x)\right\| \\
\leq& \frac{2}{750\iota^3 L_\phi}\epsilon^2 + \frac{3}{4\iota}\sqrt{\frac{\epsilon}{\rho_\phi}}\frac{1}{750\iota^2}\epsilon \\
\leq& \frac{2}{750\iota^3 L_\phi}\epsilon^2 \cdot \frac{L_\phi}{\sqrt{\rho_\phi\epsilon}} + \frac{1}{750\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} \\
=& \frac{1}{250\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}},
\end{aligned} \tag{D.20}$$

where the third inequality is due to Assumption 3 and (D.18), the fourth inequality uses (D.2), (D.3) and the parameter choice in (4.3), and the fifth inequality is due to $L_\phi > \sqrt{\rho_\phi \epsilon}$. Combining (D.19) and (D.20) we get

$$\hat{\Phi}(\tilde{x}+u_{k'}) - \hat{\Phi}(\tilde{x}) - \widehat{\nabla}\Phi(\tilde{x})^\top u_{k'} \le \left(\frac{1}{320000} - \frac{1}{25} + \frac{1}{250}\right)\frac{1}{\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} = -\frac{11519}{12800}\mathscr{F}, \quad \text{(D.21)}$$

i.e., the stopping criterion in step 14 of Algorithm 2 is satisfied. Therefore, this shows that with high probability, the Algorithm 2 terminates with the stopping criterion in step 14 being satisfied. When this happens, we have

$$\begin{aligned}
&\Phi(x+u_{k'}) - \Phi(x) - \nabla\Phi(x)^\top u_{k'} \\
\le&\hat{\Phi}(\tilde{x}+u_{k'}) - \hat{\Phi}(\tilde{x}) - \hat{\nabla}\Phi(\tilde{x})^\top u_{k'} \\
&+ \left\|\hat{\Phi}(\tilde{x}+u_{k'}) - \hat{\Phi}(\tilde{x}) - \hat{\nabla}\Phi(\tilde{x})^\top u_{k'} - \left(\Phi(x+u_{k'}) - \Phi(x) - \nabla\Phi(x)^\top u_{k'}\right)\right\| \\
\le&\left(\frac{1}{320000} - \frac{1}{25} + \frac{1}{250} + \frac{1}{250}\right)\frac{1}{\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}},
\end{aligned} \quad \text{(D.22)}$$

where we used (D.20). By the Hessian Lipschitz continuity of $\Phi(x)$, we have

$$\begin{aligned}
\frac{1}{2}u_{k'}\nabla^2\Phi(x)u_{k'} \le&\Phi(x+u_{k'}) - \Phi(x) - \nabla\Phi(x)^\top u_{k'} + \frac{\rho_\phi}{6}\|u_{k'}\|^3 \\
&\le \left(\frac{1}{320000} - \frac{1}{25} + \frac{1}{250} + \frac{1}{250}\right)\frac{1}{\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} + \frac{\rho_\phi}{6}\cdot 9\mathscr{S}^3 \\
&\le \left(\frac{1}{320000} - \frac{1}{25} + \frac{1}{250} + \frac{1}{250}\right)\frac{1}{\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} + \frac{3}{128\iota^3}\sqrt{\frac{\epsilon^3}{\rho_\phi}} \\
&\le -\frac{1}{5}\mathscr{F},
\end{aligned} \quad \text{(D.23)}$$

where the second inequality follows from (D.22) and (D.18). Finally, by (D.18) we have

$$\frac{u_{k'}\nabla^2\Phi(x)u_{k'}}{\|u_{k'}\|^2} \le -\frac{2/5\mathscr{F}}{9\mathscr{S}^2} = -\frac{32}{1125\iota}\sqrt{\rho_\phi\epsilon} \le -\frac{1}{40\iota}\sqrt{\rho_\phi\epsilon}. \quad \text{(D.24)}$$

If NEON returns 0, by Bayes theorem, we have $\lambda_{\min}(\nabla^2\Phi(x)) \ge -\sqrt{\rho_\phi\epsilon}$ with high probability $1 - O(\delta)$ for a sufficiently small δ. □

D.2 Proof of Theorem 30

We first provide several useful lemmas.

Lemma 38 *Xu et al. (2018)[Lemma 4] Suppose Assumption 28 holds. Let* $\nabla^2 F_{\mathcal{D}}(x) = \frac{1}{n}\sum_{i=1}^{D_f}\nabla^2 F(x;\xi_i)$. *For any* $\epsilon, \delta \in (0,1)$, $x \in \mathbb{R}^d$ *when* $D_f \ge \frac{16L^2\log(2d/\delta)}{\epsilon^2}$, *we have with probability at least* $1-\delta$*:*

$$\|\nabla^2 F_{\mathcal{D}}(x) - \nabla^2 F(x)\| \le \epsilon. \quad \text{(D.25)}$$

The next two lemmas mimic Lemma 38 for first-order and third-order derivatives. Their proofs are mostly identical to that of Lemma 38, and hence we omit the details for brevity.

Lemma 39 *Suppose Assumption 28 holds. Let* $\nabla F_{\mathcal{D}}(x) = \frac{1}{n}\sum_{i=1}^{D_f} F(x;\xi_i)$. *For any* $\epsilon, \delta \in (0,1)$, $x \in \mathbb{R}^d$ *when* $D_f \geq \frac{16M^2\log(2d/\delta)}{\epsilon^2}$, *we have with probability at least* $1-\delta$*:*

$$\|\nabla F_{\mathcal{D}}(x) - \nabla F(x)\| \leq \epsilon. \tag{D.26}$$

Lemma 40 *Suppose Assumption 28 holds. Let* $\nabla^3 G_{\mathcal{D}}(x) = \frac{1}{n}\sum_{i=1}^{D_g} \nabla^3 G(x;\zeta_i)$. *For any* $\epsilon, \delta \in (0,1)$, $x \in \mathbb{R}^d$ *when* $D_g \geq \frac{16\rho^2\log(2d/\delta)}{\epsilon^2}$, *we have with probability at least* $1-\delta$*:*

$$\|\nabla^3 G_{\mathcal{D}}(x) - \nabla^3 G(x)\| \leq \epsilon. \tag{D.27}$$

The following two lemmas show that with sample batch sizes $D_f, D_g = \max\{\mathcal{O}(\frac{1}{\rho_\phi \epsilon}), \mathcal{O}(\frac{1}{\rho_\phi \epsilon})\}$, we can bound the batch gradient and Hessian errors with high probability.

Lemma 41 *Suppose Assumptions 28 and 29 hold. Set batch sizes* $D_g = \tilde{\mathcal{O}}(\frac{\kappa^{10}}{\rho_\phi \epsilon} + \frac{\kappa^6}{\epsilon^2})$, $D_f = \tilde{\mathcal{O}}(\frac{\kappa^6}{\rho_\phi \epsilon} + \frac{\kappa^2}{\epsilon^2})$, *with probability at least* $1-\delta$, *we have the following inequality holds:*

$$\|\nabla^2 \Phi_{\mathcal{D}}(x) - \nabla^2 \Phi(x)\| \leq \frac{1}{80\iota}\sqrt{\rho_\Phi \epsilon}. \tag{D.28}$$

Proof. Note that the Hessian $\nabla^2\Phi(x)$ is given in (B.1). The Hessian estimation error between $\nabla^2\Phi(x)$ and $\nabla^2\Phi_{\mathcal{D}}(x)$ can be computed as

$$\begin{aligned}
&\|\nabla^2\Phi(x) - \nabla^2\Phi_{\mathcal{D}}(x)\| \\
\leq\ & \underbrace{\left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\|}_{(I)} \\
&+ \underbrace{2\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yx} f(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\cdot\nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\|}_{(II)} \\
&+ \underbrace{\left\|\frac{\partial^2 y^*(x)}{\partial^2 x}\cdot\nabla_y f(x, y^*(x)) - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x}\cdot\nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\|}_{(III)} \\
&+ \underbrace{\left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yy} f(x, y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\cdot\nabla^2_{yy} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\cdot\frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}^\top\right\|}_{(IV)},
\end{aligned} \tag{D.29}$$

where the inequality is obtained by triangle inequality and $\nabla^2_{yx} f(x,y) = \nabla^2_{xy} f(x,y)^\top$ for any smooth functions. Similar to the proof of Lemma 6, we bound the terms $(I) - (IV)$.

The first term can be bounded as follows:

$$\begin{aligned}(I) &= \left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\ &\le \left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*(x))\right\| + \left\|\nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\ &\le \left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*(x))\right\| + \rho \left\|y^*(x) - y^*_{\mathcal{D}_G}(x)\right\|,\end{aligned} \tag{D.30}$$

where we used the Assumption 3 in the last step. Secondly, we have

$$\begin{aligned}(II) =& 2\left\|\frac{\partial y^*(x)}{\partial x} \cdot \nabla^2_{yx} f(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \cdot \nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\ \le& 2\left\|\frac{\partial y^*(x)}{\partial x}\right\| \left\|\nabla^2_{yx} f(x, y^*(x)) - \nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\ &+ 2\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\right\| \left\|\nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\ \le& \frac{2\ell}{\mu}\left(\left\|\nabla^2_{yx} f(x, y^*(x)) - \nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*(x))\right\| + \rho \left\|y^*(x) - y^*_{\mathcal{D}_G}(x)\right\|\right) \\ &+ 2\ell \left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*(x')}{\partial x'}\right\|,\end{aligned} \tag{D.31}$$

where the first inequality is due to the triangle inequality and the Cauchy-Schwarz inequality and the last step uses Assumption 3 as well as Proposition 33. Furthermore, we bound $\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\right\|$ as

$$\begin{aligned}&\left\|\frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\right\| \\ =& \left\|\nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g(x, y^*(x))^{-1} - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \cdot \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))^{-1}\right\| \\ \le& \left\|\nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g(x, y^*(x))^{-1} - \nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))^{-1}\right\| \\ &+ \left\|\nabla^2_{xy} g(x, y^*(x)) \cdot \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))^{-1} - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \cdot \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))^{-1}\right\| \\ \le& \frac{\ell}{\mu^2}\left(\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \rho\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|\right) \\ &+ \frac{1}{\mu}\left(\left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \rho\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|\right),\end{aligned} \tag{D.32}$$

where the first equality is by (1.4) and the last step follows the fact that $\|X^{-1} - Y^{-1}\| \le \|X^{-1}\|\|X - Y\|\|Y^{-1}\|$, Assumption 3 and Proposition 33. Combining (D.31) and (D.32), we get

$$\begin{aligned}(II) \le& \frac{2\ell}{\mu}\left\|\nabla^2_{yx} f(x, y^*(x)) - \nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*(x))\right\| + \frac{2\ell^2}{\mu^2}\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\| \\ &+ \frac{2\ell}{\mu}\left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \left(\frac{4\ell}{\mu} + \frac{2\ell^2}{\mu^2}\right)\rho\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|.\end{aligned} \tag{D.33}$$

The third term (III) can be written as

$$\begin{aligned}(III) =& \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \cdot \nabla_y f(x, y^*(x)) - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \cdot \nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\ \leq& \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} \right\| \left\| \nabla_y f(x, y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\ &+ \left\| \nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x)) \right\| \cdot \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \right\| \\ \leq& \frac{\rho}{\mu} \left(1 + \frac{\ell}{\mu}\right)^2 \left\| \nabla_y f(x, y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\ &+ M \left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \right\|,\end{aligned} \tag{D.34}$$

where the last inequality follows (B.10) and Proposition 33. To bound $\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \right\|$, we follow the computation of the second to last inequality in (B.11) and get

$$\begin{aligned}&\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \right\| \\ \leq& \left[\frac{1}{\mu} \| \nabla^3_{xxy} g(x, y^*(x)) - \nabla^3_{xxy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \| \right. \\ &+ \frac{2}{\mu} \left\| \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\ &\left. + \frac{1}{\mu} \left\| \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \cdot \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \cdot \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}^\top \right\| \right] \\ &+ \rho \left(1 + \frac{\ell}{\mu}\right)^2 \left\| \nabla^2_{yy} g(x, y^*(x))^{-1} - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))^{-1} \right\|.\end{aligned} \tag{D.35}$$

To bound the term $\left\| \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \right\|$, we follow the computation of (II) and get

$$\begin{aligned}&\left\| \frac{\partial y^*(x)}{\partial x} \nabla^3_{yxy} g(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\ \leq& \frac{2\ell}{\mu} \left\| \nabla^3_{yxy} g(x, y^*(x)) - \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| + \frac{2\ell\rho}{\mu^2} \left\| \nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\ &+ \frac{2\rho}{\mu} \left\| \nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| + \left(\frac{2\ell\nu}{\mu} + \frac{2\ell\rho^2}{\mu^2} + \frac{2\rho^2}{\mu} \right) \| y^*(x) - y^*_{\mathcal{D}_G}(x) \|.\end{aligned} \tag{D.36}$$

For the third term, we follow similar computation in (B.13) and give the following bound

$$\begin{aligned}
&\left\| \frac{\partial y^*(x)}{\partial x} \cdot \nabla^3_{yyy} g(x, y^*(x)) \cdot \frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \cdot \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \cdot \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}^\top \right\| \\
\leq & 2 \cdot \frac{\rho \ell}{\mu} \left\| \frac{\partial y^*(x)}{\partial x} - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x} \right\| + \left(\frac{\ell}{\mu} \right)^2 \left\| \nabla^3_{yyy} g(x, y^*(x)) - \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x)) \right\| \\
\leq & \frac{2\rho\ell^2}{\mu^3} \left\| \nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{2\rho\ell}{\mu^2} \left\| \nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{\ell^2}{\mu^2} \left\| \nabla^3_{yyy} g(x, y^*(x)) - \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \left(\frac{2\rho^2\ell^2}{\mu^3} + \frac{2\rho^2\ell}{\mu^2} + \frac{\ell^2\nu}{\mu^2} \right) \| y^*(x) - y^*_{\mathcal{D}_G}(x) \|,
\end{aligned} \tag{D.37}$$

where the last inequality is by (D.32), Assumption 3 and Proposition 33. Combining (D.35) - (D.37) together yields

$$\begin{aligned}
&\left\| \frac{\partial^2 y^*(x)}{\partial^2 x} - \frac{\partial^2 y^*_{\mathcal{D}_G}(x)}{\partial^2 x} \right\| \\
\leq & \frac{1}{\mu} \| \nabla^3_{xxy} g(x, y^*(x)) - \nabla^3_{xxy} g_{\mathcal{D}_G}(x, y^*(x)) \| + \frac{4\ell}{\mu^2} \left\| \nabla^3_{yxy} g(x, y^*(x)) - \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{4\ell\rho}{\mu^3} \left\| \nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| + \frac{4\rho}{\mu^2} \left\| \nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{2\rho\ell^2}{\mu^4} \left\| \nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| + \frac{2\rho\ell}{\mu^3} \left\| \nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{\ell^2}{\mu^3} \left\| \nabla^3_{yyy} g(x, y^*(x)) - \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \frac{\rho}{\mu^2} \left(1 + \frac{\ell}{\mu} \right)^2 \left\| \nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x)) \right\| \\
& + \left(\frac{\rho^2}{\mu^2} \left(1 + \frac{\ell}{\mu} \right)^2 + \frac{\nu}{\mu} + \frac{4\ell\nu + 4\rho^2}{\mu^2} + \frac{6\ell\rho^2 + \ell^2\nu}{\mu^3} + \frac{2\rho^2\ell^2}{\mu^4} \right) \| y^*(x) - y^*_{\mathcal{D}_G}(x) \|,
\end{aligned} \tag{D.38}$$

where the last inequality uses the fact that $\|X^{-1} - Y^{-1}\| \leq \|X^{-1}\|\|X - Y\|\|Y^{-1}\|$. Plug (D.38) into (D.34) leads to

$$\begin{aligned}
(III) \leq & \frac{\rho}{\mu}\left(1+\frac{\ell}{\mu}\right)^2 \|\nabla_y f(x,y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x,y^*(x))\| \\
& + \frac{M}{\mu}\|\nabla^3_{xxy} g(x,y^*(x)) - \nabla^3_{xxy} g_{\mathcal{D}_G}(x,y^*(x))\| \\
& + \frac{4\ell M}{\mu^2}\left\|\nabla^3_{yxy} g(x,y^*(x)) - \nabla^3_{yxy} g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
& + \frac{\ell^2 M}{\mu^3}\left\|\nabla^3_{yyy} g(x,y^*(x)) - \nabla^3_{yyy} g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
& + \left(\frac{4\ell\rho M}{\mu^3} + \frac{2\rho\ell^2 M}{\mu^4} + \frac{\rho M}{\mu^2}\left(1+\frac{\ell}{\mu}\right)^2\right)\left\|\nabla^2_{yy} g(x,y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
& + \left(\frac{4\rho M}{\mu^2} + \frac{2\rho\ell M}{\mu^3}\right)\left\|\nabla^2_{xy} g(x,y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
& + \left(\left(\frac{\rho\ell}{\mu} + \frac{\rho^2 M}{\mu^2}\right)\left(1+\frac{\ell}{\mu}\right)^2 + \frac{\nu M}{\mu} + \frac{4\ell\nu M + 4\rho^2 M}{\mu^2} + \frac{6\ell\rho^2 M + \ell^2\nu M}{\mu^3}\right. \\
& \qquad \left. + \frac{2\rho^2\ell^2 M}{\mu^4}\right)\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|.
\end{aligned} \tag{D.39}$$

Finally, similar to the computation in (D.37), we bound the last term as

$$\begin{aligned}
(IV) = & \left\|\frac{\partial y^*(x)}{\partial x}\cdot\nabla^2_{yy} f(x,y^*(x))\cdot\frac{\partial y^*(x)}{\partial x}^\top - \frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}\cdot\nabla^2_{yy} f_{\mathcal{D}_F}(x,y^*_{\mathcal{D}_G}(x))\cdot\frac{\partial y^*_{\mathcal{D}_G}(x)}{\partial x}^\top\right\| \\
\leq & \frac{2\ell^3}{\mu^3}\left\|\nabla^2_{yy} g(x,y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x,y^*(x))\right\| + \frac{2\ell^2}{\mu^2}\left\|\nabla^2_{xy} g(x,y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
& + \frac{\ell^2}{\mu^2}\left\|\nabla^2_{yy} f(x,y^*(x)) - \nabla^2_{yy} f_{\mathcal{D}_F}(x,y^*(x))\right\| + \left(\frac{2\ell^3\rho}{\mu^3} + \frac{3\ell^2\rho}{\mu^2}\right)\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|.
\end{aligned} \tag{D.40}$$

Combining terms $(I)-(IV)$ together, we have

$$
\begin{aligned}
&\|\nabla^2\Phi(x)-\nabla^2\Phi_{\mathcal{D}}(x)\| \\
\le&\left\|\nabla^2_{xx}f(x,y^*(x))-\nabla^2_{xx}f_{\mathcal{D}_F}(x,y^*(x))\right\|+\frac{2\ell}{\mu}\left\|\nabla^2_{yx}f(x,y^*(x))-\nabla^2_{yx}f_{\mathcal{D}_F}(x,y^*(x))\right\| \\
&+\frac{\ell^2}{\mu^2}\left\|\nabla^2_{yy}f(x,y^*(x))-\nabla^2_{yy}f_{\mathcal{D}_F}(x,y^*(x))\right\| \\
&+\frac{\rho}{\mu}\left(1+\frac{\ell}{\mu}\right)^2\|\nabla_y f(x,y^*(x))-\nabla_y f_{\mathcal{D}_F}(x,y^*(x))\| \\
&+\left(\frac{2\ell^2}{\mu^2}+\frac{4\ell\rho M+2\ell^3}{\mu^3}+\frac{2\rho\ell^2M}{\mu^4}+\frac{\rho M}{\mu^2}\left(1+\frac{\ell}{\mu}\right)^2\right)\left\|\nabla^2_{yy}g(x,y^*(x))-\nabla^2_{yy}g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
&+\left(\frac{2\ell}{\mu}+\frac{4\rho M+2\ell^2}{\mu^2}+\frac{2\rho\ell M}{\mu^3}\right)\left\|\nabla^2_{xy}g(x,y^*(x))-\nabla^2_{xy}g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
&+\frac{M}{\mu}\|\nabla^3_{xxy}g(x,y^*(x))-\nabla^3_{xxy}g_{\mathcal{D}_G}(x,y^*(x))\|+\frac{\ell^2M}{\mu^3}\left\|\nabla^3_{yyy}g(x,y^*(x))-\nabla^3_{yyy}g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
&+\frac{4\ell M}{\mu^2}\left\|\nabla^3_{yxy}g(x,y^*(x))-\nabla^3_{yxy}g_{\mathcal{D}_G}(x,y^*(x))\right\| \\
&+\left(\rho+\left(\frac{\rho\ell}{\mu}+\frac{\rho^2M}{\mu^2}\right)\left(1+\frac{\ell}{\mu}\right)^2+\frac{\nu M+4\ell\rho}{\mu}+\frac{4\ell\nu M+4\rho^2M+5\ell^2\rho}{\mu^2}\right. \\
&\qquad\left.+\frac{6\ell\rho^2M+\ell^2\nu M+2\ell^3\rho}{\mu^3}+\frac{2\rho^2\ell^2M}{\mu^4}\right)\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|.
\end{aligned}
\tag{D.41}
$$

To deal with the last term $\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|$, we utilize the strongly convexity of $g(x,y)$ and $g_{\mathcal{D}_G}(x,y)$ with respect to y and have

$$
\begin{aligned}
0\ge&\langle\nabla_y g(x,y^*_{\mathcal{D}_G}(x))-\nabla_y g(x,y^*(x)),y^*(x)-y^*_{\mathcal{D}_G}(x)\rangle+\mu\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|^2, \\
0\ge&\langle\nabla_y g_{\mathcal{D}_G}(x,y^*_{\mathcal{D}_G}(x))-\nabla_y g_{\mathcal{D}_G}(x,y^*(x)),y^*(x)-y^*_{\mathcal{D}_G}(x)\rangle+\mu\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|^2.
\end{aligned}
\tag{D.42}
$$

Note that the optimality conditions are $\nabla_y g(x,y^*(x))=0,\nabla_y g_{\mathcal{D}_G}(x,y^*_{\mathcal{D}_G}(x))=0$, which combining with (D.42), yields

$$
0\ge\langle\nabla_y g(x,y^*_{\mathcal{D}_G}(x))-\nabla_y g_{\mathcal{D}_G}(x,y^*(x)),y^*(x)-y^*_{\mathcal{D}_G}(x)\rangle+2\mu\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|^2. \tag{D.43}
$$

Therefore, we can bound $\|y^*(x)-y^*_{\mathcal{D}_G}(x)\|$ as

$$
\begin{aligned}
&\|y^*(x)-y^*_{\mathcal{D}_G}(x)\| \\
\le&\frac{1}{2\mu}\|\nabla_y g(x,y^*_{\mathcal{D}_G}(x))-\nabla_y g_{\mathcal{D}_G}(x,y^*(x))\| \\
\le&\frac{1}{2\mu}\|\nabla_y g(x,y^*_{\mathcal{D}_G}(x))\|+\frac{1}{2\mu}\|\nabla_y g_{\mathcal{D}_G}(x,y^*(x))\| \\
\le&\frac{1}{2\mu}\|\nabla_y g(x,y^*_{\mathcal{D}_G}(x))-\nabla_y g_{\mathcal{D}_G}(x,y^*_{\mathcal{D}_G}(x))\|+\frac{1}{2\mu}\|\nabla_y g(x,y^*(x))-\nabla_y g_{\mathcal{D}_G}(x,y^*(x))\|,
\end{aligned}
\tag{D.44}
$$

where the first inequality is obtained from (D.43) and the last inequality is due to the optimality conditions. Plugging (D.44) into (D.41), we get 11 different terms in (D.41) and the major parts of them are

$$\begin{array}{ll}
\left\|\nabla^2_{xx} f(x, y^*(x)) - \nabla^2_{xx} f_{\mathcal{D}_F}(x, y^*(x))\right\|, & \left\|\nabla^2_{yx} f(x, y^*(x)) - \nabla^2_{yx} f_{\mathcal{D}_F}(x, y^*(x))\right\|, \\
\left\|\nabla^2_{yy} f(x, y^*(x)) - \nabla^2_{yy} f_{\mathcal{D}_F}(x, y^*(x))\right\|, & \left\|\nabla_{y} f(x, y^*(x)) - \nabla_{y} f_{\mathcal{D}_F}(x, y^*(x))\right\|, \\
\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\|, & \left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\|, \\
\|\nabla^3_{xxy} g(x, y^*(x)) - \nabla^3_{xxy} g_{\mathcal{D}_G}(x, y^*(x))\|, & \|\nabla^3_{yxy} g(x, y^*(x)) - \nabla^3_{yxy} g_{\mathcal{D}_G}(x, y^*(x))\|, \\
\left\|\nabla^3_{yyy} g(x, y^*(x)) - \nabla^3_{yyy} g_{\mathcal{D}_G}(x, y^*(x))\right\|, & \|\nabla_{y} g(x, y^*_{\mathcal{D}_G}(x)) - \nabla_{y} g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))\|, \\
\|\nabla_{y} g(x, y^*(x)) - \nabla_{y} g_{\mathcal{D}_G}(x, y^*(x))\|. &
\end{array} \tag{D.45}$$

Note that each of the above terms is the difference between the empirical and population derivative. Therefore, each of them can be bounded by one of Lemmas 38 - 40. For example, we consider the term with the largest coefficient, e.g.

$$\mathcal{O}(\kappa^5)\|\nabla_y g(x, y^*(x)) - \nabla_y g_{\mathcal{D}_G}(x, y^*(x))\|. \tag{D.46}$$

By Lemma 38, we can choose $D_g = \mathcal{O}\left(\kappa^{10} \cdot \frac{\log(2d/\delta')}{\rho_\phi \epsilon}\right)$ so that the above term can be bounded by $\frac{1}{880\iota}\sqrt{\rho_\phi \epsilon}$ with probability $1 - \delta'$. For other terms, we can apply the similar techniques, and select batch sizes $D_f = \mathcal{O}\left(\kappa^6 \cdot \frac{\log(2d/\delta')}{\rho_\phi \epsilon}\right), D_g = \mathcal{O}\left(\kappa^{10} \cdot \frac{\log(2d/\delta')}{\rho_\phi \epsilon}\right)$ such that each term can be bounded by $\frac{1}{880\iota}\sqrt{\rho_\phi \epsilon}$ with probability $1 - \delta'$. This completes the proof of (D.28) with probability $1 - \delta = (1 - \delta')^{11}$. □

Lemma 42 *Suppose Assumptions 28 and 29 hold. Set batch sizes $D_g = \mathcal{O}(\kappa^6 \epsilon^{-2}), D_f = \mathcal{O}(\kappa^2 \epsilon^{-2})$, with probability at least $1 - \delta$, we have the following inequality holds:*

$$\|\nabla \Phi_{\mathcal{D}}(x) - \nabla \Phi(x)\| \leq \frac{1}{10}\epsilon. \tag{D.47}$$

Proof. The gradient estimate error between $\nabla\Phi(x)$ and $\nabla\Phi_{\mathcal{D}}(x)$ can be computed as

$$
\begin{aligned}
&\|\nabla\Phi(x) - \nabla\Phi_{\mathcal{D}}(x)\| \\
\leq& \left\|\nabla_x f(x, y^*(x)) - \nabla_x f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\
&+ \left\|\frac{\partial y^*(x)}{x}\nabla_y f(x, y^*(x)) - \frac{\partial y^*_{\mathcal{D}_G}(x)}{x}\nabla_y f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| \\
\leq& \left\|\nabla_x f(x, y^*(x)) - \nabla_x f_{\mathcal{D}_F}(x, y^*_{\mathcal{D}_G}(x))\right\| + \left\|\frac{\partial y^*(x)}{x}\right\| \|\nabla_y f(x, y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x, y^*(x))\| \\
&+ \left\|\frac{\partial y^*(x)}{x} - \frac{\partial y^*_{\mathcal{D}_G}(x)}{x}\right\| \|\nabla_y f_{\mathcal{D}_F}(x, y^*(x))\| \\
\leq& \|\nabla_x f(x, y^*(x)) - \nabla_x f_{\mathcal{D}_F}(x, y^*(x))\| + \ell\|y^*(x) - y^*_{\mathcal{D}_G}(x)\| \\
&+ \frac{\ell}{\mu}\|\nabla_y f(x, y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x, y^*(x))\| + \frac{\ell^2}{\mu}\|y^*(x) - y^*_{\mathcal{D}_G}(x)\| \\
&+ \frac{\ell M}{\mu^2}\left(\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \rho\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|\right) \\
&+ \frac{M}{\mu}\left(\left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \rho\|y^*(x) - y^*_{\mathcal{D}_G}(x)\|\right) \\
\leq& \|\nabla_x f(x, y^*(x)) - \nabla_x f_{\mathcal{D}_F}(x, y^*(x))\| + \frac{\ell}{\mu}\|\nabla_y f(x, y^*(x)) - \nabla_y f(x, y^*(x))\| \\
&+ \frac{\ell M}{\mu^2}\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\| + \frac{M}{\mu}\left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\| \\
&+ \frac{1}{2\mu}\left(\ell + \frac{\ell^2 + M\rho}{\mu} + \frac{\ell M\rho}{\mu^2}\right)\left(\|\nabla_y g(x, y^*_{\mathcal{D}_G}(x)) - \nabla_y g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))\|\right. \\
&\left.+\|\nabla_y g(x, y^*(x)) - \nabla_y g_{\mathcal{D}_G}(x, y^*(x))\|\right),
\end{aligned}
\tag{D.48}
$$

where the first inequality is obtained by (1.2) and the triangle inequality, the third inequality is due to (D.32), Assumption 3 and Proposition 33. The last inequality is by (D.44). Nota that there are 6 different terms in the above gradient estimate error and they are

$$
\begin{aligned}
&\|\nabla_x f(x, y^*(x)) - \nabla_x f_{\mathcal{D}_F}(x, y^*(x))\|, \quad \|\nabla_y f(x, y^*(x)) - \nabla_y f_{\mathcal{D}_F}(x, y^*(x))\|, \\
&\left\|\nabla^2_{yy} g(x, y^*(x)) - \nabla^2_{yy} g_{\mathcal{D}_G}(x, y^*(x))\right\|, \quad \left\|\nabla^2_{xy} g(x, y^*(x)) - \nabla^2_{xy} g_{\mathcal{D}_G}(x, y^*(x))\right\|, \\
&\|\nabla_y g(x, y^*_{\mathcal{D}_G}(x)) - \nabla_y g_{\mathcal{D}_G}(x, y^*_{\mathcal{D}_G}(x))\|, \quad \|\nabla_y g(x, y^*(x)) - \nabla_y g_{\mathcal{D}_G}(x, y^*(x))\|.
\end{aligned}
\tag{D.49}
$$

For every term, we can apply one of Lemma 38 - Lemma 40 and set batch sizes

$$
D_f = \mathcal{O}\left(\kappa^2 \cdot \frac{\log(2d/\delta')}{\epsilon^2}\right), D_g = \mathcal{O}\left(\kappa^6 \cdot \frac{\log(2d/\delta')}{\epsilon^2}\right)
$$

such that each term can be bounded by $\frac{1}{60}\epsilon$ with probability $1 - \delta'$. This completes the proof of (D.47) with probability $1 - \delta = (1 - \delta')^6$. □

For the StocBiO algorithm (Algorithm 3), we have the following descent lemma.

Lemma 43 *Suppose Assumption 28 holds. We have with high probability the update* $x_{k+1} = x_k - \frac{\bar{\xi}}{80}\sqrt{\frac{\epsilon}{\rho_\phi}}u$ *in Algorithm 3 yields*

$$\mathbb{E}\Phi(x_{k+1}) \leq \mathbb{E}\Phi(x_k) - \frac{1}{3\cdot 80^3\iota^3}\cdot\sqrt{\frac{\epsilon^3}{\rho_\phi}}. \tag{D.50}$$

Proof. Combining Lemma 26 and Lemma 41, we have

$$\begin{aligned}\frac{u_{out}^\top\nabla^2\Phi(x_k)u_{out}}{\|u_{out}\|^2} \leq& \frac{u_{out}^\top\nabla^2\Phi_{\mathcal{D}}(x_k)u_{out}}{\|u_{out}\|^2} + \left|\frac{u_{out}^\top\nabla^2\Phi_{\mathcal{D}}(x_k)u_{out}}{\|u_{out}\|^2} - \frac{u_{out}^\top\nabla^2\Phi(x_k)u_{out}}{\|u_{out}\|^2}\right| \\ \leq& \frac{u_{out}^\top\nabla^2\Phi_{\mathcal{D}}(x_k)u_{out}}{\|u_{out}\|^2} + \left\|\nabla^2\Phi_{\mathcal{D}}(x_k) - \nabla^2\Phi(x_k)\right\| \\ \leq& \left(-\frac{1}{40\iota} + \frac{1}{80\iota}\right)\sqrt{\rho_\phi\epsilon} \\ =& -\frac{1}{80\iota}\sqrt{\rho_\phi\epsilon}. \end{aligned} \tag{D.51}$$

By Xu et al. (2018)[Lemma 1], we have

$$\mathbb{E}\Phi(x_{k+1}) \leq \mathbb{E}\Phi(x_k) - \frac{1}{3\cdot 80^3\iota^3}\cdot\sqrt{\frac{\epsilon^3}{\rho_\phi}}. \tag{D.52}$$

□

Lemma 44 *Suppose Assumptions 28 and 29 hold. Set parameters as*

$$\begin{aligned} S = O(\kappa^5\epsilon^{-2}), B = O(\kappa^2\epsilon^{-2}), D_f = O(\kappa^2\epsilon^{-2}), D_g = O(\kappa^2\epsilon^{-2}), \\ Q = O(\kappa\log\frac{1}{\epsilon}), D = O(\kappa\log\frac{1}{\epsilon}), \alpha = \frac{2}{\ell+\mu}, \beta = \frac{1}{4L_\phi}. \end{aligned}$$

With high probability, the update $x_{k+1} = x_k - \widehat{\nabla}\Phi(x_k)$ *in Algorithm 3 yields*

$$\mathbb{E}\left[\Phi(x_{k+1}) - \Phi(x_k)\right] \leq -\frac{1}{16L_\phi}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \frac{1}{400L_\phi}\epsilon^2. \tag{D.53}$$

To prove Lemma 44, we borrow the following two useful lemmas for stocBiO from Ji et al. (2021).

Lemma 45 *Ji et al. (2021)[Lemma 7] Suppose Assumptions 28 and 29 hold. Denote the* $\mathbb{E}_k$ *as the conditional expectation conditioning on* x_k *and* y_k^D*, i.e.,* $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid x_k, y_k^D]$*. We have*

$$\left\|\mathbb{E}_k\widehat{\nabla}\Phi(x_k) - \nabla\Phi(x_k)\right\|^2 \leq 2\Big(\ell + \frac{\ell^2}{\mu} + \frac{M\rho}{\mu} + \frac{LM\rho}{\mu^2}\Big)^2\|y_k^D - y^*(x_k)\|^2 + \frac{2\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}.$$

Lemma 46 *Ji et al. (2021)[Lemma 8] Suppose Assumptions 28 and 29 hold. Then, we have*

$$\mathbb{E}\|\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2 \leq \frac{4\ell^2M^2}{\mu^2D_g}+\Big(\frac{8\ell^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2\ell^4M^2}{\mu^2}\frac{1}{B}+\frac{16\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2} + \Big(\ell+\frac{\ell^2}{\mu}+\frac{M\rho}{\mu}+\frac{\ell M\rho}{\mu^2}\Big)^2\mathbb{E}\|y_k^D-y^*(x_k)\|^2.$$

We now prove Lemma 44.

Proof. By Assumptions 28, $\Phi(x)$ is L_ϕ smooth, which yields

$$\begin{aligned}\Phi(x_{k+1}) \leq& \Phi(x_k)+\langle\nabla\Phi(x_k),x_{k+1}-x_k\rangle+\frac{L_\phi}{2}\|x_{k+1}-x_k\|^2\\ \leq& \Phi(x_k)-\beta\langle\nabla\Phi(x_k),\widehat{\nabla}\Phi(x_k)\rangle+\beta^2L_\phi\|\nabla\Phi(x_k)\|^2+\beta^2L_\phi\|\nabla\Phi(x_k)-\widehat{\nabla}\Phi(x_k)\|^2,\end{aligned}$$

where the second inequality is by Young's inequality. Taking expectation over the above inequality, we have

$$\begin{aligned}&\mathbb{E}\Phi(x_{k+1})\\ \leq&\mathbb{E}\Phi(x_k)-\beta\mathbb{E}\langle\nabla\Phi(x_k),\mathbb{E}_k\widehat{\nabla}\Phi(x_k)\rangle+\beta^2L_\phi\mathbb{E}\|\nabla\Phi(x_k)\|^2+\beta^2L_\phi\mathbb{E}\|\nabla\Phi(x_k)-\widehat{\nabla}\Phi(x_k)\|^2\\ \leq&\mathbb{E}\Phi(x_k)+\frac{\beta}{2}\mathbb{E}\|\mathbb{E}_k\widehat{\nabla}\Phi(x_k)-\nabla\Phi(x_k)\|^2-\frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2+\frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)-\widehat{\nabla}\Phi(x_k)\|^2\\ \leq&\mathbb{E}\Phi(x_k)-\frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2+\frac{\beta\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}\\ &+\frac{\beta}{4}\left(\frac{4\ell^2M^2}{\mu^2D_g}+\Big(\frac{8\ell^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2\ell^4M^2}{\mu^2}\frac{1}{B}+\frac{16\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}\right)\\ &+\frac{5\beta}{4}\Big(\ell+\frac{L^2}{\mu}+\frac{M\tau}{\mu}+\frac{\ell M\rho}{\mu^2}\Big)^2\mathbb{E}\|y_k^D-y^*(x_k)\|^2. \qquad \text{(D.54)}\end{aligned}$$

The second inequality is by Young's inequality and $\beta=\frac{1}{4L_\phi}$. The last inequality is by Lemmas 45 and 46. Further note that for an integer $t\leq D$

$$\begin{aligned}\|y_k^{t+1}-y^*(x_k)\|^2 =&\|y_k^{t+1}-y_k^t\|^2+2\langle y_k^{t+1}-y_k^t,y_k^t-y^*(x_k)\rangle+\|y_k^t-y^*(x_k)\|^2\\ =&\alpha^2\|\nabla_yG(x_k,y_k^t;\mathcal{S}_t)\|^2-2\alpha\langle\nabla_yG(x_k,y_k^t;\mathcal{S}_t),y_k^t-y^*(x_k)\rangle\\ &+\|y_k^t-y^*(x_k)\|^2. \qquad \text{(D.55)}\end{aligned}$$

Conditioning on x_k,y_k^t and taking expectation in (D.55), we have

$$\begin{aligned}&\mathbb{E}[\|y_k^{t+1}-y^*(x_k)\|^2|x_k,y_k^t]\\ \leq&\alpha^2\Big(\frac{\sigma^2}{S}+\|\nabla_yg(x_k,y_k^t)\|^2\Big)-2\alpha\langle\nabla_yg(x_k,y_k^t),y_k^t-y^*(x_k)\rangle+\|y_k^t-y^*(x_k)\|^2\\ \leq&\frac{\alpha^2\sigma^2}{S}+\alpha^2\|\nabla_yg(x_k,y_k^t)\|^2-2\alpha\left(\frac{\ell\mu}{\ell+\mu}\|y_k^t-y^*(x_k)\|^2+\frac{\|\nabla_yg(x_k,y_k^t)\|^2}{\ell+\mu}\right)\\ &+\|y_k^t-y^*(x_k)\|^2\\ =&\frac{\alpha^2\sigma^2}{S}-\alpha\left(\frac{2}{\ell+\mu}-\alpha\right)\|\nabla_yg(x_k,y_k^t)\|^2+\left(1-\frac{2\alpha\ell\mu}{\ell+\mu}\right)\|y_k^t-y^*(x_k)\|^2, \qquad \text{(D.56)}\end{aligned}$$

where the first inequality is by Assumption 29 and the second inequality follows from the strong-convexity (with respect to y) and smoothness of the function g. Since $\alpha = \frac{2}{\ell+\mu}$, we obtain from (D.56) that

$$\mathbb{E}[\|y_k^{t+1} - y^*(x_k)\|^2 | x_k, y_k^t] \leq \left(\frac{\ell-\mu}{\ell+\mu}\right)^2 \|y_k^t - y^*(x_k)\|^2 + \frac{4\sigma^2}{(\ell+\mu)^2 S}. \tag{D.57}$$

Unconditioning on x_k, y_k^t in (D.57) and telescoping (D.57) over t from 0 to $D-1$ yield

$$\mathbb{E}\|y_k^D - y^*(x_k)\|^2 \leq \left(\frac{\ell-\mu}{\ell+\mu}\right)^{2D} \mathbb{E}\|y_k^0 - y^*(x_k)\|^2 + \frac{\sigma^2}{\ell\mu S}. \tag{D.58}$$

By setting $D > 1/2\log(1/4)/\log\left(\frac{1-\kappa}{1+\kappa}\right) = \mathcal{O}(\kappa)$, we get $\left(\frac{\ell-\mu}{\ell+\mu}\right)^{2D} < 1/4$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\|y_k^0 - y^*(x_k)\|^2 \leq & 2\mathbb{E}\|y_{k-1}^D - y^*(x_{k-1})\|^2 + 2\mathbb{E}\|y^*(x_k) - y^*(x_{k-1})\|^2 \\
\leq & 2\left(\frac{\ell-\mu}{\ell+\mu}\right)^{2D} \mathbb{E}\|y_{k-1}^0 - y^*(x_{k-1})\|^2 + 2\kappa^2\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{2\sigma^2}{\ell\mu S} \\
\leq & \frac{1}{2}\mathbb{E}\|y_{k-1}^0 - y^*(x_{k-1})\|^2 + \left(2\kappa^2\beta^2\left(M + \frac{\ell M}{\mu}\right)^2 + \frac{2\sigma^2}{\ell\mu S}\right) \\
\leq & \left(\frac{1}{2}\right)^k \mathbb{E}\|y_0^0 - y^*(x_0)\|^2 + \sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^j \left(2\kappa^2\beta^2\left(M + \frac{\ell M}{\mu}\right)^2 + \frac{2\sigma^2}{\ell\mu S}\right) \\
\leq & \mathbb{E}\|y_0^0 - y^*(x_0)\|^2 + 4\kappa^2\beta^2\left(M + \frac{\ell M}{\mu}\right)^2 + \frac{4\sigma^2}{\ell\mu S} \\
= & \widehat{\Delta} + 4\kappa^2\beta^2\left(M + \frac{\ell M}{\mu}\right)^2 + \frac{4\sigma^2}{\ell\mu S},
\end{aligned} \tag{D.59}$$

where we denoted $\widehat{\Delta} = \|y_0^0 - y^*(x_0)\|^2$, the second inequality is true as $y^*(x)$ is κ-Lipschitz continuous and (D.58), the third inequality is by the fact that $\|\widehat{\nabla}\Phi(x)\|$ can be bounded by $M + \frac{\ell M}{\mu}$. Plug (D.58) and (D.59) into (D.54) yields

$$\begin{aligned}
& \mathbb{E}\Phi(x_{k+1}) \\
\leq & \mathbb{E}\Phi(x_k) - \frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2 + \frac{\beta\ell^2 M^2(1-\kappa^{-1})^{2Q}}{\mu^2} \\
& + \frac{\beta}{4}\left(\frac{4\ell^2M^2}{\mu^2 D_g} + \left(\frac{8\ell^2}{\mu^2} + 2\right)\frac{M^2}{D_f} + \frac{16\eta^2\ell^4M^2}{\mu^2}\frac{1}{B} + \frac{16\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}\right) \\
& + \frac{5\beta}{4}\left(\ell + \frac{\ell^2}{\mu} + \frac{M\tau}{\mu} + \frac{\ell M\rho}{\mu^2}\right)^2 \left(\left(\widehat{\Delta} + 4\kappa^2\beta^2\left(M + \frac{\ell M}{\mu}\right)^2 + \frac{4\sigma^2}{\ell\mu S}\right)\left(\frac{\ell-\mu}{\ell+\mu}\right)^{2D} + \frac{\sigma^2}{\ell\mu S}\right).
\end{aligned} \tag{D.60}$$

Therefore, it suffices to choose the parameters as

$$\begin{aligned}
& S = O(\kappa^5\epsilon^{-2}), D_f = O(\kappa^2\epsilon^{-2}), D_g = O(\kappa^2\epsilon^{-2}), \\
& B = O(\kappa^2\epsilon^{-2}), Q = O(\kappa\log\frac{1}{\epsilon}), D = O(\kappa\log\frac{1}{\epsilon}),
\end{aligned} \tag{D.61}$$

such that the following inequality holds,

$$\frac{\beta\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}+\frac{\beta}{4}\left(\frac{4\ell^2M^2}{\mu^2D_g}+\Big(\frac{8\ell^2}{\mu^2}+2\Big)\frac{M^2}{D_f}+\frac{16\eta^2\ell^4M^2}{\mu^2}\frac{1}{B}+\frac{16\ell^2M^2(1-\kappa^{-1})^{2Q}}{\mu^2}\right)$$
$$+\frac{5\beta}{4}\Big(\ell+\frac{\ell^2}{\mu}+\frac{M\tau}{\mu}+\frac{\ell M\rho}{\mu^2}\Big)^2\left(\left(\widehat{\Delta}+4\kappa^2\left(M+\frac{\ell M}{\mu}\right)^2+\frac{6\sigma^2}{\ell\mu S}\right)\left(\frac{\ell-\mu}{\ell+\mu}\right)^{2D}+\frac{\sigma^2}{\ell\mu S}\right)$$
$$\leq\frac{1}{400L_\phi}\epsilon^2. \tag{D.62}$$

Finally, combining the above two inequalities yields

$$\mathbb{E}\left[\Phi(x_{k+1})-\Phi(x_k)\right]\leq-\frac{\beta}{4}\mathbb{E}\|\nabla\Phi(x_k)\|^2+\frac{1}{400L_\phi}\epsilon^2. \tag{D.63}$$

□

D.3 Proof of Theorem 30 and Corollary 31

Proof. We consider the following two possible cases

- **Case 1:** $\mathbb{E}[\|\nabla\Phi(x_k)\||x_k]>\frac{3}{5}\epsilon$. Unconditioning on x_k, we have $\mathbb{E}\|\nabla\Phi(x_k)\|>\frac{3}{5}\epsilon$. In this case, Lemma 44 yields

  $$\mathbb{E}\left[\Phi(x_{k+1})-\Phi(x_k)\right]\leq-\frac{9}{400L_\phi}\epsilon^2+\frac{1}{400L_\phi}\epsilon^2=-\frac{1}{50L_\phi}\epsilon^2$$

  and the total iteration number of **Case 1** can be bounded by

  $$\frac{50L_\phi(\Phi(x_0)-\Phi^*)}{\epsilon^2}. \tag{D.64}$$

- **Case 2:** $\mathbb{E}[\|\nabla\Phi(x_k)\||x_k]\leq\frac{3}{5}\epsilon$, which indicates

  $$\|\nabla\Phi(x_k)\|=\|\mathbb{E}[\nabla\Phi(x_k)|x_k]\|\leq\mathbb{E}[\|\nabla\Phi(x_k)\||x_k]\leq\frac{3}{5}\epsilon.$$

  Unconditioning on x_k, we have $\mathbb{E}\|\nabla\Phi(x_k)\|\leq\frac{3}{5}\epsilon$. In this case we run AID in line 12 of Algorithm 3 such that $\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)-\nabla\Phi_{\mathcal{D}}(x_k)\|\leq\frac{1}{10}\epsilon$. According to Lemma 35, it only requires us to set $D=\mathcal{O}(\kappa\log\epsilon^{-1})$ and $N=\mathcal{O}(\sqrt{\kappa}\log\epsilon^{-1})$ in the AID. Moreover, combining with Lemma 42, we have

  $$\begin{aligned}\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)\|\leq&\|\nabla\Phi_{\mathcal{D}}(x_k)\|+\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)-\nabla\Phi_{\mathcal{D}}(x_k)\|\\ \leq&\|\nabla\Phi(x_k)\|+\|\nabla\Phi_{\mathcal{D}}(x_k)-\nabla\Phi(x_k)\|+\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)-\nabla\Phi_{\mathcal{D}}(x_k)\|\\ \leq&\frac{4}{5}\epsilon.\end{aligned} \tag{D.65}$$

  Therefore, Algorithm 2 is called with high probability. When the if condition in line 13 of Algorithm 3 is satisfied, we can guarantee

  $$\|\nabla\Phi(x_k)\|\leq\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)\|+\|\nabla\Phi_{\mathcal{D}}(x_k)-\nabla\Phi(x_k)\|+\|\widehat{\nabla}\Phi_{\mathcal{D}}(x_k)-\nabla\Phi_{\mathcal{D}}(x_k)\|\leq\epsilon. \tag{D.66}$$

Moreover, when Algorithm 2 is called and it returns a nonzero vector u_{out}, combining Lemmas 26 and 43 yields

$$\mathbb{E}\Phi(x_{k+1}) - \mathbb{E}\Phi(x_k) \leq -\frac{1}{3 \cdot 80^3 \iota^3} \cdot \sqrt{\frac{\epsilon^3}{\rho_\phi}}.$$

So the total iteration number in this case is bounded by

$$\frac{3 \cdot 80^3 \iota^3 (\Phi(x_0) - \Phi(x^*)) \mathscr{T}}{\sqrt{\epsilon^3/\rho_\phi}}. \tag{D.67}$$

If Algorithm 2 returns a zero vector u_{out}, then with high probability we find an ϵ-local minimum.

Therefore, combining (D.64) and (D.67) we know that the total iteration number before we visit an ϵ-local minimum can be bounded by

$$K = \frac{3 \cdot 80^3 \iota^3 (\Phi(x_0) - \Phi(x^*)) \mathscr{T}}{\sqrt{\epsilon^3/\rho_\phi}} + \frac{50 L_\phi (\Phi(x_0) - \Phi(x^*))}{\epsilon^2} = \tilde{\mathcal{O}}\left(\kappa^3 \epsilon^{-2}\right).$$

We then prove Corollary 31. Note that we require the sample batch sizes to be

$$D_f = \tilde{\mathcal{O}}\left(\kappa^2 \epsilon^{-2}\right), \quad D_g = \tilde{\mathcal{O}}\left(\kappa^6 \epsilon^{-2}\right), \tag{D.68}$$

so that Lemma 42 and Lemma 41 hold, and

$$\begin{aligned} &S = O(\kappa^5 \epsilon^{-2}), D_f = O(\kappa^2 \epsilon^{-2}), D_g = O(\kappa^2 \epsilon^{-2}), \\ &Q = O(\kappa \log \epsilon^{-1}), D = O(\kappa \log \epsilon^{-1}), B = O(\kappa^2 \epsilon^{-2}), \end{aligned} \tag{D.69}$$

so that Lemma 44 holds. Combining (D.68) and (D.69), we have for the iNEON calls in Algorithm 3 (Line 12 - 21), the gradient, Jacobian- and Hessian-vector product complexities are

$$\begin{aligned} &G(f, \epsilon) = \tilde{\mathcal{O}}(\kappa^5 \epsilon^{-4}), \quad G(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^{10} \epsilon^{-4}), \\ &JV(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^9 \epsilon^{-4}), \quad HVc(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^{9.5} \epsilon^{-4}). \end{aligned} \tag{D.70}$$

For those stocBiO iterations in Algorithm 3 (Line 3-11), the gradient, Jacobian- and Hessian-vector product complexities are

$$\begin{aligned} &G(f, \epsilon) = \tilde{\mathcal{O}}(\kappa^5 \epsilon^{-4}), \quad G(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^9 \epsilon^{-4}), \\ &JV(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^5 \epsilon^{-4}), \quad HVc(g, \epsilon) = \tilde{\mathcal{O}}(\kappa^6 \epsilon^{-4}). \end{aligned} \tag{D.71}$$

We take the maximum between (D.70) and (D.71), which completes the proof of Corollary 31. □

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: Faster online learning of eigenvectors and faster MMWU. In *International Conference on Machine Learning*, pages 116–125. PMLR, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. NEON2: finding local minima via first-order oracles. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3720–3730, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021a.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022.

Ziyi Chen, Qunwei Li, and Yi Zhou. Escaping saddle points in nonconvex minimax optimization via cubic-regularized gradient descent-ascent. *arXiv preprint arXiv:2110.07098*, 2021b.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32:5987–5997, 2019.

Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, 162(1-2):1–32, 2017.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164. PMLR, 2018.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *32nd Annual Conference on Neural Information Processing Systems (NIPS)*, 2018.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, pages 1–46, 2021.

Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the moreau envelope in nonsmooth optimization. *arXiv preprint arXiv:2106.09815*, 2021.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

SS Du, C Jin, MI Jordan, B Póczos, A Singh, and JD Lee. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1068–1078, 2017.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234. PMLR, 2019.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

Tanner Fiez, Lillian Ratliff, Eric Mazumdar, Evan Faulkner, and Adhyyan Narang. Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. *Advances in Neural Information Processing Systems*, 34, 2021.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

Zhishuai Guo and Tianbao Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Minhui Huang. Escaping saddle points for nonsmooth weakly convex functions via perturbed proximal algorithms. *arXiv preprint arXiv:2102.02837*, 2021.

Minhui Huang, Shiqian Ma, and Lifeng Lai. On the convergence of projected alternating maximization for equitable and optimal transport. *arXiv preprint arXiv:2109.15030*, 2021a.

Minhui Huang, Shiqian Ma, and Lifeng Lai. Projection robust wasserstein barycenters. In *International Conference on Machine Learning*, pages 4456–4465. PMLR, 2021b.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A Riemannian block coordinate descent method for computing the projection robust wasserstein distance. In *International Conference on Machine Learning*, pages 4446–4455. PMLR, 2021c.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

Kaiyi Ji, Jason D. Lee, Yingbin Liang, and H. Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 2020.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of Machine Learning Research*, 23(29):1–41, 2022.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Hal Daum III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 13–18 Jul 2020.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=HjFtRc83eBB`.

Gautam Kunapuli, Kristin P Bennett, Jing Hu, and Jong-Shi Pang. Classification model selection via bilevel programming. *Optimization Methods and Software*, 23(4):475–489, 2008.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust wasserstein distance and Riemannian optimization. *Advances in Neural Information Processing Systems*, 33:9383–9397, 2020a.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020b.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020c.

Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. *Advances in Neural Information Processing Systems*, 2020, 2020a.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020b.

Luo Luo and Cheng Chen. Finding second-order stationary point for nonconvex-strongly-concave minimax problem. *arXiv preprint arXiv:2110.04814*, 2021.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Y. E. Nesterov. *Introductory lectures on convex optimization: A basic course*. Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7553-7.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32:14934–14942, 2019.

Barak A Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1): 147–160, 1994.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 2019.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30:4077–4087, 2017.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. ES-based Jacobian enables faster bilevel optimization. *arXiv preprint arXiv:2110.07004*, 2021.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 32:7276–7286, 2019.

Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *Advances in Neural Information Processing Systems*, 31:5530–5540, 2018.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=10anajdGZm`.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.