GRADIENT-BASED ALGORITHMS FOR PESSIMISTIC BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As a powerful framework for a variety of machine learning problems, bilevel optimization has attracted much attention. While many modern gradient-based algorithms have been devised for optimistic bilevel optimization, pessimistic bilevel optimization (PBO) is still less-explored and only studied under the linear settings. To fill this void, we investigate PBO with nonlinear inner- and outer-level objective functions in this work, by reformulating it into a single-level constrained optimization problem. In particular, two gradient-based algorithms are first proposed to solve the reformulated problem, i.e., the switching gradient method (SG-PBO) and the primal-dual method (PD-PBO). Through carefully handling the bias errors in gradient estimations resulted by the nature of bilevel optimization, we show that both SG-PBO and PD-PBO converge to the global minimum of the reformulated problem when it is strongly convex, which immediately implies the convergence to the original PBO. Moreover, we propose the proximal scheme (Prox-PBO) with the convergence guarantee for the nonconvex reformulated problem. To the best of our knowledge, this is the first work that investigates gradient-based algorithms and provides convergence analysis for PBO under non-linear settings. We further conduct experiments on an illustrative example and a robust hyperparameter learning problem, which clearly validate our algorithmic design and theoretical analysis.

1 INTRODUCTION

Originated from the economic and operation research studies (Stackelberg et al., 1952; Bracken & McGill, 1973), bilevel optimization has attracted extensive attention recently in the machine learning community. Many machine learning problems can be naturally captured by a bilevel optimization structure such as meta-learning (Franceschi et al., 2018; Ji et al., 2020), reinforcement learning (Hong et al., 2020; Konda & Tsitsiklis, 2000), network architecture searching (He et al., 2020), etc. Bilevel optimization typically takes the following form

$$\min_{x\in\mathcal{X}} \min_{y\in\mathcal{S}(x)} f(x,y), \quad \text{where } \mathcal{S}(x) = \arg\min_{y\in\mathcal{Y}} g(x,y). \tag{OBO}$$

Here $f(x,y)$ and $g(x,y)$ are the outer- and inner-level objective functions, respectively, and the support sets $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are convex. For a fixed $x \in \mathcal{X}$, $\mathcal{S}(x)$ is the set of all $y \in \mathcal{Y}$ that yields the minimal value of $g(x,\cdot)$. Here, the inner optimization finds a set $\mathcal{S}(x)$ that collects all points y that minimize the inner function $g(x,y)$ for any given x. Then, the outer-level function $f(x,y)$ is minimized over y in the set $\mathcal{S}(x)$ given by the inner optimization, jointly with $x \in \mathcal{X}$. The above problem is also referred to as **optimistic bilevel optimization (OBO)**, because the outer-level minimizes over $y \in \mathcal{S}(x)$, which allows the minimization over x to be over a beneficial loss value. Such OBO problems have been extensively studied in the past, see e.g., Harker & Pang (1988); Outrata (1993); Lignola & Morgan (2001); Dempe et al. (2007) and Sinha et al. (2017); Liu et al. (2021). More recently, many studies have developed various fast and scalable algorithms and provided the convergence rate guarantee for these algorithms (Li et al., 2020; Sow et al., 2022; Liu et al., 2020; Huang & Huang, 2021; Ji, 2021; Ji et al., 2022; Huang et al., 2022; Yang et al., 2021). Readers can refer to Section 1.2 for more detailed discussion of the related work.

As an equally important class of bilevel problems, **pessimistic bilevel optimization (PBO)** takes the following formulation

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{S}(x)} f(x,y), \quad \text{where } \mathcal{S}(x) \coloneqq \arg\min_{y\in\mathbb{R}^m} g(x,y). \tag{PBO}$$

For each given x, the inner optimization also collects all minima of the inner function $g(x, \cdot)$ into a set-value function $\mathcal{S}(x)$. Then the outer-level function $f(x, y)$ is first maximized over $y \in \mathcal{S}(x)$, and then minimized over the outer variable x. Intuitively, the maximization finds the worst case of the outer-level function over $y \in \mathcal{S}(x)$, and is hence called the **pessimistic** problem.

In contrast to OBO, the more challenging PBO still remains much less studied in the literature, due to its minimax nature in the outer optimization. Particularly, the previous studies of PBO mainly focused on the existence of optimal solutions and the characterization of optimality condition (Dempe et al., 2014; Wiesemann et al., 2013; Lucchetti et al., 1987), which have shown that the optimality condition of PBO is more strict than that of OBO. Besides, the design of algorithms therein was mainly restricted to linear bilevel optimization (Wiesemann et al., 2013; Zeng, 2020; Zheng et al., 2016) and lacked convergence guarantees. To the best of our knowledge, there have not been gradient-based algorithms developed for PBO. To fill this gap, the main goal of this paper is to develop principled gradient-based algorithms for PBO that enjoy convergence guarantees and are also scalable for modern machine learning applications.

1.1 Contributions

In this paper, we provide a comprehensive study on the pessimistic bilevel optimization. The specific contributions are summarized as follows.

Algorithmic design. We first propose two scalable and easy-to-implement gradient-based algorithms: switching gradient pessimistic bilevel optimizer (SG-PBO) and primal-dual pessimistic bilevel optimizer (PD-PBO). Then, we devise a proximal-point scheme (Prox-PBO) which provides a stronger theoretical guarantee but with increased complexity. Due to the page limits, we delegate the details about Prox-PBO into Appendix A. To the best of our knowledge, these are the first gradient-based algorithms for PBO. In particular, our algorithmic design features a novel reformulation of PBO to a single-level constrained optimization problem by leveraging the KKT conditions and the constraint relaxation. It can be shown that the perturbation introduced by the relaxation is bounded above by some controllable parameters, which consequently builds the equivalence between the convergence for the reformulated problem and the convergence for the original PBO.

Convergence rate analysis. We provide the first-known convergence rate analysis for PBO. Specifically, for any arbitrary $\epsilon > 0$, we show that both SG-PBO and PD-PBO converge sublinearly to an ϵ-accurate solution of the reformulated problem when it is strongly convex. Under the general nonconvex setting, Prox-PBO is proven to converge to an ϵ-KKT point of the reformulated problem with a sublinear rate in Appendix G, where the KKT condition serves as a necessary condition for the local optimality. Technically, the constrained (strongly-/non)convex optimization problem here is more challenging than the standard formulation studied in Boob et al. (2019); Ma et al. (2020) due to the nature of bilevel optimization, where careful treatments are needed to deal with the bias errors arising in gradient estimations for the updates of both the primal and dual variables. We further establish an uniform upper bound on optimal dual variables, which was taken as an assumption in the standard analysis in Boob et al. (2019); Ma et al. (2020).

Numerical verification. We evaluate SG-PBO and PD-PDO on an illustrative example and a robust hyper-representation learning problem. The experimental results show that both algorithms can converge to the global optima of the studied problems, which validates our algorithmic design and theoretical analysis. Besides, in contrast to PD-PBO, SG-PBO has a better track of the constraint violation and, as a tradeoff, the convergence of the outer-level objective may be less stable.

1.2 Related Works

Pessimistic Bilevel Optimization: On the theoretical side, previous studies focused on identifying the existence of solution (Aiyoshi & Shimizu, 1984; Aardal et al., 1996; Aboussoror & Mansouri, 2005), and characterizing the conditions of optimality (Dempe et al., 2014; 2019; Liu et al., 2014). Different reformulation schemes have also been proposed to transform PBO into some easier problems (Aboussoror & Mansouri, 2005; Zheng et al., 2013; Lampariello et al., 2019; Jia & Wan, 2013). On the numerical perspective, algorithms were only designed under restrictive settings, e.g., linear PBO (Wiesemann et al., 2013; Zeng, 2020; Zheng et al., 2016), quadratic-linear PBO (Malyshev & Strekalovskii, 2011). A finite-dimensional approximation method was also proposed to solve the

general PBO (Wiesemann et al., 2013), which restricted the support of inner-level problems to be a finite subset of $\mathbb{R}^n$, i.e. $Y_k \subseteq \mathbb{R}^n$ and $|Y_k| \leq \infty$, and enlarged the cardinalty of Y_k to gradually approximate the original problem. In this paper, we propose efficient gradient-based algorithms for PBO with non-linear objective functions by reformulating PBO into a single-level constrained optimization problem, and provide the first known convergence analysis of PBO.

Recent Advances in OBO: The gradient-based algorithms have become popular for solving the bilevel optimization problem with unique inner-minimum, due to their simplicity and scalability. For example, to compute the gradient of the outer-level optimization efficiently, both approximated implicit differentiation (AID) (Domke, 2012; Lorraine et al., 2020; Ji & Liang, 2021) and iterative differentiation (ITD) approaches (Domke, 2012; Grazzi et al., 2020b; MacKay et al., 2019) have been widely studied. Asymptotic convergence analysis was studied in, e.g., Franceschi et al. (2018); Shaban et al. (2019), and recently Grazzi et al. (2020a); Ji et al. (2020; 2021); Ji & Liang (2021) provided the non-asymptotic convergence rate analysis. Another line of studies (Sabach & Shtern, 2017; Liu et al., 2020; Li et al., 2020) utilized the gradient sequential averaging method to solve the optimistic bilevel optimization with single inner-optimum. A recent work (Sow et al., 2022) proposed gradient-based and hessian-free algorithms for the bilevel optimization problem with multiple inner-minima, and provided the first non-asymptotic analysis therein.

Generic Nonconvex Constrained Optimization: Although the convex constrained optimization problem has been extensively studied in the literature (Lan & Zhou, 2016; Nesterov et al., 2018; Lin et al., 2018; Aravkin et al., 2019; Nemirovski, 2004), the finite-time convergence analysis for the case where both objective and constraints are nonconvex is only provided recently by leveraging the proximal methods (Boob et al., 2019; Ma et al., 2020). For the constrained minmax optimization, previous studies focused on the fixed convex and closed set constraints (Tseng, 1995; Lin et al., 2020; Alkousa et al., 2019; Gidel et al., 2019). As we elaborate in Section 2.1, the PBO could be reformulated as a special case of the nonconvex objective and nonconvex constrained minmax optimization problem with variable set constraint on the inner-variable y. As a byproduct, this paper also contributes to the generic nonconvex constrained optimization.

2 Problem Formulation

To ground PBO in real-world applications, consider the following robust hyperparameter learning problem where we seek to learn the best hyperparameters that are robust to the model learning. Specifically, given a hyperparameter $x \in \mathcal{X}$, we find the optimal model on the training datasets with the training loss function $g(x, y)$

$$\min_{y \in \mathbb{R}^m} g(x, y), \tag{1}$$

where multiple optimal y, i.e., the set $\mathcal{S}(x)$, may exist. However, due to the randomness of the training dataset and the algorithm design, the validation loss of the learnt model on a different validation dataset could be as large as $\max_{y \in \mathcal{S}(x)} f(x, y)$. To guarantee a more robust learning performance, we aim to learn the hyperparameter x that excels in the worst case:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{S}(x)} f(x, y). \tag{2}$$

Clearly, the robust hyperparameter learning in the eq. (2) is a concrete example of PBO. More details on the choices of the functions f and g in practice could be found in Section 5.2.

For the problem studied in this paper, we make the following geometric assumptions. We assume that the constraint set $\mathcal{X}$ is a convex and closed subset of $\mathbb{R}^n$. Usually $\mathcal{X}$ has a simple structure, e.g., simplex or closed intervals, and the orthogonal projections to $\mathcal{X}$ are easy to compute. We also assume that functions f and g are smooth enough such that the Lipschitz smooth conditions hold.

Assumption 1. *For any given $x \in \mathcal{X}$, $f(x, y)$ is a concave function on y, and $g(x, y)$ is a convex function on y. Let $\theta = (x, y)$ and $\theta' = (x', y')$. $f(x, y)$ and $g(x, y)$ are twice continuously differentiable with Lipschitz continuous gradient and Hessian, i.e., there exist constants L_f, L_g, ρ_f and ρ_g, such that for any $x, x' \in \mathcal{X}$, $y, y' \in \mathbb{R}^m$, we have*

$$\|\nabla f(\theta) - \nabla f(\theta')\|_2 \leq L_f \|\theta - \theta'\|_2, \qquad \|\nabla g(\theta) - \nabla g(\theta')\|_2 \leq L_g \|\theta - \theta'\|_2,$$
$$\|\nabla^2 f(\theta) - \nabla^2 f(\theta')\|_F \leq \rho_f \|\theta - \theta'\|_2, \qquad \|\nabla^2 g(\theta) - \nabla^2 g(\theta')\|_F \leq \rho_g \|\theta - \theta'\|_2,$$

where ∇h and $\nabla^2 h$ denote the gradient and the Hessian matrix of a function h with respect to (w.r.t.) θ, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm of matrices.

2.1 SINGLE-LEVEL REFORMULATION

In order to solve the PBO problem, let $g^*(x) \coloneqq \min_{y\in\mathbb{R}^m} g(x,y)$ and we can first replace the set $\mathcal{S}(x)$ by its equivalent form $\mathcal{S}(x) = \{y \in \mathbb{R}^m : g(x,y) - g^*(x) \le 0\}$, such that the original PBO problem can be reformulated as the following problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathbb{R}^m} \quad f(x,y), \qquad s.t. \quad g(x,y) - g^*(x) \le 0. \tag{3}$$

For any small positive constants α and ξ, we further consider an (α,ξ)-relaxation of the problem in eq. (3) defined below:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathbb{R}^m} \quad f(x,y), \quad s.t. \quad g(x,y) - g^*_\alpha(x) - \xi \le 0, \tag{4}$$

where $g^*_\alpha(x) \coloneqq \min_{y\in\mathbb{R}^m} g_\alpha(x,y) \coloneqq g(x,y) + \frac{\alpha}{2}\|y\|_2^2$. The ℓ_2-regularization ensures $g_\alpha(x,y)$ to be strongly convex on y, and hence the solution of $\min_{y\in\mathbb{R}^m} g(x,y) + \frac{\alpha}{2}\|y\|_2^2$ is unique for any given $x \in \mathcal{X}$. It can be shown that $g^*_\alpha(x)$ is differentiable, and its gradient takes the form of $\nabla_x g^*_\alpha(x) = (\nabla_x g_\alpha(x,y))|_{y=y^*_\alpha(x)}$, where $y^*_\alpha(x) \coloneqq \arg\min_{y\in\mathbb{R}^m} g_\alpha(x,y)$. Besides, the ξ in the constraint guarantees that the relaxed problem eq. (4) has at least one strictly feasible point for any given $x \in \mathcal{X}$, which is vital for the problem to be solved efficiently. More importantly, it is worth to note that the perturbation introduced by α and ξ is indeed controllable as shown below.

Proposition 1. *For any fixed $x \in \mathcal{X}$, define $\Phi(x) = \max_{y\in\mathbb{R}^m}\{f(x,y) : g(x,y) - g^*(x) \le 0\}$, and $\Phi_{\alpha,\xi}(x) = \max_{y\in\mathbb{R}^m}\{f(x,y) : g(x,y) - g^*_\alpha(x) - \xi \le 0\}$. It can be shown that $|\Phi(x) - \Phi_{\alpha,\xi}(x)| \le \mathcal{O}(\sqrt{\alpha}) + \mathcal{O}(\sqrt{\xi})$ for every $x \in \mathcal{X}$, under some mild assumptions.*

Proposition 1 indicates that the output of eq. (4) could be set arbitrarily close to that of the original PBO problem by desire. The proof of Proposition 1 could be found in Appendix C.

Although the problem in eq. (4) enjoys properties like strict feasibility and continuously differentiable constraints, the min-max structure therein is by nature much harder to solve than the pure minimization problem. To tackle this challenge, we further convert it to a constrained minimization problem via KKT conditions as follows.

Proposition 2. *Let $w \in \mathcal{W} \coloneqq [0, \frac{\Delta_f}{\xi}]$ with $\Delta_f \coloneqq \max_{x,x'\in\mathcal{X},y,y'\in\mathcal{Y}} |f(x,y) - f(x',y')|$, and $\mathcal{Y} \coloneqq \{y \in \mathbb{R}^m : \|y\|_2 \le D_\mathcal{Y}\}$ with $D_\mathcal{Y} > 0$, such that $\{y \in \mathbb{R}^m : g(x,y) - g^*_\alpha(x) - \xi \le 0\} \subseteq \mathcal{Y}$ for all $x \in \mathcal{X}$. The minimax problem eq. (4) is equivalent to the following constrained optimization:*

$$\begin{aligned} \min_{x\in\mathcal{X},y\in\mathcal{Y},w\in\mathcal{W}} \quad & f(x,y) \\ s.t. \quad & g(x,y) - g^*_\alpha(x) - \xi \le 0 \\ & -\nabla_y f(x,y) + w\nabla_y g(x,y) \le 0 \\ & \nabla_y f(x,y) - w\nabla_y g(x,y) \le 0 \\ & w(g(x,y) - g^*_\alpha(x) - \xi) \le 0 \\ & -w(g(x,y) - g^*_\alpha(x) - \xi) \le 0. \end{aligned} \tag{5}$$

Proposition 2 eliminates the "$\max$" operation on y and simplifies the min-max problem to a minimization problem. Compared to eq. (4), we have four additional inequality constraints in eq. (5) that correspond to $-\nabla_y f(x,y) + w\nabla_y g(x,y) = 0$ and $w(g(x,y) - g^*_\alpha(x) - \xi) = 0$, i.e., the KKT conditions for y attaining the maximum of $f(x,y)$ given $g(x,y) - g^*_\alpha(x) - \xi \le 0$. The proof of Proposition 2 could be found in Appendix D.

In the rest of this paper, we solve the reformulated problem in Proposition 2, which indirectly solves the original PBO in eq. (2). To simplify the notation, let $z = (x,y,w)$, $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}\times\mathcal{W}$ and $h(z)$ be the constraints in eq. (5). Then, the reformulated problem can be written as

$$\min_{z\in\mathcal{Z}} f(z) \quad s.t. \quad h(z) \le 0. \tag{6}$$

3 ALGORITHMS FOR PESSIMISTIC BILEVEL OPTIMIZATION

In this section, we propose two gradient based methods, namely the switching gradient method (SG-PBO) and the primal-dual method (PD-PBO), to solve the PBO problem.

Algorithm 1 Switching Gradient Pessimistic Bilevel Optimization (SG-PBO)

Input: Number of iterations T and N, stepsizes $\{\gamma_t\}_{t=0}^{T-1}$, violation tolerance ϵ
Initialize feasible indices set $\mathcal{T} = \emptyset$ and $z_0 \in \mathcal{Z}$
for $t = 1, ..., T$ **do**
  Conduct projected gradient descent in eq. (7) for N times with any given $\hat{y}_0^t$ as initialization
  if $\max_j \left\{ \left(\hat{h}(z_t; \hat{y}_N^t) \right)_j \right\} \leq \frac{\epsilon}{2}$ **then**
    $\mathcal{T} = \mathcal{T} \cup \{t\}$
    $z_{t+1} = \Pi_{\mathcal{Z}} \left(z_t - \gamma_t^{-1} \nabla f(z_t) \right)$
  else
    Let $i_t = \arg\max_j \left\{ \left(\hat{h}(z_t; \hat{y}_N^t) \right)_j \right\}$.
    $z_{t+1} = \Pi_{\mathcal{Z}} \left(z_t - \gamma_t^{-1} \left(\widehat{\nabla} h(z_t; \hat{y}_N^t) \right)_{i_t} \right)$
  end if
end for
Output: $\bar{z} = \sum_{t \in \mathcal{T}} \gamma_t z_t / \left(\sum_{t \in \mathcal{T}} \gamma_t \right)$

3.1 Switching Gradient Bilevel Optimization Algorithm

Motivated by the recent advance in constrained optimization (Ma et al., 2020; Lan & Zhou, 2016), We first adopt a switching gradient method which features two different updates: either update the variable z along the gradient descent direction of the objective function if all constraints are satisfied (in order to minimize the objective), or update the variable z along the gradient descent direction of the constraint that has the maximum violation (in order to enforce the constraints).

More specifically, suppose that the variable z is updated as $z_t = (x_t, y_t, w_t)$ at iteration t. SG-PBO first runs the following gradient descent over y w.r.t. $g_\alpha(x, y)$

$$\hat{y}_{n+1}^t = \hat{y}_n^t - \tfrac{2}{L_g + 2\alpha} \left(\nabla_y g(x_t, \hat{y}_n^t) + \alpha \hat{y}_n^t \right), \tag{7}$$

such that $g_\alpha(x, \hat{y}_N^t)$ serves as a good approximation for $g_\alpha^*(x_t) := \max_y g_\alpha(x_t, y)$ in the constraint. We further denote $\hat{h}(z_t; \hat{y}_N^t)$ as the approximation of constraint $h(z)$ with $z = z_t$ and $g_\alpha^*(x_t)$ replaced by $g_\alpha(x, \hat{y}_N^t)$.

Next, if the constraint is satisfied, i.e., all components of approximated constraint is small enough ($\max_i \{\hat{h}(z_t; \hat{y}_N^t)_i\} \leq \frac{\epsilon}{2}$ for some prescribed $\epsilon > 0$), then z_t is updated along the gradient descent direction of the objective function $f(z_t)$. Otherwise, z_t is updated along the i_t-th row of $\widehat{\nabla} h(z_t; \hat{y}_N^t)$, where i_t corresponds to the maximum constraint violation component, and $\widehat{\nabla} h(z_t; \hat{y}_N^t)$ is the approximation of $\nabla h(z)$ where $\nabla h(z)$ could be derived from eq. (6) as:

$$\nabla h(z) = \begin{pmatrix} \left(\nabla_x g(\theta) - \nabla_x g_\alpha^*(x) \right)^\top & \left(\nabla_y g(\theta) \right)^\top & 0 \\ -\nabla_{yx}^2 f(\theta) + w \nabla_{yx}^2 g(\theta) & -\nabla_{yy}^2 f(\theta) + w \nabla_{yy}^2 g(\theta) & \nabla_y g(\theta) \\ \nabla_{yx}^2 f(\theta) - w \nabla_{yx}^2 g(\theta) & \nabla_{yy}^2 f(\theta) - w \nabla_{yy}^2 g(\theta) & -\nabla_y g(\theta) \\ w \left(\nabla_x g(\theta) - \nabla_x g_\alpha^*(x) \right)^\top & w \left(\nabla_y g(\theta) \right)^\top & g(\theta) - g_\alpha^*(x) - \xi \\ -w \left(\nabla_x g(\theta) - \nabla_x g_\alpha^*(x) \right)^\top & -w \left(\nabla_y g(\theta) \right)^\top & -g(\theta) + g_\alpha^*(x) + \xi \end{pmatrix}, \tag{8}$$

where $\theta = (x, y)$ for short. $\widehat{\nabla} h(z_t; \hat{y}_N^t)$ is obtained from $\nabla h(z_t)$ by replacing $g_\alpha^*(x_t)$ and $\nabla g_\alpha^*(x_t)$ with $g_\alpha(x, \hat{y}_N^t)$ and $\nabla_x g_\alpha(x_t, \hat{y}_N^t)$, respectively.

More details about SG-PBO can be found in Algorithm 1.Note that although the gradient of $\nabla h(z)$ in eq. (8) involves the calculation of the second-order Jacobian and Hessian terms of f and g, the computational complexity is not demanding since each update uses only one row of the matrix.

3.2 Primal-Dual Bilevel Optimization Algorithm

The primal-dual method solves the minimax problem over the Lagrangian function defined below:

$$\min_{z \in \mathcal{Z}} \max_{\lambda \in \mathbb{R}_+^p} \mathcal{L}(z, \lambda) := f(z) + \langle h(z), \lambda \rangle, \tag{9}$$

where $\lambda \in \mathbb{R}^p_+$ is the dual variable, by alternatively updating the primal variable z and the dual variable λ through gradient descent and gradient ascent, respectively. Because the gradients $\nabla_z \mathcal{L}(z,\lambda) = \nabla_z f(z) + (\nabla_z h(z))^\top \lambda$ and $\nabla_\lambda \mathcal{L}(z,\lambda) = h(z)$, we also need to run a subroutine to estimate $h(z)$ and $\nabla_z h(z)$, as what we have done in eq. (7). Then, the estimations of $\nabla_z \mathcal{L}(z,\lambda)$ and $\nabla_\lambda \mathcal{L}(z,\lambda)$ at the iterate (z_t, λ_{t+1}) immediately follow: $\widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t) = \nabla_z f(z_t) + (\widehat{\nabla}_z h(z_t; \hat{y}_N^t))^\top \lambda_{t+1}$ and $\widehat{\nabla}_\lambda \mathcal{L}(z_t, \lambda_{t+1}) = \hat{h}(z_t; \hat{y}_N^t)$.

We then conduct the accelerated gradient ascent and gradient descent to the Lagrangian:

$$\lambda_{t+1} = \Pi_\Lambda \left(\lambda_t + \tfrac{1}{\eta_t} \left((1+\theta_t)\hat{h}(z_t; \hat{y}_N^t) - \theta_t \hat{h}(z_{t-1}; \hat{y}_N^{t-1}) \right) \right), \tag{10}$$

$$z_{t+1} = \Pi_{\mathcal{Z}} \big(z_t - \tfrac{1}{\tau_t} \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t) \big), \tag{11}$$

where τ_t, η_t are the stepsizes, θ_t is the momentum weight, and $\Lambda \subseteq \mathbb{R}_+$ is a closed and bounded set. The entire PD-PBO method is formally described in Algorithm 2.

Algorithm 2 Primal-Dual Pessimistic Bilevel Optimization (PD-PBO)

1: **Input:** stepsizes η_t, τ_t, momentum weights θ_t, output weight γ_t, initialization z_0, λ_0, and iteration times T and N
2: **for** $t = 0, 1, ..., T-1$ **do**
3: Conduct projected gradient descent in eq. (7) for N times with any given $\hat{y}_0^t$ as initialization
4: Update λ_{t+1} according to eq. (10)
5: Update z_{t+1} according to eq. (11)
6: **end for**
7: **Output:** $\bar{z} = \frac{1}{\Gamma_T} \sum_{t=0}^{T-1} \gamma_t z_{t+1}$, where $\Gamma_T = \sum_{t=0}^{T-1} \gamma_t$

4 Theoretical Analysis of SG-PBO and PD-PBO

In this section, we provide the convergence analysis of both Algorithms 1 and 2. To this end, we first make the following assumption on the regularized objective and constrained functions.

Assumption 2. *$f(z)$ and each entry of $h(z)$ are strongly convex, Lipschitz continuous and smooth. Namely, there exist positive constants μ, L and M, such that for all $z, z' \in \mathcal{Z}$,*

$$\tfrac{L}{2}\|z'-z\|_2^2 \geq J(z') - J(z) - \langle \nabla_z J(z), z'-z \rangle \geq \tfrac{\mu}{2}\|z'-z\|_2^2, \quad |J(z) - J(z')| \leq M\|z-z'\|_2,$$

where $J(z)$ represents either $f(z)$ or one of the entry functions of $h(z)$. Further, there exists a strictly feasible point $\tilde{z} \in \mathcal{Z}$.

The assumption on convexity ensures that we could efficiently solve the problem globally. Next, we introduce the following definition to characterize the criterion of convergence.

Definition 1. *Let z^* be the solution to the constrained optimization in eq.* (6) *and $\epsilon \geq 0$ be a constant. We say $z \in \mathcal{Z}$ is an ϵ-accurate solution if $f(z) \leq f(z^*) + \epsilon$ and $h(z) \leq \epsilon$.*

We characterize the convergence performance of Algorithms 1 and 2 in the following two theorems.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Let $\gamma_t = \frac{\mu(t+1)}{2}$, $T = \mathcal{O}\left(\frac{1}{\epsilon}\right)$, $N \geq \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$, and $\bar{z}$ be the output of Algorithm 1. Then we have*

$$f(\bar{z}) - f(z^*) \leq \epsilon, \quad \textit{and} \quad \max_j \{(h(\bar{z}))_j\} \leq \epsilon,$$

which indicates that $\bar{z}$ is an ϵ-accurate solution of the problem in eq. (6)*.*

Theorem 1 shows that SG-PBO could solve the problem in eq. (6) to any arbitrary accuracy level ϵ with a gradient computation complexity of $TN = \mathcal{O}\left(\frac{1}{\epsilon}\log(\frac{1}{\epsilon})\right)$. Furthermore, the computation complexity of the second order Jacobian matrix is upper-bounded by $T/(2m+3) = \mathcal{O}(\frac{1}{m\epsilon})$, since at each iteration SG-PBO at most computes one row of the matrix in line 10 of Algorithm 1. The formal statement of Theorem 1 is provided in Appendix. Compared to the standard analysis for constrained convex optimization (Boob et al., 2019; Ma et al., 2020), our analysis needs to carefully deal with the

bias error of the function estimation $\hat{h}(z_t; \hat{y}_N^t)$ and the bias error of the Jocobian matrix estimation $\widehat{\nabla} h(z_t; \hat{y}_N^t)$, which are resulted by the fact that $\hat{y}_N^t$ is only an approximation of a minimum point of the inner function of PBO.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Let* $\gamma_t = \mathcal{O}(t)$, $\eta_t = \mathcal{O}(t)$, $\tau_t = \mathcal{O}\left(\frac{1}{t}\right)$, $\theta_t = \frac{\gamma_{t+1}}{\gamma_t}$, *and* $\Lambda = \{\lambda \in \mathbb{R}^p : 0 \leq \lambda \leq B\}$ *for some positive constant* B. *Moreover, let* $T = \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$, $N = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\bar{z}$ *be the output of Algorithm 2. Then we have*

$$f_\phi(\bar{z}) - f(z^*) \leq \epsilon, \quad \textit{and} \quad \max_j\{(h(\bar{z}))_j\} \leq \epsilon,$$

which indicates that $\bar{z}$ *is an* ϵ*-accurate solution of the problem in eq.* (6).

Theorem 2 shows that PD-PBO could solve the problem in eq. (6) to any prescribed ϵ with a gradient computation complexity of $TN = \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$. Moreover, because in line 5 of Algorithm 2 (i.e., eq. (43)) PD-PBO needs the information of the entire Jacobian matrix, the computation complexity of second order oracle equals to $T = \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$. The formal statement of Theorem 2 is given in Appendix. Due to the nature of pessimistic bilevel optimization, we also have to treat the bias error of the function estimation $\hat{h}(z_t; \hat{y}_N^t)$ and the bias error of the Jacobian matrix estimation $\widehat{\nabla} h(z_t; \hat{y}_N^t)$, which makes our analysis more challenging than that in generic optimization (Boob et al., 2019).

Remark. We provide the comparison of SG-PBO and PD-PBO in Table 1. It can be seen that PD-PBO has a lower complexity on the first-order orcale compared to SG-PBO. The complexity on second-order computation depends on the dimension m and the accuracy level ϵ, i.e., if $m\sqrt{\epsilon} > 1$, SG-PBO has a better rate; otherwise, PD-PBO would be a better choice.

	first-order orcale	second-order orcale
SG-PBO	$\mathcal{O}\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{1}{m\epsilon}\right)$
PD-PBO	$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$

Table 1: Comparison between SG-PBO and PD-PBO on the first- and second-order oracle computation

5 Experiments

In this section, we conduct experimental studies on two specific problems to verify that the proposed SG-PBO and PD-PBO indeed solve the PBO.

5.1 Illustrative Example

Consider the following example:

$$\min_{x\in\mathbb{R}} \max_{y\in\mathcal{S}(x)} \quad -xy \quad s.t. \quad x^2 + y^2 - 1 \leq 0, \tag{12}$$

where $\mathcal{S}(x)$ is the set of solutions to the following inner-level optimization with a fixed $x \in \mathbb{R}$,

$$\min_{y\in\mathbb{R}} g(x,y) \coloneqq \begin{cases} |y-|x||^3, & |y| \geq |x| \\ 0, & -|x| \leq y \leq |x| \end{cases}.$$

It is clear that $\mathcal{S}(x) = \{y \in \mathbb{R} : |y| \leq |x|\}$ and $g^*(x) = 0$. For any fixed α, $g_\alpha^*(x) = 0$. More details like the KKT reformulation and the exact forms of gradients could be find in Appendix H.

Figure 1 shows the performance of both SG-PBO (Algorithm 1) and PD-PBO (Algorithm 2) in solving the problem eq. (12), where the x-axis denotes the iteration number. It is clear that both algorithms could solve the objective function to its global minimum efficiently. Besides, as illustrated in the left figure in Figure 1, SG-PBO converges at a faster rate than PD-PBO. This is because SG-PBO enforces the constraints only when the threshold ϵ is violated and will focus solely in

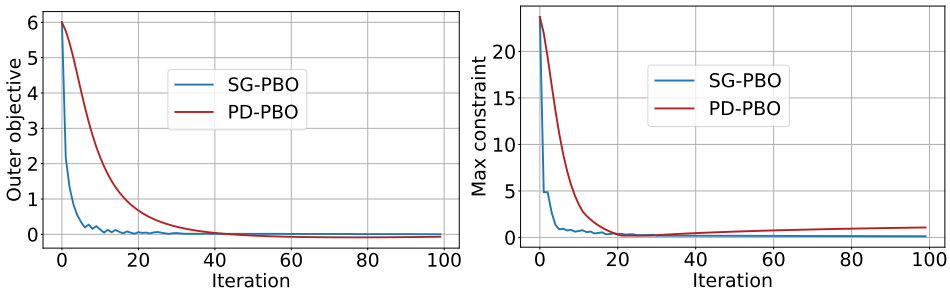

Figure 1: Comparison of PD-PBO and SG-PBO for the illustrative example in eq. (12)

minimizing the outer objective f when all the constraints are less than ϵ. Whereas PD-PBO will always minimize the Lagrangian, which may result in unnecessary delays in minimizing f when all the constraints are satisfied. Moreover, the left figure of Figure 1 indicates that the constraint violation in SG-PBO decreases much faster than that in PD-PBO. Recall that the update direction of PD-PBO is $\nabla f(z_t) + \langle \nabla h(z_t), \lambda_{t+1} \rangle$, where the i-th constraint gets penalized when the i-th entry of λ is large enough. Since PD-PBO updates the primal variables based on the constraints' value after observing the updates of λ, it is not hard to tell that the decrease of constraint violation would be slow if the stepsize for updating λ is small.

5.2 Learning Robust Hyper-representation

In the hyper-representation (HR) (Grazzi et al., 2020a; Franceschi et al., 2017) problem, the goal is to find good representations of the data that can be used for subsequent regression/classification problem by following a two-phase optimization process. The PBO framework can be used to robustly learn such representations. More specifically, we consider the following formulation:

$$\min_{\Lambda \in \mathbb{R}^{d \times m}} \max_{w^* \in \mathcal{S}_\Lambda} \mathcal{L}\left(h_\Lambda(X_1) w^*, Y_1\right) \tag{13}$$
$$\text{with } \mathcal{S}_\Lambda = \operatorname*{argmin}_{w \in \mathbb{R}^m} \mathcal{L}(h_\Lambda(X_2) w, Y_2)$$

where $h_\Lambda(\cdot)$ is the embedding model (linear transformation in this case) parameterized by the matrix Λ, and the vector w corresponds to the parameters of a linear regression/classification model. $X_1 \in \mathbb{R}^{n_1 \times d}$ and $X_2 \in \mathbb{R}^{n_2 \times d}$ are the matrices of outer (validation) and inner (training) data. $Y_1 \in \mathbb{R}^{n_1}$ and $Y_2 \in \mathbb{R}^{n_2}$ are the corresponding label vectors, respectively.

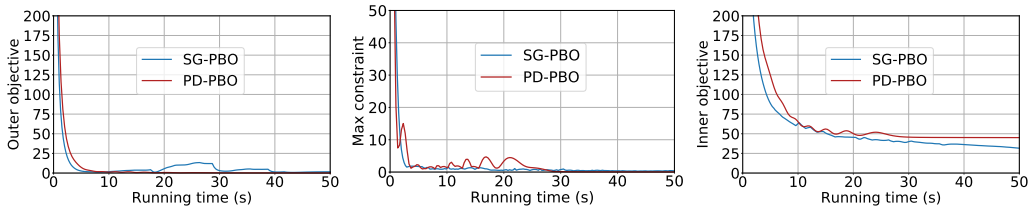

Figure 2: Comparison of PD-PBO and SG-PBO for the robust HR problem in eq. (13) with $m = 512$

Intuitively, the inner problem in eq. (13) finds the set $\mathcal{S}_\Lambda$ of best model parameters w^*, and the upper problem optimizes Λ so that the *worst* performing w^* in $\mathcal{S}_\Lambda$ yields minimal validation error. Representations learned this way are robust as they allow all minimizers in $\mathcal{S}_\Lambda$ to achieve low validation error. Note that this problem is intrinsically hard because one needs to compute the set $\mathcal{S}_\Lambda$, which can be untractable. Fortunately, our proposed algorithms SG-PBO and PD-PBO provide a way to solve problem eq. (13) without having to explicitely find the set $\mathcal{S}_\Lambda$.

In our experiments, we consider regression problems where the loss function $\mathcal{L}(\cdot, \cdot)$ corresponds to the squared ℓ_2-norm. We conduct the experiments on synthetic random data as in Grazzi et al. (2020a). The input matrices X_1 and X_2 are well conditioned and Gaussian with zero mean and unit

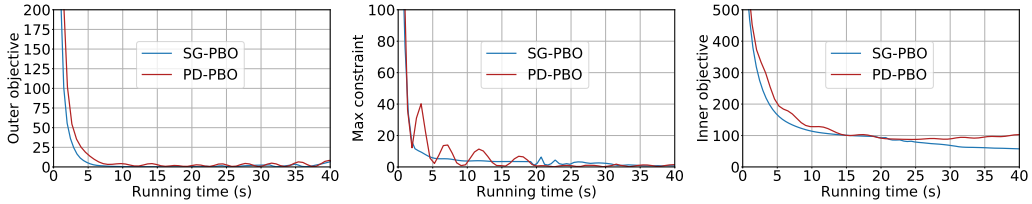

Figure 3: Comparison of PD-PBO and SG-PBO for the robust HR problem in eq. (13) with $m = 1024$

Algo	$m = 512$	$m = 1024$
SG-PBO	0.29	0.57
PD-PBO	0.32	0.65

Table 2: Iteration time (s). The running time of each iteration of SG-PBO and PD-PBO scale similarly w.r.t. the dimension m but with SG-PBO slightly faster.

variance. We generate the outputs Y_1 and Y_2 by applying a linear model on a subset of the features (20% of the features) and adding a random Gaussian noise term.

Figures 2 and 3 show the performance comparisons between SG-PBO and PD-PBO w.r.t. the running time for solving the HR problem, when the representation dimension is set to $m = 512$ and $m = 1024$, respectively. As depicted, both algorithms solve the problem within a comparable time frame, while SG-PBO is slightly faster. We note the following remarks about the plots in Figures 2 and 3, which are intuitively expected. (a) SG-PBO by design tries to minimize the maximum constraint violation and hence is more stable at achieving this goal compared to PD-PBO (middle plots in Figures 2 and 3), but this can come with a less stable minimization of the outer objective (left plot in Figure 2). (b) Because SG-PBO enforces the constraints more effectively, it also achieves a better optimization of the inner problem, which is just one of the constraints in our reformulation. The fact that SG-PBO algorithm is more sensitive to the constraint violations is intuitively expected. Indeed, during the algorithm running, whenever some certain constraints are not satisfied, then SG-PBO directly penalizes the maximum violation with no delay in line 10 of Algorithm 1. However, the PD-PBO algorithm penalizes the violated constraints through increasing the corresponding Lagrangian terms in λ, i.e. push the updating direction of z closer to the directions alleviating the violation. We provide the iteration time comparison of SG-PBO and PD-PBO in Table 2, where SG-PBO and PD-PBO scale similarly with the problem dimension m and SG-PBO is slightly faster.

6 Conclusions

In this paper, we provide the first gradient-based algorithms, namely SG-PBO, PD-PDO and Prox-PBO (in Appendix A), for pessimistic bilevel optimization. Our algorithmic design features a novel single-level reformulation which has controllable gap from the original PBO. Later, we provide the convergence rate analysis of the SG-PBO and PD-PBO and show that they both could converge sublinearly to the ϵ-accurate solution of the reformulated problem when it is strongly convex. For the general nonconvex reformulated problem, we further show the convergence of Prox-PBO to the an ϵ-KKT point in Appendix G. Since the reformulated problem can be arbitrarily close to the original PBO given suitable parameters, all the convergence guarantees for the reformulated problem immediately apply to the original PBO. Our experiments on an illustrative example and the robust hyper-representation learning problem clearly validate our algorithmic design and theoretical analysis. More importantly, beyond the novel contribution in PBO, the techniques proposed in this paper could also be applied in other areas including optimistic bilevel optimization, constrained minimax problems, and constrained bilevel optimization. We refer the reader to Appendix B for elaborated discussions.

REFERENCES

Karen Aardal, Martine Labbé, Janny Leung, and Maurice Queyranne. On the two-level uncapacitated facility location problem. *INFORMS Journal on Computing*, 8(3):289–301, 1996.

Abdelmalek Aboussoror and Abdelatif Mansouri. Weak linear bilevel programming problems: existence of solutions via a penalty method. *Journal of Mathematical Analysis and Applications*, 304(1):399–408, 2005.

Eitaro Aiyoshi and Kiyotaka Shimizu. A solution method for the static constrained stackelberg problem via penalty method. *IEEE Transactions on Automatic Control*, 29(12):1111–1114, 1984.

Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, Alexander Gasnikov, and Dmitry Kovalev. Accelerated methods for composite non-bilinear saddle point problem. *arXiv preprint arXiv:1906.03620*, 2019.

Aleksandr Y Aravkin, James V Burke, Dmitry Drusvyatskiy, Michael P Friedlander, and Scott Roy. Level-set methods for convex optimization. *Mathematical Programming*, 174(1):359–390, 2019.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *arXiv preprint arXiv:1908.02734*, 2019.

Digvijay Boob, Qi Deng, Guanghui Lan, and Yilin Wang. A feasible level proximal point method for nonconvex sparse constrained optimization. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 33:16773–16784, 2020.

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

S Dempe, J Dutta, and BS Mordukhovich. New necessary optimality conditions in optimistic bilevel programming. *Optimization*, 56(5-6):577–604, 2007.

Stephan Dempe, Boris S Mordukhovich, and Alain B Zemkoho. Necessary optimality conditions in pessimistic bilevel programming. *Optimization*, 63(4):505–533, 2014.

Stephan Dempe, Boris S Mordukhovich, and Alain B Zemkoho. Two-level value function approach to non-smooth optimistic and pessimistic bilevel programs. *Optimization*, 68(2-3):433–455, 2019.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 318–326, 2012.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, pp. 1165–1173, 2017.

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 (ICML)*, pp. 1568–1577, 2018.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. *International Conference on Machine Learning (ICML))*, 2020a.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proc. International Conference on Machine Learning (ICML)*, 2020b.

Patrick T Harker and Jong-Shi Pang. Existence of optimal solutions to mathematical programs with equilibrium constraints. *Operations research letters*, 7(2):61–64, 1988.

Chaoyang He, Haishan Ye, Li Shen, and Tong Zhang. Milenas: Efficient neural architecture search via mixed-level reformulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11993–12002, 2020.

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.

Feihu Huang and Heng Huang. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

Minhui Huang, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684*, 2022.

Kaiyi Ji. Bilevel optimization for machine learning: Algorithm design and convergence analysis. *arXiv preprint arXiv:2108.00330*, 2021.

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 parameter. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiyi Ji, Mingrui Liu, Yingbin Liang, and Lei Ying. Will bilevel optimizers benefit from loops. *arXiv preprint arXiv:2205.14224*, 2022.

Kayi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. *International Conference on Machine Learning (ICML))*, 2021.

Shihui Jia and Zhongping Wan. A penalty function method for solving ill-posed bilevel programming problem via weighted summation. *Journal of Systems Science and Complexity*, 26(6):1019–1027, 2013.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems (NeurIPS)*, pp. 1008–1014, 2000.

Lorenzo Lampariello, Simone Sagratella, and Oliver Stein. The standard pessimistic bilevel problem. *SIAM Journal on Optimization*, 29(2):1634–1656, 2019.

Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with functional or expectation constraints. *arXiv preprint arXiv:1604.03887*, 2016.

Junyi Li, Bin Gu, and Heng Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.

Maria Beatrice Lignola and Jacqueline Morgan. Existence of solutions to bilevel variational problems in banach spaces. In *Equilibrium Problems: Nonsmooth Optimization and Variational Inequality Models*, pp. 161–174. Springer, 2001.

Qihang Lin, Selvaprabu Nadarajah, and Negar Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization*, 28(4):3290–3311, 2018.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Proc. Annual Conference on Learning Theory (COLT)*, pp. 2738–2779. PMLR, 2020.

Bingbing Liu, Zhongping Wan, Jiawei Chen, and Guangmin Wang. Optimality conditions for pessimistic semivectorial bilevel programming problems. *Journal of Inequalities and Applications*, 2014(1):1–26, 2014.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning (ICML)*, 2020.

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3132674.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1540–1552. PMLR, 2020.

Roberto Lucchetti, F Mignanego, and G Pieri. Existence theorems of equilibrium points in stackelberg. *Optimization*, 18(6):857–866, 1987.

Runchao Ma, Qihang Lin, and Tianbao Yang. Proximally constrained methods for weakly convex optimization with weakly convex constraints. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Matthew MacKay, Paul Vicol, Jon Lorraine, David Duvenaud, and Roger Grosse. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. *arXiv preprint arXiv:1903.03088*, 2019.

Anton Valentinovich Malyshev and Aleksandr Sergeevich Strekalovskii. Global search for guaranteed solutions in quadratic-linear bilevel optimization problems. *The Bulletin of Irkutsk State University. Series Mathematics*, 4(1):73–82, 2011.

Olvi L Mangasarian and Stan Fromovitz. The fritz john necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and applications*, 17(1):37–47, 1967.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

JV Outrata. Necessary optimality conditions for stackelberg problems. *Journal of optimization theory and applications*, 76(2):305–320, 1993.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1723–1732, 2019.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22 (2):276–295, 2017.

Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

Heinrich von Stackelberg et al. *Theory of the market economy*. Oxford University Press, 1952.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Wolfram Wiesemann, Angelos Tsoukalas, Polyxeni-Margarita Kleniati, and Berç Rustem. Pessimistic bilevel optimization. *SIAM Journal on Optimization*, 23(1):353–380, 2013.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, 2021.

Bo Zeng. A practical scheme to compute the pessimistic bilevel optimization problem. *INFORMS Journal on Computing*, 32(4):1128–1142, 2020.

Yue Zheng, Zhongping Wan, Kangtai Sun, and Tao Zhang. An exact penalty method for weak linear bilevel programming problem. *Journal of Applied Mathematics and Computing*, 42(1):41–49, 2013.

Yue Zheng, Debin Fang, and Zhongping Wan. A solution approach to the weak linear bilevel programming problems. *Optimization*, 65(7):1437–1449, 2016.

A PROXIMAL METHOD FOR PESSIMISTIC BILEVEL OPTIMIZATION

Building on the foundation of Sections 3 and 4 where we study the pessimistic bilevel optimization in eq. (6) when it is strongly convex (Assumption 2), we next propose a proximal point method (Ma et al., 2020; Boob et al., 2019), so as to solve the constrained optimization problem eq. (5) (which is nonconvex optimization with a nonconvex constraint) with stronger convergence guarantee. More specifically, at each iteration k, we add regularizers centered at the current solution ($\tilde{z}_k$) to both the objective and constrained functions, resulting in a subproblem P_k which satisfies Assumption 2 when regularization parameter σ is large enough. Next, we call a solver to solve P_k to a $\frac{\beta}{4K}$-accurate solution, denoted by $\tilde{z}_{k+1}$, which will serve as the regularizers for the next subproblem P_{k+1}. After K iterations, the algorithm picks uniformly at random one of the $\tilde{z}_k$ as the output. More details are provided in Algorithm 3.

Algorithm 3 Proximal Pessimistic Bilevel Optimization (Prox-PBO)

1: **Input:** Number of iterations K, T, relaxation level β, regularization parameter σ, and initial point $\tilde{z}_1$.
2: **for** $k = 1, ..., K$ **do**
3: Set the kth subproblem (P_k) as

$$\begin{aligned} \min_{z\in\mathcal{Z}} \quad & f(z) + \tfrac{\sigma}{2}\|z - \tilde{z}_k\|_2^2 \\ s.t. \quad & h(z) + \tfrac{\sigma}{2}\|z - \tilde{z}_k\|_2^2 - \tfrac{k\beta}{K} \leq 0. \end{aligned} \tag{14}$$

4: Call solver to solve P_k to a $\frac{\beta}{4K}$-accurate solution
5: **end for**
6: Pick $\hat{k}$ from $\{1, \ldots, K\}$ uniformly at random
7: **Output:** $\tilde{z}_{\hat{k}}$

Remark. (1) As shown in Section 4, both SG-PBO and PD-PBO could solve P_k to any prescribed accuracy efficiently with suitable hyperparameters. By setting the accurate-level being $\frac{\beta}{4K}$, $\tilde{z}_{k+1}$ will violates the constraints of P_k by no more than $\frac{\beta}{4K}$. (2) To ensure that the next subproblem is solvable, previous studies (Ma et al., 2020; Boob et al., 2019) made strong assumptions on the original nonconvex problem, such as uniformly Slater condition and strong feasibility. However, the fact that the constraints in PBO are related to the derivative of g makes these assumptions impossible to be satisfied in general. To address this challenge, we devise a gradually enlarged relaxation scheme on the constraints in Algorithm 3, i.e., the $-\frac{k\beta}{K}$ term in the constraints in eq. (14). Through gradually increasing the relaxation by $\frac{\beta}{K}$ in each iteration, $\tilde{z}_{k+1}$ is still $\frac{\beta}{2K}$ strictly feasible for constraints in next subproblem, even if it may violate the current constraints by $\frac{\beta}{2K}$. The analysis of Algorithm 3 is postponed to Appendix G, after the proofs of convergences of SG-PBO and PD-PBO.

B DISCUSSIONS AND EXTENSIONS

In this section, we discuss a few other problems, for which our proposed approach can serve as either a new or the first-known algorithm with the convergence rate characterization.

B.1 PESSIMISTIC VS OPTIMISTIC BILEVEL OPTIMIZATION PROBLEMS

We compare the two main classes of bilevel problems, i.e., optimistic bilevel optimization (OBO) and pessimistic bilevel optimization (PBO). Recall that the optimistic bilevel optimization is given by the following form:

$$\min_{x\in\mathcal{X}, y\in\mathcal{S}(x)} f(x,y), \quad \text{with} \quad \mathcal{S}(x) = \arg\min_{y\in\mathbb{R}^m} g(x,y).$$

One approach to solve such a problem is to reformulate it as in eq. (3) as follows:

$$\min_{x\in\mathcal{X}, y\in\mathbb{R}^m} f(x,y) \quad \text{s.t.} \quad g(x,y) - g^*(x) \leq 0.$$

It is clear that the above form is already a single-level constrained optimization, because the outer-level is taken as $\min_{y\in\mathcal{Y}}$, which is the same as $\min_{x\in\mathcal{X}}$, so that (x, y) can be viewed as a joint vector to be minimized.

However, the pessimistic bilevel optimization studied in this paper is very different. Even in the reformulation in eq. (3), the outer-level is a minimax form over $\min_{x\in\mathcal{X}}$ and $\max_{y\in\mathcal{Y}}$, and hence (x, y) cannot be treated as one joint vector yet. Instead, KKT-condition is applied to reformulate the constrained maximization over y, so as to change the problem to a single-level constrained optimization. Thus, the final structure of such a reformulated problem is very different from that resulted from the OBO problem.

The Prox-PBO algorithm we provide here can be applied to solve the reformulated problem in OBO, and the convergence rate can be developed in a similar way. This serves as an alternative approach to the primal-dual approach recently developed in Sow et al. (2022) for OBO.

B.2 Connection to Constrained Minimax Problems

As we stated in Section 2.1, the original bilevel optimization is equivalent to the **constrained minimax problem** defined in eq. (3). We write such a constrained minimax optimization in a generic form as follows:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y) \qquad s.t. \quad h(x, y) \leq 0. \tag{15}$$

Suppose for any given x, $f(x, y)$ is concave on y, and $h(x, y)$ is convex on y. The problem in eq. (15) is an open problem, and no algorithm has been proposed with theoretical guarantee. Interestingly, it is clear that our approach in this paper solves the above problem by first reformulating it into a single-level optimization problem as in Proposition 2 and then applying our Prox-PBO algorithm. This also serves as the **first-known approach** with provable convergence rate guarantee for solving the problem.

We next briefly comment that it appears challenging to solve the above constrained minimax optimization directly (without the reformulation) in eq. (15) by applying the conventional primal-dual method for constrained optimization, even if $f(x, y)$ is convex-concave, and $h(x, y)$ is convex on y for any given x. Namely, define the Lagrangian function as $\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda h(x, y)$, and define the dual function as $\psi(x, \lambda) := \max_{y\in\mathcal{Y}} f(x, y) - \lambda h(x, y)$. Based on the convexity of $h(x, y)$ and concavity of $f(x, y)$ on y, the strong duality still holds here, and hence $\min_{x,\lambda} \psi(x, \lambda)$ will provide the solution of the original problem. However, it is easy to show that $\mathcal{L}(x, y, \lambda)$ is not convex jointly on (x, λ) for each fixed y, and hence $\psi(x, y)$ is in general nonconvex. It appears challenging here to design an efficient algorithm with certain convergence guarantee for solving such a dual problem.

B.3 Constrained Bilevel Optimization

Our approach can also be applied to the constrained bilevel problem, where the outer-level problem has additional constraint, as described below:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{S}(x)} f(x, y), \quad s.t. \quad F(x, y) \leq 0, \text{ where } \mathcal{S}(x) := \arg\min_{y\in\mathbb{R}^m} g(x, y).$$

For such a more general problem, we can still reformulate the problem as in eq. (3), and then simply treat the constraint on $F(x, y)$ as an additional constraint and further reformulate the problem as in Proposition 2. We can finally apply our Prox-PBO algorithm to solve the reformulated problem. The convergence guarantee can be similarly derived.

C Proof of Proposition 1

To prove the proposition, we may further assume that, for every $x \in \mathcal{X}$ and $y \in \mathbb{R}^m$ such that $\nabla_y g(x, y) \neq 0$, $\lambda_{min}(\nabla^2_{yy} g(x, y)) > \kappa$, where $\lambda_{min}(\cdot)$ denotes the minimum eigenvalue of a matrix, and $\kappa > 0$ is a constant. This assumption requires that the function $g(x, y)$ is discriminative enough. Such an assumption is standard in the literature (Li et al., 2020).

Given an $x \in \mathcal{X}$, it is clear that

$$\{y \in \mathbb{R}^m : g(x, y) - g^*(x) \leq 0\} \subseteq \{y \in \mathbb{R}^m : g(x, y) - g^*_\alpha(x) - \xi \leq 0\}.$$

Thus, we have

$$\Phi(x) = \max_{y\in\mathbb{R}^m}\{f(x,y) : g(x,y)-g^*(x)\leq 0\} \leq \max_{y\in\mathbb{R}^m}\{f(x,y) : g(x,y)-g^*_\alpha(x)-\xi\leq 0\} = \Phi_{\alpha,\xi}(x). \tag{16}$$

Moreover, suppose $y^*(x) \in \{y\in\mathbb{R}^m : g(x,y)-g^*(x)\leq 0\}$ and $y^*_{\alpha,\xi}(x) \in \{y\in\mathbb{R}^m : g(x,y)-g^*_\alpha(x)-\xi\leq 0\}$ satisfying $f(x,y^*(x)) = \Phi(x)$ and $f(x,y^*_{\alpha,\xi}(x)) = \Phi_{\alpha,\xi}(x)$. Then, there exist two conditions:

(a). Suppose $y^*_{\alpha,\xi}(x) \in \{y\in\mathbb{R}^m : g(x,y)-g^*(x)\leq 0\}$. Then, by the definition of $\Phi(x)$, we have

$$\Phi_{\alpha,\xi}(x) = f(x,y^*_{\alpha,\xi}(x)) \leq \max_{y\in\mathbb{R}^m}\{f(x,y) : g(x,y)-g^*(x)\leq 0\} = \Phi(x). \tag{17}$$

(b). Suppose $y^*_{\alpha,\xi}(x) \notin \{y\in\mathbb{R}^m : g(x,y)-g^*(x)\leq 0\}$. Because $g(x,y)$ is convex on y, $\mathcal{S}(x) = \{y\in\mathbb{R}^m : g(x,y)-g^*(x)\leq 0\}$ is a convex set. Let $\tilde{y}$ be the orthogonal projection of $y^*_{\alpha,\xi}(x)$ on $\mathcal{S}(x)$. Since $\tilde{y}\in\mathcal{S}(x)$, we have

$$\begin{aligned}
g(x,y^*_{\alpha,\xi}(x)) - g^*(x) &= g(x,y^*_{\alpha,\xi}(x)) - g(x,\tilde{y}) \\
&= \int_{t=0}^{1} \langle \nabla_y g(x,\tilde{y}+t(y^*_{\alpha,\xi}(x)-\tilde{y})), y^*_{\alpha,\xi}(x)-\tilde{y}\rangle dt \\
&= \int_{t=0}^{1} \left\langle \int_{s=0}^{t} \nabla^2_{yy} g(x,\tilde{y}+s(y^*_{\alpha,\xi}(x)-\tilde{y}))ds, y^*_{\alpha,\xi}(x)-\tilde{y}\right\rangle dt \\
&= \int_{t=0}^{1}\int_{s=0}^{t} (y^*_{\alpha,\xi}(x)-\tilde{y})^\top \nabla^2_{yy} g(x,\tilde{y}+s(y^*_{\alpha,\xi}(x)-\tilde{y}))(y^*_{\alpha,\xi}(x)-\tilde{y}))dsdt \\
&\overset{(i)}{\geq} \tfrac{\kappa}{2}\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2^2,
\end{aligned} \tag{18}$$

where (i) follows from the facts that, for any $s\in[0,t]\subseteq[0,1]$, $\nabla_y g(x,y(s)) \neq 0$, where we denote $y(s) \coloneqq \tilde{y} + s(y^*_{\alpha,\xi}(x)-\tilde{y})$ for short, and thus

$$(y^*_{\alpha,\xi}(x)-\tilde{y})^\top \nabla^2_{yy} g(x,y(s))(y^*_{\alpha,\xi}(x)-\tilde{y}) \geq \lambda_{min}(\nabla^2_{yy} g(x,y(s))\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2^2 > \kappa\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2^2.$$

Moreover, it is clear that

$$g(x,y^*_{\alpha,\xi}(x)) \overset{(i)}{\leq} g^*_\alpha(x)+\xi \overset{(ii)}{\leq} g^*(x) + \tfrac{\alpha}{2}D_{\mathcal{Y}}^2 + \xi, \tag{19}$$

where (i) follows from the fact that $y^*_{\alpha,\xi}(x) \in \{y\in\mathbb{R}^m : g(x,y)-g^*_\alpha(x)-\xi\leq 0\}$, and (ii) follows from $g^*_\alpha(x) \leq g_\alpha(x,y^*(x)) = g(x,y^*(x)) + \frac{\alpha}{2}\|y^*(x)\|_2^2 \leq g^*(x) + \frac{\alpha}{2}D_{\mathcal{Y}}^2$.

Combining eqs. (18) and (19), we obtain

$$\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2 \leq \sqrt{\tfrac{2}{\kappa}(\tfrac{D_{\mathcal{Y}}^2}{2}\alpha+\xi)}. \tag{20}$$

By the Lipschitz continuity of $f(x,y)$, there exists $M>0$ such that

$$\begin{aligned}
f(x,y^*_{\alpha,\xi}(x)) &\leq f(x,\tilde{y}) + M\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2 \\
&\overset{(i)}{=} f(x,y^*(x)) + M\|y^*_{\alpha,\xi}(x)-\tilde{y}\|_2 \\
&\overset{(ii)}{\leq} f(x,y^*(x)) + M\sqrt{\tfrac{2}{\kappa}(\tfrac{D_{\mathcal{Y}}^2}{2}\alpha+\xi)},
\end{aligned} \tag{21}$$

where (i) follows from $\tilde{y}\in\{y\in\mathbb{R}^n : g(x,y)-g^*(x)\leq 0\}$, and (ii) follows from eq. (20).

Equation (21) implies that $\Phi_{\alpha,\xi}(x) \leq \Phi(x) + \mathcal{O}(\sqrt{\xi}) + \mathcal{O}(\sqrt{\alpha})$. Together with eqs. (16) and (17), we complete the proof.

D Proof of Proposition 2

Lemma 1. *For any given $x\in\mathcal{X}$, consider the following constrained optimization problem.*

$$\min_{y\in\mathbb{R}^m} \quad -f(x,y)$$

$$s.t. \quad g(x,y) - g_\alpha^*(x) - \xi \le 0. \tag{22}$$

There exists $y^*(x) \in \mathcal{Y}$ that attains the solution of the above problem. Moreover, there exists $w^*(x) \ge 0$, such that the following KKT condition holds.

$$\begin{aligned} -\nabla_y f(x, y^*(x)) + w^*(x)\nabla_y g(x,y) = 0 \\ w^*(x) g(x, y^*(x)) - g_\alpha^*(x) - \xi) = 0. \end{aligned} \tag{23}$$

For all $w^*(x)$ satisfying the above KKT condition, we have $w^*(x) \le \frac{\Delta_f}{\xi}$ with

$$\Delta_f \coloneqq \max_{x,x' \in \mathcal{X}, y, y' \in \mathcal{Y}} |f(x,y) - f(x',y')|.$$

Proof. Given $x \in \mathcal{X}$, let $\tilde{y} \in \mathcal{S}(x)$. We have $g(x,\tilde{y}) - g_\alpha^*(x) - \xi \le -\xi$. Thus, $\tilde{y}$ is a strictly feasible point with margin ξ for the problem in eq. (22).

Define the dual function $d(w) = \min_{y \in \mathbb{R}^m} -f(x,y) + w(g(x,y) - g_\alpha^*(x) - \xi)$. By its definition, we have, for any $w \in \mathbb{R}_+$ and $y \in \mathbb{R}^m$,

$$d(w) \le -f(x,\tilde{y}) + w(g(x,y) - g_\alpha^*(x) - \xi) = -f(x,\tilde{y}) - w\xi, \tag{24}$$

Moreover, it is known that convex constrained optimization has no duality gap Lan (2020). And the existence of $\tilde{y}$ ensures the Slater's condition holds. Therefore, the existence of $y^*(x)$ and $w^*(x)$ is ensured. And, eq. (23) is the necessary and sufficient condition for the optimality of eq. (22). In the other words, $d(w^*(x)) = d^* = p^* = -f(x, y^*(x))$. Taking $w = w^*(x)$ in eq. (24), we obtain

$$-f(x, y^*(x)) = d(w^*(x)) \le -f(x,\tilde{y}) - w^*(x)\xi.$$

Rearranging terms in the above inequality, we have

$$w^*(x) \le \frac{f(x, y^*(x)) - f(x,\tilde{y})}{\xi} \overset{(i)}{\le} \frac{\Delta_f}{\xi}.$$

where (i) follows from the definition of Δ_f. □

Proposition 3 (Restatement of Proposition 2)**.** *The minimax problem eq.* (4) *is equivalent to the following constrained optimization:*

$$\begin{aligned} \min_{z \in \mathcal{Z}} \quad & f(z) \\ s.t. \quad & h(z) \coloneqq \begin{pmatrix} g(x,y) - g_\alpha^*(x) - \xi \\ -\nabla_y f(x,y) + w\nabla_y g(x,y) \\ \nabla_y f(x,y) - w\nabla_y g(x,y) \\ w(g(x,y) - g_\alpha^*(x) - \xi) \\ -w(g(x,y) - g_\alpha^*(x) - \xi) \end{pmatrix} \le 0, \end{aligned} \tag{25}$$

where $z = (x,y,z)$, $\mathcal{W} \coloneqq [0, \frac{\Delta_f}{\xi}]$, with $\Delta_f \coloneqq \max_{x,x' \in \mathcal{X}, y, y' \in \mathcal{Y}} |f(x,y) - f(x',y')|$, and $\mathcal{Y} \coloneqq \{y \in \mathbb{R}^m : \|y\|_2 \le D_\mathcal{Y}\}$ with $D_\mathcal{Y} > 0$, such that, for all $x \in \mathcal{X}$, $\{y \in \mathbb{R}^m : g(x,y) - g_\alpha^*(x) - \xi \le 0\} \subseteq \mathcal{Y}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \times \mathcal{W}$, and $f(z) = f(x,y)$.

Proof. Let $p^* = \min_{x \in \mathcal{X}} \{\psi(x) \coloneqq \max_{y \in \mathbb{R}^m} \{f(x,y) : g(x,y) - g_\alpha^*(x) - \xi \le 0\}\}$ be the solution of eq. (4). And let $p_r^* = \min_{x \in \mathcal{X}} \psi_r(x)$ be the solution of eq. (25), with

$$\begin{aligned} \psi_r(x) \coloneqq \min_{y \in \mathcal{Y}, w \in \mathcal{W}} \quad & f(x,y) \\ s.t. \quad & g(x,y) - g_\alpha^*(x) - \xi \le 0 \\ & -\nabla_y f(x,y) + w\nabla_y g(x,y) = 0 \\ & w(g(x,y) - g_\alpha^*(x) - \xi) = 0. \end{aligned} \tag{26}$$

By Lemma 1, the feasible set for a given $x \in \mathcal{X}$ of eq. (26) is non-empty, i.e., there exist at least $(y^*(x), w^*(x)) \in \mathcal{Y} \times \mathcal{W}$ satisfying all three constraints, which implies $\psi_r(x) \le +\infty$. Moreover, for all (y,w) in the feasible set of eq. (26), we have it satisfies the KKT condition and $g(x,y) - g_\alpha^*(x) - \xi \le 0$, which the sufficient condition for y to be the solution of eq. (4), i.e., $f(x,y) = \psi(x)$. Therefore, we have $\psi(x) = \psi_r(x)$ for all $x \in \mathcal{X}$, which complete the proof. □

E PROOF OF THEOREM 1

Lemma 2 (Theorem 2.2.14 Nesterov et al. (2018))**.** *Suppose Assumption 1 holds Consider the gradient descent in eq.* (7)*. We have*

$$\|\hat{y}_N^t - y_\alpha^*(x_t)\|_2 \leq \left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N \|\hat{y}_0 - \tilde{y}^*(x_t)\|_2.$$

Lemma 3. *(Three-point lemma, (Lan, 2020, Lemma 3.5)). Given $\mathcal{Z} \subseteq \mathbb{R}^q$ is a convex an closed set, let $z_{t+1} = \Pi_{\mathcal{Z}}(z_t - G)$, where $G \in \mathbb{R}^q$. Then, for any point $z \in \mathcal{Z}$, we have*

$$\langle G, z - z_{t+1}\rangle \geq \tfrac{1}{2}\|z - z_{t+1}\|_2^2 + \tfrac{1}{2}\|z_{t+1} - z_t\|_2^2 - \tfrac{1}{2}\|z - z_t\|_2^2.$$

Lemma 4. *Suppose Assumptions 1 and 2 hold. Let $H(z) \coloneqq \max_j \left\{(h(z))_j\right\}$. Consider i_t, $\widehat{h}(z_t; \hat{y}_N^t)$ and $\widehat{\nabla} h(z_t; \hat{y}_N^t)$ specified in Algorithm 1. We have*

$$\left|(\widehat{h}(z_t; \hat{y}_N^t))_{i_t} - H(z_t)\right| \leq (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N.$$

Moreover, let $\widehat{\partial} H(z_t) = (\widehat{\nabla} h(z_t; \hat{y}_N^t))_{i_t}$, we have for all $z \in \mathcal{Z}$,

$$H(z) \geq H(z_t) + \langle \widehat{\partial} H(z_t), z - z_t\rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 - 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N.$$

Proof. Because each entry of $h(z)$ is a strongly convex function, we have $H(z)$ is also a strongly convex function. Moreover, for any given $z \in \mathcal{Z}$, let $I(z) \coloneqq \arg\max_j \left\{(h(z))_j\right\}$, we have $\nabla(h(z))_{I(z)} \in \partial H(z)$.

(a). Suppose $I(z_t) = i_t$.

Observing the form of $\hat{h}(z_t; \hat{y}_N^t)$, only its first and last two entries do not equal to $h(z_t)$. Thus, we have

$$\begin{aligned}
\left|\left(\widehat{h}(z_t; \hat{y}_N^t)\right)_{i_t} - H(z_t)\right| &\leq \max\left\{|g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t)|, |w_t(g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t))|\right\} \\
&\overset{(i)}{\leq} D_{\mathcal{Z}} |g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t)| \\
&\overset{(ii)}{\leq} (L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}} \|\hat{y}_N^t - y_\alpha^*(x_t)\|_2 \\
&\overset{(iii)}{\leq} (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N,
\end{aligned} \tag{27}$$

where (i) follows from $w_t \leq D_{\mathcal{Z}}$, (ii) follows from that $g_\alpha(z)$ is $(L_g + \alpha)D_{\mathcal{Y}}$ Lipschitz continuous, and (iii) follows from Lemma 2.

It is clear that $\widehat{\partial} H(z_t) - \partial H(z_t) \neq 0$ if and only if i_t selects the first or the last two constraints, i.e., $\|\widehat{\partial} H(z_t) - \partial H(z_t)\|_2$ equals one of the following three: 0, $\|((\nabla_x g_\alpha(x_t, \hat{y}_N^t) - \nabla_x g_\alpha^*(x))^\top, 0, 0)\|_2$, or $\|(w_t(\nabla_x g_\alpha(x_t, \hat{y}_N^t) - \nabla_x g_\alpha^*(x_t))^\top, 0, g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t))\|_2$. Thus, we have

$$\begin{aligned}
\|\widehat{\partial} H(z_t) - \partial H(z_t)\|_2 &\leq \sqrt{w_t^2 \|\nabla_x g_\alpha(x_t, \hat{y}_N^t) - \nabla_x g_\alpha^*(x_t)\|_2^2 + \|g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t)\|_2^2} \\
&\overset{(i)}{\leq} w_t \|\nabla_x g_\alpha(x_t, \hat{y}_N^t) - \nabla_x g_\alpha^*(x_t)\|_2 + \|g_\alpha(x_t, \hat{y}_N^t) - g_\alpha^*(x_t)\|_2 \\
&\overset{(ii)}{\leq} D_{\mathcal{Z}}(L_g + \alpha)\|y_\alpha^*(x_t) - \hat{y}_N^t\|_2 + (L_g + \alpha) D_{\mathcal{Z}} \|y_\alpha^*(x_t) - \hat{y}_N^t\|_2 \\
&\overset{(iii)}{\leq} 2(L_g + \alpha) D_{\mathcal{Z}} D_{\mathcal{Y}} \left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N,
\end{aligned} \tag{28}$$

where (i) follows from the $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$, (ii) follows from $\nabla_x g(x, y)$ is $L_g + \alpha$ gradient Lipschitz, $w_t \leq D_{\mathcal{Z}}$, and $g_\alpha(x, y)$ is $(L_g + \alpha)D_{\mathcal{Z}}$ Lipschitz continuous, and (iii)

follows from Lemma 2. Following the definition of $\partial H(z_t)$, strong convexity, and Cauchy Schwartz inequality, we obtain

$$H(z) \geq H(z_t) + \langle \widehat{\partial} H(z_t), z - z_t \rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 - 2(L_g + \alpha) D_{\mathcal{Z}}^2 D_{\mathcal{Y}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N. \quad (29)$$

(b). Suppose $I(z_t) \neq i_t$.

Similar to eq. (27), we have $\left|(\widehat{h}(z_t; \hat{y}_N^t))_{i_t} - (h(z_t))_{i_t}\right| \leq (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N$ and

$$\left|(\widehat{h}(z_t; \hat{y}_N^t))_{I(z_t)} - H(z_t)\right| \leq (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N.$$

Together with the facts that $(\widehat{h}(z_t; \hat{y}_N^t))_{I(z)} \leq (\widehat{h}(z_t; \hat{y}_N^t))_{i_t}$ and $H(z_t) \geq (h(z_t))_{i_t}$, we have

$$\left|(\widehat{h}(z_t; \hat{y}_N^t))_{i_t} - H(z_t)\right| \leq (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N \quad (30)$$

$$H(z_t) - 2(L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N \leq (h(z_t))_{i_t} \leq H(z_t). \quad (31)$$

Given $z \in \mathcal{Z}$, following the strong convexity of $(h(z))_{i_t}$, we have

$$\begin{aligned}
H(z) \geq (h(z))_i &\geq (h(z_t))_{i_t} + \langle \nabla (h(z_t))_{i_t}, z - z_t \rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 \\
&\overset{(i)}{\geq} H(z_t) - 2(L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N + \langle \widehat{\partial} H(z_t), z - z_t \rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 \\
&\qquad + \langle \nabla (h(z_t))_{i_t} - \widehat{\partial} H(z_t), z - z_t \rangle \\
&\overset{(ii)}{\geq} H(z_t) + \langle \widehat{\partial} H(z_t), z - z_t \rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 - 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N, \quad (32)
\end{aligned}$$

where (i) follows from eq. (31) and (ii) follow from eq. (28), Cauchy-Schwartz inequality and $D_{\mathcal{Y}} \leq D_{\mathcal{Z}}$.

Thus, from eqs. (27) and (30), we conclude

$$\left|(\widehat{h}(z_t; \hat{y}_N^t))_{i_t} - H(z_t)\right| \leq (L_g + \alpha) D_{\mathcal{Y}}^2 D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N.$$

From eqs. (29) and (32), we conclude

$$H(z) \geq H(z_t) + \langle \widehat{\partial} H(z_t), z - z_t \rangle + \tfrac{\mu}{2}\|z - z_t\|_2^2 - 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N.$$

□

Theorem 3 (Formal Statement of Theorem 1)**.** *Suppose Assumptions 1 and 2 holds. Consider Algorithm 1. Let* $\gamma_t = \frac{\mu(t+1)}{2}$, $T \geq \frac{4M^2}{\mu\epsilon}$, *and* $N \geq \log\left(\frac{\epsilon}{4(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2}\right) / \log(1 - \frac{\alpha}{L_g+2\alpha})$. *Then, we have*

$$f(\bar{z}) - f(z^*) \leq \epsilon, \quad \textit{and} \quad \max_j \left\{(h(\bar{z}))_j\right\} \leq \epsilon.$$

In the other words, we have $\bar{z}$ *is an* ϵ*-accurate solution of eq.* (6)*.*

Proof. We let $H(z) = \max_j \{(h(z))_j\}$ for short.

(a). Suppose $t \in \mathcal{T}$, we have $\hat{h}(z_t; \hat{y}_N^t) \leq \frac{\epsilon}{2}$. Applying Lemma 3 the update with respect to the $\nabla f(z)$ ensures that, for any given $z \in \mathcal{Z}$,

$$\gamma_t^{-1} \langle \nabla f(z_t), z - z_{t+1} \rangle \geq \tfrac{1}{2}\|z - z_{t+1}\|_2^2 + \tfrac{1}{2}\|z_{t+1} - z_t\|_2 - \tfrac{1}{2}\|z - z_t\|_2^2. \quad (33)$$

Moreover, using the strongly convexity of $f(z)$ (Assumption 2), we obtain

$$f(z^*) \geq f(z_t) + \langle \nabla f(z_t), z^* - z_t \rangle + \tfrac{\mu}{2}\|z^* - z_t\|_2^2. \quad (34)$$

Taking $z = z^*$ in eq. (33) and using eq. (34), we have

$$f(z_t) - f(z^*) \leq \langle \nabla f(z_t), z_t - z_{t+1} \rangle - \frac{\gamma_t}{2}\|z_{t+1} - z_t\|_2^2 + \frac{\gamma_t - \mu}{2}\|z^* - z_t\|_2^2 - \frac{\gamma_t}{2}\|z^* - z_{t+1}\|_2^2$$
$$\overset{(i)}{\leq} \frac{\|\nabla f(z_t)\|_2^2}{2\gamma_t} + \frac{\gamma_t - \mu}{2}\|z^* - z_t\|_2^2 - \frac{\gamma_t}{2}\|z^* - z_{t+1}\|_2^2. \tag{35}$$

where (i) follows from the Young's inequality, $\langle \nabla f(z_t), z_t - z_{t+1} \rangle \leq \frac{\|\nabla f(z_t)\|_2^2}{2\gamma_t} + \frac{\gamma_t}{2}\|z_t - z_{t+1}\|_2^2$.

(b). Suppose $t \notin \mathcal{T}$, we have $\hat{h}(z_t; \hat{y}_N^t) > \frac{\epsilon}{2}$, Applying Lemma 3 the update with respect to the $\widehat{\nabla}(h(z_t; \hat{y}_N^t))_{i_t}$ (we denote as $\widehat{\partial} H(z_t)$ for short) ensures that, for any given $z \in \mathcal{Z}$,

$$\gamma_t^{-1}\langle \widehat{\partial} H(z), z - z_{t+1} \rangle \geq \tfrac{1}{2}\|z - z_{t+1}\|_2^2 + \tfrac{1}{2}\|z_{t+1} - z_t\|_2 - \tfrac{1}{2}\|z - z_t\|_2^2. \tag{36}$$

Moreover, applying Lemma 4 with $z = z^*$, we obtain

$$H(z^*) \geq H(z_t) + \langle \widehat{\partial} H(z_t), z^* - z_t \rangle + \tfrac{\mu}{2}\|z^* - z_t\|_2^2 - 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N. \tag{37}$$

Take $z = z^*$ in eq. (36) and recall eq. (37). We have

$$\begin{aligned}
&H(z_t) - H(z^*) \\
&\quad \leq \langle \widehat{\partial} H(z_t), z_t - z_{t+1} \rangle + \tfrac{\gamma_t - \mu}{2}\|z^* - z_t\|_2^2 - \tfrac{\gamma_t}{2}\|z^* - z_{t+1}\|_2^2 - \tfrac{\gamma_t}{2}\|z_{t+1} - z_t\|_2^2 \\
&\qquad + 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N \\
&\quad \overset{(i)}{\leq} \tfrac{\|\widehat{\partial} H(z_t)\|_2^2}{2\gamma_t} + \tfrac{\gamma_t - \mu}{2}\|z^* - z_t\|_2^2 - \tfrac{\gamma_t}{2}\|z^* - z_{t+1}\|_2^2 + 4(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N,
\end{aligned} \tag{38}$$

where (i) follows from applying Young's inequality.

Proceeding with the following inductions.

$$\begin{aligned}
&\sum_{t \in \mathcal{T}} \gamma_t (f(z_t) - f(z^*)) + \sum_{t \in [T], t \notin \mathcal{T}} \gamma_t H(z_t) \\
&\overset{(i)}{\leq} \sum_{t \in \mathcal{T}} \gamma_t (f(z_t) - f(z^*)) + \sum_{t \in [T], t \notin \mathcal{T}} \gamma_t (H(z_t) - H(z^*)) \\
&\overset{(ii)}{\leq} \sum_{t \in \mathcal{T}} \tfrac{1}{2}\|\nabla f(z_t)\|_2^2 + \sum_{t \in [T], t \neq \mathcal{T}} \left(\tfrac{1}{2}\|\widehat{\partial} H(z_t)\|_2^2 + 4\gamma_t (L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N \right) \\
&\qquad + \sum_{t=1}^{T} \left(\tfrac{\mu^2 (t-1) t}{8}\|z^* - z_t\|_2^2 - \tfrac{\mu^2 t (t+1)}{8}\|z^* - z_{t+1}\|_2^2 \right) \\
&= \sum_{t \in \mathcal{T}} \tfrac{1}{2}\|\nabla f(z_t)\|_2^2 + \sum_{t \in [T], t \neq \mathcal{T}} \left(\tfrac{1}{2}\|\widehat{\partial} H(z_t)\|_2^2 + 2\mu(t+1)(L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N \right) \\
&\leq \frac{M^2 T}{2} + \mu (T+2)^2 (L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N,
\end{aligned} \tag{39}$$

where (i) follows from $H(z^*) \leq 0$, (ii) follows from eqs. (35) and (38), and (iii) follows from $\gamma_t = \frac{\mu(t+1)}{2}$, $(t-1)(t+1) \leq t(t-1)$ and $t(t+1) \leq (t+1)^2$.

Recall that, for all $t \in \mathcal{T}$, we have $\hat{h}(z_t; \hat{y}_N^t) \leq \frac{\epsilon}{2}$. Applying Lemma 4, we have, for all $t \in \mathcal{T}$,

$$H(z_t) \leq \frac{\epsilon}{2} + (L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N. \tag{40}$$

Applying Lemma 4, we have, for all $t \notin \mathcal{T}$,

$$H(z_t) \geq \frac{\epsilon}{2} - (L_g + \alpha) D_{\mathcal{Y}} D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N.$$

Multiplying γ_t on both sides of the above inequality and telescoping, we obtain

$$\begin{aligned}\sum_{t\in[T],t\neq\mathcal{T}}\gamma_t H(z_t) &\geq \sum_{t\in[T],t\notin\mathcal{T}}\gamma_t\left(\epsilon-(L_g+\alpha)D_{\mathcal{Y}}^2D_{\mathcal{Z}}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N\right)\\ &\geq \frac{\epsilon}{2}\sum_{t\in[T],t\notin\mathcal{T}}\gamma_t-\mu(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N.\end{aligned}$$

Substituting the above inequality into eq. (39), we obtain

$$\begin{aligned}\sum_{t\in\mathcal{T}}\gamma_t(f(z_t)-f(z^*)) &\leq -\frac{\epsilon}{2}\sum_{t\in[T],t\notin\mathcal{T}}\gamma_t+\tfrac{M^2T}{2}+2\mu(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N\\ &\overset{(i)}{\leq}\frac{\epsilon}{2}\sum_{t\in\mathcal{T}}\gamma_t-\frac{\epsilon\mu T^2}{8}+\frac{M^2T}{2}+2\mu(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N,\end{aligned}$$

where (i) follows from $-\sum_{t\in[T],t\notin\mathcal{T}}\gamma_t=\sum_{t\in\mathcal{T}}\gamma_t-\sum_{t\in[T]}\gamma_t$ and $\sum_{t\in[T]}\gamma_t\geq\frac{\mu T^2}{4}$.

Dividing $\sum_{t\in\mathcal{T}}\gamma_t$ on both side of the above inequality and using the fact $\sum_{t\in\mathcal{T}}\gamma_t\geq\mu$, we obtain

$$\frac{\sum_{t\in\mathcal{T}}\gamma_t(f(z_t)-f(z^*))}{\sum_{t\in\mathcal{T}}\gamma_t}\leq\frac{\epsilon}{2}+\frac{\frac{M^2T}{2}-\frac{\mu\epsilon T^2}{8}}{\sum_{t\in\mathcal{T}}\gamma_t}+2(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N. \tag{41}$$

By the convexity of $f(z)$ and eq. (41), we have

$$f(\bar{z})-f(z^*)\leq\frac{\epsilon}{2}+\frac{\frac{M^2T}{2}-\frac{\mu\epsilon T^2}{8}}{\sum_{t\in\mathcal{T}}\gamma_t}+2(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N.$$

Finally using the convexity of $H(z)$, and eq. (40), we obtain

$$\max_j\left\{\left(h(\bar{z})\right)_j\right\}=H(\bar{z})\leq\frac{\epsilon}{2}+(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N.$$

Recall $N\geq\log\left(\frac{\epsilon}{4(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2}\right)/\log(1-\frac{\alpha}{L_g+2\alpha})$ and $T\geq\frac{4M^2}{\mu\epsilon}$, we have $f(\bar{z})-f(z^*)\leq\epsilon$, and $\max_j\left\{\left(h(\bar{z})\right)_j\right\}\leq\epsilon$. □

F Proof of Theorem 2

Before the proof of Theorem 2, we first prove that the optimal dual variable is upper-bounded.

Lemma 5. *Suppose Assumption 2 holds. We have the optimal dual λ^* exists and $\|\lambda^*\|_1$ satisfies $\|\lambda^*\|_1\leq\frac{f(\tilde{z})-f(z^*)}{-\max_i\{(h(\tilde{z}))_i\}}\coloneqq B_0$.*

Proof. Recall that convex constrained optimization has no duality gap Lan (2020). Then the existence of $\tilde{z}$ ensures that the Slater's condition holds. Therefore, the existence of λ^* is ensured, and the following inequality holds

$$f(z^*)=f(z^*)+\langle h(z^*),\lambda^*\rangle\leq f(\tilde{z})+\langle h(\tilde{z}),\lambda^*\rangle\leq f(\tilde{z})+\|\lambda^*\|_1\max_i\{(h(\tilde{z}))_i\}.$$

Rearranging terms in the above inequality, we have

$$\|\lambda^*\|_1\leq\frac{f(\tilde{z})-f(z^*)}{-\max_i\{(h(\tilde{z}))_i\}}.$$

□

We first provide the formal statement of the theorem and then provide the convergence.

Theorem 4 (Formal Statement of Theorem 2)**.** *Suppose Assumptions 1 and 2 hold. Consider Algorithm 2. Let* $\gamma_t = t + t_0 + 3$, $\eta_t = \frac{\rho_f(t+t_0+1)}{2}$, $\tau_t = \frac{4(L_g+2\rho_h D_{\mathcal{Z}})^2}{\rho_f(t+1)}$, $\theta_t = \frac{t+t_0+2}{t+t_0+3}$, *where* $t_0 = \frac{\rho_f + B\rho_h}{\rho_f} + 1$, $B = B_0 + 1$ *and* B_0 *defined in Lemma 5. We have*

$$f(\bar{z}) - f(z^*) \leq \frac{2L_g BD_{\mathcal{Z}}}{\Gamma_T}\left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{(L_g + 2\rho_h D_{\mathcal{Z}})BD_{\mathcal{Z}}}{\Gamma_T} + (L_g D_{\mathcal{Z}} + 3L_g)BD_{\mathcal{Z}}\left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{\gamma_0(\eta_0 - \rho_f)\|z^* - z_0\|_2^2}{2\Gamma_T},$$

$$\max_j \{(h(\bar{z}))_j\} \leq \left((L_g D_{\mathcal{Z}} + 3L_g)BD_{\mathcal{Z}} + \frac{2L_g BD_{\mathcal{Z}}}{\Gamma_T}\right)\left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{(L_g + 2\rho_h D_{\mathcal{Z}})BD_{\mathcal{Z}}}{\Gamma_T} + + \frac{\gamma_0\tau_0}{2\Gamma_T}(\lambda^* - \lambda_0)^2 + \frac{\gamma_0(\eta_0 - \rho_f)}{2\Gamma_T}\|z_k^* - z_0\|_2^2,$$

and

$$\|\tilde{z} - z^*\|_2^2 \leq \frac{1}{\gamma_{T-1}(\eta_{T-1} - 3(\rho_f + B\rho_h))}\left(4L_g BD_{\mathcal{Z}}\left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N + 2(L_g + 2\rho_h D_{\mathcal{Z}})BD_{\mathcal{Z}}\right) + \frac{1}{\gamma_{T-1}(\eta_{T-1} - 3(\rho_f + B\rho_h))}\left(\gamma_0\tau_0(\lambda^* - \lambda_0)^2 + \gamma_0(\eta_0 - \rho_f)\|z^* - z_0\|_2^2\right) + \frac{\Gamma_T}{\gamma_{T-1}(\eta_{T-1} - 3(\rho_f + B\rho_h))}(L_g D_{\mathcal{Z}} + 3L_g)BD_{\mathcal{Z}}\left(1 - \tfrac{\alpha}{L_g+2\alpha}\right)^N.$$

The proof is as follow.

We first define some notations that will be used later. Let $\hat{d}_t = (1+\theta_t)\hat{h}(z_t; \hat{y}_N^t) - \theta_t \hat{h}(z_{t-1}; \hat{y}_N^{t-1})$, $d_t = (1+\theta_t)h(z_t) - \theta_t h(z_{t-1})$, and $\xi_t = \hat{h}(z_t; \hat{y}_N^t) - \hat{h}(z_{t-1}; \hat{y}_N^{t-1})$. Moreover, we specify $\mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t) = f(z_t) + \langle \lambda_{t+1}, \hat{h}(z_t; \hat{y}_N^t)$ and the gradient of Laguragian as $\widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t) = \nabla f(z_t) + \langle \lambda_{t+1}, \widehat{\nabla} h(z_t; \hat{y}_N^t)\rangle$. Further define the primal-dual gap function as

$$Q(w, \tilde{w}) \coloneqq f(z) + \tilde{\lambda}h(z) - \left(f(\tilde{z}) + \lambda h(\tilde{z})\right),$$

where $w = (z, \lambda)$, $w = (\tilde{z}, \tilde{\lambda}) \in \mathcal{Z} \times \Lambda$ are primal-dual pairs.

Consider the update of λ in eq. (10). Applying Lemma 3 with $G = -\hat{d}_t/\tau_t$, $\mathcal{Z} = \Lambda$, $\bar{z} = \lambda_{t+1}$, $z = \lambda_t$ and letting $\tilde{z} = \lambda$ be an arbitrary point inside Λ, we have

$$-(\lambda_{t+1} - \lambda)\hat{d}_t \leq \frac{\tau_t}{2}\left((\lambda - \lambda_t)^2 - (\lambda_{t+1} - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right). \tag{42}$$

Similarly, consider the update of z in eq. (11). Applying Lemma 3 with $G = \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t)/\eta_t$, we obtain

$$\langle \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t), z_{t+1} - z\rangle \leq \frac{\eta_t}{2}\left((z - z_t)^2 - (z_{t+1} - z_t)^2 - (z - z_{t+1})^2\right). \tag{43}$$

Recall that $f(z)$ and $h(z)$ are L-gradient Lipschitz. This implies

$$\langle \nabla f(z_t), z_{t+1} - z_t\rangle \geq f(z_{t+1}) - f(z_t) - \frac{L\|z_t - z_{t+1}\|_2^2}{2}, \tag{44}$$

$$\langle \nabla h(z_t), z_{t+1} - z_t\rangle \geq h(z_{t+1}) - h(z_t) - \frac{L\|z_t - z_{t+1}\|_2^2}{2}. \tag{45}$$

Moreover, recall that $f(z)$ and $h(z)$ are μ-strongly convex functions. These two properties yield

$$\langle \nabla f(z_t), z_t - z\rangle \geq f(z_t) - f(z) + \frac{\mu\|z - z_t\|_2^2}{2}, \tag{46}$$

$$\langle \nabla h(z_t), z_t - z\rangle \geq h(z_t) - h(z) + \frac{\mu\|z - z_t\|_2^2}{2}. \tag{47}$$

Consider the exact gradient of Lagrangian with respect to the primal variable, we have

$$\begin{aligned}
&\langle \nabla_z \mathcal{L}(z_t, \lambda_{t+1}), z_{t+1} - z \rangle \\
&\quad = \langle \nabla f(z_t) + \lambda_{t+1} \nabla h(z_t), z_{t+1} - z_t \rangle \\
&\quad = \langle \nabla f(z_t), z_{t+1} - z \rangle + \langle \nabla f(z_t), z - z_t \rangle + \lambda_{t+1} \langle \nabla h(z_t), z_{t+1} - z \rangle + \lambda_{t+1} \langle \nabla h(z_t), z - z_t \rangle \\
&\quad \overset{(i)}{\geq} f(z_{t+1}) - f(z) + \lambda_{t+1}(h(z_{t+1}) - h(z)) - \frac{L(1+\lambda_{t+1})\|z_{t+1} - z_t\|_2^2}{2} + \frac{\mu(1+\lambda_{t+1})\|z - z_t\|_2^2}{2},
\end{aligned} \tag{48}$$

where (i) follows from combining eqs. (44) to (47).

Combining eqs. (43) and (48) yields

$$\begin{aligned}
f(z_{t+1}) - f(z) \leq \langle \nabla_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t), z_{t+1} - z \rangle + \lambda_{t+1}(h(z) - h(z_{t+1})) \\
+ \frac{\eta_t - \mu(1+\lambda_{t+1})}{2}\|z - z_t\|_2^2 - \frac{\eta_t - L(1+\lambda_{t+1})}{2}\|z_{t+1} - z_t\|_2^2 \\
- \frac{\eta_t}{2}\|z - z_{t+1}\|_2^2.
\end{aligned} \tag{49}$$

Recall the definition of $\xi_t = \hat{h}(z_t; \hat{y}_N^t) - \hat{h}(z_{t-1}; \hat{y}_N^{t-1})$. Substituting it into eq. (42) yields

$$\begin{aligned}
0 \leq -(\lambda - \lambda_{t+1})\hat{h}(z_t; \hat{y}_N^t) - (\lambda_{t+1} - \lambda)\xi_{t+1} + \theta_t(\lambda_{t+1} - \lambda)\xi_t \\
+ \frac{\tau_t}{2}\left((\lambda - \lambda_t)^2 - (\lambda_{t+1} - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right).
\end{aligned} \tag{50}$$

Let $w = (z, \lambda)$ and $w_{t+1} = (z_{t+1}, \lambda_{t+1})$. By the definition of the primal-dual gap function, we have

$$\begin{aligned}
&Q(w_{t+1}, w) \\
&\quad = f(z_{t+1}) + \lambda h(z_{t+1}) - f(z) - \lambda_{t+1} h(z) \\
&\quad \overset{(i)}{\leq} \langle \nabla_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t), z_{t+1} - z \rangle + (\lambda - \lambda_{t+1}) h(z_{t+1}) \\
&\qquad + \frac{\eta_t - \mu(1+\lambda_{t+1})}{2}\|z - z_t\|_2^2 - \frac{\eta_t - L(1+\lambda_{t+1})}{2}\|z_{t+1} - z_t\|_2^2 - \frac{\eta_t}{2}\|z - z_{t+1}\|_2^2 \\
&\quad \overset{(ii)}{\leq} \langle \nabla_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t), z_{t+1} - z \rangle + (\lambda - \lambda_{t+1})(h(z_{t+1}) - \hat{h}(z_{t+1}; \hat{y}_N^{t+1}) \\
&\qquad - (\lambda_{t+1} - \lambda)\xi_{t+1} + \theta_t(\lambda_{t+1} - \lambda)\xi_t + \frac{\tau_t}{2}\left((\lambda - \lambda_t)^2 - (\lambda_{t+1} - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right) \\
&\qquad + \frac{\eta_t - \mu}{2}\|z - z_t\|_2^2 - \frac{\eta_t - L(B+1)}{2}\|z_{t+1} - z_t\|_2^2 - \frac{\eta_t}{2}\|z - z_{t+1}\|_2^2,
\end{aligned} \tag{51}$$

where (i) follows from eq. (49) and (ii) follows from eq. (50) and $0 \leq \lambda_{t+1} \leq B$.

Now we proceed with $|h(z_t) - \hat{h}(z_t; \hat{y}_N^t)|$.

$$|h(z_t) - \hat{h}(z_t; \hat{y}_N^t)| = |g(x_t, y_\alpha^*) - g(x_t, \hat{y}_N^t)| \overset{(i)}{\leq} 2L_g\|y_\alpha^* - \hat{y}_N^t\|_2 \overset{(ii)}{\leq} L_g D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N, \tag{52}$$

where (i) follows from Assumption 1 and (ii) follows from the following Lemma 2 and $\|\hat{y}_0^t - y_\alpha^*(x_t)\|_2 \leq D_{\mathcal{Z}}$.

The following inequality follows immediately from the above eq. (52).

$$(\lambda - \lambda_{t+1})(h(z_t) - \hat{h}(z_t; \hat{y}_N^t)) \leq |\lambda - \lambda_{t+1}||h(z_t) - \hat{h}(z_t; \hat{y}_N^t)| \leq L_g B D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha}\right)^N. \tag{53}$$

By the definitions of $\nabla_z \mathcal{L}(z_t, \lambda_{t+1})$ and $\widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t)$, we have

$$\|\nabla_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t)\|_2$$

$$
\begin{aligned}
&= \left\| \nabla f(z_t) + \lambda_{t+1} \nabla h(z_t) - \left(\nabla f(z_t) + \lambda_{t+1} \widehat{\nabla} h(z_t; \hat{y}_N^t) \right) \right\|_2 \\
&= \lambda_{t+1} \left\| \nabla g(x_t, y_\alpha^*(x_t)) - g(x_t, \hat{y}_N^t) \right\|_2 \overset{(i)}{\leq} \lambda_{t+1} L_g \| y_\alpha^*(x_t) - \hat{y}_N^t \|_2 \\
&\overset{(ii)}{\leq} B L_g D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N, \qquad (54)
\end{aligned}
$$

where (i) follows from Assumption 1 and (ii) follows from Lemma 2, and because $\lambda_{t+1} \leq B$ and $\|\hat{y}_0 - \tilde{y}^*(x_t)\|_2 \leq D_{\mathcal{Z}}$.

By Cauchy-Schwartz inequality and eq. (54), we have

$$
\begin{aligned}
&\langle \nabla_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t), z_{t+1} - z \rangle \\
&\quad \leq \| \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}) - \widehat{\nabla}_z \mathcal{L}(z_t, \lambda_{t+1}; \hat{y}_N^t) \|_2 \| z_{t+1} - z \|_2 \leq B L_g D_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N. \qquad (55)
\end{aligned}
$$

By the definition of ξ_t, we have

$$
\begin{aligned}
\theta_t(\lambda_{t+1} - \lambda_t)\xi_t &= \theta_t(\lambda_{t+1} - \lambda_t)(\hat{h}(z_t; \hat{y}_N^t) - \hat{h}(z_{t-1}; \hat{y}_N^{t-1})) \\
&= \theta_t(\lambda_{t+1} - \lambda_t)(\hat{h}(z_t; \hat{y}_N^t) - h(z_t) - \hat{h}(z_{t-1}; \hat{y}_N^{t-1}) + h(z_{t-1}) + h(z_t) - h(z_{t-1})) \\
&\leq \theta_t |\lambda_{t+1} - \lambda_t| \left(|\hat{h}(z_t; \hat{y}_N^t) - h(z_t)| + |\hat{h}(z_{t-1}; \hat{y}_N^{t-1}) - h(z_{t-1})| + |h(z_t) - h(z_{t-1})| \right) \\
&\overset{(i)}{\leq} |\lambda_{t+1} - \lambda_t| \left(2 L_g D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N + M \| z_t - z_{t-1} \|_2 \right) \\
&\overset{(ii)}{\leq} 2 B L_g D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N + M |\lambda_{t+1} - \lambda_t| \| z_t - z_{t-1} \|_2 \\
&\overset{(iii)}{\leq} 2 B L_g D_{\mathcal{Z}} \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N + \frac{\tau_t}{2} (\lambda_{t+1} - \lambda_t)^2 + \frac{M^2}{2\tau_t} \| z_t - z_{t-1} \|_2^2, \qquad (56)
\end{aligned}
$$

where (i) follows from eq. (52), $\theta_t \leq 1$, and $h(z)$ is M Lipschitz continuous, (ii) follows from $0 \leq \lambda_t, \lambda_{t+1} \leq B$, and (iii) follows from Young's inequality.

Substituting eqs. (53), (55) and (56) into eq. (51) yields

$$
\begin{aligned}
Q(w_{t+1}, w) \leq &-(\lambda_{t+1} - \lambda)\xi_{t+1} + \theta_t(\lambda_t - \lambda)\xi_t + 4LBD_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N \\
&+ \frac{\tau_t}{2} \left((\lambda - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2 \right) + \frac{\eta_t - \mu}{2} \| z - z_t \|_2^2 - \frac{\eta_t}{2} \| z - z_{t+1} \|_2^2 \\
&+ \frac{M^2}{2\tau_t} \| z_t - z_{t-1} \|_2^2 - \frac{\eta_t - L(1 + B)}{2} \| z_{t+1} - z_t \|_2^2. \qquad (57)
\end{aligned}
$$

Recall that γ_t, θ_t, η_t and τ_t are set to satisfy $\gamma_{t+1}\theta_{t+1} = \gamma_t$, $\gamma_t \tau_t \geq \gamma_{t+1}\tau_{t+1}$ and

$$
\gamma_t (L(1 + B) - \eta_t) + \frac{\gamma_{t+1} M^2}{\tau_{t+1}} \leq 0.
$$

Multiplying γ_t on both sides of eq. (57) and telescoping from $t = 0, 1, \ldots T - 1$ yield

$$
\begin{aligned}
\sum_{t=0}^{T-1} \gamma_t Q(w_{t+1}, w) \leq &-\gamma_{T-1}(\lambda_T - \lambda)\xi_T + 4LBD_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N \sum_{t=0}^{T-1} \gamma_t \\
&+ \frac{\gamma_0 \tau_0}{2} (\lambda - \lambda_0)^2 + \frac{\gamma_0 (\eta_0 - \mu)}{2} \| z - z_0 \|_2^2 \\
&- \frac{\gamma_{T-1}(\eta_{T-1} - L(B + 1))}{2} \| z - z_T \|_2^2.
\end{aligned}
$$

Divide both sides of the above inequality by $\Gamma_T = \sum_{t=0}^{T-1} \gamma_t$. We obtain

$$
\frac{1}{\Gamma_T} \sum_{t=0}^{T-1} \gamma_t Q(w_{t+1}, w) \leq -\tfrac{\gamma_{T-1}(\lambda_T - \lambda)\xi_T}{\Gamma_T} + 4LBD_{\mathcal{Z}}^2 \left(1 - \tfrac{\alpha}{L_g + 2\alpha} \right)^N
$$

$$+ \frac{\gamma_0\tau_0}{2\Gamma_T}(\lambda-\lambda_0)^2 + \frac{\gamma_0(\eta_0-\rho_f)}{2\Gamma_T}\|z-z_0\|_2^2 - \frac{\gamma_{T-1}(\eta_{T-1}-3(\rho_f+B\rho_h))}{2\Gamma_T}\|z-z_T\|_2^2. \tag{58}$$

Similarly to the steps in eq. (56), we have

$$\begin{aligned}|(\lambda_T-\lambda)\xi_T| &\le |\lambda_T-\lambda|\left(2L_gD_{\mathcal{Z}}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + M\|z_T-z_{T-1}\|_2\right)\\ &\le 2L_gBD_{\mathcal{Z}}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + MBD_{\mathcal{Z}}.\end{aligned}$$

Define $\bar{w} \coloneqq \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t w_{t+1}$. Noting that $Q(\cdot, w)$ is a convex function and substituting the above inequality into eq. (58) yield

$$\begin{aligned}Q(\bar{w},w) &\le \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t Q(w_{t+1},w)\\ &\le \frac{2L_gBD_{\mathcal{Z}}}{\Gamma_T}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{(L_g+2\rho_hD_{\mathcal{Z}})BD_{\mathcal{Z}}}{\Gamma_T}\\ &\quad + (L_gD_{\mathcal{Z}}+3L_g)BD_{\mathcal{Z}}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{\gamma_0(\eta_0-\rho_f)}{2\Gamma_T}\|z-z_0\|_2^2\\ &\quad + \frac{\gamma_0\tau_0}{2\Gamma_T}(\lambda-\lambda_0)^2 - \frac{\gamma_{T-1}(\eta_{T-1}-3(\rho_f+B\rho_h))}{2\Gamma_T}\|z-z_T\|_2^2.\end{aligned} \tag{59}$$

Let $w = (z_k^*, 0)$. Then, we have

$$Q(\tilde{w}_k, w) = f_k(\tilde{z}_k) - f_k(z_k^*) - \bar{\lambda}_T h_k(z_T^*) \overset{(i)}{\ge} f_k(\tilde{z}_k) - f_k(z_k^*),$$

where (i) follows from the fact $h_k(z_T^*) \le 0$ and $\bar{\lambda}_T = \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t\lambda_{t+1} \ge 0$.

Substituting the above inequality into eq. (59) yields

$$\begin{aligned}f_k(\tilde{z}_k) - f_k(z_k^*) &\le \frac{2L_gBD_{\mathcal{Z}}}{\Gamma_T}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{(L_g+2\rho_hD_{\mathcal{Z}})BD_{\mathcal{Z}}}{\Gamma_T}\\ &\quad + (L_gD_{\mathcal{Z}}+3L_g)BD_{\mathcal{Z}}\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{\gamma_0(\eta_0-\rho_f)\|z_k^*-z_0\|_2^2}{2\Gamma_T}.\end{aligned}$$

Recall that (z_k^*, λ_k^*) is a Nash equilibrium of $\mathcal{L}_k(z,\lambda)$, we have

$$\mathcal{L}_k(\tilde{z}_k, \lambda_k^*) \ge \mathcal{L}_k(z_k^*, \lambda_k^*) \overset{\text{by def.}}{\iff} f_k(\tilde{z}_k) + \lambda_k^* h_k(\tilde{z}_k) - f_k(z_k^*) \ge 0 \tag{60}$$

Let $w = (z_k^*, (\lambda_k^*+1)\mathbf{I}(h_k(\tilde{z}_k))$, where $\mathbf{I}(x) = 0$ if $x \le 0$ and $\mathbf{I}(x) = 1$ otherwise. If $h_k(\tilde{z}_k) \le 0$, the constraint is satisfied. If $h_k(\tilde{z}_k) > 0$, we have

$$Q(\tilde{w}_k, w) = f_k(\tilde{z}_k) + (\lambda_k^*+1)h_k(\tilde{z}_k) - f_k(z_k^*) - \lambda_k^* h_k(\tilde{z}_k). \tag{61}$$

Recall that (z_k^*, λ_k^*) satisfies the KKT condition of (P_k), i.e. $\lambda_k^* h_k(z_k^*) = 0$. Equations (59) to (61) together yield,

$$\begin{aligned}h_k(\tilde{z}_k) &= Q(\tilde{w}_k, w) - (f_k(\tilde{z}_k) + \lambda_k^* h_k(\tilde{z}_k) - f_k(z_k^*)) \le Q(\tilde{w}_k, w)\\ &\le \left(\frac{2L_gBD_{\mathcal{Z}}}{\Gamma_T} + (L_gD_{\mathcal{Z}}+3L_g)BD_{\mathcal{Z}}\right)\left(1-\tfrac{\alpha}{L_g+2\alpha}\right)^N + \frac{(L_g+2\rho_hD_{\mathcal{Z}})BD_{\mathcal{Z}}}{\Gamma_T}\\ &\quad + \frac{\gamma_0\tau_0}{2\Gamma_T}(\lambda_k^*+1)^2 + \frac{\gamma_0(\eta_0-\rho_f)}{2\Gamma_T}\|z_k^*\|_2^2.\end{aligned}$$

G Analysis of Prox-PBO

Since the problem in eq. (5) generally has a nonconvex objective function and nonconvex constraints, we aim to provide the convergence guarantee to an ϵ-KKT point (Ma et al., 2020; Boob et al., 2019) as defined below.

Definition 2. *Consider the constrained optimization problem in eq.* (5)*. Let q be the dimension of $h(z)$ and $\mathcal{N}(z;\mathcal{Z})$ be the normal cone to $\mathcal{Z}$ at z. Denote $\operatorname{dist}(z,\mathcal{N}) \coloneqq \min_{z'\in\mathcal{N}}\{\|z-z'\|_2\}$. A point $\hat{z}\in\mathcal{Z}$ is an ϵ-KKT point if and only if there exist $z\in\mathcal{Z}$ and $\lambda\in\mathbb{R}^q_+$, such that $h(z)\le\epsilon$, $\|z-\hat{z}\|_2^2\le\epsilon$, $\sum_{i=1}^q|\lambda_i h_i(z)|\le\epsilon$, and $\operatorname{dist}\left(\nabla f(z)+\langle\nabla h(z),\lambda\rangle,-\mathcal{N}(z;\mathcal{Z})\right)^2\le\epsilon$. Further, a random $\hat{z}\in\mathcal{Z}$ is a stochastic ϵ-KKT point if there exist $z\in\mathcal{Z}$ and $\lambda\ge 0$ such that the same requirements of ϵ-KKT hold in expectation.*

The KKT condition is the necessary condition for local optimality (Bertsekas, 1997; Mangasarian & Fromovitz, 1967) for constrained optimization. Here, we will show that Algorithm 3 converges to an ϵ-KKT point in expectation taken over the randomness of the algorithm (the randomness of generation index $\hat{k}$) for constrained nonconvex optimization problems. Before the analysis, we make the following boundedness assumption on the optimal dual variable, which is standard in the literature (Boob et al., 2020; 2019).

Assumption 3. *For each subproblem P_k, the optimal dual variable λ_k^* is uniformly bounded, i.e., there exists a constant $B\ge 0$ such that $\|\lambda_k^*\|_1\le B$ holds for all $k=1,\dots,K$.*

Theorem 5. *Suppose Assumption 1 holds. Given $\tilde{z}_1$ that is $\frac{\beta}{2K}$ strictly feasible of* (P_1)*. Let $\sigma=2\max\{L_f,L_c\}$, where L_c is defined in Lemma 6. Let $\mu=\frac{\sigma}{2}$, $D_{\mathcal{Z}}=\max_{z,z'\in\mathcal{Z}}\|z-z'\|_2$, and $M=\max\{M_f,M_h\}+\sigma D_{\mathcal{Z}}$, where $M_f=\max_{z\in\mathcal{Z}}\{\|\nabla f(z)\|_2\}$ and $M_h=\max_{z\in\mathcal{Z},i\in[q]}\{\|\nabla(h(z))_i\|_2\}$. Set $K=\frac{3B\sigma}{\mu\epsilon}$, $\beta=\min\{\frac{\epsilon}{4B},\frac{\Delta_f}{\mu}\}$, $T=\frac{8KM^2}{\mu\epsilon}$, $\gamma_t=\frac{\mu(t+1)}{2}$, and $N=\log\left(\frac{\epsilon}{4(T+2)^2(L_g+\alpha)D_{\mathcal{Y}}D_{\mathcal{Z}}^2}\right)/\log(1-\frac{\alpha}{L_g+2\alpha})$. Then we have $\tilde{z}_{\hat{k}}$ is an ϵ-KKT point of eq.* (5) *in expectation that takes the randomness over $\hat{k}$.*

Theorem 5 shows that Algorithm 3 is capable to solve the problem in eq. (5) to arbitrarily prescribed accuracy ϵ with $\mathcal{O}(\frac{1}{\epsilon})$ times call of the subproblem solver. Since Assumption 2 holds for each subproblem, the first- and second-order oracle consumption immediately follows by applying Theorems 1 and 2. Compared with results in standard constrained optimization (Ma et al., 2020; Boob et al., 2019), the problem here in eq. (5) is more challenging and our proof features the following three ingredients: (a) dealing with the bias error of the Jacobian matrix estimation, (b) proving that the objective and constrained functions are weakly convex functions, and (c) designing the gradually relaxed scheme so as to ensure the strict feasibility of each subproblem.

G.1 Supporting Lemmas

Lemma 6. *Given a function $J:\mathbb{R}^n\to\mathbb{R}$, which is twice differentiable and is a L_J-gradient Lipschitz function on the bounded support $\mathcal{X}\subseteq\mathbb{R}^n$, and for all $x\in\mathcal{X}$, $\|\nabla J(x)\|_2\le M_J$. Then, define a new function $I:\mathcal{X}\times[0,B]\to\mathbb{R}$ as $I(x,y)=yJ(x)$. We have $I(x,y)$ is a (BL_J+M_J)- gradient Lipschitz function.*

Proof. By the definition of $I(x,y)$, we have its gradient $\nabla I(x,y)=[\nabla_x I(x,y);\frac{\partial I(x,y)}{\partial y}]$ equals $[y\nabla_x J(x);J(x)]$. And its Hessian equals

$$\nabla^2 I(x,y)=\begin{pmatrix} y\nabla_{xx}^2 J(x) & \nabla_x J(x)\\ (\nabla_x J(x))^\top & 0\end{pmatrix},$$

where we let $\nabla^2=\nabla^2_{(x,y),(x,y)}$.

Let $z=(a,b)^\top\in\mathbb{R}^n$, with $a\in\mathbb{R}^n$ and $b\in\mathbb{R}$, for any $x\in\mathcal{X}$ and $0\le y\le B$, we have

$$\begin{aligned} z^\top\nabla^2 I(x,y)z &= ya^\top\nabla_{xx}^2 J(x)a+2b\cdot a^\top\nabla_x J(x)\\ &\overset{(i)}{\le} yL_J\|a\|_2^2+2b\|a\|_2\|\nabla_x J(x)\|_2\\ &\overset{(ii)}{\le} yL\|z\|_2^2+(\|a\|_2^2+b^2)\|\nabla_x J(x)\|_2\\ &\le (BL_J+M_J)\|z\|_2^2, \end{aligned}$$

where (i) follows from the L_J gradient Lipschitz condition of $J(x)$ and Cauchy-Schwartz inequality and (ii) follows from the Young's inequality. □

Lemma 7. *Suppose Assumption 1 holds. Consider the constrained function $h(z)$ in eq.* (5)*. Each entry of $h(z)$ is a L_c gradient Lipschitz function with*

$$L_c = \max\left\{2\mu_f + \tfrac{4\mu_g\Delta_f}{\xi} + 4M_{Hg} + L_g, \tfrac{\Delta_f}{\xi}(2L_g + \tfrac{L_g^2}{\alpha}) + M_g\right\},$$

where $M_g = \sup_{z\in\mathcal{Z}} \|\nabla_z(g(x,y) - g_\alpha^*(x))\|_2$ and $M_{Hg} = \sup_{z\in\mathcal{Z}} \|\nabla^2_{zz} g(z)\|_F$.

Proof. In the following proof, we consider each component of the $h(z)$ and prove that they are L_c gradient Lipschitz.

For the first component $g(x,y) - g_\alpha^*(x) - \xi$, it has been shown to be $(2L_g + L_g^2/\alpha)$-gradient Lipschitz (Sow et al., 2022, Lemma 1). The next m components of $h(z)$ is the entries of $-\nabla_y f(x,y) + w\nabla_y g(x,y)$. Consider the ith entry. For any given z and $z' \in \mathcal{Z}$, let $e_i^+(z) = (-\nabla_y f(x,y) + w\nabla_y g(x,y))_i$, we have

$$\begin{aligned}
&\|\nabla e_i^+(z) - \nabla e_i^+(z')\|_2^2 \\
&\quad= \left\|\nabla\left(-\nabla_y f(x,y) + w\nabla_y g(x,y)\right)_i - \nabla\left(-\nabla_y f(x',y') + w'\nabla_y g(x',y')\right)_i\right\|_2^2 \\
&\quad\overset{(i)}{=} \left\|\left(-\nabla^2_{yx} f(x,y) + w\nabla^2_{yx} g(x,y)\right)_{(i,\cdot)} - \left(-\nabla^2_{yx} f(x',y') + w'\nabla^2_{yx} g(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad+ \left\|\left(-\nabla^2_{yy} f(x,y) + w\nabla^2_{yy} g(x,y)\right)_{(i,\cdot)} - \left(-\nabla^2_{yy} f(x',y') + w'\nabla^2_{yy} g(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad\quad+ \left((\nabla_y g(x,y))_i - (\nabla_y g(x',y'))_i\right)^2 \\
&\quad\overset{(ii)}{\leq} 2\left\|\left(\nabla^2_{yx} f(x,y) - \nabla^2_{yx} f(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad+ 2\left\|\left(\nabla^2_{yx} g(x,y) - \nabla^2_{yx} g(x',y')\right)_{(i,\cdot)} + (w-w')\left(\nabla^2_{yx} g(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad\quad+ 2\left\|\left(\nabla^2_{yy} f(x,y) - \nabla^2_{yy} f(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad\qquad+ 2\left\|\left(\nabla^2_{yy} g(x,y) - \nabla^2_{yy} g(x',y')\right)_{(i,\cdot)} + (w-w')\left(\nabla^2_{yy} g(x',y')\right)_{(i,\cdot)}\right\|_2^2 \\
&\qquad\qquad\quad+ \left((\nabla_y g(x,y))_i - (\nabla_y g(x',y'))_i\right)^2
\end{aligned} \tag{62}$$

where (i) follows from $\|\nabla_z h\|_2^2 = \|\nabla_x h\|_2^2 + \|\nabla_y h\|_2^2 + (\frac{\partial h}{\partial w})^2$ and $(M)_{(i,\cdot)}$ denotes the ith row of the matrix M, and (ii) follows from the fact $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$.

Using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$, eq. (62) induces

$$\begin{aligned}
&\|\nabla e_i^+(z) - \nabla e_i^+(z')\|_2 \\
&\quad\leq 2\left\|\left(\nabla^2_{yx} f(x,y) - \nabla^2_{yx} f(x',y')\right)_{(i,\cdot)}\right\|_2 \\
&\qquad+ 4\left\|v\left(\nabla^2_{yx} g(x,y) - \nabla^2_{yx} g(x',y')\right)_{(i,\cdot)}\right\|_2 + 4\left\|(w-w')\left(\nabla^2_{yx} g(x',y')\right)_{(i,\cdot)}\right\|_2 \\
&\qquad\quad+ 2\left\|\left(\nabla^2_{yy} f(x,y) - \nabla^2_{yy} f(x',y')\right)_{(i,\cdot)}\right\|_2 \\
&\qquad\qquad+ 4\left\|w\left(\nabla^2_{yy} g(x,y) - \nabla^2_{yy} g(x',y')\right)_{(i,\cdot)}\right\|_2 + 4\left\|(w-w')\left(\nabla^2_{yy} g(x',y')\right)_{(i,\cdot)}\right\|_2 \\
&\qquad\qquad\quad+ \left|(\nabla_y g(x,y))_i - (\nabla_y g(x',y'))_i\right| \\
&\quad\leq \left(2\mu_f + \tfrac{4\mu_g\Delta_f}{\xi} + 4M_{Hg} + L_g\right)\|z - z'\|_2,
\end{aligned} \tag{63}$$

where $M_{Hg} = \sup_{z\in\mathcal{Z}} \|\nabla^2 g\|_F$.

Next, let $e_i^-(z) = (\nabla_y f(x,y) - w\nabla_y g(x,y))_i$. Following the same steps in eqs. (62) and (63), we also obtain

$$\|\nabla e_i^-(z) - \nabla e_i^-(z')\|_2 \leq \left(2\mu_f + \tfrac{4\mu_g\Delta_f}{\xi} + 4M_{Hg} + L_g\right)\|z - z'\|_2 \tag{64}$$

For the last two components, $w(g(x,y) - g_\alpha^*(x) - \xi)$ and $-w(g(x,y) - g_\alpha^*(x) - \xi)$, because $g(x,y) - g_\alpha^*(x)$ is $(2L_g + \frac{L_g^2}{\alpha})$-gradient Lipschitz. Moreover, since the support $\mathcal{Z}$ is bounded, there exist M_g, such that $\|\nabla(g(x,y) - g_\alpha^*(x) - \xi)\|_2 \leq M_g$, and w is bounded in interval $[0, \frac{\Delta_f}{\xi}]$. Applying Lemma 6, we have $w(g(x,y) - g_\alpha^*(x) - \xi)$ and $-w(g(x,y) - g_\alpha^*(x) - \xi)$ are $\frac{\Delta_f}{\xi}(2L_g + \frac{L_g^2}{\alpha}) + M_g$ gradient Lipschitz. □

Lemma 8. *Suppose Assumptions 1 and 3 hold. And suppose the input $\tilde{z}_1$ is strictly feasible with respect to P_k with margin $\frac{\beta}{2K}$. Moreover, suppose $\tilde{z}_{k+1}$ is $\frac{\beta}{2K}$-accurate solution of P_k. Then, we have*

$$\frac{1}{K}\sum_{k=1}^{K} \|\tilde{z}_k - z_k^*\|_2^2 \leq \frac{2\Delta_f}{\mu K} + \frac{\beta}{K},$$

with $\Delta_f = \max_{z,z' \in \mathcal{Z}} |f(z) - f(z')|$.

Proof. For $\tilde{z}_{k+1}$ with $k \geq 1$, the $\frac{\beta}{2K}$-accuracy implies that

$$f_k(\tilde{z}_{k+1}) - f_k(z_k^*) \leq \tfrac{\beta}{2K}, \tag{65}$$
$$h_k(\tilde{z}_{k+1}) \leq \tfrac{\beta}{2K}. \tag{66}$$

Then, we have $h(\tilde{z}_{k+1}) = h_k(\tilde{z}_{k+1}) + \frac{k\beta}{K} - \frac{\sigma}{2}\|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2 \leq \frac{(2k+1)\beta}{2K}$, which immediately implies that $h_{k+1}(\tilde{z}_{k+1}) = h(\tilde{z}_{k+1}) - \frac{(k+1)\beta}{K} \leq -\frac{\beta}{2K}$. Thus, given that $\tilde{z}_1$ is $\frac{\beta}{2K}$ strictly feasible of problem P_1, we conclude that $\tilde{z}_k$ is $\frac{\beta}{2K}$ strictly feasible of problem P_k through induction.

Let $\mathcal{L}_k(z) = f_k(z) + (\lambda_k^*)^\top h_k(z) + \mathbb{1}_{\mathcal{Z}}(z)$, where $\mathbb{1}_{\mathcal{Z}}(z)$ is the indicator function. We have $\mathcal{L}_k(z)$ is a strongly convex function over $\mathbb{R}^{n+m+2}$. Given any $\zeta \in \mathcal{N}_{\mathcal{Z}}(z)$, we have $\nabla f_k(z) + \langle \nabla h_k(z), \lambda_k^* \rangle + \zeta \in \partial\mathcal{L}(z)$ for all $z \in \mathcal{Z}$. Clearly $z_k^* \in \arg\min_{z \in \mathcal{Z}} \mathcal{L}_k(z)$. The optimality gives us that $0 \in \partial\mathcal{L}_k(z_k^*)$. And, due to the strong convexity of $f_k(z)$ and $h_k(z)$ and $\lambda_k^* \geq 0$, $\mathcal{L}_k(z)$ is μ strongly convex function. Thus, we have

$$\begin{aligned}
\frac{\mu}{2}\|\tilde{z}_k - z_k^*\|_2^2 &\overset{(i)}{\leq} \mathcal{L}_k(\tilde{z}_k) - \mathcal{L}_k(z_k^*) \\
&= f_k(\tilde{z}_k) + (\lambda_k^*)^\top h_k(\tilde{z}_k) - \big(f_k(z_k^*) + (\lambda_k^*)^\top h_k(z_k^*)\big) \\
&\overset{(ii)}{\leq} f_k(\tilde{z}_k) - f_k(z_k^*),
\end{aligned} \tag{67}$$

where (i) follows from the strong convexity of $\mathcal{L}_k$ and $0 \in \partial\mathcal{L}_k(z_k^*)$, and (ii) follows from the complementary slackness $(\lambda_k^*)^\top h_k(\tilde{z}_k) = 0$ and $\tilde{z}_k$ is feasible for $h_k(z)$.

Combining eqs. (65) and (67), we have

$$\begin{aligned}
\frac{\mu}{2}\|\tilde{z}_k - z_k^*\|_2^2 &\leq f_k(\tilde{z}_k) - f_k(\tilde{z}_{k+1}) + \tfrac{\beta}{2K} \\
&\overset{(i)}{\leq} f_k(\tilde{z}_k) - f_{k+1}(\tilde{z}_{k+1}) + \tfrac{\beta}{2K} \\
&\overset{(ii)}{=} f(\tilde{z}_k) - f(\tilde{z}_{k+1}) + \tfrac{\beta}{2K},
\end{aligned} \tag{68}$$

where (i) follows from the fact that $f_k(\tilde{z}_{k+1}) = f_{k+1}(\tilde{z}_{k+1}) + \frac{\sigma}{2}\|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2$, and (ii) follows from $f_k(\tilde{z}_k) = f(\tilde{z}_k)$, $k \in \mathbb{N}$.

Telescoping eq. (68) and utilizing the definition of $\hat{k}$, we obtain

$$\mathbb{E}\left[\|\tilde{z}_{\hat{k}} - z_{\hat{k}}^*\|_2^2\right] = \frac{1}{K}\sum_{k=1}^{K}\|\tilde{z}_k - z_k^*\|_2^2 \leq \frac{2}{\mu K}\left(f_1(z_1) - f_{K+1}(\tilde{z}_{K+1}) + \tfrac{\beta}{2}\right) \overset{(i)}{\leq} \frac{2\Delta_f}{\mu K} + \frac{\beta}{K} \tag{69}$$

where (i) follows from the definition $\Delta_f = \max_{z,z'} |f(z) - f(z')|$. □

G.2 PROOF OF THEOREM 5

Since $\sigma = 2\max\{L_f, L_c\}$, by Assumption 1 and lemma 7, we have both $f(z)$ and each entry of $h(z)$ are μ strongly convex function with $\mu \geq \frac{\sigma}{2}$. Moreover, Assumption 1 ensures that there exists M_f and M_h, such that $\|\nabla f(z)\|_2 \leq M_f$ and $\|\nabla(h(z))_i\|_2 \leq M_h$, thus we have $\|\nabla f_k(z)\|_2 \leq M_f + \sigma D_{\mathcal{Z}}$ and $\|\nabla(h_k(z))_i\|_2 \leq M_h + \sigma D_{\mathcal{Z}}$. Thus, Assumption 2 holds with $f(z) = f_k(z)$ and $h(z) = h_k(z)$. Applying Theorem 3 with the ϵ replaced by $\frac{\epsilon}{2K}$ there, we have, for each $k = 1, \ldots, K$,

$$\begin{aligned} f_k(\tilde{z}_k) - f_k(z_k^*) &\leq \tfrac{\epsilon}{2K}, \\ \max\left\{(h_k(\tilde{z}_k))_i\right\} &\leq \tfrac{\epsilon}{2K}. \end{aligned}$$

Applying Lemma 8, we have

$$\frac{1}{K}\sum_{k=1}^{K} \|\tilde{z}_k - z_k^*\|_2^2 \leq \frac{2\Delta_f}{\mu K} + \frac{\beta}{K}, \tag{70}$$

Moreover, the optimality of (z_k^*, λ_k^*) for subproblem P_k shows that, there exists $\zeta_k \in \mathcal{N}_{\mathcal{Z}}(z_k^*)$ such that

$$\nabla f_k(z_k^*) + \langle \nabla h_k(z), \lambda_k^* \rangle + \zeta_k = 0. \tag{71}$$

Using the facts, $\nabla f_k(z_k^*) = \nabla f(z_k^*) + \sigma(z_k^* - \tilde{z}_k)$ and $\nabla h(z_k^*) + \sigma \mathbf{1}(z_k^* - \tilde{z}_k)^\top$, eq. (71) implies

$$\nabla f(z_k^*) + \langle \nabla h(z_k^*), \lambda_k^* \rangle + \zeta_k = -(\|\lambda_k^*\|_1 + 1)\sigma(\tilde{z}_k - z_k^*).$$

Taking ℓ_2-norm on both sides of the above equality, and using the upper bound of $\|\lambda_k^*\|_2$ in Assumption 3, we have

$$\|\nabla f(z_k^*) + \langle \lambda_k^* \nabla h(z_k^*) \rangle + \zeta_k\|_2 \leq (B+1)\sigma\|\tilde{z}_k - z_k^*\|_2. \tag{72}$$

Telescoping eq. (72) and applying eq. (70), we have

$$\mathbb{E}\left[\left\|\nabla f(z_{\hat{k}}^*) + \langle \lambda_{\hat{k}}^* \nabla h(z_{\hat{k}}^*) \rangle + \zeta_{\hat{k}}\right\|_2\right] \leq \frac{(B+1)\sigma(2\Delta_f + \mu\beta)}{\mu K},$$

Using the fact that $\zeta_k \in \mathcal{N}_{\mathcal{Z}}(z_k^*)$, we have

$$\mathbb{E}\left[\mathrm{dist}\left(\nabla f(z_{\hat{k}}^*) + \langle \lambda_{\hat{k}}^* \nabla h(z_{\hat{k}}^*) \rangle, -\mathcal{N}_{\mathcal{Z}}(z_{\hat{k}}^*)\right)\right] \leq \frac{(B+1)\sigma(2\Delta_f + \mu\beta)}{\mu K}. \tag{73}$$

Moreover we have,

$$\sum_{i=1}^{q} |(\lambda_k^*)_i (h(z_k^*))_i| = \sum_{i=1}^{q} \left|(\lambda_k^*)_i ((h_k(z_k^*))_i - \tfrac{\sigma}{2}\|\tilde{z}_k - z_k^*\|_2^2 + \tfrac{k\beta}{K})\right| \overset{(i)}{\leq} \tfrac{B\sigma}{2}\|\tilde{z}_k - z_k^*\|_2^2 + \tfrac{kB\beta}{K},$$

where (i) follows from the complementary slackness of z_k^*. Telescoping the above inequality, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{q} \left|\left(\lambda_{\hat{k}}^*\right)_i \left(h(z_{\hat{k}}^*)\right)_i\right|\right] = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{q} |(\lambda_k^*)_i (h(z_k^*))_i| \leq \frac{B\sigma(2\Delta_f + \mu\beta)}{2\mu K} + \frac{K(K+1)B}{K^2} \cdot \beta \tag{74}$$

And $\mathbb{E}[h(z_{\hat{k}}^*)] = \frac{1}{K}\sum_{k=1}^{K} h(z_k^*) \leq \frac{(K+1)\beta}{K} \leq 2\beta$. Using the facts, $K \geq \frac{3(B+1)\Delta_f\sigma}{\mu\epsilon}$, $\beta = \min\{\frac{\epsilon}{4B}, \frac{\Delta_f}{\mu}\}$, eq. (58) induces $\mathbb{E}\left[\|\tilde{z}_{\hat{k}} - z_{\hat{k}}^*\|_2^2\right] \leq \epsilon$, eqs. (73) and (74) imply that

$$\mathbb{E}\left[\mathrm{dist}\left(\nabla f(z_{\hat{k}}^*) + \langle \lambda_{\hat{k}}^* \nabla h(z_{\hat{k}}^*) \rangle, -\mathcal{N}_{\mathcal{Z}}(z_{\hat{k}})\right)\right] \leq \epsilon, \mathbb{E}\left[\sum_{i=1}^{q} \left|(\lambda_{\hat{k}}^*)_i (h(z_{\hat{k}}^*))_i\right|\right] \leq \epsilon.$$

H GRADIENTS OF THE RELAXED PROBLEM IN ILLUSTRATIVE EXAMPLE

The KKT reformulation of the problem in eq. (12) is

$$\begin{aligned}\min_{x,y,w,v\in\mathbb{R}} \quad & -xy \\ s.t. \quad & x^2+y^2-1-\xi \le 0 \\ & g(x,y)-\xi \le 0 \\ & x+2wy+vG(x,y)=0 \\ & w(x^2+y^2-1-\xi)=0 \\ & v(g(x,y)-\xi)=0,\end{aligned}$$

where $G(x,y) \coloneqq \nabla_y g(x,y)$ and it equals

$$G(x,y)=\begin{cases}3(y-|x|)^2 & y \ge |x| \\ 0 & -|x| \le y \le |x| \\ -3(y+|x|)^2 & y \le -|x|\end{cases}.$$

The final relaxed problem is

$$\begin{aligned}\min_{x,y,w,v\in\mathbb{R}} \quad & -xy \\ s.t. \quad & x^2+y^2-1-\xi \le 0 \\ & g(x,y)-\xi \le 0 \\ & x+2wy+vG(x,y)-\beta \le 0 \\ & -x-2wy-vG(x,y)-\beta \le 0 \\ & w(x^2+y^2-1-\xi)-\beta \le 0 \\ & -w(x^2+y^2-1-\xi)-\beta \le 0 \\ & v(g(x,y)-\xi)-\beta \le 0 \\ & -v(g(x,y)-\xi)-\beta \le 0.\end{aligned}$$

Denote $h(z)$ as

$$h(z)=\begin{pmatrix} x^2+y^2-1-\xi \\ g(x,y)-\xi \\ x+2wy+vG(x,y)-\beta \\ -x-2wy-vG(x,y)-\beta \\ w(x^2+y^2-1-\xi)-\beta \\ -w(x^2+y^2-1-\xi)-\beta \\ v(g(x,y)-\xi)-\beta \\ -v(g(x,y)-\xi)-\beta \end{pmatrix}.$$

$\nabla f(z)=[-y;-x;0;0]$, and for the constrained function $h(z)$, we have

$$\nabla h(z)=\begin{pmatrix} 2x & 2y & 0 & 0 \\ \frac{\partial g(x,y)}{\partial x} & \frac{\partial g(x,y)}{\partial y} & 0 & 0 \\ 1+v\frac{\partial G(x,y)}{\partial x} & 2w+v\frac{\partial G(x,y)}{\partial y} & 2y & G(x,y) \\ -1-v\frac{\partial G(x,y)}{\partial x} & -2w-v\frac{\partial G(x,y)}{\partial y} & -2y & -G(x,y) \\ 2wx & 2wy & x^2+y^2-1-\xi & 0 \\ -2wx & -2wy & -(x^2+y^2-1-\xi) & 0 \\ v\frac{\partial g(x,y)}{\partial x} & v\frac{\partial g(x,y)}{\partial y} & 0 & g(x,y)-\xi \\ -v\frac{\partial g(x,y)}{\partial x} & -v\frac{\partial g(x,y)}{\partial y} & 0 & -g(x,y)+\xi \end{pmatrix}, \tag{75}$$

where the gradient of the ith entry equals to the ith row of the above matrix and we have

$$g(x,y) \coloneqq \begin{cases} (y-|x|)^3, & y \geq |x| \\ 0, & -|x| \leq y \leq |x| \\ -(y+|x|)^3, & y \leq -|x| \end{cases}.$$

$$\frac{\partial g(x,y)}{\partial x} = \begin{cases} -3\text{sgn}(x)(|x|-y)^2 & y \geq |x| \\ 0 & |y| \leq |x| \\ -3\text{sgn}(x)(|x|+y)^2 & y \leq -|x| \end{cases},$$

with $\text{sgn}(x) = 1$ if $x \geq 0$ and $\text{sgn}(x) = -1$ otherwise.

$$\frac{\partial g(x,y)}{\partial y} = G(x,y) = \begin{cases} 3(y-|x|)^2 & y \geq |x| \\ 0 & |y| \leq |x| \\ -3(y+|x|)^2 & y \leq -|x| \end{cases},$$

$$\frac{\partial G(x,y)}{\partial x} = \begin{cases} 6\text{sgn}(x)(|x|-y) & y \geq |x| \\ 0 & |y| \leq |x| \\ -6\text{sgn}(x)(y+|x|) & y \leq -|x| \end{cases},$$

$$\frac{\partial G(x,y)}{\partial y} = \begin{cases} 6(y-|x|) & y \geq |x| \\ 0 & |y| \leq |x| \\ -6(y+|x|) & y \leq -|x| \end{cases}.$$