On Penalty-based Bilevel Gradient Descent Method

Han Shen [1] **Tianyi Chen** [1]

Abstract

Bilevel optimization enjoys a wide range of applications in hyper-parameter optimization, meta-learning and reinforcement learning. However, bilevel problems are difficult to solve and recent progress on scalable bilevel algorithms mainly focuses on bilevel optimization problems where the lower-level objective is either strongly convex or unconstrained. In this work, we tackle the bilevel problem through the lens of the penalty method. We show that under certain conditions, the penalty reformulation recovers the solutions of the original bilevel problem. Further, we propose the penalty-based bilevel gradient descent algorithm and establish its finite-time convergence for the constrained bilevel problem under some lower-level error bound conditions weaker than strong convexity. The experimental results showcase the efficiency of the proposed algorithm. The code is available on GitHub (link).

1. Introduction

Bilevel optimization plays an increasingly important role in machine learning as it has a wide range of applications including hyper-parameter optimization (Maclaurin et al., 2015; Franceschi et al., 2018), meta-learning (Finn et al., 2017; Rajeswaran et al., 2019), reinforcement learning (Cheng et al., 2022) and adversarial learning (Jiang et al., 2021; Zhang et al., 2022).

Define $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \mapsto \mathbb{R}$ and $g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \mapsto \mathbb{R}$. In this paper, we consider the following bilevel problem:

$$\mathcal{BP}: \min_{x,y} f(x,y) \text{ s.t. } x \in \mathcal{C}, \qquad y \in \mathcal{S}(x) := \arg\min_{y\in\mathcal{U}(x)} g(x,y)$$

where $\mathcal{C}$, $\mathcal{U}(x)$ and $\mathcal{S}(x)$ are non-empty and closed sets given any $x \in \mathcal{C}$. We call f and g respectively as the upper-level and lower-level objective.

The bilevel optimization problem $\mathcal{BP}$ can be extremely difficult to solve due to the tangling between the upper-level and lower-level problems. Even for the simpler case where $g(x,\cdot)$ is strongly-convex, $\mathcal{U}(x) = \mathbb{R}^{d_y}$ and $\mathcal{C} = \mathbb{R}^{d_x}$, it was not until recently that the veil of an efficient method was partially lifted. Under the strong convexity of $g(x,\cdot)$, the lower-level solution set $\mathcal{S}(x)$ is a singleton. In this case, $\mathcal{BP}$ reduces to minimizing $f(x,\mathcal{S}(x))$, the gradient of which can be calculated with the implicit gradient (IG) method (Pedregosa, 2016; Ghadimi & Wang, 2018). It has been later shown by (Chen et al., 2021) that the IG method converges almost as fast as the gradient-descent method. However, existing IG methods cannot handle either the lower-level constraint $\mathcal{U}(x)$ or the non-strong convexity of $g(x,\cdot)$ due to the difficulty of computing the implicit gradient and thus can not be applied to more complicated bilevel problems.

To overcome the above challenges, recent work aims to develop gradient-based methods without lower-level strong convexity. A prominent branch of algorithms are based on the iterative differentiation method; see e.g., (Franceschi et al., 2017; Liu et al., 2021b). In this case, the lower-level solution set $\mathcal{S}(x)$ is replaced by the output of an iterative optimization algorithm that solves the lower-level problem (e.g., gradient descent (GD)) which allows for explicit differentiation. However, these methods are typically restricted to the unconstrained case since the lower-level algorithm with the projection operator is difficult to differentiate. Furthermore, the algorithm has high memory and computational costs when the lower-level iteration number is large.

On the other hand, it is tempting to penalize certain optimality metric of the lower-level problem (e.g., $\|\nabla_y g(x,y)\|^2$) to the upper-level objective, leading to the single-level optimization problem. The high-level idea is that minimizing the optimality metric guarantees the lower-level optimality condition $y \in \mathcal{S}(x)$ and as long as the optimality metric admits simple gradient evaluation, then the penalized objective can be optimized via gradient-based algorithms. However, as we will show in the next example, GD with a straightforward penalization mechanism may not lead to the desired solution of the original bilevel problems.

[1] Department of ECSE, Rensselaer Polytechnic Institute, Troy, NY, USA. Correspondence to: Han Shen <shenh5@rpi.edu>, Tianyi Chen <chent18@rpi.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Table 1: Comparison of this work (V-PBGD) and IAPTT-GM (Liu et al., 2021b), BOME (Ye et al., 2022), AiPOD (Xiao et al., 2023). In the table, $\mathcal{U}$ is a convex compact set.

	V-PBGD	BOME	IAPTT-GM	AiPOD
Upper-level constraint	✓	✗	✓	✓
Lower-level constraint	$\mathcal{U}(x)=\mathbb{R}^{d_x}$ or $\mathcal{U}$	✗	$\mathcal{U}(x)=\mathcal{U}$	equality constraint
Lower-level non-strongly-convex	✓	✓	✓	✗
Non-singleton $\mathcal{S}(x)$	✓	✗	✓	✗
First-order	✓	✓	✗	✗
Convergence	finite-time	finite-time	asymptotic	finite-time

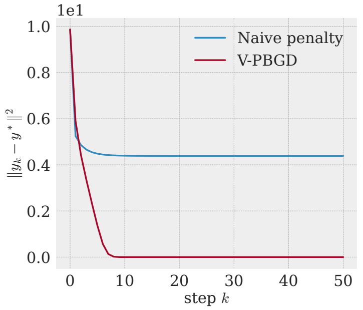


Figure 1: Naive penalty yields suboptimal points, while the proposed algorithm finds the solution.

Example 1. Consider the following special case of $\mathcal{BP}$:

$$\min_{y\in\mathbb{R}} \; f(y) \; := \sin^2(y - \frac{2\pi}{3}),$$
$$\text{s.t. } y \in \arg\min_{y\in\mathbb{R}} g(y) := y^2 + 2\sin^2 y. \qquad (1)$$

The only solution of (1) is $y^* = 0$. In this example, it can be checked that $\|\nabla_y g(y)\|^2 = 0$ if only if $y \in \arg\min_{y\in\mathbb{R}} g(y)$ and thus $\|\nabla_y g(y)\|^2 = 0$ is a lower-level optimality metric. Penalizing $f(y)$ with $\|\nabla g(y)\|^2$ and $\gamma > 0$ gives $\min_y f(y) + \gamma(y + \sin(2y))^2$. For any γ, $y = \frac{2\pi}{3}$ is a local solution of the penalized problem while it is neither a global nor a local solution of (1).

In Figure 1, we show that in Example 1 the naive penalty method, i.e. solving $\min f(x,y) + \gamma(y + \sin(2y))^2$, can get stuck at sub-optimal points. To tackle such issues, it is crucial to study the relation between the bilevel problem and its penalized problem. Specifically, *what impact do different penalty terms, penalization constants and problem properties have on this relation*? Through studying this relation, we aim to develop an efficient penalty-based bilevel gradient descent method.

Our contributions. In this work, we will first consider the following penalty reformulation of $\mathcal{BP}$, given by

$$\mathcal{BP}_{\gamma p} : \min_{x,y} f(x,y) + \gamma p(x,y), \text{ s.t. } x \in \mathcal{C}, \; y \in \mathcal{U}(x)$$

where $p(x,y)$ is a certain penalty function that will be specified in Section 3. Our first result shows that under certain generic conditions on $g(x,y)$, one can recover approximate global (local) solutions of $\mathcal{BP}$ by solving $\mathcal{BP}_{\gamma p}$ globally (locally). Further, we show that these generic conditions hold without the strong convexity of $g(x,\cdot)$. We then propose the penalized bilevel GD (PBGD) method and establish its finite-time convergence when $\mathcal{U}(x) = \mathbb{R}^{d_y}$ or $\mathcal{U}(x) = \mathcal{U}$ which is a compact convex set. We summarize the convergence results of our algorithm and compare them with several related works in Table 1. Finally, we empirically showcase the performance, computation and memory efficiency of the proposed algorithm in comparison with several competitive baselines.

Related works. The bilevel optimization problem can be dated back to (Stackelberg, 1952). Recently, the gradient-based bilevel optimization methods have gained growing popularity in the machine learning area; see, e.g., (Sabach & Shtern, 2017; Franceschi et al., 2018; Liu et al., 2020). A branch of gradient-based methods belongs to the IG method (Pedregosa, 2016). The finite-time convergence was first established in (Ghadimi & Wang, 2018) for the unconstrained strongly-convex lower-level problem. Later, the convergence was improved in (Hong et al., 2023; Ji et al., 2021; Chen et al., 2022; 2021; Khanduri et al., 2021; Shen & Chen, 2022; Li et al., 2022; Sow et al., 2022). Recent works extend IG to constrained strongly-convex lower-level problems; see, e.g., the equality-constrained IG method (Xiao et al., 2023) and a 2nd-derivative-free approach (Giovannelli et al., 2022).

Another branch of methods is based on the iterative differentiation (ITD) methods (Maclaurin et al., 2015; Franceschi et al., 2017; Nichol et al., 2018; Shaban et al., 2019). Later, (Liu et al., 2021b) proposes an ITD method with initialization optimization and shows asymptotic convergence. Another work (Liu et al., 2022b) develops an ITD method where each lower-level iteration uses a combination of upper-level and lower-level gradients. Recently, the iterative differentiation of non-smooth lower-level algorithms has been studied in (Bolte et al., 2022). The ITD methods generally lack finite-time guarantee unless restrictive assumptions are made for the iteration mapping (Grazzi et al., 2020; Ji et al., 2022).

Recently, bilevel optimization methods have also been studied in distributed learning (Tarzanagh et al., 2022; Lu et al., 2022; Yang et al., 2022), corset selection (Zhou et al., 2022), overparametrized setting (Vicol et al., 2022), multi-block min-max (Hu et al., 2022), game theory (Arbel & Mairal, 2022) and several acceleration methods have been proposed (Huang et al., 2022; Dagréou et al., 2022). The works (Liu et al., 2021a) and (Mehra & Hamm, 2021) propose penalty-based methods respectively with log-barrier and gradient norm penalty, and establish their asymptotic convergence. Another work (Gao et al., 2022) develops a method based on the difference-of-convex algorithm. In preparing our final version, a concurrent work (Chen et al., 2023) studies the bilevel problem with convex lower-level objectives and proposes a zeroth-order optimization method with finite-time convergence to the Goldstein stationary point. Another concurrent work (Lu & Mei, 2023) proposes a penalty method for the bilevel problem with a convex lower-level objective $g(x,\cdot)$. It shows convergence to a weak KKT point of the bilevel problem while does not study the relation between the bilevel problem and its penalized problem.

The relation between the bilevel problem and its penalty reformulation has been first studied in (Ye et al., 1997) under the calmness condition paired with other conditions such as the 2-Hölder continuity, which are difficult to satisfy. A recent work (Ye et al., 2022) proposes a novel first-order method that is termed BOME. By assuming the constant rank constraint qualification (CRCQ), (Ye et al., 2022) shows convergence of BOME to a KKT point of the bilevel problem. However, it is unclear when CRCQ can be satisfied and the convergence relies on restrictive assumptions like the uniform boundedness of $\|\nabla g\|$, $\|\nabla f\|$, $|f|$ and $|g|$. It is also difficult to argue when the KKT point is a solution of the bilevel problem under lower-level non-convexity.

Notations. We use $\|\cdot\|$ to denote the l^2-norm. Given $r>0$ and $z\in\mathbb{R}^d$, define $\mathcal{N}(z,r):=\{z'\in\mathbb{R}^d:\|z-z'\|\leq r\}$. Given vectors x and y, we use (x,y) to indicate the concatenated vector of x,y. Given a non-empty closed set $\mathcal{S}\subseteq\mathbb{R}^d$, define the distance of $y\in\mathbb{R}^d$ to the set $\mathcal{S}$ as $d_{\mathcal{S}}(y):=\min_{y'\in\mathcal{S}}\|y-y'\|$. We use $\mathrm{Pj}_{\mathcal{Z}}$ to denote the projection to the set $\mathcal{Z}$.

2. Penalty Reformulation of Bilevel Problems

In this section, we study the relationship between the bilevel problem $\mathcal{BP}$ and its penalty reformulation $\mathcal{BP}_{\gamma p}$.

Since $\mathcal{S}(x)$ is closed, $y\in\mathcal{S}(x)$ is equivalent to $d_{\mathcal{S}(x)}(y)=0$. We therefore rewrite $\mathcal{BP}$ as

$$\min_{x,y} f(x,y) \ \text{ s.t. } x\in\mathcal{C},\ d^2_{\mathcal{S}(x)}(y)=0. \tag{2}$$

The squared distance function is non-differentiable, and thus penalizing $d^2_{\mathcal{S}(x)}(y)$ to the upper-level objective is computationally-intractable. Instead, we consider its parametric upper bounds defined as follows.

Definition 1 (Squared-distance bound function). A function $p:\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\mapsto\mathbb{R}$ is a ρ-squared-distance-bound if there exists $\rho>0$ such that for any $x\in\mathcal{C},y\in\mathcal{U}(x)$, it holds

$$p(x,y)\geq 0,\ \rho p(x,y)\geq d^2_{\mathcal{S}(x)}(y) \tag{3a}$$

$$p(x,y)=0 \ \text{ if and only if } \ d_{\mathcal{S}(x)}(y)=0. \tag{3b}$$

Suppose $p(x,y)$ is a squared-distance bound function. Given $\epsilon>0$, we define the following problem:

$$\mathcal{BP}_\epsilon:\ \min_{x,y} f(x,y)\ \text{ s.t. } x\in\mathcal{C},\ y\in\mathcal{U}(x),\ p(x,y)\leq\epsilon.$$

It is clear that $\mathcal{BP}_\epsilon$ with $\epsilon=0$ recovers $\mathcal{BP}$. For $\epsilon>0$, $\mathcal{BP}_\epsilon$ is an ϵ-approximate problem of $\mathcal{BP}$ since $p(x,y)$ is an upper bound of $d^2_{\mathcal{S}(x)}(y)$ and is smaller than ϵ.

We start by considering the relation between the global solutions of $\mathcal{BP}_\epsilon$ and $\mathcal{BP}_{\gamma p}$. Before we introduce the theorem, we first give the following definition and assumption.

Definition 1 (Lipschitz continuity). Given $L>0$, a function $\ell:\mathbb{R}^d\mapsto\mathbb{R}^{d'}$ is said to be L-Lipschitz-continuous on $\mathcal{X}\subseteq\mathbb{R}^d$ if it holds for any $x,x'\in\mathcal{X}$ that $\|\ell(x)-\ell(x')\|\leq L\|x-x'\|$. A function ℓ is said to be L-Lipschitz-smooth if its gradient is L-Lipschitz-continuous.

Assumption 1. *There exists constant L such that given any $x\in\mathcal{C}$, $f(x,\cdot)$ is L-Lipschitz continuous on $\mathcal{U}(x)$.*

The above assumption is standard and has been made in several other works studying bilevel optimization; see, e.g., (Ghadimi & Wang, 2018; Chen et al., 2021; 2023). In order to establish relation between the solutions of $\mathcal{BP}_{\gamma p}$ and those of $\mathcal{BP}$, a crucial step is to guarantee that (x_γ,y_γ), which is a solution of $\mathcal{BP}_{\gamma p}$, is feasible for $\mathcal{BP}_\epsilon$, i.e., to guarantee $p(x_\gamma,y_\gamma)$ is small. Under Assumption 1, the growth of $f(x_\gamma,\cdot)$ is controlled. Then an important intuition is that increasing γ in $\mathcal{BP}_{\gamma p}$ likely makes $p(x_\gamma,\cdot)$ more dominant, and thus decreases $p(x_\gamma,y_\gamma)$.

With this intuition, we introduce the theorem as follows.

Theorem 1 (Relation on global solutions). *Assume $p(x,y)$ is a ρ-squared-distance-bound function and blueAssumption 1 holds. Given any $\epsilon_1>0$, any global solution of $\mathcal{BP}$ is an ϵ_1-global-minimum point of $\mathcal{BP}_{\gamma p}$ with any $\gamma\geq\gamma^*=\frac{L^2\rho}{4}\epsilon_1^{-1}$. Conversely, given $\epsilon_2\geq 0$, if (x_γ,y_γ) achieves ϵ_2-global-minimum of $\mathcal{BP}_{\gamma p}$ with $\gamma>\gamma^*$, (x_γ,y_γ) is the global solution of $\mathcal{BP}_{\epsilon_\gamma}$ with some $\epsilon_\gamma\leq(\epsilon_1+\epsilon_2)/(\gamma-\gamma^*)$.*

The proof of Theorem 1 can be found in Appendix A.1. In Example 1, $\|\nabla_y g(x,y)\|^2$ is actually a squared-distance-bound and the above theorem regarding global solutions holds for this example. However, as illustrated in Example 1, a penalized problem with any γ always admits a local solution meaningless to the original problem. In fact, the relationship between local solutions is more intricate than that between global solutions. Nevertheless, we prove in the following theorem that under some verifiable conditions, the local solutions of $\mathcal{BP}_{\gamma p}$ are local solutions of the $\mathcal{BP}_{\epsilon}$.

Theorem 2 (Relation on local solutions)**.** *Assume $p(x,\cdot)$ is continuous given any $x \in \mathcal{C}$ and $p(x,y)$ is ρ-squared-distance-bound function. Given $\gamma > 0$, let (x_γ, y_γ) be a local solution of $\mathcal{BP}_{\gamma p}$ on $\mathcal{N}((x_\gamma, y_\gamma), r)$. Assume $f(x_\gamma, \cdot)$ is L-Lipschitz-continuous on $\mathcal{N}(y_\gamma, r)$. Assume either one of the following is true:*

(i) There exists $\bar{y} \in \mathcal{N}(y_\gamma, r)$ such that $\bar{y} \in \mathcal{U}(x_\gamma)$ and $p(x_\gamma, \bar{y}) \le \epsilon$ for some $\epsilon \ge 0$. Define $\bar{\epsilon}_\gamma = \frac{L^2\rho}{\gamma^2} + 2\epsilon$.

(ii) The set $\mathcal{U}(x_\gamma)$ is convex and the function $p(x_\gamma, \cdot)$ is convex. Define $\bar{\epsilon}_\gamma = \frac{L^2\rho}{\gamma^2}$.

Then (x_γ, y_γ) is a local solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\epsilon_\gamma \le \bar{\epsilon}_\gamma$.

The proof of Theorem 2 can be found in Appendix A.2.

Remark 1 (Intuition of the conditions)**.** In (i), we need an approximate global minimizer of $p(x_\gamma, \cdot)$; and in (ii), we assume $p(x_\gamma, \cdot)$ is convex. Loosely speaking, these conditions essentially require $\min_{\mathcal{U}(x)} p(x, \cdot)$ to be globally solvable. Such a requirement is natural since finding a feasible point in $\mathcal{S}(x)$ is possible only if one can solve for $p(x, \cdot) = 0$ on $\mathcal{U}(x)$. While they appear to be abstract, we will show how Conditions (i) and (ii) in Theorem 2 can be verified in the following sections.

3. Solving Bilevel Problems with Non-convex Lower-level Objectives

To develop algorithms with non-asymptotic convergence, we consider $\mathcal{BP}$ with $\mathcal{U}(x) = \mathbb{R}^{d_y}$, given by

$$\mathcal{UP}: \min_{x,y} f(x,y) \;\; \text{s.t. } x \in \mathcal{C},\; y \in \arg\min_{y \in \mathbb{R}^{d_y}} g(x,y)$$

where we assume $\mathcal{C}$ is a closed convex set and f, g are continuously differentiable.

3.1. Candidate penalty terms

Following Section 2, to reformulate $\mathcal{UP}$, we first seek a squared-distance bound function p that satisfies Definition 1. For a non-convex function $g(x, \cdot)$, an interesting property is the Polyak-Łojasiewicz (PL) inequality defined as follows.

Assumption 2 (Polyak-Lojasiewicz function)**.** *The lower-level function $g(x, \cdot)$ satisfies the $\frac{1}{\mu}$-PL inequality; that is, there exists $\mu > 0$ such that given any $x \in \mathcal{C}$, it holds for any $y \in \mathbb{R}^{d_y}$ that*

$$\|\nabla_y g(x,y)\|^2 \ge \frac{1}{\mu}(g(x,y) - v(x)) \tag{4}$$

where $v(x) := \min_{y \in \mathbb{R}^{d_y}} g(x,y)$.

In reinforcement learning, it has been proven in (Mei et al., 2020, Lemma 8&9) that the non-convex discounted return objective satisfies the PL inequality under certain parameterization. Moreover, recent studies found that over-parameterized neural networks can lead to losses that satisfy the PL inequality (Liu et al., 2022a).

We consider the following potential penalty functions:

$$p(x,y) = g(x,y) - v(x) \tag{5a}$$
$$p(x,y) = \mu\|\nabla_y g(x,y)\|^2 \tag{5b}$$

Under the PL inequality, the following lemma shows that the above penalty functions are squared-distance bound functions.

Lemma 1. *Suppose Assumption 2 holds. Then* (5a) *and* (5b) *are μ-squared-distance-bound functions.*

The proof of Lemma 1 is deferred to Appendix B.1.

3.2. Penalty reformulation

Given $p(x,y)$ and $\gamma > 0$, define the penalized $\mathcal{UP}$ as

$$\mathcal{UP}_{\gamma p}: \min_{x,y} F_\gamma(x,y) := f(x,y) + \gamma p(x,y) \;\; \text{s.t. } x \in \mathcal{C}.$$

It remains to show that the solutions of $\mathcal{UP}_{\gamma p}$ are meaningful to $\mathcal{UP}$. Starting with the global solutions, we give the following proposition.

Proposition 1 (Relation on global solutions)**.** *Assume Assumption 1 and 2 hold. Choose $p(x,y)$ as either* (5a) *or* (5b)*. Suppose $\gamma \ge L\sqrt{\mu\delta^{-1}}$ with some $\delta > 0$. Let (x_γ, y_γ) be a global solution of $\mathcal{UP}_{\gamma p}$. Then (x_γ, y_γ) is a global solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\mathcal{U}(x) = \mathbb{R}^{d_y}$ and $\epsilon_\gamma \le \delta$.*

Proposition 1 follows directly from Theorem 1 with $\epsilon_1 = L\sqrt{\rho\delta}/2$, $\gamma \ge 2\gamma^* = L\sqrt{\mu\delta^{-1}}$ and $\epsilon_2 = 0$. By the above proposition, the global solution of $\mathcal{UP}_{\gamma p}$ solves an approximate bilevel problem of $\mathcal{UP}$. However, since $\mathcal{UP}_{\gamma p}$ is generally non-convex, it is also important to consider the local solutions. Following Theorem 2, the following proposition captures the meaning of the local solutions.

Proposition 2 (Relation on local solutions)**.** *Assume Assumption 1, 2 and either one of the following hold:*

(a) With some $\delta > 0$, choose

$$p(x,y) = g(x,y) - v(x) \quad \text{and} \quad \gamma \ge L\sqrt{3\mu\delta^{-1}}.$$

(b) With some $\delta > 0$, choose $p(x,y) = \mu\|\nabla_y g(x,y)\|^2$ and

$$\gamma \ge \max\{L\sqrt{2\mu\delta^{-1}}, L\sigma^{-1}\sqrt{\mu^{-1}\delta^{-1}}\}$$

where $\sigma > 0$ is the lower-bound of the singular values of $\nabla_{yy} g(x,y)$ on $\{(x,y) \in \mathcal{C} \times \mathbb{R}^{d_y} : y \notin \mathcal{S}(x)\}$.

Algorithm 1 PBGD: Penalized bilevel GD

```
Select $(x_1, y_1) \in \mathcal{Z} := \mathcal{C} \times \mathcal{U}(x)$. Select $\alpha$, $\gamma$, $K$.
for $k = 1$ to $K$ do
    Compute $h_k = \nabla p(x_k, y_k)$ or its estimate.
    $(x_{k+1}, y_{k+1}) = \mathrm{Pj}_{\mathcal{Z}}\Big((x_k, y_k) - \alpha\big(\nabla f(x_k, y_k) + \gamma h_k\big)\Big)$.
end for
```

Let (x_γ, y_γ) be a local solution of $\mathcal{UP}_{\gamma p}$. Then (x_γ, y_γ) is a local solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\mathcal{U}(x) = \mathbb{R}^{d_y}$ and $\epsilon_\gamma \leq \delta$.

The proof of Proposition 2 can be found in Appendix B.2. Proposition 2 explains the observations in Figure 1 and Example 1. The failing point $y = \frac{2\pi}{3}$ mentioned in Example 1 yields $\nabla_{yy} g(x, y) = 0$ which violates (b) in Proposition 2. It can be checked that (a) in Proposition 2 holds.

Propositions 1 and 2 suggest $\gamma = \Omega(\delta^{-0.5})$ in order to achieve $\epsilon_\gamma \leq \delta$. Next we show this is tight.

Corollary 1. *In Proposition 1 or 2, to guarantee $\epsilon_\gamma = \mathcal{O}(\delta)$, the lower-bound $\gamma = \Omega(\delta^{-0.5})$ is tight.*

Proof. Consider the following special case of $\mathcal{UP}$:

$$\min_{y \in \mathbb{R}} f(x, y) = y, \text{ s.t. } y \in \arg\min_{y \in \mathbb{R}} g(x, y) = y^2. \tag{6}$$

In this example, the two penalty terms $\frac{1}{4}\|\nabla_y g(x, y)\|^2$ and $g(x, y) - v(x)$ coincide to be $p(x, y) = y^2$. Accordingly, the solution of $\mathcal{UP}_{\gamma p}$ is $-\frac{1}{2\gamma}$. It can be checked that $-\frac{1}{2\gamma}$ is a local solution of $\mathcal{BP}_{\epsilon_\gamma}$ ($\mathcal{U}(x) = \mathbb{R}^{d_y}$) with $\epsilon_\gamma = 1/(4\gamma^2)$. To ensure $\epsilon_\gamma = \mathcal{O}(\delta)$, $\gamma = \Omega(\delta^{-0.5})$ is required in this example. Then the poof is complete by the fact that the assumptions in Proposition 1 and 2 hold in this example. □

Propositions 1 and 2 imply that $\mathcal{UP}$ and $\mathcal{UP}_{\gamma p}$ are related in the sense that one can globally/locally solve an approximate bilevel problem of $\mathcal{UP}$ by globally/locally solving $\mathcal{UP}_{\gamma p}$ instead. A natural approach to solving $\mathcal{UP}_{\gamma p}$ is the projected gradient method described in Algorithm 1.

When $p(x, y) = \mu\|\nabla_y g(x, y)\|^2$, $\nabla p(x, y)$ can be exactly evaluated. In this case, Algorithm 1 is a standard projected gradient method and the convergence property directly follows from the existing literature. In the next subsection, we focus on the other penalty function $p(x, y) = g(x, y) - v(x)$ and discuss when $\mathcal{UP}_{\gamma p}$ can be efficiently solved.

3.3. PBGD with function value gap

We consider solving $\mathcal{UP}_{\gamma p}$ with $p(x, y)$ chosen as the function value gap penalty (5a). To solve $\mathcal{UP}_{\gamma p}$ with the gradient-based method, the obstacle is that $\nabla p(x, y)$ requires $\nabla v(x)$. For one, $v(x)$ is not necessarily smooth. Even if $v(x)$ is differentiable, $\nabla v(x) \neq \nabla_x g(x, y^*)$ in general, where $y^* \in \mathcal{S}(x)$. However, it is possible to compute $\nabla v(x)$ efficiently under some relatively mild assumptions.

Algorithm 2 V-PBGD: Function value gap based PBGD

```
Select $(x_1, y_1) \in \mathcal{Z} = \mathcal{C} \times \mathbb{R}^{d_y}$. Select $\alpha, \beta$, $\gamma$, $T_k$,$K$.
for $k = 1$ to $K$ do
    Obtain $\hat{y}_k = \omega^{(k)}_{T_k+1}$ following (7a).
    Update $(x_k, y_k)$ following (7b).
end for
```

Lemma 2 ((Nouiehed et al., 2019, Lemma A.5))**.** *Assume Assumption 2 holds, and g is L_g-Lipschitz-smooth with constant L_g. Then $\nabla v(x) = \nabla_x g(x, y^*)$ for any $y^* \in \mathcal{S}(x)$, and $v(x)$ is $(L_g + L_g^2\mu)$-Lipschitz-smooth.*

Under the conditions in Lemma 2, $\nabla v(x)$ can be evaluated directly at any optimal solution of the lower-level problem. This suggests one find a lower-level optimal solution $y^* \in \mathcal{S}(x)$, and evaluate the penalized gradient $\nabla F_\gamma(x, y)$ with $\nabla v(x) = \nabla_x g(x, y^*)$. Following this idea, given outer iteration k and x_k, we run T_k steps of inner GD update to solve the lower-level problem:

$$\omega^{(k)}_{t+1} = \omega^{(k)}_t - \beta\nabla_y g(x_k, \omega^{(k)}_t),\ t = 1, \ldots, T_k \tag{7a}$$

where $\omega^{(k)}_1 = y_k$. Update (7a) yields an approximate lower-level solution $\hat{y}_k = \omega^{(t)}_{T_k+1}$. Then we can approximate $\nabla F_\gamma(x_k, y_k)$ with $\hat{y}_k$ and update (x_k, y_k) via:

$$\begin{aligned}(x_{k+1}, y_{k+1}) = \mathrm{Pj}_{\mathcal{Z}}\Big((x_k, y_k) - \alpha\big(\nabla f(x_k, y_k) \\ + \gamma(\nabla g(x_k, y_k) - \overline{\nabla}_x g(x_k, \hat{y}_k))\big)\Big)\end{aligned} \tag{7b}$$

where $\mathcal{Z} = \mathcal{C} \times \mathbb{R}^{d_y}$, $\overline{\nabla}_x g(x, y) := (\nabla_x g(x, y), \mathbf{0})$ with $\mathbf{0} \in \mathbb{R}^{d_y}$. The update is summarized in Algorithm 2, which is a function value gap-based special case of PBGD (Algorithm 1) with $h_k = \nabla g(x_k, y_k) - \overline{\nabla}_x g(x_k, \hat{y}_k)$.

Notice that only the first-order derivatives are required in update (7), which is in contrast to the implicit gradient methods or some iterative differentiation methods where higher-order derivatives are required; see, e.g., (Ghadimi & Wang, 2018; Franceschi et al., 2017; Liu et al., 2021b). In modern machine learning applications, this could substantially save computational cost since the dimension of the parameter is often large, making higher-order derivatives particularly costly.

3.4. Analysis of PBGD with function value gap

We first introduce the following regularity assumption commonly made in the convergence analysis of the projected GD method (Chen et al., 2021; Grazzi et al., 2020).

Assumption 3 (smoothness)**.** *There exist constants L_f and L_g such that $f(x, y)$ and $g(x, y)$ are respectively L_f-Lipschitz-smooth and L_g-Lipschitz-smooth in (x, y).*

Define the projected gradient of $\mathcal{UP}_{\gamma p}$ at $(x_k, y_k) \in \mathcal{Z}$ as

$$G_\gamma(x_k, y_k) \coloneqq \frac{1}{\alpha}\big((x_k, y_k) - (\bar{x}_{k+1}, \bar{y}_{k+1})\big), \tag{8}$$

where $(\bar{x}_{k+1}, \bar{y}_{k+1}) \coloneqq \mathrm{Pj}_{\mathcal{Z}}((x_k, y_k) - \alpha \nabla F_\gamma(x_k, y_k))$. This definition (8) is commonly used as the convergence metric for the projected gradient methods. It is known that given a convex $\mathcal{Z}$, $G_\gamma(x, y) = 0$ if and only if (x, y) is a stationary point of $\mathcal{UP}_{\gamma p}$ (Ghadimi et al., 2016). We provide the following theorem on the convergence of V-PBGD.

Theorem 3. *Consider V-PBGD (Algorithm 2). Suppose Assumption 1,2 and 3 hold. Select $\omega_1^{(k)} = y_k$ and*

$$\alpha \in (0, (L_f + \gamma(2L_g + L_g^2 \mu))^{-1}],\ \beta \in (0, L_g^{-1}],$$
$$\gamma \geq L\sqrt{3\mu\delta^{-1}} \text{ with some } \delta > 0,\ T_k = \Omega(\log(\alpha k)).$$

i) With $C_f = \inf_{(x,y)\in\mathcal{Z}} f(x, y)$, it holds that

$$\frac{1}{K}\sum_{k=1}^{K} \|G_\gamma(x_k, y_k)\|^2 \leq \frac{18\big(F_\gamma(x_1, y_1) - C_f\big)}{\alpha K} + \frac{10L^2L_g^2}{K}.$$

ii) Suppose $\lim_{k\to\infty}(x_k, y_k) = (x^*, y^*)$, then (x^*, y^*) is a stationary point of $\mathcal{UP}_{\gamma p}$. If (x^*, y^*) is a local/global solution of $\mathcal{UP}_{\gamma p}$, it is also a local/global solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\mathcal{U}(x) = \mathbb{R}^{d_y}$ and $\epsilon_\gamma \leq \delta$.

The proof of Theorem 3 is deferred to Appendix B.3. Theorem 3 implies an iteration complexity of $\widetilde{\mathcal{O}}(\gamma\epsilon^{-1})$ to find an ϵ-stationary-point of $\mathcal{UP}_{\gamma p}$. This recovers the iteration complexity of the projected GD method (Nesterov, 2013) with a smoothness constant of $\Theta(\gamma)$. If we choose $\delta = \epsilon$, then the iteration complexity is $\widetilde{\mathcal{O}}(\epsilon^{-1.5})$. In (ii), under no stronger conditions needed for the projected GD method to yield meaningful solutions, the V-PBGD algorithm finds a local/global solution of the approximate $\mathcal{UP}$.

Remark 2 (On the computational and memory complexity.)**.** Regarding computational compexity, V-PBGD requires the addition/subtraction of gradients and parameters, but previous methods based on implicit differentiation and iterative differentiation in addition require computing the Hessian-vector products. In this sense, when the parameter dimension d_x, d_y are large, the per-iteration computational complexity of PBGD will be lower than those methods.

Regarding memory complexity, PBGD requires storing the parameters $y, \omega \in \mathbb{R}^{d_y}, x \in \mathbb{R}^{d_x}$ and their gradients. Thus PBGD requires $\mathcal{O}(d_x + d_y)$ space. Previous methods (e.g., implicit gradient methods) that utilize Hessian-vector products to compute Neuman series approximation requires $\mathcal{O}(d_x + d_y)$ space. Other methods (e.g., RHG) that requires storing the history parameters to compute hyper-gradient requires $\mathcal{O}(d_x + Td_y)$ space where t is the lower-level iteraiton number. BVFIM [3] requires $\mathcal{O}(d_x + d_y)$ space.

4. Solving Bilevel Problems with Lower-level Constraints

In the previous section, we have introduced the PBGD method to solve a class of non-convex bilevel problems with only upper-level constraints. When the lower-level constraints are involved, it becomes more difficult to develop a gradient-based algorithm with finite-time guarantees.

In this section, under assumptions on the lower-level objective that are jointly weaker than the commonly used smooth strong-convexity assumption, we propose an algorithm with finite-time convergence guarantee. Specifically, consider the following special case of $\mathcal{BP}$ with $\mathcal{U}(x) = \mathcal{U}$:

$$\mathcal{CP}:\ \min_{x,y} f(x, y) \text{ s.t. } x \in \mathcal{C},\ y \in \operatorname*{argmin}_{y\in\mathcal{U}} g(x, y)$$

where we assume $\mathcal{C}, \mathcal{U}$ are convex compact in this section.

4.1. Penalty reformulation

Following Section 2, we seek to reformulate $\mathcal{CP}$ with a suitable penalty function $p(x, y)$ in this section.

We first list some conditions that will be repeatedly used in this section.

Assumption 4. *Consider the following conditions:*

(i) There exists $\mu > 0$ such that given any $x \in \mathcal{C}$, $g(x, \cdot)$ has $\frac{1}{\mu}$-quadratic-growth, that is, $\forall y \in \mathcal{U}$, it holds that

$$g(x, y) - v(x) \geq \frac{1}{\mu} d^2_{\mathcal{S}(x)}(y)$$

where $v(x) \coloneqq \min_{y\in\mathcal{U}} g(x, y)$.

(ii) There exists $\bar{\mu} > 0$ such that given $x \in \mathcal{C}$, $g(x, \cdot)$ satisfies $\frac{1}{\bar{\mu}}$-proximal-error-bound, i.e., $\forall y \in \mathcal{U}$, it holds

$$\frac{1}{\beta}\big\|y - \mathrm{Pj}_{\mathcal{U}}\left(y - \beta\nabla_y g(x, y)\right)\big\| \geq \frac{1}{\bar{\mu}} d_{\mathcal{S}(x)}(y)$$

where β is specified in Theorem 4.

(iii) Given any $x \in \mathcal{C}$, $g(x, \cdot)$ is convex.

Note that the above assumptions do not need to hold simultaneously. Define the penalty function as the lower-level function value gap $g(x, y) - v(x)$ where $v(x) = \min_{y\in\mathcal{U}} g(x, y)$.

Lemma 3. *Assume (i) in Assumption 4 holds. Then $p(x, y) = g(x, y) - v(x)$ is a μ-squared-distance-bound.*

The proof is similar to that of Lemma 1 and thus will be omitted. Given a penalization constant $\gamma > 0$, we can define the penalized $\mathcal{CP}$ as follows:

$$\mathcal{CP}_{\gamma p}:\ \min_{x,y} F_\gamma(x, y) \coloneqq f(x, y) + \gamma\big(g(x, y) - v(x)\big)$$
$$\text{s.t. } x \in \mathcal{C},\ y \in \mathcal{U}.$$

It remains to show that the solutions of $\mathcal{CP}_{\gamma p}$ are meaningful to $\mathcal{CP}$. In the following proposition, we show that the solutions of $\mathcal{CP}_{\gamma p}$ approximately solve $\mathcal{CP}$.

Proposition 3 (Relation on local/global solutions)**.** *Assume Assumption 1 and either one of the following hold:*

(a) Conditions (i) and (ii) in Assumption 4 hold. Choose $\gamma \geq \max\{L\sqrt{\mu\delta^{-1}}, L\sqrt{3\bar{\mu}\delta^{-1}}, 3\bar{L}\bar{\mu}L\delta^{-1}\}$ with $\bar{L} = \max_{x\in\mathcal{C},y\in\mathcal{U}} \|\nabla_y g(x,y)\|$ and some $\delta > 0$;
(b) Conditions (i) and (iii) in Assumption 4 hold. Choose $\gamma \geq L\sqrt{\mu\delta^{-1}}$ with some $\delta > 0$.

If (x_γ, y_γ) is a local/global solution of $\mathcal{CP}_{\gamma p}$, it is also a local/global solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\mathcal{U}(x) = \mathcal{U}$ and $\epsilon_\gamma \leq \delta$.

The proof of Proposition 3 can be found in Appendix C.1. In the above proposition, $\bar{L}$ is finite since $\mathcal{C} \times \mathcal{U}$ is compact and $\nabla_y g(x,y)$ is continuous.

4.2. PBGD under lower-level constraints

To study the gradient-based method for solving $\mathcal{CP}_{\gamma p}$, it is crucial to identify when $v(x)$ is Lipschitz-smooth. In the unconstrained lower-level case, we have shown that $v(x)$ is Lipschitz-smooth by Lemma 2. However, the proof of Lemma 2 relies on the condition that $\nabla_y g(x,y) = 0$ for any $y \in \mathcal{S}(x)$ which is not necessarily true under lower-level constraint.

To characterize $v(x)$, we introduce the following lemma.

Lemma 4. *(Dem'yanov & Malozemov, 1971, Theorem 2.2.1) Given any direction $d \in \mathbb{R}^{d_x}$ such that $\|d\| = 1$, $\nabla_d v(x)$ which is the directional derivative of $v(x)$ along d exists at any $x \in \mathcal{C}$. It holds that*

$$\nabla_d v(x) = \min_{y^* \in \mathcal{S}(x)} \langle \nabla_x g(x, y^*),\, d\rangle,\ \forall x \in \mathcal{C}. \tag{9}$$

In order for $v(x)$ to have Lipschitz continuous gradient, all its directional derivatives are necessarily Lipschitz-continuous. By Lemma 4, one will need $\nabla_x g(x,y)$ and $\mathcal{S}(x)$ to be Lipschitz-continuous in some sense, which is formalized next.

Lemma 5 (Lipschitz-continuity of $\mathcal{S}(x)$)**.** *Assume there exists L_g such that g is L_g-Lipschitz-smooth. Assume either one of the following is true:*

(a) Condition (ii) in Assumption 4 holds;
(b) Conditions (i) and (iii) in Assumption 4 hold.

Then there exists $L_S > 0$ that given any $x_1, x_2 \in \mathcal{C}$, for any $y_1 \in \mathcal{S}(x_1)$ there exists $y_2 \in \mathcal{S}(x_2)$ such that

$$\|y_1 - y_2\| \leq L_S \|x_1 - x_2\|.$$

The proof of Lemma 5 can be found in Appendix C.2. It is worth noting that Lemma 5 is not sufficient for $v(x)$ to be smooth. Given any d in Lemma 4, consider $\nabla_d v(x_1)$ and $\nabla_d v(x_2)$ at different points x_1, x_2 in a small neighborhood. Suppose in Lemma 4, the minimum are achieved at $y_1 \in \mathcal{S}(x_1)$ and $y_2 \in \mathcal{S}(x_2)$. Since y_1 and y_2 are not necessarily close in any sense, $\nabla_d v(x)$ is not guaranteed to be continuous. To address this issue, we provide a sufficient condition under which $v(x)$ is (Lipschitz) smooth in the following lemma.

Lemma 6 (Lipschitz-smoothness of $v(x)$)**.** *Suppose the conditions in Lemma 5 hold. Given any $x \in \mathcal{C}$, if $\nabla_x g(x, y_1) = \nabla_x g(x, y_2)$ for any $y_1, y_2 \in \mathcal{S}(x)$, then it holds*

$$\nabla v(x) = \nabla_x g(x, y^*),\ \forall y^* \in \mathcal{S}(x) \tag{10}$$

and $v(x)$ is L_v-Lipschitz-smooth with $L_v = L_g(1+L_S)$.

The proof of Lemma 6 can be found in Appendix C.2. In Lemma 6 , a sufficient condition for the assumption on $\nabla_x g(x,y)$ to hold is $\mathcal{S}(x)$ is a singleton. Alternatively, the assumption holds if the constraint set $\mathcal{U}$ allows $\nabla_y g(x, y^*) = 0$ for any $y^* \in \mathcal{S}(x)$ by following the proof of Lemma 2. Below we give a simple example where Lemma 6 holds without a singleton $\mathcal{S}(x)$.

Example 2. Suppose $\mathcal{C} = [0, \infty)$ and $\mathcal{U} = [2, 3]$. Let

$$g(x,y) = \begin{cases} \frac{1}{2}(y - x - 1)^2 & y \geq x + 1 \\ 0 & x - 1 < y < x + 1 \\ \frac{1}{2}(y - x + 1)^2 & y \leq x - 1. \end{cases}$$

It can be checked that Lemma 5 holds for this example. When $0 \leq x \leq 1$ or $x \geq 4$, $\mathcal{S}(x)$ is a singleton. Otherwise, $\nabla_x g(x,y) = 0$ on a non-singleton $\mathcal{S}(x)$. Thus for any $x \in \mathcal{C}$, we have $\nabla_x g(x, y_1) = \nabla_x g(x, y_2)$ for any $y_1, y_2 \in \mathcal{S}(x)$, further indicating Lemma 6 holds for this example.

Lemma 6 suggests one evaluate $\nabla v(x)$ with any solution of the lower-level problem. Given iteration k and x_k, it is then natural to run the projected GD method to find the solution of the lower-level problem:

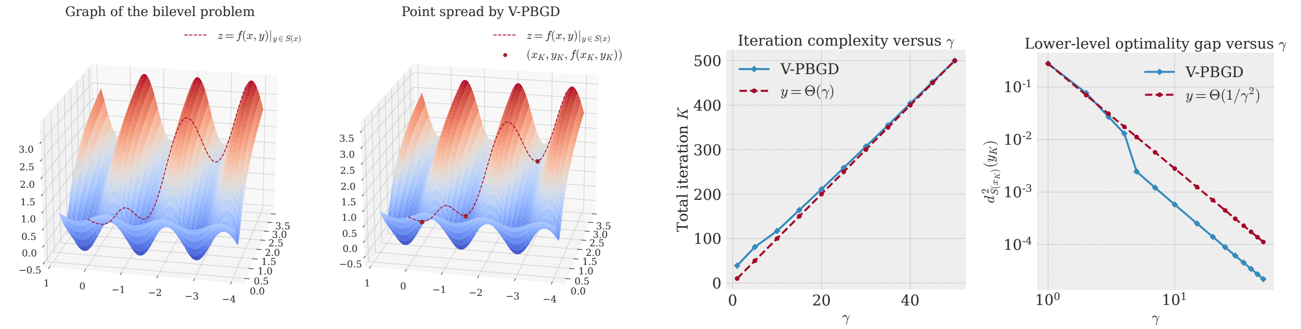
$$\omega_{t+1}^{(k)} = \mathrm{Pj}_{\mathcal{U}}\left(\omega_t^{(k)} - \beta \nabla_y g(x_k, \omega_t^{(k)})\right),\ t = 1, \ldots, T_k. \tag{11}$$

We can then calculate $\nabla v(x_k) \approx \nabla_x g(x_k, \hat{y}_k)$ with $\hat{y}_k = \omega_{T_k+1}^{(k)}$ and update (x_k, y_k) following (7b) with $\mathcal{Z} = \mathcal{C} \times \mathcal{U}$. The V-PBGD update for the constrained lower-level problem is summarized in Algorithm 3.

We provide the convergence result for Algorithm 3 next.

Theorem 4. *Consider V-PBGD with lower-level constraint (Algorithm 3). Suppose Assumption 1, Assumption 3, and either (i)&(ii) or (i)&(iii) in Assumption 4 hold. Also assume the condition in Lemma 6 holds. With a prescribed accuracy $\delta > 0$, select*

$$\alpha = (L_f + \gamma(L_g + L_v))^{-1},\ \beta = L_g^{-1},$$
$$\gamma \textit{ chosen by Proposition 3},\ T_k = \Omega(\log(\alpha\gamma k)).$$



(a) Left figure: the graph of $z=f(x,y)$ and $z=f(x,y)|_{y\in\mathcal{S}(x)}$; right figure: the plot of the last iterates generated by 1000 runs of V-PBGD with random initial points.

(b) Test of V-PBGD with different constant γ. The figure is generated by running V-PBGD (Algorithm 2) for K steps such that $\|G_\gamma(x_K,y_K)\|^2 \leq 10^{-4}$ given γ.

Figure 2: Test of V-PBGD in the bilevel problem (12).

Algorithm 3 V-PBGD under lower-level constraint

```
1: Select $(x_1, y_1) \in \mathcal{Z} = \mathcal{C} \times \mathcal{U}$. Select $\alpha, \beta, \gamma, T_k, K$.
2: for $k = 1$ to $K$ do
3:     Obtain $\hat{y}_k = \omega^{(k)}_{T_k+1}$ following (11).
4:     Update $(x_k, y_k)$ following (7b) with $\mathcal{Z}=\mathcal{C}\times\mathcal{U}$.
5: end for
```

i) With $C_f = \inf_{(x,y)\in\mathcal{Z}} f(x,y)$, it holds that

$$\frac{1}{K}\sum_{k=1}^{K}\|G_\gamma(x_k,y_k)\|^2 \leq \frac{8\big(F_\gamma(x_1,y_1)-C_f\big)}{\alpha K} + \frac{3L_g^2\mu C_g}{K}.$$

ii) Suppose $\lim_{k\to\infty}(x_k,y_k) = (x^*,y^*)$, then (x^*,y^*) is a stationary point of $\mathcal{CP}_{\gamma p}$. If (x^*,y^*) is a local/global solution of $\mathcal{CP}_{\gamma p}$, then (x^*,y^*) is a local/global solution of $\mathcal{BP}_{\epsilon_\gamma}$ with $\mathcal{U}(x)=\mathcal{U}$ and $\epsilon_\gamma \leq \delta$.

The proof of Theorem 4 can be found in Appendix C.3. The above theorem implies an iteration complexity of $\widetilde{\mathcal{O}}(\gamma\epsilon^{-1})$ to find an ϵ-stationary-point of $\mathcal{CP}_{\gamma p}$. This recovers the iteration complexity of the projected GD method (Nesterov, 2013) with a smoothness constant of $\Theta(\gamma)$. If choosing $\delta=\epsilon$, the iteration complexity is $\widetilde{\mathcal{O}}(\epsilon^{-1.5})$ under Condition (b) or $\widetilde{\mathcal{O}}(\epsilon^{-2})$ under Condition (a) in Proposition 3.

5. Simulations

In this section, we first verify our main theoretical results in a toy problem and then compare the PBGD[1] algorithm with several other baselines on the data hyper-cleaning task.

5.1. Numerical verification

We first consider the following non-convex bilevel problem:

$$\min_{x,y} f(x,y) = \frac{\cos(4y+2)}{1+e^{2-4x}} + \frac{1}{2}\ln((4x-2)^2+1) \quad (12)$$
$$\text{s.t. } x\in[0,3],\ y \in \arg\min_{y\in\mathbb{R}} \underbrace{(y+x)^2 + x\sin^2(y+x)}_{g(x,y)}$$

which is a special case of $\mathcal{UP}$. It can be checked that the assumptions in Theorem 3 are satisfied in (12).

We plot the graph of $z = f(x,y)$ in Figure 2a (left). Notice that given any $x \in [0,3]$, we have $\mathcal{S}(x) = \{-x\}$. Thus the the bilevel problem in (12) can be reduced to $\min_{x\in\mathcal{C}} f(x,y)|_{y=-x}$. We plot the single-level objective function $z = f(x,y)|_{y=-x}$ in Figure 2a (left) as the intersected line of the surface $z = f(x,y)$ and the plane $y=-x$. We then run V-PBGD with $\gamma = 10$ for 1000 random initial points (x_1,y_1) and plot the last iterates in Figure 2a (right). It can be observed that V-PBGD consistently finds the local solutions of the bilevel problem (12).

Next we test the impact of γ on the performance of V-PBGD, and report the results in Figure 2b. From Figure 2b, the iteration complexity is $\Theta(\gamma)$, while the lower-level accuracy is $\Theta(1/\gamma^2)$, consistent with Theorem 3.

5.2. Data hyper-cleaning

In this section, we test PBGD in the data hyper-cleaning task (Franceschi et al., 2017; Shaban et al., 2019). In this task, one is given a set of polluted training data, along with a set of clean validation and test data. The goal is to train a data cleaner that assigns smaller weights to the polluted data to improve the generalization in unseen clean data.

We evaluate the performance of our algorithm in terms of speed, memory usage and solution quality in comparison with several competitive baseline algorithms including the

[1] The code is available on github (link).

Table 2: Comparison of solution quality. The results are averaged over 20 runs and ± is followed by an estimated margin of error under 95% confidence. F1 score measures the quality of the data cleaner (Franceschi et al., 2017).

Method	Linear model		2-layer MLP	
	Test accuracy	F1 score	Test accuracy	F1 score
RHG	87.64 ± 0.19	89.71 ± 0.25	87.50 ± 0.23	89.41 ± 0.21
T-RHG	87.63 ± 0.19	89.04 ± 0.24	87.48 ± 0.22	89.20 ± 0.21
BOME	87.09 ± 0.14	89.83 ± 0.18	87.42 ± 0.16	89.26 ± 0.17
G-PBGD	90.09 ± 0.12	90.82 ± 0.19	92.17 ± 0.09	90.73 ± 0.27
IAPTT-GM	90.44 ± 0.14	91.89 ± 0.15	91.72 ± 0.11	91.82 ± 0.19
V-PBGD	90.48 ± 0.13	91.99 ± 0.14	94.58 ± 0.08	93.16 ± 0.15

Table 3: Comparison of GPU memory usage and the runtime to reach the highest test accuracy averaged over 20 runs.

	RHG	T-RHG	BOME	**G-PBGD**	IAPTT-GM	**V-PBGD**
GPU memory (MB) linear	1369	1367	1149	1149	1237	1149
GPU memory (MB) MLP	7997	7757	1201	1235	2613	1199
Runtime (second) linear	73.21	32.28	5.92	7.72	693.65	9.12
Runtime (second) MLP	94.78	54.96	39.78	185.08	1310.63	207.53

IAPTT-GM (Liu et al., 2021b), BOME (Ye et al., 2022), RHG (Franceschi et al., 2017) and T-RHG (Shaban et al., 2019). In addition to V-PBGD, we also test G-PBGD which is a special case of PBGD with $p(x, y) = \|\nabla_y g(x, y)\|^2$.

Adopting the settings in (Franceschi et al., 2017; Liu et al., 2021b; Shaban et al., 2019), we randomly split the MNIST data-set into a training data-set of size 5000, a validation set of size 5000 and a test set of size 10000; and pollute 50% of the training data with uniformly drawn labels. Then we run the algorithms with a linear model and an MLP network.

We report the solution quality in Table 2. It can be observed that both PBGD algorithms achieve competitive performance, and V-PBGD achieves the best performance among the baselines. We also evaluate PBGD in terms of convergence speed and memory usage, which is reported in Table 3. It can be observed that PBGD does not have a steep increase in memory or runtime as compared to the ITD baselines, indicating PBGD is potentially more scalable.

6. Conclusions

In this work, we study the bilevel optimization problem through the lens of the penalty method. We prove that the solutions of the penalized problem approximately solve the original bilevel problem under certain generic conditions verifiable with commonly made assumptions. To solve the penalized problem, we propose the penalty-based bilevel GD method and establish its finite-time convergence under unconstrained and constrained lower-level problems. Experiments verify the effectiveness of the proposed algorithm.

Acknowledgements

The work was partially supported by NSF MoDL-SCALE Grant 2134168 and the Rensselaer-IBM AI Research Collaboration (`http://airc.rpi.edu`), part of the IBM AI Horizons Network (`http://ibm.biz/AIHorizons`).

References

Arbel, M. and Mairal, J. Non-convex bilevel games with critical point selection maps. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Bolte, J., Pauwels, E., and Vaiter, S. Automatic differentiation of nonsmooth iterative algorithms. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Chen, L., Xu, J., and Zhang, J. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.

Chen, T., Sun, Y., and Yin, W. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. In *Proc. of Advances Neural Information Processing Systems*, 2021.

Chen, T., Sun, Y., Xiao, Q., and Yin, W. A single-timescale method for stochastic bilevel optimization. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2022.

Cheng, C., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *Proc. of International Conference on Machine Learning*, 2022.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Dem'yanov, V. F. and Malozemov, V. N. On the theory of non-linear minimax problems. *Russian Mathematical Surveys*, 26(3):57, 1971.

Drusvyatskiy, D. and Lewis, A. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of International Conference on Machine Learning*, 2017.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *Proc. of International Conference on Machine Learning*, 2017.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. of International Conference on Machine Learning*, 2018.

Gao, L., Ye, J., Yin, H., Zeng, S., and Zhang, J. Value function based difference-of-convex algorithm for bilevel hyperparameter selection problems. In *Proc. of International Conference on Machine Learning*, 2022.

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

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1): 267–305, 2016.

Giovannelli, T., Kent, G., and Vicente, L. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *arXiv preprint arXiv:2110.00604*, 2022.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *Proc. of International Conference on Machine Learning*, pp. 3748–3758, 2020.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1), 2023.

Hu, Q., Zhong, Y., and Yang, T. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization. *Proc. of Advances Neural Information Processing Systems*, 2022.

Huang, F., Li, J., Gao, S., and Huang, H. Enhanced bilevel optimization via bregman distance. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Ji, K., Yang, J., and Liang, Y. Provably faster algorithms for bilevel optimization and applications to meta-learning. In *Proc. of International Conference on Machine Learning*, 2021.

Ji, K., Liu, M., Liang, Y., and Ying, L. Will bilevel optimizers benefit from loops. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Jiang, H., Chen, Z., Shi, Y., Dai, B., and Zhao, T. Learning to defend by learning to attack. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2021.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *Proc. of Joint European conference on machine learning and knowledge discovery in databases*, 2016.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Proc. of Advances Neural Information Processing Systems*, 2021.

Li, J., Gu, B., and Huang, H. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proc. of AAAI Conference on Artificial Intelligence*, 2022.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022a.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proc. of International Conference on Machine Learning*, 2020.

Liu, R., Liu, X., Yuan, X., Zeng, S., and Zhang, J. A value-function-based interior-point method for non-convex bi-level optimization. In *Proc. of International Conference on Machine Learning*, 2021a.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. In *Proc. of Advances Neural Information Processing Systems*, 2021b.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):38–57, 2022b.

Lu, S., Cui, X., Squillante, M., Kingsbury, B., and Horesh, L. Decentralized bilevel optimization for personalized client learning. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Lu, Z. and Mei, S. First-order penalty methods for bilevel optimization. *arXiv preprint arXiv:2301.01716*, 2023.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *Proc. of International Conference on Machine Learning*, 2015.

Mehra, A. and Hamm, J. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, 2021.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *Proc. of International Conference on Machine Learning*, 2020.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Nesterov, Y. and Polyak, B. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *Proc. of Advances Neural Information Processing Systems*, 2019.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *Proc. of International Conference on Machine Learning*, 2016.

Rajeswaran, A., Finn, C., Kakade, S., and Levine, S. Meta-learning with implicit gradients. In *Proc. of Advances Neural Information Processing Systems*, 2019.

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

Shaban, A., Cheng, C., Hatch, N., and Boots, B. Truncated back-propagation for bilevel optimization. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2019.

Shen, H. and Chen, T. A single-timescale analysis for stochastic approximation with multiple coupled sequences. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Sow, D., Ji, K., and Liang, Y. On the convergence theory for hessian-free bilevel algorithms. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Stackelberg, H. *The Theory of Market Economy*. Oxford University Press, 1952.

Tarzanagh, D., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. In *Proc. of International Conference on Machine Learning*, 2022.

Vicol, P., Lorraine, J., Pedregosa, F., Duvenaud, D., and Grosse, R. On implicit bias in overparameterized bilevel optimization. In *Proc. of International Conference on Machine Learning*, 2022.

Xiao, Q., Shen, H., Yin, W., and Chen, T. Alternating implicit projected sgd and its efficient variants for equality-constrained bilevel optimization. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2023.

Yang, S., Zhang, X., and Wang, M. Decentralized gossip-based stochastic bilevel optimization over communication networks. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Ye, J., Zhu, D., and Zhu, Q. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on Optimization*, 7 (2), 1997.

Ye, M., Liu, B., Wright, S., Stone, P., and Liu, Q. Bome! bilevel optimization made easy: A simple first-order approach. In *Proc. of Advances Neural Information Processing Systems*, 2022.

Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *Proc. of International Conference on Machine Learning*, 2022.

Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., and Zhang, T. Probabilistic bilevel coreset selection. In *Proc. of International Conference on Machine Learning*, 2022.

Appendix for "On Penalty-based Bilevel Gradient Descent Method"

A. Proof in Section 2

For ease of reading, we restate $\mathcal{BP}$ and $\mathcal{BP}_{\gamma p}$ below.

$$\mathcal{BP}: \min_{x,y} f(x,y) \text{ s.t. } x \in \mathcal{C},\ y \in \mathcal{S}(x) := \arg\min_{y \in \mathcal{U}(x)} g(x,y),$$

and

$$\mathcal{BP}_{\gamma p}: \min_{x,y} f(x,y) + \gamma p(x,y), \text{ s.t. } x \in \mathcal{C},\ y \in \mathcal{U}(x).$$

A.1. Proof of Theorem 1

Theorem 1. *Assume $p(x,y)$ is an ρ-squared-distance-bound function and $f(x,\cdot)$ is L-Lipschitz continuous on $\mathcal{U}(x)$ for any $x \in \mathcal{C}$. Given any $\epsilon_1 > 0$, any global solution of $\mathcal{BP}$ is an ϵ_1-global-minimum point of $\mathcal{BP}_{\gamma p}$ with any $\gamma \geq \gamma^* = \frac{L^2\rho}{4}\epsilon_1^{-1}$. Conversely, given $\epsilon_2 \geq 0$, let (x_γ, y_γ) achieves ϵ_2-global-minimum of $\mathcal{BP}_{\gamma p}$ with $\gamma > \gamma^*$. Then (x_γ, y_γ) is the global solution of the following approximate problem of $\mathcal{BP}$ with some $0 \leq \epsilon_\gamma \leq (\epsilon_1+\epsilon_2)/(\gamma-\gamma^*)$; given by*

$$\begin{aligned} \min_{x,y} f(x,y) \text{ s.t. } & x \in \mathcal{C},\ y \in \mathcal{U}(x), \\ & p(x,y) \leq \epsilon_\gamma. \end{aligned} \tag{13}$$

Proof. Given any $x \in \mathcal{C}$ and $y \in \mathcal{U}(x)$, since $\mathcal{S}(x)$ is closed and non-empty, we can find $y_x \in \arg\min_{y' \in \mathcal{S}(x)} \|y' - y\|$. By Lipschitz continuity assumption on $f(x,\cdot)$, given any $x \in \mathcal{C}$, it holds for any $y \in \mathcal{U}(x)$ that

$$f(x,y) - f(x,y_x) \geq -L d_{\mathcal{S}(x)}(y) \quad \text{by } d_{\mathcal{S}(x)}(y) = \|y_x - y\|.$$

Then it follows that

$$\begin{aligned} f(x,y) + \gamma^* p(x,y) - f(x,y_x) &\geq -L d_{\mathcal{S}(x)}(y) + \gamma^* p(x,y) \\ &\geq -L d_{\mathcal{S}(x)}(y) + \frac{\gamma^*}{\rho} d^2_{\mathcal{S}(x)}(y) \\ &\geq \min_{z \in \mathbb{R}_{\geq 0}} -Lz + \frac{\gamma^*}{\rho} z^2 = -\epsilon_1 \quad \text{with } \gamma^* = \frac{L^2\rho}{4}\epsilon_1^{-1}. \end{aligned} \tag{14}$$

Since $y_x \in \mathcal{S}(x)$ (thus $y_x \in \mathcal{U}(x)$) and $x \in \mathcal{C}$, (x, y_x) is feasible for $\mathcal{BP}$. Let f^* be the optimal objective value for $\mathcal{BP}$, we know $f(x,y_x) \geq f^*$. This along with (14) indicates

$$f(x,y) + \gamma^* p(x,y) - f^* \geq -\epsilon_1,\ \forall x \in \mathcal{C},\ y \in \mathcal{U}(x). \tag{15}$$

Let (x^*, y^*) be a global solution of $\mathcal{BP}$ so that $f(x^*,y^*) = f^*$. Since $y^* \in \mathcal{S}(x^*)$, $p(x^*,y^*) = 0$. By (15), we have

$$f(x^*,y^*) + \gamma^* p(x^*,y^*) \leq f(x,y) + \gamma^* p(x,y) + \epsilon_1,\ \forall x \in \mathcal{C},\ \forall y \in \mathcal{U}(x). \tag{16}$$

Inequality (16) along with the fact that the global solution of $\mathcal{BP}$ is feasible for $\mathcal{BP}_{\gamma p}$ prove that the global solution of $\mathcal{BP}$ achieves ϵ_1-global-minimum for $\mathcal{BP}_{\gamma p}$.

Now for the converse part. Since (x_γ, y_γ) achieves ϵ_2-global-minimum, it holds for any x, y feasible for $\mathcal{BP}_{\gamma p}$ that

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) - \epsilon_1 \leq f(x,y) + \gamma p(x,y) - \epsilon_1 + \epsilon_2. \tag{17}$$

In (17), choosing $(x,y) = (x^*,y^*)$ which is a global solution of $\mathcal{BP}$ yields

$$\begin{aligned} f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) - \epsilon_1 &\leq f(x^*,y^*) - \epsilon_1 + \epsilon_2 \quad \text{since } p(x^*,y^*) = 0 \\ &\leq f(x_\gamma, y_\gamma) + \gamma^* p(x_\gamma, y_\gamma) + \epsilon_2 \quad \text{by (15).} \end{aligned}$$

Then we have

$$(\gamma-\gamma^*)p(x_\gamma,y_\gamma)\le \epsilon_1+\epsilon_2 \Rightarrow p(x_\gamma,y_\gamma)\le(\epsilon_1+\epsilon_2)/(\gamma-\gamma^*).$$

Define $\epsilon_\gamma=p(x_\gamma,y_\gamma)$, then $\epsilon_\gamma\le(\epsilon_1+\epsilon_2)/(\gamma-\gamma^*)$. By (17), it holds for any (x,y) feasible for problem (13) that

$$f(x_\gamma,y_\gamma)+\gamma p(x_\gamma,y_\gamma)\le f(x,y)+\gamma p(x,y)\Rightarrow f(x_\gamma,y_\gamma)-f(x,y)\le\gamma(p(x,y)-\epsilon_\gamma)\le 0,$$

where the last inequality follows from the feasibility of (x,y). This along with the fact that (x_γ,y_γ) is feasible for problem (13) prove that (x_γ,y_γ) is a global solution for problem (13). □

A.2. Proof of Theorem 2

In this section, we give the proof of a stronger version of Theorem 2 in the sense of weaker assumptions.

We first define a new class of functions as follows.

Definition 2 (Restricted α-sublinearity)**.** Let $\mathcal{X}\subseteq\mathbb{R}^d$. We say a function $\ell:\mathcal{X}\mapsto\mathbb{R}$ is restricted α-sublinear on $x\in\mathcal{X}$ if there exists $\alpha\in[0,1]$ and $x^*\in\mathcal{X}$ which is the projection of x onto the minimum point set of ℓ such that the following inequality holds.

$$\ell\big((1-\alpha)x+\alpha x^*\big)\le(1-\alpha)\ell(x)+\alpha\ell(x^*).$$

Suppose ℓ is a continuous convex or more generally a star-convex function (Nesterov & Polyak, 2006, Definition 1) defined on a closed convex set $\mathcal{X}$ and ℓ has a non-empty minimum point set, then ℓ is restricted α-sublinear for any $\alpha\in[0,1]$ on every $x\in\mathcal{X}$.

Now we are ready to give the stronger version of Theorem 2.

Theorem 5 (Stronger version of Theorem 2)**.** *Assume $p(x,\cdot)$ is continuous given any $x\in\mathcal{C}$ and $p(x,y)$ is ρ-squared-distance-bound function. Given $\gamma>0$, let (x_γ,y_γ) be a local solution of $\mathcal{BP}_{\gamma p}$ on $\mathcal{N}((x_\gamma,y_\gamma),r)$. Assume $f(x_\gamma,\cdot)$ is L-Lipschitz-continuous on $\mathcal{N}(y_\gamma,r)$.*

Assume either one of the following is true:

(i) There exists $\bar y\in\mathcal{N}(y_\gamma,r)$ such that $\bar y\in\mathcal{U}(x_\gamma)$ and $p(x_\gamma,\bar y)\le\epsilon$ for some $\epsilon\ge 0$. Define $\bar\epsilon_\gamma=\frac{L^2\rho}{\gamma^2}+2\epsilon$.

(ii) The set $\mathcal{U}(x_\gamma)$ is convex and $p(x_\gamma,\cdot)$ is restricted α-sublinear on y_γ with some $\alpha\in(0,\min\{r/d_{\mathcal{S}(x_\gamma)}(y_\gamma),1\}]$. Define $\bar\epsilon_\gamma=\frac{L^2\rho}{\gamma^2}$.

Then (x_γ,y_γ) is a local solution of the following approximate problem of $\mathcal{BP}$ with $0\le\epsilon_\gamma\le\bar\epsilon_\gamma$.

$$\min_{x,y} f(x,y)\ \text{s.t.}\ x\in\mathcal{C},\ y\in\mathcal{U}(x) \qquad (18)$$
$$p(x,y)\le\epsilon_\gamma.$$

The above theorem is stronger than Theorem 2 in the sense that the condition (ii) in above theorem is weaker than (ii) in Theorem 2 since the continuity and convexity of $p(x_\gamma,\cdot)$ implies the restricted α-sublinearity of $p(x_\gamma,\cdot)$ on y_γ with any $\alpha\in[0,1]$.

Proof of Theorem 5. We will prove the theorem for two cases separately.

(i). Assume (i) is true. For $\delta\ge 0$, define

$$\mathcal{S}_\delta(x):=\{y\in\mathcal{U}(x):p(x,y)\le\delta\},\ x\in\mathcal{C}.$$

Since $p(x,y)=0$ if and only if (iff) $y\in\mathcal{S}(x)=\arg\min_{y\in\mathcal{U}(x)}g(x,y)$, it follows that $\mathcal{S}(x)=\{y\in\mathcal{U}(x):p(x,y)=0\}$. Then $\mathcal{S}_\delta(x)\supseteq\mathcal{S}(x)$, and thus $\mathcal{S}_\delta(x)\ne\emptyset$. Moreover, $\mathcal{S}_\delta(x)$ is closed by continuity of $p(x,\cdot)$ and closeness of $\mathcal{U}(x)$ for $x\in\mathcal{C}$.

Since (x_γ, y_γ) is a local solution of $\mathcal{BP}_{\gamma p}$ on $\mathcal{N}((x_\gamma, y_\gamma), r)$, it holds for any $(x, y) \in \mathcal{N}((x_\gamma, y_\gamma), r)$ that is feasible for $\mathcal{BP}_{\gamma p}$ that

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) \leq f(x, y) + \gamma p(x, y). \tag{19}$$

Since $\mathcal{S}_\epsilon(x_\gamma)$ is closed and non-empty, we can find $y_x \in \arg\min_{y' \in \mathcal{S}_\epsilon(x_\gamma)} \|y' - y_\gamma\|$. Since $\bar{y} \in \mathcal{N}(y_\gamma, r) \cap \mathcal{S}_\epsilon(x_\gamma)$, we have $\|y_x - y_\gamma\| \leq \|\bar{y} - y_\gamma\| \leq r$. This indicates $y_x \in \mathcal{N}(y_\gamma, r)$ and $(x_\gamma, y_x) \in \mathcal{N}((x_\gamma, y_\gamma), r)$. Moreover, since $y_x \in \mathcal{U}(x_\gamma)$, (x_γ, y_x) is feasible for $\mathcal{BP}_{\gamma p}$. This allows to choose $(x, y) = (x_\gamma, y_x)$ in (19), leading to

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) \leq f(x_\gamma, y_x) + \gamma\epsilon \quad \text{since } y_x \in \mathcal{S}_\epsilon(x_\gamma).$$

By Lipschitz continuity of $f(x_\gamma, \cdot)$ on $\mathcal{N}(y_\gamma, r)$, we further have

$$\gamma p(x_\gamma, y_\gamma) - L\|y_x - y_\gamma\| - \gamma\epsilon \leq 0. \tag{20}$$

Since $\mathcal{S}_\epsilon(x_\gamma) \supseteq \mathcal{S}(x_\gamma)$, we have $\|y_x - y_\gamma\| = d_{\mathcal{S}_\epsilon(x_\gamma)}(y_\gamma) \leq d_{\mathcal{S}(x_\gamma)}(y_\gamma) \leq \sqrt{\rho p(x_\gamma, y_\gamma)}$. Plugging this into (20) yields

$$\gamma p(x_\gamma, y_\gamma) - L\sqrt{\rho p(x_\gamma, y_\gamma)} - \gamma\epsilon \leq 0$$

which implies $p(x_\gamma, y_\gamma) \leq \bar{\epsilon}_\gamma = \frac{L^2\rho}{\gamma^2} + 2\epsilon$. Let $\epsilon_\gamma = p(x_\gamma, y_\gamma)$, then $\epsilon_\gamma \leq \bar{\epsilon}_\gamma$ and (x_γ, y_γ) is feasible for problem (18). By (19), it holds for any $(x, y) \in \mathcal{N}((x_\gamma, y_\gamma), r)$ that are feasible for problem (18) that

$$f(x_\gamma, y_\gamma) - f(x, y) \leq \gamma(p(x, y) - \epsilon_\gamma) \leq 0.$$

This and the fact that (x_γ, y_γ) is feasible for (18) imply (x_γ, y_γ) is a local solution of (18).

(ii). Assume (ii) is true. Since $\mathcal{S}(x_\gamma)$ is closed and non-empty, we can find y_x such that $y_x \in \arg\min_{y \in \mathcal{S}(x_\gamma)} \|y - y_\gamma\|$. Let $\bar{y} = (1 - \alpha)y_\gamma + \alpha y_x$. Since $0 < \alpha \leq \min\{r/\|y_\gamma - y_x\|, 1\}$, we know $\bar{y} \in \mathcal{N}(y_\gamma, r)$ and $(x_\gamma, \bar{y}) \in \mathcal{N}((x_\gamma, y_\gamma), r)$. Moreover, since $\mathcal{U}(x_\gamma)$ is convex, we have $\bar{y} \in \mathcal{U}(x_\gamma)$ and $(x_\gamma, \bar{y})$ is feasible for $\mathcal{BP}_{\gamma p}$.

Since (x_γ, y_γ) is a local solution of $\mathcal{BP}_{\gamma p}$ on $\mathcal{N}((x_\gamma, y_\gamma), r)$, we have

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) \leq f(x_\gamma, \bar{y}) + \gamma p(x_\gamma, \bar{y}). \tag{21}$$

Since $p(x_\gamma, y) \geq 0$ and $p(x_\gamma, y) = 0$ iff $d_{\mathcal{S}(x_\gamma)}(y) = 0$, we know the minimum point set of $p(x_\gamma, \cdot)$ is $\mathcal{S}(x_\gamma)$. Then by the restricted α-sublinearity of $p(x_\gamma, \cdot)$ on y_γ, we have

$$p(x_\gamma, \bar{y}) \leq \alpha p(x_\gamma, y_x) + (1 - \alpha)p(x_\gamma, y_\gamma) = (1 - \alpha)p(x_\gamma, y_\gamma).$$

Substituting the above inequality into (21) yields

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) \leq f(x_\gamma, \bar{y}) + \gamma(1 - \alpha)p(x_\gamma, y_\gamma).$$

Re-arranging the above inequality and using the Lipschitz continuity of $f(x_\gamma, \cdot)$ on $\mathcal{N}(y_\gamma, r)$ yield

$$\gamma\alpha p(x_\gamma, y_\gamma) \leq L\alpha d_{\mathcal{S}(x_\gamma)}(y_\gamma) \Rightarrow \gamma\alpha p(x_\gamma, y_\gamma) \leq L\alpha\sqrt{\rho p(x_\gamma, y_\gamma)}$$

which implies $p(x_\gamma, y_\gamma) \leq \bar{\epsilon}_\gamma = \frac{L^2\rho}{\gamma^2}$. Let $\epsilon_\gamma = p(x_\gamma, y_\gamma)$, then $\epsilon_\gamma \leq \bar{\epsilon}_\gamma$ and (x_γ, y_γ) is feasible for problem (18). Since (x_γ, y_γ) is a local solution of $\mathcal{BP}_{\gamma p}$ on $\mathcal{N}((x_\gamma, y_\gamma), r)$, it holds for any $(x, y) \in \mathcal{N}((x_\gamma, y_\gamma), r)$ that is feasible for $\mathcal{BP}_{\gamma p}$ that

$$f(x_\gamma, y_\gamma) + \gamma p(x_\gamma, y_\gamma) \leq f(x, y) + \gamma p(x, y).$$

Following from the above inequality, it holds for any $(x, y) \in \mathcal{N}((x_\gamma, y_\gamma), r)$ that are feasible for problem (18) that

$$f(x_\gamma, y_\gamma) - f(x, y) \leq \gamma(p(x, y) - \epsilon_\gamma) \leq 0.$$

This and the fact that (x_γ, y_γ) is feasible for (18) imply (x_γ, y_γ) is a local solution of (18). □

B. Proof in Section 3

B.1. Proof of Lemma 1

(i). We first consider $p(x,y)=g(x,y)-v(x)$. By the definition of $v(x)$, it is clear that $g(x,y)-v(x)\geq 0$ for any $x\in\mathcal{C}$ and $y\in\mathbb{R}^{d_y}$. Since $\mathcal{S}(x)$ is closed, $y\in\mathcal{S}(x)$ iff $d_{\mathcal{S}(x)}(y)=0$. Then by the definition of $\mathcal{S}(x)$, it holds for any $x\in\mathcal{C}$ and $y\in\mathbb{R}^{d_y}$ that

$$g(x,y)-v(x)=0 \text{ iff } y\in\mathcal{S}(x)\Rightarrow g(x,y)-v(x)=0 \text{ iff } d_{S(x)}(y)=0.$$

It then suffices to check whether $g(x,y)-v(x)$ is an upper-bound of $d_{S(x)}(y)$. By $\frac{1}{\mu}$-PL condition of $g(x,\cdot)$ and (Karimi et al., 2016, Theorem 2), $g(x,\cdot)$ satisfies the $\frac{1}{\mu}$-quadratic-growth condition, and thus for any $x\in\mathcal{C}$ and $y\in\mathbb{R}^{d_y}$, it holds

$$g(x,y)-v(x)\geq\frac{1}{\mu}d^2_{S(x)}(y). \tag{22}$$

This completes the proof.

(ii). Consider $p(x,y)=\mu\|\nabla_y g(x,y)\|^2$. When $g(x,\cdot)$ satisfies PL condition given any $x\in\mathcal{C}$. By the PL inequality, it is clear that $\|\nabla_y g(x,y)\|^2=0$ is equivalent to $g(x,y)=\min_{y\in\mathbb{R}^{d_y}}g(x,y)$ given any $x\in\mathcal{C}$, thus $\|\nabla_y g(x,y)\|^2=0$ iff $d_{\mathcal{S}(x)}(y)=0$ for any $x\in\mathcal{C}$.

By $\frac{1}{\mu}$-PL condition of $g(x,\cdot)$, we have $\|\nabla_y g(x,y)\|^2\geq\frac{1}{\mu}(g(x,y)-v(x))$. By (22), we have $g(x,y)-v(x)\geq\frac{1}{\mu}d^2_{S(x)}(y)$. Thus it holds that

$$\mu^2\|\nabla_y g(x,y)\|^2\geq d^2_{S(x)}(y),$$

which completes the proof.

B.2. Proof of Proposition 2

We prove the proposition from the two conditions separately.

(a) Since (x_γ,y_γ) is a local solution of $\mathcal{UP}_{\gamma p}$, y_γ is a local solution of $\mathcal{UP}_{\gamma p}$ with $x=x_\gamma$. By the first-order stationary condition, it holds that

$$\nabla_y f(x_\gamma,y_\gamma)+\gamma\nabla_y g(x_\gamma,y_\gamma)=0\Rightarrow\|\nabla_y g(x_\gamma,y_\gamma)\|\leq L/\gamma.$$

Since $g(x_\gamma,\cdot)$ satisfies $1/\mu$-PL inequality, it holds that

$$\|\nabla_y g(x_\gamma,y_\gamma)\|^2\geq\frac{1}{\mu}p(x_\gamma,y_\gamma).$$

The above two inequalities imply $p(x_\gamma,y_\gamma)\leq\frac{L^2\mu}{\gamma^2}$. Further notice that $\mathcal{UP}$ and $\mathcal{UP}_{\gamma p}$ are respectively special cases of $\mathcal{BP}$ and $\mathcal{BP}_{\gamma p}$ with $\mathcal{U}(x)=\mathbb{R}^{d_y}$; and $p(x,y)$ is a squared distance bound by Lemma 1, then the result directly follows from Theorem 2 where condition (i) is met with $\bar{y}=y_\gamma$, $\rho=\mu$ and $\epsilon=\frac{L^2\mu}{\gamma^2}$ with $\gamma\geq L\sqrt{3\mu\delta^{-1}}$.

(b) Suppose $y_\gamma\notin\mathcal{S}(x_\gamma)$. Since (x_γ,y_γ) is a local solution of $\mathcal{UP}_{\gamma p}$, y_γ is a local solution of $\mathcal{UP}_{\gamma p}$ with $x=x_\gamma$. By the first-order stationary condition, it holds that

$$\nabla_y f(x_\gamma,y_\gamma)+2\mu\gamma\nabla_{yy}g(x_\gamma,y_\gamma)\nabla_y g(x_\gamma,y_\gamma)=0\Rightarrow\|\nabla_{yy}g(x_\gamma,y_\gamma)\nabla_y g(x_\gamma,y_\gamma)\|\leq L/2\mu\gamma$$

which along with the assumption that the singular values of $\nabla_{yy}g(x,y)$ on $\{x\in\mathcal{C},y\in\mathcal{U}(x):y\notin\mathcal{S}(x)\}$ are lower bounded by $\sigma>0$ gives

$$p(x_\gamma,y_\gamma)=\mu\|\nabla_y g(x_\gamma,y_\gamma)\|^2\leq\mu\frac{\|\nabla_{yy}g(x_\gamma,y_\gamma)\nabla_y g(x_\gamma,y_\gamma)\|^2}{\sigma^2}\leq L^2/(4\gamma^2\mu\sigma^2). \tag{23}$$

When $y_\gamma\in\mathcal{S}(x_\gamma)$, we know $p(x_\gamma,y_\gamma)=0$ and thus (23) still holds. Further notice that $\mathcal{UP}$ and $\mathcal{UP}_{\gamma p}$ are special cases of $\mathcal{BP}$ and $\mathcal{BP}_{\gamma p}$ with $\mathcal{U}(x)=\mathbb{R}^{d_y}$; and $p(x,y)$ is a squared distance bound by Lemma 1, then the result directly follows from Theorem 2 where condition (i) is met with $\bar{y}=y_\gamma$, $\rho=\mu$ and $\epsilon=L^2/(4\gamma^2\mu\sigma^2)$ with $\gamma\geq\max\{L\sqrt{2\mu\delta^{-1}},L\sqrt{\mu^{-1}\delta^{-1}}/\sigma\}$.

B.3. Proof of Theorem 3

We first provide the convergence theorem on the sequence $\{\omega_t^{(k)}\}$ given the outer iteration k.

Theorem 6. *Assume there exist $L_g > 0$ and $\mu > 0$ such that $g(x,\cdot)$ is L_g-Lipschitz-smooth and $\frac{1}{\mu}$-PL given any $x \in \mathcal{C}$. Choose $\beta \in (0, \frac{1}{L_g}]$. Given any $x_k \in C$, $\omega_1^{(k)} \in \mathbb{R}^{d_y}$, running T_k steps of inner GD updates* (7a) *gives $\hat{y}_k$ satisfying*

$$d^2_{S(x_k)}(\hat{y}_k) \leq \mu\big(1 - \frac{\beta}{2\mu}\big)^{T_k}\big(g(x_k, \omega_1^{(k)}) - v(x_k)\big). \tag{24}$$

Proof of Theorem 6. We omit all index k since this proof holds given any k. By smoothness of $g(x,\cdot)$, it holds that

$$\begin{aligned} g(x, \omega_{t+1}) &\leq g(x,\omega_t) - \beta\|\nabla_y g(x,\omega_t)\|^2 + \frac{L_g\beta^2}{2}\|\nabla_y g(x,\omega_t)\|^2 \\ &\leq g(x,\omega_t) - \frac{\beta}{2}\|\nabla_y g(x,\omega_t)\|^2 \quad \text{since } L_g\beta \leq 1. \end{aligned}$$

By $\frac{1}{\mu}$-PL condition of $g(x,\cdot)$, we further have

$$\begin{aligned} g(x,\omega_{t+1}) - v(x) &\leq g(x,\omega_t) - v(x) - \frac{\beta}{2\mu}\big(g(x,\omega_t) - v(x)\big) \\ &\leq \big(1 - \frac{\beta}{2\mu}\big)\big(g(x,\omega_t) - v(x)\big). \end{aligned}$$

Iteratively applying the above inequality for $t = 1, \ldots, T$ yields

$$g(x,\omega_{T+1}) - v(x) \leq \big(1 - \frac{\beta}{2\mu}\big)^T\big(g(x,\omega_1) - v(x)\big). \tag{25}$$

By (Karimi et al., 2016, Theorem 2), the $\frac{1}{\mu}$-PL condition of $g(x,\cdot)$ also implies the error bound of $g(x,\cdot)$, which leads to

$$g(x,\omega_{T+1}) - v(x) \geq \frac{1}{\mu}d^2_{S(x)}(\omega_{T+1}),\; x \in \mathcal{C}$$

which along with (25) proves the result in (24). □

Notice the term $g(x_k,\omega_1^{(k)}) - v(x_k)$ in (24) depends on the drifting variable x_k. If $\omega_1^{(k)}$ is not carefully chosen, $g(x_k,\omega_1^{(k)}) - v(x_k)$ can grow unbounded with k and hence hinder the convergence. To prevent this, we choose $\omega_1^{(k)} = y_k$ in the analysis. Since $g(x,\cdot)$ is $\frac{1}{\mu}$-PL for any $x \in \mathcal{C}$, it holds that

$$\begin{aligned} g(x_k,\omega_1^{(k)}) - v(x_k) \leq \frac{1}{\mu}\|\nabla_y g(x_k,\omega_1^{(k)})\|^2 &= \frac{1}{\mu}\left\|\frac{y_k - y_{k+1} - \alpha\nabla_y f(x_k,y_k)}{\alpha\gamma}\right\|^2 \\ &\leq \frac{2}{\mu\gamma^2\alpha^2}\|y_{k+1} - y_k\|^2 + \frac{2L^2}{\mu\gamma^2} \end{aligned} \tag{26}$$

where we have used Young's inequality and the condition that $f(x_k,\cdot)$ is L-Lipschitz-continuous. Later we will show that the inexact gradient descent update (7b) decreases $\|(x_{k+1},y_{k+1}) - (x_k,y_k)\|$ and therefore upper-bounds $g(x_k,\omega_1^{(k)}) - v(x_k)$.

Next we give the proof of Theorem 3.

Proof of Theorem 3. In this proof, we write $z = (x,y)$. Update (7b) can be written as

$$z_{k+1} = \mathrm{Pj}_{\mathcal{Z}}\left(z_k - \alpha\hat{\nabla}F_\gamma(z_k;\hat{y}_k)\right)$$

where $\hat{\nabla}F_\gamma(z_k;\hat{y}_k) := \nabla f(z_k) + \gamma(\nabla g(z_k) - \overline{\nabla}_x g(x_k,\hat{y}_k))$.

By the assumptions made in this theorem and Lemma 2, F_γ is L_γ-Lipschitz-smooth with $L_\gamma = L_f + \gamma(2L_g + L_g^2\mu)$. Then by Lipschitz-smoothness of F_γ, it holds that

$$\begin{aligned} F_\gamma(z_{k+1}) &\le F_\gamma(z_k) + \langle \nabla F_\gamma(z_k),\, z_{k+1}-z_k\rangle + \frac{L_\gamma}{2}\|z_{k+1}-z_k\|^2 \\ &\overset{\alpha\le\frac{1}{L_\gamma}}{\le} F_\gamma(z_k) + \langle \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z_{k+1}-z_k\rangle + \frac{1}{2\alpha}\|z_{k+1}-z_k\|^2 + \langle \nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z_{k+1}-z_k\rangle. \end{aligned} \tag{27}$$

Consider the second term in the RHS of (27). By Lemma 7, z_{k+1} can be written as

$$z_{k+1} = \arg\min_{z\in\mathcal{Z}} \langle \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z\rangle + \frac{1}{2\alpha}\|z-z_k\|^2.$$

By the first-order optimality condition of the above problem, it holds that

$$\langle \hat{\nabla} F_\gamma(z_k;\hat{y}_k) + \frac{1}{\alpha}(z_{k+1}-z_k),\, z_{k+1}-z\rangle \le 0,\ \forall z\in\mathcal{Z}.$$

Since $z_k\in\mathcal{Z}$, we can choose $z=z_k$ in the above inequality and obtain

$$\langle \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z_{k+1}-z_k\rangle \le -\frac{1}{\alpha}\|z_{k+1}-z_k\|^2. \tag{28}$$

Consider the last term in the RHS of (27). By Young's inequality, we first have

$$\langle \nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z_{k+1}-z_k\rangle \le \alpha\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 + \frac{1}{4\alpha}\|z_{k+1}-z_k\|^2 \tag{29}$$

where the first term in the above inequality can be bounded as

$$\begin{aligned} \|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 &= \gamma^2\|\nabla v(x_k) - \overline{\nabla}_x g(x_k,\hat{y}_k)\|^2 \\ &= \gamma^2\|\nabla g(x_k,y^*)|_{y^*\in\mathcal{S}(x_k)} - \overline{\nabla}_x g(x_k,\hat{y}_k)\|^2 \quad \text{by Lemma 2,} \\ &\le \gamma^2 L_g^2 d^2_{\mathcal{S}(x_k)}(\hat{y}_k) \quad \text{by choosing } y^*\in\arg\min_{y'\in\mathcal{S}(x_k)}\|y'-\hat{y}_k\|, \\ &\le \gamma^2 L_g^2\mu\big(1-\frac{\beta}{2\mu}\big)^{T_k}\big(g(x_k,\omega_1^{(k)}) - v(x_k)\big) \quad \text{by Theorem 6,} \\ &\le \big(1-\frac{\beta}{2\mu}\big)^{T_k}\big(\frac{2L_g^2}{\alpha^2}\|z_{k+1}-z_k\|^2 + 2L^2L_g^2\big) \quad \text{by (26),} \\ &\le \frac{1}{8\alpha^2}\|z_{k+1}-z_k\|^2 + \frac{L^2L_g^2}{2\alpha^2k^2} \end{aligned} \tag{30}$$

where the last inequality requires $T_k \ge \max\{-\log_{c_\beta}(16L_g^2), -2\log_{c_\beta}(2\alpha k)\}$ with $c_\beta = 1-\frac{\beta}{2\mu}$.

Plugging the inequality (30) into (29) yields

$$\langle \nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\, z_{k+1}-z_k\rangle \le \frac{3}{8\alpha}\|z_{k+1}-z_k\|^2 + \frac{L^2L_g^2}{2\alpha k^2}. \tag{31}$$

Substituting (31) and (28) into (27) and rearranging the resulting inequality yield

$$\frac{1}{8\alpha}\|z_{k+1}-z_k\|^2 \le F_\gamma(z_k) - F_\gamma(z_{k+1}) + \frac{L^2L_g^2}{2\alpha k^2}. \tag{32}$$

With $\bar{z}_{k+1}$ defined in (8), we have

$$\begin{aligned} \|\bar{z}_{k+1}-z_k\|^2 &\le 2\|\bar{z}_{k+1}-z_{k+1}\|^2 + 2\|z_{k+1}-z_k\|^2 \\ &\le 2\alpha^2\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 + 2\|z_{k+1}-z_k\|^2 \\ &\le \frac{9}{4}\|z_{k+1}-z_k\|^2 + \frac{L^2L_g^2}{k^2} \end{aligned} \tag{33}$$

where the second inequality uses non-expansiveness of $\mathrm{Pj}_{\mathcal{Z}}$ and the last one follows from (30).

Together (32) and (33) imply

$$\|\bar{z}_{k+1} - z_k\|^2 \leq 18\alpha\big(F_\gamma(z_k) - F_\gamma(z_{k+1})\big) + \frac{10L^2L_g^2}{k^2}.$$

Since $f(z) \geq C_f$ for any $z \in \mathcal{Z}$ and $g(x,y) - v(x) \geq 0$, $F_\gamma(z) \geq C_f$ for any $z \in \mathcal{Z}$. Taking a telescope sum of the above inequality and using $G_\gamma(z_k) = \frac{1}{\alpha}(z_k - \bar{z}_{k+1})$ yield

$$\sum_{k=1}^{K} \|G_\gamma(z_k)\|^2 \leq \frac{18\big(F_\gamma(z_1) - C_f\big)}{\alpha} + \sum_{k=1}^{K} \frac{10L^2L_g^2}{k^2}$$

which along with the fact $\sum_{k=1}^{K} \frac{1}{k^2} \leq \int_1^K \frac{1}{x^2}dx = 1 - \frac{1}{K}$ implies

$$\sum_{k=1}^{K} \|G_\gamma(z_k)\|^2 \leq \frac{18\big(F_\gamma(z_1) - C_f\big)}{\alpha} + 10L^2L_g^2. \tag{34}$$

This proves (i) in Theorem 3.

Suppose $\lim_{k\to\infty} z_k = z^*$. Since $\nabla F_\gamma(z)$ is continuous, $G_\gamma(z)$ is continuous and thus $\lim_{k\to\infty} G_\gamma(z_k) = G_\gamma(z^*)$. By (34), $G_\gamma(z^*) = 0$, that is $z^* = \mathrm{Pj}_{\mathcal{Z}}\left(z^* - \alpha\nabla F_\gamma(z^*)\right)$. This further implies

$$\langle \nabla F_\gamma(z^*),\, z^* - z\rangle \leq 0,\ \forall z \in \mathcal{Z}$$

which indicates z^* is a stationary point of $\mathcal{UP}_{\gamma p}$. If z^* is a local/global solution, it follows from Proposition 1 and 2 that the rest of the result holds. □

Lemma 7. *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be a closed convex set. Given any $z \in \mathcal{Z}$, $q \in \mathbb{R}^d$ and $\alpha > 0$, it holds that*

$$\mathrm{Pj}_{\mathcal{Z}}\left(z - \alpha q\right) = \arg\min_{z' \in \mathcal{Z}} \langle q,\, z'\rangle + \frac{1}{2\alpha}\|z - z'\|^2.$$

Proof. Given $z \in \mathbb{R}^d$, define $z^* = \arg\min_{z' \in \mathbb{R}^d} E(z')$ where

$$E(z') := \langle q,\, z' - z\rangle + \frac{1}{2\alpha}\|z' - z\|^2. \tag{35}$$

By the optimality condition, it follows $z^* = z - \alpha q$. For any $z' \in \mathbb{R}^d$, it follows that

$$\begin{aligned} E(z') - E(z^*) &= \langle q,\, z' - z\rangle + \frac{1}{2\alpha}\|z' - z\|^2 - \langle q,\, -\alpha q\rangle - \frac{\alpha}{2}\|q\|^2 \\ &= \frac{1}{2\alpha}\|z' - z\|^2 + \langle q,\, z' - z\rangle + \frac{\alpha}{2}\|q\|^2 \\ &= \frac{1}{2\alpha}\|z' - (z - \alpha q)\|^2. \end{aligned} \tag{36}$$

Then we have

$$\begin{aligned} \arg\min_{z' \in \mathcal{Z}} \langle q,\, z'\rangle + \frac{1}{2\alpha}\|z - z'\|^2 &= \arg\min_{z' \in \mathcal{Z}} E(z') \\ &= \arg\min_{z' \in \mathcal{Z}} \|z' - (z - \alpha q)\|^2 \ \text{ by (36)} \\ &= \mathrm{Pj}_{\mathcal{Z}}\left(z - \alpha q\right). \end{aligned} \tag{37}$$

This proves the result. □

B.4. Extension to the stochastic case

In Seciton 3, we have studied the deterministic V-PBGD algorithm. In this section, we extend the V-PBGD algorithm to the stochastic case.

With random variables ξ, ψ, we assume access to $\nabla g(x, y; \psi)$ and $\nabla f(x, y; \xi)$ which are respectively stochastic versions of $\nabla g(x, y)$ and $\nabla f(x, y)$. Following the idea of V-PBGD, given iteration k and x_k, we first solve the lower-level problem with the stochastic gradient descent method:

$$\omega_{t+1}^{(k)} = \omega_t^{(k)} - \beta_t \nabla_y g(x_k, \omega_t^{(k)}; \psi_t^{(k)}) \text{ for } t = 1, \dots, T_k \tag{38}$$

Then we choose the approximate lower-level solution $\hat{y}_k = \omega_i^{(k)}$ where i is drawn from a step-size weighted distribution specified by $\mathbf{P}(i = t) = \beta_t / \sum_{t=1}^{T_k} \beta_t,\ t = 1, ..., T_k$. Given $\hat{y}_k$ and the batch size M, (x_k, y_k) is updated with the approximate stochastic gradient of $F_\gamma(x_k, y_k)$ as follows:

$$(x_{k+1}, y_{k+1}) = \mathrm{Pj}_{\mathcal{Z}} \Big((x_k, y_k) - \frac{\alpha_k}{M} \sum_{i=1}^{M} \big(\nabla f(x_k, y_k; \xi_k^i) + \gamma (\nabla g(x_k, y_k; \psi_k^i) - \overline{\nabla}_x g(x_k, \hat{y}_k; \psi_k^i)) \big) \Big).$$

The update is summarized in Algorithm 4.

Algorithm 4 V-PBSGD: Value-gap based penalized bilevel stochastic gradient descent.

```
Select $(x_1, y_1) \in \mathcal{Z} = \mathcal{C} \times \mathbb{R}^{d_y}$. Select $\gamma, K, T_k, \alpha_k, \beta_t$ and $M$.
for $k = 1$ to $K$ do
    Choose $\omega_1^{(k)} = y_k$, do $\omega_{t+1}^{(k)} = \omega_t^{(k)} - \beta_t \nabla_y g(x_k, \omega_t^{(k)}; \psi_t^{(k)})$ for $t = 1, \dots, T_k$
    Choose $\hat{y}_k = \omega_i^{(k)}$, $i \sim \mathbf{P}$ where $\mathbf{P}(i = t) = \beta_t / \sum_{t=1}^{T_k} \beta_t,\ t = 1, ..., T_k$.
    $(x_{k+1}, y_{k+1}) = \mathrm{Pj}_{\mathcal{Z}} \Big( (x_k, y_k) - \frac{\alpha_k}{M} \sum_{i=1}^{M} \big( \nabla f(x_k, y_k; \xi_k^i) + \gamma (\nabla g(x_k, y_k; \psi_k^i) - \overline{\nabla}_x g(x_k, \hat{y}_k; \psi_k^i)) \big) \Big)$.
end for
```

We make the following standard assumption commonly used in the analysis for stochastic gradient methods.

Assumption 5. *There exists constant $c > 0$ such that given any k, the stochastic gradients in Algorithm 4 are unbiased and have variance bounded by c.*

With the above assumption, we provide the convergence result as follows.

Theorem 7. *Consider V-PBSGD (Algorithm 4). Assume Assumption 5 and the assumptions in Theorem 3 hold. Choose $\alpha_k = \alpha \leq (L_f + \gamma(2L_g + L_g^2 \mu))^{-1}$, $\beta_t = 1/(L_g \sqrt{t})$ and $T_k = T$ for any k. It holds that*

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| G_\gamma(x_k, y_k) \|^2 = \mathcal{O}\Big(\frac{1}{\alpha K} \Big) + \mathcal{O}\Big(\frac{\gamma^2 c^2}{M} \Big) + \mathcal{O}\Big(\frac{\gamma^2 \ln T}{\sqrt{T}} \Big). \tag{39}$$

Proof. **Convergence of ω.** We omit the superscription (k) of $\omega_t^{(k)}$ and $\psi_t^{(k)}$ since the proof holds for any k. We write $\mathbb{E}_t[\cdot]$ as the conditional expectation given the filtration of samples before iteration (k, t). By the L_g-Lipschitz-smoothness of $g(x, \cdot)$, it holds that

$$\begin{aligned} \mathbb{E}_t[g(x_k, \omega_{t+1})] &\leq g(x_k, \omega_t) + \langle \nabla_y g(x_k, \omega_t),\, \mathbb{E}_t[\omega_{t+1} - \omega_t] \rangle + \frac{L_g}{2} \mathbb{E}_t \| \omega_{t+1} - \omega_t \|^2 \\ &\leq g(x_k, \omega_t) - \beta_t \| \nabla_y g(x_k, \omega_t) \|^2 + \frac{L_g \beta_t^2}{2} \mathbb{E}_t \| \nabla_y g(x_k, \omega_t; \psi_t) \|^2 \end{aligned} \tag{40}$$

which follows $\nabla_y g(x_k, \omega_t; \psi_t)$ is unbiased.

The last term of (40) can be bounded as

$$\begin{aligned} \mathbb{E}_t \| \nabla_y g(x_k, \omega_t; \psi_t) \|^2 &= \mathbb{E}_t \| \nabla_y g(x_k, \omega_t; \psi_t) - \nabla_y g(x_k, \omega_t) + \nabla_y g(x_k, \omega_t) \|^2 \\ &= \mathbb{E}_t \| \nabla_y g(x_k, \omega_t; \psi_t) - \nabla g(x_k, \omega_t) \|^2 + \| \nabla_y g(x_k, \omega_t \|^2 \\ &\leq c^2 + \| \nabla_y g(x_k, \omega_t) \|^2. \end{aligned} \tag{41}$$

Substituting the above inequality back to (40) yields

$$\begin{aligned}\mathbb{E}_t[g(x_k,\omega_{t+1})] &\leq g(x_k,\omega_t) - \frac{\beta_t}{2}\|\nabla_y g(x_k,\omega_t)\|^2 + \frac{L_g c^2}{2}\beta_t^2. \\ &\leq g(x_k,\omega_t) - \frac{\beta_t}{2\mu^2} d^2_{\mathcal{S}(x_k)}(\omega_t) + \frac{L_g c^2}{2}\beta_t^2 \end{aligned} \tag{42}$$

where the first inequality requires $\beta_t \leq L_g^{-1}$ and the last one follows from the fact that Lipschitz-smooth $1/\mu$-PL function $g(x,\cdot)$ satisfies $1/\mu^2$-error bound (Karimi et al., 2016, Theorem 2).

We write $\mathbb{E}_k[\cdot]$ as the conditional expectation given the filtration of samples before iteration k. Taking $\mathbb{E}_k$ and a telescope sum over both sides of (42) yields

$$\begin{aligned}\sum_{t=1}^{T_k}\beta_t\mathbb{E}_k[d^2_{\mathcal{S}(x_k)}(\omega_t^{(k)})] &\leq 2\mu^2(g(x_k,\omega_1^{(k)}) - g(x_k,\omega_{T_k+1}^{(k)})) + \frac{2L_g c^2\mu^2}{2}\sum_{t=1}^{T_k}\beta_t^2 \\ &\leq 2\mu^2(g(x_k,\omega_1^{(k)}) - v(x_k)) + L_g c^2\mu^2\sum_{t=1}^{T_k}\beta_t^2. \end{aligned}\tag{43}$$

Convergence of (x,y). In this proof, we write $z=(x,y)$. Given z_k, define $\bar{z}_{k+1} = \mathrm{Pj}_{\mathcal{Z}}(z_k - \alpha_k\nabla F_\gamma(z_k))$. For convenience, we also write

$$\begin{aligned}\overline{\nabla}_k F_\gamma &= \frac{1}{M}\sum_{i=1}^{M}\left(\nabla f(x_k,y_k;\xi_k^i) + \gamma(\nabla g(x_k,y_k;\psi_k^i) - \overline{\nabla}_x g(x_k,\hat{y}_k;\psi_k^i)\right) \\ \hat{\nabla}F_\gamma(z_k;\hat{y}_k) &= \mathbb{E}_k[\overline{\nabla}_k F_\gamma] = \nabla f(x_k,y_k) + \gamma(\nabla g(x_k,y_k) - \overline{\nabla}_x g(x_k,\hat{y}_k)). \end{aligned}\tag{44}$$

By the assumptions made in this theorem, F_γ is L_γ-Lipschitz-smooth with $L_\gamma = L_f + \gamma(2L_g + L_g^2\mu)$. Then by Lipschitz-smoothness of F_γ, it holds that

$$\begin{aligned}\mathbb{E}_k[F_\gamma(z_{k+1})] &\leq F_\gamma(z_k) + \mathbb{E}_k\langle\nabla F_\gamma(z_k),\, z_{k+1}-z_k\rangle + \frac{L_\gamma}{2}\mathbb{E}_k\|z_{k+1}-z_k\|^2 \\ &\overset{L_\gamma\leq\frac{1}{\alpha_k}}{\leq} F_\gamma(z_k) + \mathbb{E}_k\langle\overline{\nabla}_k F_\gamma,\, z_{k+1}-z_k\rangle + \mathbb{E}_k\langle\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma,\, z_{k+1}-\bar{z}_{k+1}\rangle \\ &\quad + \mathbb{E}_k\langle\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma,\, \bar{z}_{k+1}-z_k\rangle + \frac{1}{2\alpha_k}\mathbb{E}_k\|z_{k+1}-z_k\|^2. \end{aligned}\tag{45}$$

Consider the second term in the RHS of (45). By Lemma 7, z_{k+1} can be written as

$$z_{k+1} = \arg\min_{z\in\mathcal{Z}}\langle\overline{\nabla}_k F_\gamma,\, z\rangle + \frac{1}{2\alpha_k}\|z-z_k\|^2.$$

By the first-order optimality condition of the above problem, it holds that

$$\langle\overline{\nabla}_k F_\gamma + \frac{1}{\alpha_k}(z_{k+1}-z_k),\, z_{k+1}-z\rangle \leq 0,\ \forall z\in\mathcal{Z}.$$

Since $z_k\in\mathcal{Z}$, we can choose $z=z_k$ in the above inequality and obtain

$$\langle\overline{\nabla}_k F_\gamma,\, z_{k+1}-z_k\rangle \leq -\frac{1}{\alpha_k}\|z_{k+1}-z_k\|^2. \tag{46}$$

The third term in the RHS of (45) can be bounded as

$$\begin{aligned}\mathbb{E}_k\langle\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma,\, z_{k+1}-\bar{z}_{k+1}\rangle &\leq \mathbb{E}_k[\|\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma\|\|z_{k+1}-\bar{z}_{k+1}\|] \\ &\leq \alpha_k\mathbb{E}_k\|\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma\|^2 \end{aligned}\tag{47}$$

where the second inequality follows from the non-expansiveness of the projection operator.

The fourth term in the RHS of (45) can be bounded as

$$\begin{aligned}\mathbb{E}_k\langle \nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma,\ \bar{z}_{k+1} - z_k\rangle &= \langle \nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k),\ \bar{z}_{k+1} - z_k\rangle \\ &\le 2\alpha_k \mathbb{E}_k\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 + \frac{1}{8\alpha_k}\|\bar{z}_{k+1} - z_k\|^2 \end{aligned} \tag{48}$$

where the last inequality follows from Young's inequality. In addition, we have

$$\begin{aligned}&\|\bar{z}_{k+1} - z_k\|^2 \\ &\le 2\mathbb{E}_k\|\bar{z}_{k+1} - z_{k+1}\|^2 + 2\mathbb{E}_k\|z_{k+1} - z_k\|^2 \\ &\le 2\alpha_k^2\mathbb{E}_k\|\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma\|^2 + 2\mathbb{E}_k\|z_{k+1} - z_k\|^2 \\ &\le 4\alpha_k^2\mathbb{E}_k\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 + 4\alpha_k^2\mathbb{E}_k\|\hat{\nabla} F_\gamma(z_k;\hat{y}_k) - \overline{\nabla}_k F_\gamma\|^2 + 2\mathbb{E}_k\|z_{k+1} - z_k\|^2\end{aligned}$$

which after rearranging gives

$$\begin{aligned}&\mathbb{E}_k\|z_{k+1} - z_k\|^2 \\ &\ge \frac{1}{2}\|\bar{z}_{k+1} - z_k\|^2 - 2\alpha_k^2\mathbb{E}_k\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 - 2\alpha_k^2\mathbb{E}_k\|\hat{\nabla} F_\gamma(z_k;\hat{y}_k) - \overline{\nabla}_k F_\gamma\|^2.\end{aligned} \tag{49}$$

Substituting (46)–(49) into (45) and rearranging yields

$$\begin{aligned}\frac{1}{8\alpha_k}\mathbb{E}_k\|\bar{z}_{k+1} - z_k\|^2 \ &\le F_\gamma(z_k) - \mathbb{E}_k[F_\gamma(z_{k+1})] + 2\alpha_k\mathbb{E}_k\|\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma\|^2 \\ &\quad + 3\alpha_k\mathbb{E}_k\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2.\end{aligned} \tag{50}$$

Under Assumption 5, the third term in the RHS of (50) is bounded by the $\mathcal{O}(1/M)$ dependence of variance as follows

$$\mathbb{E}_k\|\nabla F_\gamma(z_k) - \overline{\nabla}_k F_\gamma\|^2 \le \frac{3(2\gamma^2+1)c^2}{M}. \tag{51}$$

The fourth term in the RHS of (50) can be bounded by

$$\begin{aligned}\mathbb{E}_k\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k;\hat{y}_k)\|^2 &= \gamma^2\mathbb{E}_k\|\nabla v(x_k) - \nabla_x g(x_k,\hat{y}_k)\|^2 \\ &\overset{\text{Lemma 2}}{=} \gamma^2\mathbb{E}_k\|\nabla_x g(x_k,y)|_{y\in\mathcal{S}(x_k)} - \nabla_x g(x_k,\hat{y}_k)\|^2 \\ &\le \gamma^2 L_g^2\mathbb{E}_k[d^2_{\mathcal{S}(x_k)}(\hat{y}_k)] \\ &= \gamma^2 L_g^2\sum_{t=1}^{T}\frac{\beta_t\mathbb{E}_k[d^2_{\mathcal{S}(x_k)}(\omega_t^{(k)})]}{\sum_{i=1}^{T}\beta_i}\end{aligned} \tag{52}$$

where the last equality follows from the distribution of $\hat{y}_k$.

By (43), it holds that

$$\sum_{t=1}^{T}\beta_t\mathbb{E}_k[d^2_{\mathcal{S}(x_k)}(\omega_t^{(k)})] \le 2\mu^2(g(x_k,\omega_1^{(k)}) - v(x_k)) + L_g c^2\mu^2\sum_{t=1}^{T}\beta_t^2. \tag{53}$$

In the above inequality, we can further bound the initial gap as (cf. $\omega_1^{(k)} = y_k$)

$$\begin{aligned}g(x_k,\omega_1^{(k)}) - v(x_k) \le \frac{1}{\mu}\|\nabla_y g(x_k,\omega_1^{(k)})\|^2 &= \frac{1}{\mu}\|\frac{y_k - \bar{y}_{k+1} - \alpha_k\nabla_y f(x_k,y_k)}{\alpha_k\gamma}\|^2 \\ &\le \frac{2}{\mu\gamma^2\alpha_k^2}\|\bar{z}_{k+1} - z_k\|^2 + \frac{2L^2}{\mu\gamma^2}\end{aligned} \tag{54}$$

where the first inequality follows from $g(x,\cdot)$ is $1/\mu$-PL; the equality follows from the definition of $\bar{z}_{k+1}$; and the last one follows from Young's inequality and the Lipschitz continuity of $f(x,\cdot)$.

Substituting (53) and (54) into (52) yields

$$\mathbb{E}_k\|\nabla F_\gamma(z_k)-\hat{\nabla}F_\gamma(z_k;\hat{y}_k)\|^2\leq\frac{1}{48\alpha_k^2}\|\bar{z}_{k+1}-z_k\|^2+\frac{4\mu L^2L_g^2}{\sum_{i=1}^T\beta_i}+\gamma^2L_g^3c^2\mu^2\frac{\sum_{t=1}^T\beta_t^2}{\sum_{t=1}^T\beta_t} \tag{55}$$

where we have also used $\sum_{i=1}^T\beta_i^2\geq 192\mu L_g^2$ to simplify the first term. This can always be satisfied by a large enough T. Substituting (55) and (51) into (50), rearranging and taking total expectation yield

$$\begin{aligned}\frac{1}{16\alpha_k}\mathbb{E}\|\bar{z}_{k+1}-z_k\|^2 &\leq\mathbb{E}[F_\gamma(z_k)-F_\gamma(z_{k+1})]+\frac{6(2\gamma^2+1)c^2}{M}\alpha_k\\ &\quad+\Big(\frac{12\mu L^2L_g^2}{\sum_{i=1}^T\beta_i}+3\gamma^2L_g^3c^2\mu^2\frac{\sum_{t=1}^T\beta_t^2}{\sum_{t=1}^T\beta_t}\Big)\alpha_k.\end{aligned}$$

Using $\|\bar{z}_{k+1}-z_k\|^2=\alpha_k^2\|G_\gamma(z_k)\|^2$ in the LHS of the above inequality and taking telescope sum over $k=1,\dots,K$ yields

$$\sum_{k=1}^K\alpha_k\mathbb{E}\|G_\gamma(z_k)\|^2=\mathcal{O}(F_\gamma(z_1)-C_f)+\mathcal{O}\Big(\frac{\gamma^2c^2}{M}\alpha_k\Big)+\mathcal{O}\Big(\frac{\gamma^2\sum_{t=1}^T\beta_t^2}{\sum_{i=1}^T\beta_i}\alpha_k\Big) \tag{56}$$

By the choice of step size, we have in the RHS $\sum_{i=1}^T\beta_i\geq\sum_{i=1}^T\beta_T=\Theta(\sqrt{T})$ and $\sum_{t=1}^T\beta_t^2\leq 1+\int_1^T\frac{\beta_1}{x}dx=\beta_1\ln T+1$. This proves the result. □

C. Proof in Section 4

C.1. Proof of Proposition 3

We prove the proposition from the two conditions separately.

(a) Suppose condition (a) holds. Given $x\in\mathcal{C}$, define the projected gradient of $g(x,\cdot)$ as

$$G(y;x)=\frac{1}{\beta}\big(y-\mathrm{Pj}_{\mathcal{U}}(y-\beta\nabla_y g(x,y))\big).$$

Since y_γ is a local solution of $\mathcal{CP}_{\gamma p}$ given $x=x_\gamma$, we have

$$\frac{1}{\beta}\Big[y_\gamma-\mathrm{Pj}_{\mathcal{U}}\Big(y_\gamma-\beta\big(\frac{1}{\gamma}\nabla_y f(x,y_\gamma)+\nabla_y g(x_\gamma,y_\gamma)\big)\Big)\Big]=0. \tag{57}$$

Then we have

$$\begin{aligned}\|G(y_\gamma;x_\gamma)\|&=\frac{1}{\beta}\Big\|\mathrm{Pj}_{\mathcal{U}}\Big(y_\gamma-\beta\big(\frac{1}{\gamma}\nabla_y f(x,y)+\nabla_y g(x_\gamma,y_\gamma)\big)\Big)-\mathrm{Pj}_{\mathcal{U}}(y_\gamma-\beta\nabla_y g(x_\gamma,y_\gamma))\Big\|\\ &\leq\frac{1}{\gamma}\|\nabla_y f(x,y)\|\leq\frac{L}{\gamma}.\end{aligned} \tag{58}$$

By the proximal error bound inequality, we further have

$$d_{\mathcal{S}(x_\gamma)}(y_\gamma)\leq\bar{\mu}\|G(y_\gamma;x_\gamma)\|\leq\frac{\bar{\mu}L}{\gamma}.$$

Since g is continuously differentiable and $\mathcal{C}\times\mathcal{U}$ is compact, we can define $L_{1,g}=\max_{x\in\mathcal{C},y\in\mathcal{U}}\|\nabla_y g(x,y)\|$. Then $g(x,\cdot)$ is $L_{1,g}$-Lipschitz-continuous on $\mathcal{U}$ given any $x\in\mathcal{C}$, which yields

$$p(x_\gamma,y_\gamma)=g(x_\gamma,y_\gamma)-v(x_\gamma)\leq L_{1,g}d_{\mathcal{S}(x_\gamma)}\leq\frac{L_{1,g}\bar{\mu}L}{\gamma}. \tag{59}$$

In addition, Lemma 3 holds under condition (a) so $p(x,y)$ is a squared distance bound. Further notice that $\mathcal{CP}$ and $\mathcal{CP}_{\gamma p}$ are special cases of $\mathcal{BP}$ and $\mathcal{BP}_{\gamma p}$ with $\mathcal{U}(x)=\mathcal{U}$, then the rest of the result follows from Theorem 1 with $\epsilon_1 = L\sqrt{\rho\delta}/2$, $\gamma \geq 2\gamma^* = L\sqrt{\mu\delta^{-1}}$, $\epsilon_2 = 0$ and Theorem 2 where condition (i) holds with (59).

(b). Under condition (b), Lemma 3 holds so $p(x,y)$ is a squared distance bound. Further notice that $\mathcal{CP}$ and $\mathcal{CP}_{\gamma p}$ are special cases of $\mathcal{BP}$ and $\mathcal{BP}_{\gamma p}$ with $\mathcal{U}(x)=\mathcal{U}$, then the result follows directly from Theorem 1 with $\epsilon_1 = L\sqrt{\rho\delta}/2$, $\gamma \geq 2\gamma^* = L\sqrt{\mu\delta^{-1}}$, $\epsilon_2 = 0$ and Theorem 2 where condition (ii) holds by the convexity of $g(x,\cdot)$.

C.2. Proof of Lemma 5 and 6

Lemma 5. *Assume there exists $L_g > 0$ such that $\nabla g(x,y)$ is L_g-Lipschitz-continuous. Assume either one of the following is true:*

(a) Condition (ii) in Assumption 4 holds. Let $L_S = L_g\bar{\mu}$.

(b) Conditions (i) and (iii) in Assumption 4 hold. Let $L_S = L_g(\mu{+}1)(L_g{+}1)$.

Then given any $x_1, x_2 \in \mathcal{C}$, for any $y_1 \in \mathcal{S}(x_1)$ there exists $y_2 \in \mathcal{S}(x_2)$ such that

$$\|y_1 - y_2\| \leq L_S\|x_1 - x_2\|.$$

Proof. **(a).** Given x, define the projected gradient of $g(x,\cdot)$ at point y as

$$G(y;x) = \frac{1}{\beta}\big(y - \mathrm{Pj}_{\mathcal{U}}\left(y - \beta\nabla_y g(x,y)\right)\big).$$

By the assumption, the proximal-error-bound inequality holds, that is

$$\bar{\mu}\|G(y;x)\|^2 \geq d^2_{\mathcal{S}(x)}(y),\ \forall y \in \mathcal{U} \text{ and } x \in \mathcal{C}.$$

Therefore, given $x_1, x_2 \in \mathcal{C}$, we have for any $y_1 \in \mathcal{S}(x_1)$ there exists $y_2 \in \mathcal{S}(x_2)$ such that

$$\begin{aligned}\|y_1 - y_2\| &\leq \bar{\mu}\|G(y_1;x_2) - G(y_1;x_1)\|^2 \quad \text{since } G(y_1;x_1)=0\\ &= \frac{\bar{\mu}}{\beta}\big\|\,\mathrm{Pj}_{\mathcal{U}}\left(y_1 - \beta\nabla_y g(x_2,y_1)\right) - \mathrm{Pj}_{\mathcal{U}}\left(y_1 - \beta\nabla_y g(x_1,y_1)\right)\big\|\\ &\leq \bar{\mu}\|\nabla g(x_1,y_1) - \nabla g(x_2,y_1)\| \leq L_g\bar{\mu}\|x_1 - x_2\|. \end{aligned} \tag{60}$$

This completes the proof for condition (a).

(b). By the $1/\mu$-quadratic-growth of $g(x,\cdot)$ and (Drusvyatskiy & Lewis, 2018, Corrolary 3.6), the proximal-error-bound inequality holds, that is

$$(\mu+1)(L_g+1)\|G(y;x)\|^2 \geq d^2_{\mathcal{S}(x)}(y),\ \forall y \in \mathcal{U} \text{ and } x \in \mathcal{C}.$$

where we set $\beta = 1$ to simplify the constant. The result then directly follows from case (a). □

Proof of Lemma 6. Given any $x_1, x_2 \in \mathcal{C}$, for any $y_1 \in \mathcal{S}(x_1)$ there exists a $y_2 \in \mathcal{S}(x_2)$ such that

$$\begin{aligned}\|\nabla_x g(x_1,y_1) - \nabla_x g(x_2,y_2)\| &\leq L_g(\|x_1 - x_2\| + \|y_1 - y_2\|)\\ &\leq L_g(1{+}L_S)\|x_1 - x_2\| \quad \text{by Lemma 5.}\end{aligned} \tag{61}$$

By lemma 4 and the condition that $\nabla_x g(x,y_1) = \nabla_x g(x,y_2)$ for any $y_1, y_2 \in \mathcal{S}(x)$, it holds that

$$\nabla_d v(x) = \langle \nabla_x g(x,y^*),\, d\rangle,\ \forall y^* \in \mathcal{S}(x). \tag{62}$$

This along with (61) gives

$$|\nabla_d v(x_1) - \nabla_d v(x_2)| \leq L_g(1{+}L_S)\|x_1 - x_2\|,\ \forall x_1, x_2 \in \mathcal{C}.$$

Thus $\nabla v(x)$ exists by the continuity of $\nabla_d v(x)$ for any d. Further, $\nabla v(x) = \nabla_x g(x,y^*)$ for any $y^* \in \mathcal{S}(x)$ by (62) and is Lipschitz-continuous by (61). □

C.3. Proof of Theorem 4

Convergence of ω. Given any $x \in \mathcal{C}$, by the $1/\mu$-quadratic-growth of $g(x, \cdot)$ and (Drusvyatskiy & Lewis, 2018, Corrolary 3.6), there exists some constant $\bar{\mu}$ such that the proximal-error-bound inequality holds. Thus under the either condition of Proposition 3, there exists $\bar{\mu} > 0$ such that $1/\bar{\mu}$-proximal-error-bound condition holds for $g(x, \cdot)$. This along with the Lipschitz-smoothness of $g(x, \cdot)$ implies the proximal PL condition by (Karimi et al., 2016, Appendix G).

We state the proximal PL condition below. Defining

$$\mathcal{D}(\omega; x) := -\frac{2}{\beta} \min_{\omega' \in \mathcal{X}} \{\langle \nabla_\omega g(x, \omega),\, \omega' - \omega \rangle + \frac{1}{2\beta} \|\omega' - \omega\|^2\} \tag{63}$$

there exists some constant $\tilde{\mu} > 0$ such that

$$\tilde{\mu} \mathcal{D}(\omega; x) \geq (g(x, \omega) - v(x)),\ \forall \omega \in \mathcal{U} \text{ and } x \in \mathcal{C}. \tag{64}$$

We omit index k since the proof holds for any k. By the Lipschitz gradient of $g(x, \cdot)$, we have

$$\begin{aligned} g(x, \omega_{t+1}) &\leq g(x, \omega_t) + \langle \nabla_y g(x, \omega_t),\, \omega_{t+1} - \omega_t \rangle + \frac{L_g}{2} \|\omega_{t+1} - \omega_t\|^2 \\ &= g(x, \omega_t) - \frac{\beta}{2} \mathcal{D}(\omega_t; x) \end{aligned} \tag{65}$$

where in the last equality we have used Lemma 7 that

$$\omega_{t+1} = \arg\min_{\omega \in \mathcal{U}} \langle \nabla_y g(x, \omega_t),\, \omega - \omega_t \rangle + \frac{1}{2\beta} \|\omega - \omega_t\|^2.$$

Using (64) in (65) yields

$$g(x, \omega_{t+1}) - v(x) \leq (1 - \frac{\beta}{2\tilde{\mu}})(g(x, \omega_t) - v(x)).$$

Repeatedly applying the last inequality for $t = 1, ..., T$ yields

$$g(x, \omega_{T+1}) - v(x) \leq (1 - \frac{\beta}{2\tilde{\mu}})^T (g(x, \omega_1) - v(x)).$$

This along with the $1/\mu$-quadratic-growth property of $g(x, \cdot)$ yields

$$d^2_{\mathcal{S}(x)}(\omega_{T+1}) \leq \mu(1 - \frac{\beta}{2\tilde{\mu}})^T (g(x, \omega_1) - v(x)) \leq \mu(1 - \frac{\beta}{2\tilde{\mu}})^T C_g,\ \forall x \in \mathcal{C} \tag{66}$$

where $C_g = \max_{x \in \mathcal{C}, y \in \mathcal{U}} (g(x, y) - v(x))$ is a constant.

Convergence of (x, y). The proof is similar to that of Theorem 3. We write the only step that is different here. In deriving (30), instead we have

$$\begin{aligned} \|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k; \hat{y}_k)\|^2 \leq \gamma^2 L_g^2 d^2_{\mathcal{S}(x_k)}(\hat{y}_k) &\leq \gamma^2 L_g^2 \mu C_g (1 - \frac{\beta}{2\tilde{\mu}})^{T_k} \quad \text{by (66)} \\ &\leq \frac{L_g^2 \mu C_g}{4\alpha^2 k^2} \end{aligned} \tag{67}$$

where the last inequality requires $T_k \geq -2 \log_{c_\beta}(2\alpha\gamma k)$ with $c_\beta = 1 - \frac{\beta}{2\tilde{\mu}}$.

Then (31) is replaced with

$$\|\nabla F_\gamma(z_k) - \hat{\nabla} F_\gamma(z_k; \hat{y}_k)\|^2 \leq \frac{1}{4\alpha} \|z_{k+1} - z_k\|^2 + \frac{L_g^2 \mu C_g}{4\alpha k^2}. \tag{68}$$

Result i) in this theorem then follows from the rest of the proof of i) in Theorem 3. Result ii) in this theorem follows similarly from the proof of ii) in Theorem 3 under Proposition 3.