LINEARLY CONSTRAINED BILEVEL OPTIMIZATION: A SMOOTHED IMPLICIT GRADIENT APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

This work develops analysis and algorithms for solving a class of bilevel optimization problems where the lower-level (LL) problems have linear constraints. Most of the existing approaches for constrained bilevel problems rely on value function based approximate reformulations, which suffer from issues such as non-convex and non-differentiable constraints. In contrast, in this work, we develop an implicit gradient-based approach, which is easy to implement, and is suitable for machine learning applications. We first provide in-depth understanding of the problem, by showing that the implicit objective for such problems is in general non-differentiable. However, if we add some small (linear) perturbation to the LL objective, the resulting problem becomes differentiable almost surely. This key observation opens the door for developing (deterministic and stochastic) gradient-based algorithms similar to the state-of-the-art ones for unconstrained bi-level problems. We show that when the implicit function is assumed to be strongly-convex, convex and non-convex, the resulting algorithms converge with guaranteed rate. Finally, we experimentally corroborate the theoretical findings and evaluate the performance of the proposed framework on numerical and adversarial learning problems. To our knowledge, this is the first time that (implicit) gradient-based methods have been developed and analyzed for the considered class of bilevel problems.

1 INTRODUCTION

Bilevel optimization problems (Colson et al., 2005; Dempe & Zemkoho, 2020) can be used to model an important class of hierarchical optimization tasks with two levels of hierarchy, the *upper-level (UL)* and the *lower-level (LL)*. The key characteristics of bilevel problems are: 1) the solution of the UL problem requires access to the solution of the LL problem and, 2) the LL problem is parametrized by the UL variable. Bilevel optimization problems arise in a wide range of machine learning applications, such as meta-learning (Rajeswaran et al., 2019; Franceschi et al., 2018), data hypercleaning (Shaban et al., 2019), hyperparameter optimization (Sinha et al., 2020; Franceschi et al., 2018; 2017; Pedregosa, 2016), adversarial learning (Li et al., 2019; Liu et al., 2021a; Zhang et al., 2021), as well as in other application domains such as network optimization (Migdalas, 1995), economics (Cecchini et al., 2013), and transport research (Didi-Biha et al., 2006; Kalashnikov et al., 2010). In this work, we focus on a special class of stochastic bilevel optimization problems, where the LL problem involves the minimization of a *strongly convex* objective over a set of *linear inequality constraints*. More precisely, we consider the following formulation:

$$\min_{\mathbf{x}\in\mathcal{X}}\Big\{G(\mathbf{x}) \coloneqq f(\mathbf{x},\mathbf{y}^*(\mathbf{x})) \coloneqq \mathbb{E}_\xi[f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\xi)]\Big\}, \tag{1a}$$

$$\text{s.t. } \mathbf{y}^*(\mathbf{x}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_\ell}}\Big\{h(\mathbf{x},\mathbf{y}) \;\big|\; A\mathbf{y}\le\mathbf{b}\Big\}, \tag{1b}$$

where $\xi\sim\mathcal{D}$ represents a stochastic sample of the objective $f(\cdot,\cdot)$. $\mathcal{X}\subseteq\mathbb{R}^{d_u}$ is a convex and closed set, $f:\mathcal{X}\times\mathbb{R}^{d_\ell}\to\mathbb{R}$ is the UL objective, $h:\mathcal{X}\times\mathbb{R}^{d_\ell}\to\mathbb{R}$ is the LL objective, and f,h are smooth functions. We focus on the problems where $h(\mathbf{x},\mathbf{y})$ is strongly convex with respect to $\mathbf{y}$. The matrix $A\in\mathbb{R}^{k\times d_u}$, and vector $\mathbf{b}\in\mathbb{R}^k$ define the linear constraints. In the following, we refer to (1a) as the UL problem, and to (1b) as the LL one.

The success of the bilevel formulation and its algorithms in many machine learning applications can be attributed to the use of the efficient (stochastic) gradient-based methods (Liu et al., 2021a). These methods take the following form, in which an (approximate) gradient direction of the UL problem is computed (using chain rule), and then the UL variable is updated using gradient descent (GD):

$$\widehat{\nabla} G(\mathbf{x}) \approx \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \ \texttt{GD Update:} \mathbf{x}^+ = \mathbf{x} - \beta \widehat{\nabla} G(\mathbf{x}). \quad (2)$$

The gradient of $G(\mathbf{x})$ is often referred as the *implicit gradient*. However, computing this implicit gradient not only requires access to the optimal $\mathbf{y}^*(\mathbf{x})$, but also assumes differentiability of the mapping $\mathbf{y}^*(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{d_\ell}$. One can potentially solve the LL problem approximately and obtain an approximation $\widehat{\mathbf{y}}(\mathbf{x})$ such that $\widehat{\mathbf{y}}(\mathbf{x}) \approx \mathbf{y}^*(\mathbf{x})$, and use it to compute the implicit gradient (Ghadimi & Wang, 2018). Unfortunately, not all solutions $\mathbf{y}^*(\mathbf{x})$ are differentiable, and when they are not, the above approach cannot be applied.

It is known that when the LL problem is strongly convex and *unconstrained*, then $\nabla \mathbf{y}^*(\mathbf{x})$ (and thereby the implicit gradient in (2)) can be easily evaluated using the implicit function theorem (Ghadimi & Wang, 2018). This is the reason that the majority of recent works have focused on developing algorithms for the class of bilevel problems with unconstrained LL (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Khanduri et al., 2021b; Franceschi et al., 2018; Chen et al., 2021a). However, when the LL problem is constrained, $\nabla \mathbf{y}^*(\mathbf{x})$ might not even exist. Therefore, the majority of works adopt a value function-based approach to solve problems with LL constraints (Liu et al., 2021b; Sow et al., 2022; Liu et al., 2021c). Value function-based methods typically transform the original problem into a single-level problem with non-convex and non-differentiable constraints. To resolve this, these approaches regularize the problem by adding a strongly-convex penalty term, as well as a constant error term to make the constraint set feasible, altering the problem's structure. In contrast, we introduce a perturbation-based smoothing technique, which makes $\mathbf{y}^*(\mathbf{x})$ differentiable almost surely, without practically changing the landscape of the original problem (please see (Lu et al., 2020, pg. 5)). This allows us to compute $\nabla \mathbf{y}^*(\mathbf{x})$ in closed-form, and to develop gradient-based algorithms to solve (1). Finally, it is important to note that the value function-based approaches are more suited for deterministic implementations, and therefore it is difficult to use such algorithms for large scale applications and/or when the data sizes are large. On the other hand, the gradient-based algorithms developed in our work can easily handle stochastic problems where the full deterministic gradient cannot be computed. For detailed literature review see Appendix A.

Contributions. In this work, we study a class of bilevel optimization problems with strongly convex LL objective and linear constraints in the LL problem. Major challenges for solving such problems are the following: 1) How to ensure that the implicit function $G(\mathbf{x})$ is differentiable? and 2) Even if the implicit function is differentiable, how to compute its (approximate) gradient in order to develop first-order methods? Our work addresses these challenges and develops first-order methods to tackle such constrained bilevel problems. Specifically, our contributions are the following:

- We provide and in-depth understanding of bilevel problems with strongly convex linearly constrained LL problems. Specifically, we first illustrate through an example that the implicit objective $G(\mathbf{x})$, for such problems is in general non-differentiable. To address the non-differentiability, we then propose a perturbation-based smoothing technique that makes the implicit objective $G(\mathbf{x})$ differentiable in an almost sure sense, and we provide a closed-form expression for the (approximate) implicit gradient.
- The smoothed problem we obtain is challenging, since its objective function does not have Lipschitz continuous gradients. Therefore, conventional gradient based algorithms may no longer work. To address this issue, we propose the Deterministic Smoothed Implicit Gradient ([D]SIGD) method that utilizes an (approximate) line search-based algorithm and establish asymptotic convergence guarantees. We also analyze [S]SIGD for the stochastic version of problem (1) (with fixed/diminishing step-sizes) and establish finite-time convergence guarantees for the cases when the implicit function is weakly-convex, strongly-convex, and convex (but not Lipschitz smooth).
- Finally, we evaluate the performance of the proposed algorithmic framework via experiments on quadratic bilevel and adversarial learning problems.

Motivating Applications. Bilevel problems of the form 1 capture several important applications in the domains of machine learning and information processing. Below we provide two such applications.

Adversarial Learning. Consider the problem of robustly training a model $\phi(\mathbf{x}; \mathbf{c})$, where $\mathbf{x}$ denotes the model parameters and $\mathbf{c}$ the input to the model; let $\{(\mathbf{c}_i, \mathbf{d}_i)\}_{i=1}^N$ with $\mathbf{c}_i \in \mathbb{R}^{d_{\ell_i}}, \mathbf{d}_i \in \mathbb{R}$ be the

training set. Then an *adversarial training* problem aims to solve the following problem (Zhang et al., 2021; Goodfellow et al., 2014):

$$\min_{\mathbf{x}\in\mathbb{R}^{d_u}} \Big\{ \sum_{i=1}^{N} f_i(\phi(\mathbf{x};\mathbf{c}_i+\mathbf{y}_i^*(\mathbf{x})),\mathbf{d}_i)\Big\} \; s.t. \; \mathbf{y}^*(\mathbf{x}) \in \begin{cases} \arg\min_{\mathbf{y}_i\in\mathbb{R}^{d_{\ell_i}}} \sum_{i=1}^{N} h_i(\phi(\mathbf{x};\mathbf{c}_i+\mathbf{y}_i),\mathbf{d}_i) \\ \text{s.t.} \quad -\mathbf{b}\le \mathbf{y}\le \mathbf{b} \end{cases}, \tag{3}$$

where $\mathbf{y}=[\mathbf{y}_1^T,\dots,\mathbf{y}_N^T]^T\in\mathbb{R}^{d_\ell}$; with $\mathbf{y}_i\in\mathbb{R}^{d_{\ell_i}}$ denotes the attack on the ith example and we have $\sum_{i=1}^N d_{\ell_i}=d_\ell$. Moreover, $f_i:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ denotes the loss function for learning the model parameter $\mathbf{x}$, while $h_i:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ denotes the adversarial objective used to design the optimal attack $\mathbf{y}$. Note that the linear constraints in the LL problem $-\mathbf{b}\le\mathbf{y}\le\mathbf{b}$ models the attack budget.

Distributed Optimization. In distributed optimization (Chang et al., 2020; Yang et al., 2019), a set of N agents aim to jointly minimize an objective function $G(\mathbf{x})$ over an undirected graph $\mathcal{G}=(V,E)$. We consider the following distributed bilevel problem

$$\min_{\{\mathbf{x}_i\in\mathcal{X}\,|\,A\mathbf{x}=0\}} \Big\{G(\mathbf{x}):=\sum_{i=1}^{N} f_i(\mathbf{x}_i,\mathbf{y}_i^*(\mathbf{x}_i))\Big\} \;\text{ s.t. }\; \mathbf{y}^*(\mathbf{x})\in\arg\min_{\mathbf{y}\in\mathbb{R}^{d_\ell}}\Big\{\sum_{i=1}^{N} h_i(\mathbf{x}_i,\mathbf{y}_i) \text{ s.t. } A\mathbf{y}=\mathbf{0}\Big\},$$

where $\mathbf{x}=[\mathbf{x}_1,\dots,\mathbf{x}_N]$ and $\mathbf{y}=[\mathbf{y}_1,\dots,\mathbf{y}_N]$. Each agent $i\in[N]$ has access to f_i and h_i. The constraint $A\mathbf{y}=0$ (resp. $A\mathbf{x}=0$) is introduced to ensure the consensus of LL (resp. UL) variables. Such problems arise in many areas of signal processing and sensor networks (Yousefian, 2021). The above formulation also models a decentralized meta learning problem where the training and validation data is distributed among agents while each agent aims to solve the meta learning problem globally (Ji et al., 2021).

2 Properties of the Linearly Constrained Bilevel Problem

In this section we study the properties of problem 1. First, let us define the necessary notations. Let $\overline{A}(\mathbf{y})$ be the matrix that contains the rows $S(\mathbf{y})\subseteq\{1,\dots,k\}$ of A that correspond to the active constraints of inequality $A\mathbf{y}\le\mathbf{b}$ in the LL problem, that is we have $\overline{A}(\mathbf{y})\mathbf{y}=\overline{\mathbf{b}}(\mathbf{y})$, where $\overline{\mathbf{b}}(\mathbf{y})$ contains the elements of $\mathbf{b}$ with indices in $S(\mathbf{y})$. Also, we denote with $\overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ the Lagrange multipliers vector that corresponds to the active constraints at $\mathbf{y}^*(\mathbf{x})$; it always holds that $\overline{\boldsymbol{\lambda}}^*(\mathbf{x})\ge 0$. Next, we introduce some fundamental assumptions.

Assumption 1. *We assume that the following conditions hold for problem* (1)*:*

- *(a)* $f(\mathbf{x},\mathbf{y})$ *is continuously differentiable, and* $h(\mathbf{x},\mathbf{y})$ *is twice continuously differentiable.*
- *(b)* $\mathcal{X}$ *and* $\big\{\mathbf{y}\in\mathbb{R}^{d_\ell}\;\big|\;A\mathbf{y}\le\mathbf{b}\big\}$ *are closed and convex sets.*
- *(c)* $h(\mathbf{x},\mathbf{y})$ *is strongly convex in* $\mathbf{y}$*, for every* $\mathbf{x}\in\mathcal{X}$*, with modulus* μ_h*.*
- *(d) There exists* $\mathbf{y}\in\mathbb{R}^{d_\ell}$ *such that* $A\mathbf{y}<\mathbf{b}$*.*
- *(e)* $\overline{A}(\mathbf{y}^*(\mathbf{x}))$ *is full row rank, for every* $\mathbf{x}\in\mathcal{X}$[1]*.*

The Assumptions 1(a), (b) and (c) are standard assumptions in bilevel optimization literature and are required to ensure the continuity of the implicit function (Proposition 1). Assumption 1(c) ensures that the implicit function $G(\mathbf{x})$ is well defined as the LL problem returns a single point. Assumption 1(d) ensures strict feasibility of the LL problem, while Assumption 1(e) implies that the rows of A corresponding to the active constraints are linearly independent. Note that this assumption allows us to ensure the differentiability of the implicit function (Lemma 1). Also note that there are some special cases in which Assumption 1(e) is automatically satisfied. For instance, consider a problem where the LL problem has box constraints, i.e., $\mathbf{a}\le\mathbf{y}\le\mathbf{b}$. Then for any $\mathbf{y}\in\mathcal{Y}$ the only possible non-zero values in the matrix $\overline{A}(\mathbf{y})$ are $+1,-1$, and there is only one non-zero value at each column. Therefore, $\overline{A}(\mathbf{y})$ is full row rank.

Assumption 2. *We assume that the following holds for problem* (1)*,* $\forall\mathbf{x},\overline{\mathbf{x}}\in\mathcal{X}$ *and* $\mathbf{y},\overline{\mathbf{y}}\in\mathbb{R}^{d_\ell}$*:*

[1] This is the LICQ condition (Bertsekas, 1998) of the LL problem. In literature it is used to ensure that the optimal solutions satisfy the KKT conditions.

(a) f *has bounded gradients, i.e.,* $\|\nabla f(\mathbf{x},\mathbf{y})\| \leq \overline{L}_f$.

(b) f *has Lipschitz continuous gradients, i.e.,* $\|\nabla f(\mathbf{x},\mathbf{y}) - \nabla f(\overline{\mathbf{x}},\overline{\mathbf{y}})\| \leq L_f\|[\mathbf{x};\mathbf{y}] - [\overline{\mathbf{x}};\overline{\mathbf{y}}]\|$.

(c) h *has Lipschitz continuous gradient in* $\mathbf{y}$*, i.e.,* $\|\nabla_y h(\mathbf{x},\mathbf{y}) - \nabla_y h(\mathbf{x},\overline{\mathbf{y}})\| \leq L_h\|\mathbf{y} - \overline{\mathbf{y}}\|$.

(d) h *has Lipschitz continuous Hessian in* $\mathbf{y}$*, i.e.,* $\|\nabla^2_{yy} h(\mathbf{x},\mathbf{y}) - \nabla^2_{yy} h(\mathbf{x},\overline{\mathbf{y}})\| \leq L_{h_{yy}}\|\mathbf{y} - \overline{\mathbf{y}}\|$.

(e) h *has Lipschitz continuous Jacobian, i.e.,* $\|\nabla^2_{xy} h(\mathbf{x},\mathbf{y}) - \nabla^2_{xy} h(\mathbf{x},\overline{\mathbf{y}})\| \leq L_{h_{xy}}\|\mathbf{y} - \overline{\mathbf{y}}\|$.

(f) h *has a bounded Jacobian,* $\|\nabla^2_{xy} h(\mathbf{x},\mathbf{y})\| \leq \overline{L}_{h_{xy}}$.

Assumption 2 is standard in bilevel optimization literature (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2021a; Ji et al., 2021) and is used to derive some useful properties of the (approximate) implicit gradient (Lemma 3, Appendix E.2). Next, we utilize the above assumptions to analyze the properties of mapping $\mathbf{y}^*(\mathbf{x})$.

Proposition 1. *Under Assumption 1, the mapping* $\mathbf{y}^*(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{d_\ell}$ *and the implicit function* $G(\mathbf{x})$ *are both continuous.*

Proposition 1 ensures that $\mathbf{y}^*(\mathbf{x})$ and $G(\cdot)$ are both continuous. Now if we can ensure differentiability of $G(\cdot)$, then we should be able to implement a gradient-based update rule to solve (1). However, as the following example illustrates, $\mathbf{y}^*(\mathbf{x})$ and thus $G(\mathbf{x})$ are not differentiable in general.

Example. Consider the following problem

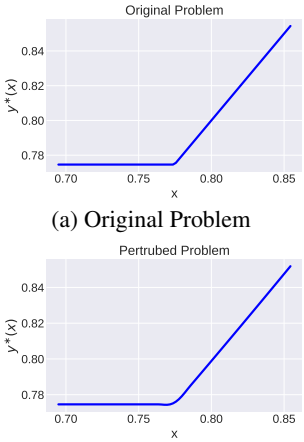
$$\min_{x\in[0,1]} x + y^*(x) \text{ s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}} \left\{(y^2 - x^2)^2 \;\middle|\; \sqrt{3/5} \leq y \leq 1\right\}. \tag{4}$$

The mapping $y^*(x)$ is $y^*(x) = x$, if $x \in [\sqrt{3/5}, 1]$, and $y^*(x) = \sqrt{3/5}$, if $x \in [0, \sqrt{3/5})$. In Figure 1a, we plot this mapping. Notice that at the point $\mathbf{x} = \sqrt{3/5}$ the mapping (and thus the implicit function) is non-differentiable. □

The example above suggests that the considered class of problem is challenging. A natural approach to address such non-differentiability is to *smooth* the non-differentiable points. To this end, we consider a perturbed version of the problem where we add a linear perturbation to the LL problem, i.e., $\arg\min_{y\in\mathbb{R}}\{(y^2 - x^2)^2 + q \cdot y,\ q \sim \mathcal{N}(0,\sigma^2) \mid \sqrt{3/5} \leq y \leq 1\}$. It turns out that the optimal solution $y^*(x)$ obtained in this way is now differentiable at $\mathbf{x} = \sqrt{3/5}$; Figure 1b. Motivated by this interesting observation, in the rest of the paper we consider a perturbed version of the original problem (1).

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{F(\mathbf{x}) := f(\mathbf{x},\mathbf{y}^*(\mathbf{x})) := \mathbb{E}_\xi[f(\mathbf{x},\mathbf{y}^*(\mathbf{x});\xi)]\right\} \tag{5a}$$

$$\text{s.t. } \mathbf{y}^*(\mathbf{x}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_\ell}} \left\{g(\mathbf{x},\mathbf{y}) := h(\mathbf{x},\mathbf{y}) + \mathbf{q}^T\mathbf{y} \;\middle|\; A\mathbf{y} \leq \mathbf{b}\right\}, \tag{5b}$$



(a) Original Problem

(b) Perturbed Problem

Figure 1: Plot of $y^*(x)$.

where $\mathbf{q} \in \mathbb{R}^{d_\ell}$ is generated from a continuous measure. Note that Assumptions 1(a)-(d) and 2 hold directly for the perturbed problem (5) with constants $\mu_g = \mu_h, L_h = L_g, L_{g_{yy}} = L_{h_{yy}}, L_{g_{xy}} = L_{h_{xy}}, \overline{L}_{g_{xy}} = \overline{L}_{h_{xy}}$; we also assume that Assumption 1(e) holds for problem 5a-5b. Next, we show that adding this perturbation suffices to ensure (almost sure) differentiability of the implicit function in (5) at each $\mathbf{x} \in \mathcal{X}$. This is achieved by invoking an existing result from (Lu et al., 2020), which shows that the perturbed problem (5) satisfies the Strict Complementarity (SC) condition, i.e., $\overline{\boldsymbol{\lambda}}^*(\mathbf{x}) > 0$, almost surely.

Lemma 1. *((Lu et al., 2020, Proposition 1)) If* $\mathbf{y}^*(\mathbf{x})$ *is a KKT point of problem* (5b)*,* $\mathbf{q}$ *is generated from a continuous measure, and* $\overline{A}(\mathbf{y}^*(\mathbf{x}))$ *is full row rank, then the SC condition holds with probability one.*

First, note that Proposition 1 also holds for the implicit function $F(\mathbf{x})$ of the perturbed problem. Moreover, the SC condition ensured by Lemma 1 in conjunction with Assumption 1 implies that the mapping $\mathbf{y}^*(\mathbf{x})$ and thereby the implicit function $F(\mathbf{x})$ is differentiable almost surely at any given $\mathbf{x} \in \mathcal{X}$ (Friesz & Bernstein, 2015, Theorem 2.22). This result will be leveraged shortly to derive a

closed-form expression for the gradient of the implicit function of problem (5). Next, we show that the implicit functions of the two bilevel problems, the exact one (1) and the perturbed one (5), differ by a quantity that is controlled by the perturbation vector $\mathbf{q}$.

Proposition 2. *Under Assumptions 1 and 2, we have:* $|G(\mathbf{x}) - F(\mathbf{x})| \leq \overline{L}_f \frac{\|\mathbf{q}\|}{\mu_g} \; \forall \mathbf{x} \in \mathcal{X}$.

Clearly, the size of perturbation $\|\mathbf{q}\|$ determines the difference between the original and the perturbed implicit functions. Also note that the only requirement on $\mathbf{q}$ is that it is generated from a continuous measure. Therefore we can always choose a distribution such that $\mathbf{q}$ will be arbitrarily small without essentially changing the landscape of the original problem (1) (Lu et al., 2020, pg. 5). As a result, in the subsequent discussion, we will focus on optimizing the perturbed problem 5.

3 Implicit Gradient Computation

In this section, we derive the expression for the gradient of the implicit function $F(\mathbf{x})$ and propose the Smoothed Implicit Gradient Descent (**SIGD**) algorithm for solving the perturbed problem (5).

Lemma 2 (Implicit Gradient)**.** *Under Assumption 1, for any given* $\mathbf{x} \in \mathcal{X}$*, we have*

$$\nabla \mathbf{y}^*(\mathbf{x}) = \big[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big]^{-1} \big[- \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\big] \tag{6}$$

$$\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = -\big[\overline{A}\big[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big]^{-1} \overline{A}^T\big]^{-1} \big[\overline{A}\big[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big]^{-1} \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\big], \tag{7}$$

where we set $\overline{A} := \overline{A}(\mathbf{y}^*(\mathbf{x}))$.

Note that when problem (5) does not have the LL constraints, the implicit gradient derived in Lemma 2 becomes exactly same as the one in Ghadimi & Wang (2018); Ji et al. (2021); Hong et al. (2020); Chen et al. (2021a). Moreover, note that if the LL problem has only linear equality constraints, the differentiability of $\mathbf{y}^*(\mathbf{x})$ can be easily verified using the implicit function theorem under Assumptions 1(a) and (c) along with full row rankness of A. The expression of the implicit gradient stays the same as in Lemma 2 with $\overline{A}$ and $\overline{\boldsymbol{\lambda}^*}(\mathbf{x})$ replaced by A and $\boldsymbol{\lambda}^*(\mathbf{x})$, respectively (i.e., we use the full matrix A). Using Lemma 2 above we now have an expression of the implicit gradient as

$$\nabla F(\mathbf{x}) = \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})). \tag{8}$$

Note however, that computing $\nabla F(\mathbf{x})$ requires the precise knowledge of $\mathbf{y}^*(\mathbf{x})$ which is not possible for many problems of interest. Therefore, we approximate the implicit gradient by replacing $\widehat{\mathbf{y}}(\mathbf{x})$ in place of $\mathbf{y}^*(\mathbf{x})$ in (6), (7) and (8), and define the approximate implicit gradient as

$$\widehat{\nabla} F(\mathbf{x}) = \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) + [\widehat{\nabla} \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})), \tag{9}$$

where $\widehat{\nabla} \mathbf{y}^*(\mathbf{x})$ is defined by substituting $\widehat{\mathbf{y}}(\mathbf{x})$ in place of $\mathbf{y}^*(\mathbf{x})$ in expressions (6) and (7). In order to ensure that (9) returns a useful approximation of the (exact) implicit gradient, we impose a few assumptions on the quality of the estimate $\widehat{\mathbf{y}}(\mathbf{x})$.

Assumption 3. *The approximate solution of problem* (5b) $\widehat{\mathbf{y}}(\mathbf{x})$ *satisfies the following* $\forall \mathbf{x} \in \mathcal{X}$*:*

(a) $\|\widehat{\mathbf{y}}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\| \leq \delta$ *for* $\delta > 0$*,*

(b) $\widehat{\mathbf{y}}(\mathbf{x})$ *is a feasible point, i.e.,* $A\widehat{\mathbf{y}}(\mathbf{x}) \leq \mathbf{b}$*,*

(c) It holds that $\overline{A}(\mathbf{y}^*(\mathbf{x})) = \overline{A}(\widehat{\mathbf{y}}(\mathbf{x}))$.

The LL problem requires the solution of a strongly convex linearly constrained task. As a result, Assumptions3(a),(b) can be easily satisfied. Specifically, we can obtain approximate feasible solutions of given accuracy with known methods, such as projected gradient descent, or by using some convex optimization solver; in section B of the Appendix we provide one such method. Moreover, Assumption 3(c) will be satisfied if we find a "sufficiently accurate" solution $\widehat{\mathbf{y}}(\mathbf{x})$. Specifically, from Calamai & Moré (1987, Theorem 4.1) we know that if $\widehat{\mathbf{y}}^k(\mathbf{x}) \in \mathcal{Y}$ is an arbitrary sequence that converges to a non-degenerate (i.e., Assumption 1(e) and SC holds) stationary solution $\mathbf{y}^*(\mathbf{x})$, then there exists an integer k_0 such that $\overline{A}(\mathbf{y}^*(\mathbf{x})) = \overline{A}(\widehat{\mathbf{y}}^k(\mathbf{x}))$, $\forall k > k_0$.

Algorithm 1 [Deterministic] Smoothed Implicit Gradient Descent (**[D]SIGD**)

Input: Initial parameter $\mathbf{x}^0$, # of iteration T, LL solution accuracy, δ^r, σ, measure $\mathcal{Q}$
Sample $\mathbf{q} \sim \mathcal{Q}$ and perturb LL problem
for $r = 0, 1, \ldots, T-1$ **do**
  Find an approximate solution $\widehat{\mathbf{y}}(\mathbf{x}^r)$ s.t. Assumption 3 is satisfied.
  Compute $\widehat{\nabla} F(\mathbf{x}^r)$ using (9), $\widehat{\mathbf{d}}^r = \widetilde{\mathbf{x}}^r - \mathbf{x}^r$ with $\widetilde{\mathbf{x}}^r = \text{proj}_{\mathcal{X}}(\mathbf{x}^r - \widehat{\nabla} F(\mathbf{x}^r))$
  Select a^r s.t. the following Armijo-type rule condition is satisfied

$$\widehat{F}(\mathbf{x}^r) - \widehat{F}(\mathbf{x}^r + a^r \widehat{\mathbf{d}}^r) \geq -\sigma a^r [\widehat{\nabla} F(\mathbf{x}^r)]^T \widehat{\mathbf{d}}^r - \epsilon(\delta; r) \tag{11}$$

  where $\epsilon(\delta; r)$ depends on δ^r, α^r and problem-dependent parameters; $\widehat{F}(\cdot) = f(\cdot, \widehat{\mathbf{y}}(\cdot))$.
  Perform one projected gradient step: $\mathbf{x}^{r+1} = \mathbf{x}^r + a^r \widehat{\mathbf{d}}^r$
end for

Remark 1. *There are certain special cases where we can obtain an upper bound for k_0. For instance, in the case of non-negative constraints $\mathbf{y} \geq \mathbf{0}$ it can be shown*[2] *that $\frac{L_h}{\mu_h} \log\left(2L_h \|\mathbf{y}^0 - \mathbf{y}^*(\mathbf{x})\| / \tau\right)$ iterations of the projected gradient descent method suffice to ensure that the active set of the approximate solution $\widehat{\mathbf{y}}(\mathbf{x})$ coincides with the active set of the exact one $\mathbf{y}^*(\mathbf{x})$ (see Nutini et al. (2019, Corollary 1)), where $\tau = \min_{i \in S(\mathbf{y}^*(\mathbf{x}))} \nabla_{y_i} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ and $\mathbf{y}^0$ is the algorithm's initialization. A similar result can be derived for the case with bound constraints $\mathbf{a} \leq \mathbf{y} \leq \mathbf{b}$.*

Next, we analyze the properties of the approximate implicit gradient.

Lemma 3. *Suppose that Assumptions 1,2 and 3 hold. Then, for every $\mathbf{x} \in \mathcal{X}$ the following holds*

$$\|\widehat{\nabla} F(\mathbf{x}) - \nabla F(\mathbf{x})\| \leq L_F \delta, \quad \|\nabla F(\mathbf{x})\| \leq \overline{L}_F \quad \textit{and} \quad \|\widehat{\nabla} F(\mathbf{x})\| \leq \overline{L}_F,$$

where $L_F = L_f + L_{\mathbf{y}^*} \overline{L}_f + L_f \overline{L}_{\mathbf{y}^*}$, and $\overline{L}_F = \left(1 + \overline{L}_{\mathbf{y}^*}\right) \overline{L}_f$; the constants $\overline{L}_{\mathbf{y}^*}, L_{\mathbf{y}^*}$ are defined in Lemmas 6,8, respectively, provided in the Appendix.

Note that in contrast to bilevel programs with unconstrained LL problems (see Lemma 2.2(c) in Ghadimi & Wang (2018)), the implicit gradient $\nabla F(\mathbf{x})$ in (8) is not Lipschitz smooth, i.e., we do not have $\|\nabla F(\mathbf{x}_1) - \nabla F(\mathbf{x}_2)\| \leq L \|\mathbf{x}_1 - \mathbf{x}_2\|$ for $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ in general. This implies that for problem (5), standard smoothness based analyses will not work, particularly for the case when the implicit function is non-convex. To address this challenge we use a line search-based method, which does not require smoothness or another special structure (such as convexity), and show asymptotic convergence. Then the cases when the implicit function is weakly-convex, convex and strongly convex are considered, and a gradient-based method is developed. For the latter cases, we consider the stochastic setting and evaluate the stochastic gradient using (9) as:

$$\widehat{\nabla} F(\mathbf{x}; \xi) = \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}); \xi) + [\widehat{\nabla} \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}); \xi). \tag{10}$$

Also, we make the following assumption on the stochastic gradients of the UL problem.

Assumption 4. *We assume that the stochastic gradients are unbiased, i.e. $\mathbb{E}_\xi[\nabla f(\mathbf{x}, \mathbf{y}; \xi)] = \nabla f(\mathbf{x}, \mathbf{y})$ and have bounded variance, i.e., $\mathbb{E}_\xi \|\nabla f(\mathbf{x}, \mathbf{y}; \xi) - \nabla f(\mathbf{x}, \mathbf{y})\|^2 = \sigma_f^2$ for some $\sigma_f > 0$.*

Assumption 4 is a typical assumption required to ensure that the approximate implicit stochastic gradient is also unbiased and has finite variance (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2021a) as shown in Lemma 4 below.

Lemma 4. *Under Assumptions 1, 2, 3 and 4, the stochastic gradient estimate in* (10) *is unbiased, i.e., $\mathbb{E}_\xi[\widehat{\nabla} F(\mathbf{x}; \xi)] = \widehat{\nabla} F(\mathbf{x})$ and has bounded variance , i.e., $\mathbb{E}_\xi \|\widehat{\nabla} F(\mathbf{x}; \xi) - \widehat{\nabla} F(\mathbf{x})\| \leq \sigma_F^2$ where $\sigma_F^2 = 2\sigma_f^2 + 2\overline{L}_{\mathbf{y}^*} \sigma_f^2$; $\overline{L}_{\mathbf{y}^*}$ is defined in Lemma 6 in the Appendix.*

4 The SIGD Algorithm and Convergence Analysis

To solve problem (5), we propose the [Deterministic] Smoothed Implicit Gradient Descent ([D]SIGD) method that utilizes line search to choose an appropriate step-size; the pseudo code is given in

[2] Under the assumption that $\nabla_{y_i} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) > 0, \forall i \in S(\mathbf{y}^*(\mathbf{x}))$.

Algorithm 2 [Stochastic] Smoothed Implicit Gradient Descent (**[S]SIGD**)

1: **Input:** Initial parameter $\mathbf{x}^0$, # of iterations T, step-sizes $\{\beta^r\}_{r=0}^{T-1}$, LL solution accuracy δ
2: Sample $\mathbf{q} \sim \mathcal{Q}$ and perturb LL problem
3: **for** $r = 0, 1, \ldots, T-1$ **do**
4: Find an approximate solution $\widehat{\mathbf{y}}(\mathbf{x}^r)$ s.t. Assumption 3 is satisfied.
5: Compute $\widehat{\nabla} F(\mathbf{x}^r; \xi^r)$ using (10)
6: Perform one stochastic projected gradient descent step: $\mathbf{x}^{r+1} = \text{proj}_{\mathcal{X}}(\mathbf{x}^r - \beta^r \widehat{\nabla} F(\mathbf{x}^r; \xi^r))$
7: **end for**

Algorithm 1. Moreover, for the cases where the implicit function is either weakly-convex, strongly convex or convex, we propose the [Stochastic] Smoothed Implicit Gradient Descent ([S]SIGD) method with pseudo code given in Algorithm 2. Note that in Algorithms 1 and 2 we sample the perturbation $\mathbf{q} \sim \mathcal{Q}$ ($\mathcal{Q}$ is a continuous measure) only once[3] at the beginning of the algorithm (in Line 2), although Lemma 2 suggests that for each given $\mathbf{x} \in \mathcal{X}$ we need to sample a different $\mathbf{q} \sim \mathcal{Q}$ to ensure differentiability. We note that if the points of non-differentiability in the original implicit function $G(\mathbf{x})$ are countable, then a single perturbation $\mathbf{q}$ suffices to ensure the differentiability of $F(\mathbf{x})$ at every $\mathbf{x} \in \mathcal{X}$ with probability 1, since the countable intersection of almost sure events is an almost sure event. Even if the non-differentiable points are uncountable, our algorithms will encounter only a countable number of points during their execution, again a single $\mathbf{q}$ suffices to ensure differentiability of $F(\mathbf{x})$ on those (countable) points using a similar argument as above. Below, we provide convergence guarantees for the [D]SIGD and the [S]SIGD algorithms.

4.1 Asymptotic Convergence Guarantees With the [D]SIGD Method

Theorem 1. *Suppose that Assumptions 1, 2 hold. At each iteration r of Algorithm 1 we find $0 < a^r < 1$ such that the Armijo-type condition* (11) *is satisfied with $\epsilon(\delta; r) = L_f \delta^r + \overline{L}_F L_F a^r \delta^r + L_f \delta^{r+1} + L_F^2 \sigma a^r (\delta^r)^2 + 2 L_F \overline{L}_F \sigma a^r \delta^r$. Further, we select δ^r such that Assumption 3 is satisfied, $\lim_{r\to\infty} \delta^r = 0$, and it holds that $\delta^r / a^r \sim \mathcal{O}(c^r)$, where c^r is some sequence with $\lim_{r\to\infty} c^r = 0$. In addition, the sequence $\widehat{\mathbf{d}}^r$ is selected such that it is gradient related to $\widehat{\nabla} F(\mathbf{x}^r)$, i.e., "for any subsequence $\{\mathbf{x}^r\}_{r\in\mathcal{R}}$ converging to a non-stationary point, the corresponding subsequence $\{\widehat{\mathbf{d}}^r\}_{r\in\mathcal{R}}$ is bounded and satisfies $\limsup_{r\to\infty, r\in\mathcal{R}} \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r < 0$" (Bertsekas, 1998, eq. 1.13),(Bertsekas & Rheinboldt, 2014, sec. 1.3.1). Then with probability 1 every limit point $\bar{\mathbf{x}}$ of the sequence of iterates generated by the [D]SIGD Algorithm 1 is a stationary point.*

Note that in Theorem 1 only asymptotic convergence is guaranteed. However, this is the best we can do since we do not impose any Lipschitz smoothness or convexity assumptions. In the special cases where the implicit function is weakly-convex, strongly convex or convex (but still not Lipschitz smooth), it is possible to derive finite-time convergence guarantees as presented next.

4.2 Finite Time Convergence Guarantees with the [S]SIGD

To obtain finite time convergence guarantees we need first to impose the additional assumption that the set $\mathcal{X}$ is bounded; combining this property with Assumption 1 implies that $\mathcal{X}$ is compact. So, in the following results we assume that $\mathcal{X}$ is a compact set with diameter $\mathcal{D}_{\mathcal{X}} := \sup_{\mathbf{x},\bar{\mathbf{x}}\in\mathcal{X}} \|\mathbf{x} - \bar{\mathbf{x}}\|$.

Weakly Convex Objective. We make the following assumption on the implicit function $F(\cdot)$.

Assumption 5. *We assume that for some $\rho > 0$ the implicit function $F(\mathbf{x})$ satisfies: $F(\mathbf{z}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle - \frac{\rho}{2}\|\mathbf{z} - \mathbf{x}\|^2 \; \forall \, \mathbf{x}, \mathbf{z} \in \mathbb{R}^{d_u}$.*

Assumption 5 implies that the function $F(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x}\|^2$ for $\hat{\rho} = \rho$ is convex while for $\hat{\rho} > \rho$ is strongly convex with modulus $\hat{\rho} - \rho$. Many problems of practical interest satisfy the weak-convexity, for example, phase retrieval (Davis et al., 2020), covariance matrix estimation (Chen et al., 2015), dictionary learning (Davis & Drusvyatskiy, 2019), Robust PCA (Candès et al., 2011) etc. (please see (Davis & Drusvyatskiy, 2019) and (Drusvyatskiy, 2017) for more details). For providing guarantees for the [S]SIGD algorithm we utilize a Moreau envelope based analysis. For this purpose, we first

[3] In the opposite case (i.e., we sample $\mathbf{q}$ for every different $\mathbf{x}$) the problem we are solving would change at each iteration and the subsequent analysis would become intractable.

rephrase problem (5) as an unconstrained problem: $\min_{\mathbf{x}\in\mathbb{R}^{d_u}} H(\mathbf{x}) := F(\mathbf{x}) + \mathbf{I}_{\mathcal{X}}(\mathbf{x})$, where the function $F(\mathbf{x})$ is as defined in (5) and $\mathbf{I}_{\mathcal{X}}(\mathbf{x})$ is the indicator function of set $\mathcal{X}$ defined as: $\mathbf{I}_{\mathcal{X}}(\mathbf{x}) := 0$ if $\mathbf{x} \in \mathcal{X}$ and $\mathbf{I}_{\mathcal{X}}(\mathbf{x}) := \infty$ if $\mathbf{x} \notin \mathcal{X}$. Below we define the Moreau envelope of $H(\mathbf{x})$.

Definition 1. *Given $\lambda > 0$, the Moreau envelope of $H(\mathbf{x})$ is defined as*

$$H_\lambda(\mathbf{x}) := \min_{\mathbf{z}\in\mathbb{R}^{d_u}} \left\{ H(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{x}-\mathbf{z}\|^2 \right\} = \min_{\mathbf{z}\in\mathcal{X}} \left\{ F(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{x}-\mathbf{z}\|^2 \right\},$$

where the second equality follows from the definition of $H(\mathbf{x})$. Moreover, we denote the proximal map of $H(\mathbf{x})$ as $\hat{\mathbf{x}} := prox_{\lambda H}(x)$ which is defined as

$$\hat{\mathbf{x}} := \arg\min_{\mathbf{z}\in\mathbb{R}^{d_u}} \left\{ H(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{x}-\mathbf{z}\|^2 \right\} = \arg\min_{\mathbf{z}\in\mathcal{X}} \left\{ F(\mathbf{z}) + \frac{1}{2\lambda}\|\mathbf{x}-\mathbf{z}\|^2 \right\}.$$

The norm of the gradient of the Moreau envelope satisfies the following:

$$\|\mathbf{x}-\hat{\mathbf{x}}\| = \lambda\|\nabla H_\lambda(\mathbf{x})\|, \quad H(\hat{\mathbf{x}}) \le H(\mathbf{x}), \quad \text{and} \quad \text{dist}(0;\partial H(\hat{\mathbf{x}})) \le \|\nabla H_\lambda(\mathbf{x})\|, \tag{12}$$

where $\text{dist}(0;\partial H(\hat{\mathbf{x}})) = -\inf_{\mathbf{v}:\|\mathbf{v}\|\le 1} H'(\mathbf{x};\mathbf{v})$ and $H'(\mathbf{x};\mathbf{v})$ denotes the directional derivative of H at $\mathbf{x}$ in direction $\mathbf{v}$. Note that a small gradient $\|\nabla H_\lambda(\mathbf{x})\|$ implies that $\mathbf{x}$ is near some point $\hat{\mathbf{x}}$ that is nearly stationary (Davis & Drusvyatskiy, 2019). Then we have the following result.

Theorem 2. *Under Assumptions 1, 2, 3, 4 and 5, with step-sizes $\beta^r = \beta$ for all $r \in \{0,\dots,T-1\}$ and for any constant $\hat{\rho} > \frac{3\rho}{2}$, the iterates generated by Algorithm 2 satisfy (with probability 1)*

$$\frac{1}{T}\sum_{r=0}^{T-1} \mathbb{E}\|\nabla H_{1/\hat{\rho}}(\mathbf{x}^r)\|^2 \le \frac{2\hat{\rho}}{2\hat{\rho}-3\rho}\left[\frac{H_{1/\hat{\rho}}(\mathbf{x}^0)-H^*}{\beta T} + \beta\hat{\rho}\big(\sigma_F^2+\overline{L}_F^2\big) + \frac{\hat{\rho}}{2\rho}L_F^2\delta^2\right].$$

Theorem 2 implies that with the choice of $\beta = \mathcal{O}(1/\sqrt{T})$, the [S]SIGD algorithm converges to a stationary point of problem (5) at a rate of $\mathcal{O}(1/\sqrt{T})$ with an additive error determined by the accuracy of the LL problem's solution δ (see Assumption 3).

Strongly Convex and Convex Objective. Next, we provide the guarantees for the case when the implicit function is strongly convex. We make the following assumption.

Assumption 6. *We assume that the objective $F(\mathbf{x})$ is μ_F-strongly convex, i.e., $F(\mathbf{z}) \ge F(\mathbf{x}) + \langle\nabla F(\mathbf{x}), \mathbf{z}-\mathbf{x}\rangle + \frac{\mu_F}{2}\|\mathbf{x}-\mathbf{z}\|^2\ \forall \mathbf{z},\mathbf{x}\in\mathcal{X}$. Note that for $\mu_F = 0$, the objective becomes convex.*

Theorem 3. *Under the Assumptions 1, 2, 3, 4 and 6, with $\mu_F > 0$ and the choice of step-sizes $\beta^r = \frac{1}{\mu_F(r+1)}$ the iterates generated by Algorithm 2 satisfy the following (with probability 1),*

$$\mathbb{E}[F(\underline{\mathbf{x}}) - F^*] \le \frac{(\sigma_F^2+\overline{L}_F^2)}{\mu_F}\frac{\log(T)}{T} + D_{\mathcal{X}}L_F\delta.$$

Theorem 4. *Under Assumption 1, 2, 3, 4 and 6, with $\mu_F = 0$, and step-sizes $\beta^r = \beta$ for $r \in \{0,\dots,T-1\}$, the iterates generated by Algorithm 2 satisfy the following (with probability 1),*

$$\mathbb{E}[F(\underline{\mathbf{x}}) - F^*] \le \frac{\|\mathbf{x}^1-\mathbf{x}^*\|^2}{\beta T} + 2\beta(\sigma_F^2+\overline{L}_F^2) + D_{\mathcal{X}}L_F\delta, \quad \textit{where } \underline{\mathbf{x}} = \frac{1}{T}\sum_{r=0}^{T-1}\mathbf{x}^r.$$

The results of Theorems 3 and 4 imply that the implicit function of problem (5) converges to the optimal value at a rate of $\mathcal{O}(1/T)$ for strongly-convex objectives with diminishing step-sizes, and at a rate of $\mathcal{O}(1/\sqrt{T})$ for convex objectives with $\beta = \mathcal{O}(1/\sqrt{T})$. Again, note that convergence is shown to a neighborhood of the optimal solution with the size of the neighborhood determined by the size of the LL error δ.

5 EXPERIMENTS

In this section, we evaluate the performance of Algorithms 1 and 2 via numerical experiments. First, we compare the performance of [D]SIGD to the recently proposed PDBO (Sow et al., 2022) for constrained bilevel optimization on a quadratic bilevel problem. Then in the second set of experiments, we evaluate the performance of [S]SIGD against popular adversarial training algorithms.

Quadratic Bilevel Optimization. Consider the quadratic bilevel problem of the form (5) with

$$f(\mathbf{x},\mathbf{y}) = \frac{1}{4}\|\mathbf{x}\|^2 + 10\mathbf{x}^T\mathbf{y} - \frac{1}{4}\|\mathbf{y}\|^2 + \mathbf{1}^T\mathbf{x} + \mathbf{1}^T\mathbf{y} + 1 \text{ and } h(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y} + \frac{1}{2}\|\mathbf{y}\|^2 + x_1 + y_2, \tag{13}$$

and linear constraints of the form $|y_i| \leq 1, i \in \{1, 2\}$. Here, $\mathbf{x} = [x_1, x_2]^T$, $\mathbf{y} = [y_1, y_2]^T$ with $x_i, y_i \in \mathbb{R}$ for $i \in \{1, 2\}$, and $\mathbf{1} = [1, 1]^T$. The evaluation criterion is the stationarity gap $\|\nabla F(\mathbf{x})\|$.

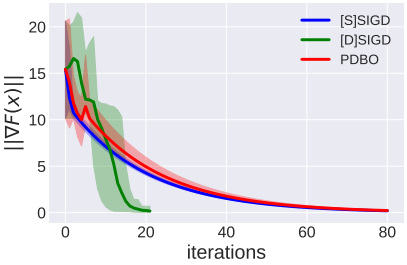


Figure 2: $\|\nabla F(\mathbf{x})\|$ vs # of iterations.

On this problem we execute [D]SIGD (Algorithm 1), [S]SIGD (Algorithm 2), and PDBO (Sow et al., 2022). In the first two cases, we solve the inner-level problem using 10 steps of projected gradient descent with stepsize 10^{-1}. For the stepsize of [S]SIGD, we choose $\beta = 0.1$, while in [D]SIGD we find the proper Armijo step-size by successively adapting (by increasing m) the quantity $a_r = (0.9)^m$ until condition (11) is met. In PDBO we select 10^{-1} for the stepsizes of both the primal and dual steps, and the number of inner iterations is set to 10. In Figure 2, we plot the convergence curves for the three algorithms with respect to number of iterations; the results are averaged over 10 runs. Note that the line search [D]SIGD method outperforms the fixed step-size [S]SIGD and PDBO while [S]SIGD performs similar to PDBO.

Adversarial Learning. We consider an adversarial learning problem of the form given in (3). For the perturbation we focus on the ϵ-tolerant ℓ_∞-norm attack constraint, namely $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{d_\ell} \,|\, \|\mathbf{y}\|_\infty \leq \epsilon\}$, which can easily be expressed as a linear inequality constraint as in the LL problem of (3). We consider two widely accepted adversarial learning method as our baselines, namely AT (Madry et al., 2017) and TRADES (Zhang et al., 2019b). Also, we consider two representative datasets CIFAR-10/100 (Krizhevsky et al., 2009) and adopt the ResNet-18 (He et al., 2016) model; the results for CIFAR-10 are provided in Appendix C. In particular, we studied two widely used (Madry et al., 2017; Zhang et al., 2019b; Wong et al., 2020; Andriushchenko & Flammarion, 2020; Zhang et al., 2019a) attack budget choices $\epsilon \in \{8/255, 16/255\}$. In the implementation of our [S]SIGD method, we adopt a perturbation generated by a Gaussian random vector $\mathbf{q}$ with variances from the following list $\sigma^2 \in \{2\text{e}{-5}, 4\text{e}{-5}, 6\text{e}{-5}, 8\text{e}{-5}, 1\text{e}{-4}, \}$, in order to study different levels of smoothness. Moreover, for solving the LL problem in each iteration we select a fixed batch of samples. We choose f_i to be cross-entropy loss and $h_i = -f_i + \lambda\|\mathbf{y}_i\|^2$ for hyper-parameter $\lambda > 0$. For [S]SIGD, we follow the implementation of (Zhang et al., 2021) but with perturbations in the LL problem. We evaluate the robustly trained model with two metrics, namely the standard accuracy (SA) and robust accuracy (RA), where we evaluate the accuracy of the robustified model on the clean and attacked test set, respectively; the attacked set is generated using PGD-50-10 (Madry et al., 2017) (i.e., 50-step PGD attack with 10 restarts). Desirably, a well trained model possesses high RA while maintaining simultaneously the SA at a high level. Table 1 shows the performance overview of our experiments. We make the following observations. First, a low level of perturbation variance (*e.g.*, $\sigma^2 \in \{2\text{e}{-5}, 4\text{e}{-5}\}$) in general improves both SA as well as RA, which presents an enhanced RA-SA trade-off. For example, in the setting (CIFAR-100, $\epsilon = 16/255$), our algorithm boosts the RA by over 0.3% and the SA by 2%. Second, a high level of perturbation variance harms the robustness but results in high SA. This is reasonable, since the stochastic gradient becomes too noisy with large variances. Third, our method outperforms AT and closely matches the performance of the stronger baseline TRADES. However, we would like to stress that the intent of our work is not to design a specialized adversarial learning method, and thus robustness gap between our method and the strong baseline does not diminish the value of our method. Additional details are provided in Appendix C.2.

Table 1: Performance overview of different methods on CIFAR-100 Krizhevsky et al. (2009) with ResNet-18 He et al. (2016). The result $a_{\pm b}$ represents the mean value a with a standard deviation of b over 5 random trials.

CIFAR-100, $\epsilon = 8/255$							
Metrics	AT	TRADES	[S]SIGD (Gaussian variance σ^2)				
			2e−5	4e−5	6e−5	8e−5	1e−4
SA	$53.83_{\pm 0.19}$	$53.33_{\pm 0.18}$	$53.88_{\pm 0.22}$	$54.01_{\pm 0.24}$	$53.79_{\pm 0.14}$	$54.44_{\pm 0.18}$	$57.74_{\pm 0.22}$
RA	$27.36_{\pm 0.24}$	$28.44_{\pm 0.17}$	$27.43_{\pm 0.12}$	$28.22_{\pm 0.10}$	$28.12_{\pm 0.14}$	$27.14_{\pm 0.21}$	$25.22_{\pm 0.15}$
$\epsilon = 16/255$							
SA	$42.06_{\pm 0.17}$	$42.19_{\pm 0.23}$	$44.06_{\pm 0.19}$	$45.66_{\pm 0.25}$	$46.57_{\pm 0.22}$	$47.11_{\pm 0.32}$	$47.46_{\pm 0.44}$
RA	$15.10_{\pm 0.28}$	$16.59_{\pm 0.26}$	$15.51_{\pm 0.17}$	$14.18_{\pm 0.22}$	$13.92_{\pm 0.25}$	$13.54_{\pm 0.18}$	$13.42_{\pm 0.26}$

REFERENCES

Amin Abedi, Mohammad Reza Hesamzadeh, and Franco Romerio. An acopf-based bilevel optimization approach for vulnerability assessment of a power system. *International Journal of Electrical Power & Energy Systems*, 125:106455, 2021. ISSN 0142-0615.

G. B. Allende and G. Still. Solving bilevel programs with the kkt-approach. *Mathematical Programming*, 138:309–332, 2013.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

D. A. Arias, A. A. Mota, L. T. M. Mota, and C. A. Castro. A bilevel programming approach for power system operation planning considering voltage stability and economic dispatch. In *2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America*, pp. 1–6, 2008. doi: 10.1109/TDC-LA.2008.4641718.

Dimitri P Bertsekas. *Nonlinear programming, 2nd ed.* Athena Scientific Belmont, MA, 1998.

D.P. Bertsekas and W. Rheinboldt. *Constrained Optimization and Lagrange Multiplier Methods*. Computer science and applied mathematics. Elsevier Science, 2014. ISBN 9781483260471. URL `https://books.google.com/books?id=j6LiBQAAQBAJ`.

Paul H Calamai and Jorge J Moré. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Mark Cecchini, Joseph Ecker, Michael Kupferschmid, and Robert Leitch. Solving nonlinear principal-agent problems using bilevel programming. *European Journal of Operational Research*, 230(2): 364–373, 2013.

Tsung-Hui Chang, Mingyi Hong, Hoi-To Wai, Xinwei Zhang, and Songtao Lu. Distributed learning in the nonconvex world: From batch data to streaming and beyond. *IEEE Signal Processing Magazine*, 37(3):26–38, 2020. doi: 10.1109/MSP.2020.2970170.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method, 2021a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021b.

Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7): 4034–4059, 2015.

Benoît Colson, Patrice Marcotte, and Gilles Savard. Bilevel programming: A survey. *4or*, 3(2): 87–107, 2005.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

Stephan Dempe and Alain Zemkoho. *Bilevel optimization*. Springer, 2020.

Mohamed Didi-Biha, Patrice Marcotte, and Gilles Savard. *Path-based formulations of a bilevel toll setting problem*, pp. 29–50. Springer US, Boston, MA, 2006. ISBN 978-0-387-34221-4. doi: 10.1007/0-387-34221-4_2.

Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

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

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

T.L. Friesz and D. Bernstein. *Foundations of Network Optimization and Games*. Complex Networks and Dynamic Systems. Springer US, 2015. ISBN 9781489975942.

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

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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

Dr. Vyacheslav Kalashnikov, José-Fernando Camacho-Vallejo, Ronald Askin, and Nataliya Kalashnykova. Comparison of algorithms for solving a bi-level toll setting problem. *International journal of innovative computing, information & control: IJICIC*, 6:3529–3549, 08 2010.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization, 2021a.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34, 2021b.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Gérard Lebourg. Generic differentiability of lipschitzian functions. *Transactions of the American Mathematical Society*, 256:125–144, 1979.

Lecture. 5: Correspondences and berge's maximum theorem. *Math Camp Notes, Department of Economics, Yale University*, 2017.

Yi Li, Lingxiao Song, Xiang Wu, Ran He, and Tieniu Tan. Learning a bi-level adversarial network with global and local perception for makeup-invariant face verification. *Pattern Recognition*, 90: 99–108, 2019.

Gui-Hua Lin, Mengwei Xu, and Jane J. Ye. On solving simple bilevel programs with a nonconvex lower level program. *Math. Program.*, 144(1–2):277–305, April 2014.

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *arXiv*, 2021a.

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International Conference on Machine Learning*, pp. 6882–6892. PMLR, 2021b.

Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, and Yixuan Zhang. Value-function-based sequential minimization for bi-level optimization. *arXiv preprint arXiv:2110.04974*, 2021c.

Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. *Advances in Neural Information Processing Systems*, 33:2811–2822, 2020.

Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Patrice Marcotte, Gilles Savard, and D. Zhu. A trust region algorithm for nonlinear bilevel programming. *Oper. Res. Lett.*, 29:171–179, 11 2001. doi: 10.1016/S0167-6377(01)00092-X.

Athanasios Migdalas. Bilevel programming in traffic planning: Models, methods and challenge. *Journal of global optimization*, 7(4):381–405, 1995.

James A Mirrlees. The theory of moral hazard and unobservable behaviour: Part i. *The Review of Economic Studies*, 66(1):3–21, 1999.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Julie Nutini, Mark Schmidt, and Warren Hare. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13(4):645–655, 2019.

Francesca Parise and Asuman Ozdaglar. Sensitivity analysis for network aggregative games. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 3200–3205. IEEE, 2017.

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

Arvind U Raghunathan and Lorenz T Biegler. Mathematical programs with equilibrium constraints (mpecs) in process engineering. *Computers & Chemical Engineering*, 27(10):1381–1392, 2003. ISSN 0098-1354.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

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

Ankur Sinha, Tanmay Khandait, and Raja Mohanty. A gradient-based bilevel optimization approach for tuning hyperparameters in machine learning. *arXiv preprint arXiv:2007.11022*, 2020.

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

Heinrich Von Stackelberg and Stackelberg Heinrich Von. *The theory of the market economy*. Oxford University Press, 1952.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

JJ Ye and DL Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33(1): 9–27, 1995.

Farzad Yousefian. Bilevel distributed optimization in directed networks. In *2021 American Control Conference (ACC)*, pp. 2230–2235. IEEE, 2021.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in Neural Information Processing Systems*, 32, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019b.

Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. *arXiv: 2112.12376*, 2021.

A RELATED LITERATURE

Bilevel optimization problems, initially encountered in the context of Stackelberg (leader-follower) games (Von Stackelberg & Von, 1952), find applications in a multitude of areas, including machine learning (Liu et al., 2021a), economics (Mirrlees, 1999), power systems (Abedi et al., 2021; Arias et al., 2008), chemical industry (Raghunathan & Biegler, 2003), transport research (Didi-Biha et al., 2006; Kalashnikov et al., 2010); see (Colson et al., 2005; Dempe & Zemkoho, 2020; Sinha et al., 2017; Liu et al., 2021a) for a number of survey papers. The "classical" approaches for solving bilevel problems include the use of approximate descent methods (Shaban et al., 2019; Ghadimi & Wang, 2018; Franceschi et al., 2017), penalty methods (Lin et al., 2014), KKT reformulations-based approaches (Allende & Still, 2013), value function-based methods (Ye & Zhu, 1995; Sow et al., 2022), and trust-region algorithms (Marcotte et al., 2001). In addition, bilevel problems are known to be related to mathematical programs with equilibrium constraints (MPEC) (Luo et al., 1996).

Recently, motivated by machine learning applications, gradient-based approaches have gained popularity for solving bilevel optimization problems (Liu et al., 2021a), e.g., in hyperparameter optimization (Shaban et al., 2019; Franceschi et al., 2017; 2018), and meta learning (Rajeswaran et al., 2019; Franceschi et al., 2018). The majority of those works are focused on solving bilevel problems with unconstrained strongly convex LL problem, for both stochastic and deterministic objectives (Ghadimi & Wang, 2018; Hong et al., 2020; Khanduri et al., 2021a;b; Chen et al., 2021a; Ji et al., 2021; Chen et al., 2021b; Yang et al., 2021). An attractive property of such problems is the existence and easy computability of the implicit gradient. Moreover, under mild assumptions, the implicit gradient for these problems can be shown to be Lipschitz smooth (e.g., see (Ghadimi & Wang, 2018, Lemma 2.2) and (Khanduri et al., 2021b, Lemma 3.1)). In contrast, for bilevel problems with linear LL constraints the implicit gradient in general might not exist, and even if it exists computing it in closed-form is a challenging task. As discussed earlier, we develop a perturbation-based smoothing framework for the constrained LL problem that ensures the existence of the implicit gradient in an almost sure sense, and allows us to compute an explicit expression for the implicit gradient.

In Liu et al. (2021c) and Sow et al. (2022) the authors have considered bilevel optimization with (general) constraints in the LL problem. Both papers develop a value function-based framework that leads to a single level problem with non-convex constraints. In Liu et al. (2021c) a sequential minimization approach is followed where the value-function and the LL constraints are incorporated into the objective using penalty or barrier functions. In Sow et al. (2022) a primal-dual-based framework is proposed in which the problem is regularized with the addition of a strongly-convex penalty term, while a constant error term is added to make the constraint set strictly feasible. In contrast, our approach relies only on a small linear perturbation which can be made arbitrarily small without practically changing the landscape of the LL problem.

B SOLUTION METHODS FOR THE LL PROBLEM

The LL problem is a strongly convex linearly constrained optimization task. As a result, there exist many efficient ways to find its solutions. In order to discuss about them, we consider two different classes of problems depending on the exact form of the linear constraints and the difficulty of computing the respective projection operator: 1) the projection has a closed-form solution, 2) the projection requires the solution of an optimization problem. Before we proceed, we would like to stress that the problem we are solving, i.e., the bilevel problem with linear constraints in the LL, is a very challenging one, regardless of the specific form and the exact way we approach the solution of the LL problem.

In the first class of problems, where the projection can be computed in closed form, we have problems with special linear constraints. One characteristic example is box constraints, i.e. constraints of the form $\mathbf{a} \leq \mathbf{y} \leq \mathbf{b}$, where the inequalities apply in a component-wise manner. These constraints appear in applications, such as adversarial learning (see the motivating applications in the main text). In this case, we can use some first-order iterative algorithm to solve the LL problem and project each iterate onto the constraint set using the closed-form expression (which only incurs a constant cost per iteration). For instance, we can use the projected gradient descent method which probably converges to the optimal solution with a linear rate.

In the second class of problems, the projection operator does not possess a closed-form expression. In this case we can approach the LL problem as a convex optimization task, and solve it using some convex optimization solver (e.g. employing interior-point methods) to obtain a highly accurate solution with a complexity of $\mathcal{O}\left(p(d_\ell, k) log(d_\ell/\epsilon)\right)$, where $p(\cdot)$ is some polynomial and ϵ is solution accuracy. Alternatively, as mentioned in the previous case, we can use a projected gradient descent-type method that enjoys a linear convergence rate guarantee. Differently from the previous case though the projection operator computed at each iteration requires the solution of an optimization problem. Nonetheless, the projection task we are referring to is a (strongly convex) quadratic linearly constrained problem, that is a special quadratic programming task, which is easy to solve in practice. In algorithm 3 we describe the solution of the LL using a projected gradient descent algorithm.

Algorithm 3 Projected Gradient Descent (**PGD**)

1: **Input:** Initial parameter $\mathbf{y}^0$, Current iterate $\mathbf{x}$, # iter T, step-sizes $\{\gamma^r\}_{r=0}^{T-1}$, Constraints $A, \mathbf{b}$
2: **for** $r = 0, 1, \ldots, T-1$ **do**
3: $\quad \mathbf{y}^{r+1} = \mathbf{y}^r - \gamma^r \nabla_y g(\mathbf{x}, \mathbf{y}^r)$
4: $\quad$ Project $\mathbf{y}^{r+1}$ to $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{d_\ell} | A\mathbf{y} \leq \mathbf{b}\}$ by solving the following QP:

$$\min_{\mathbf{y} \in \mathbb{R}^{d_\ell}} \|\mathbf{y} - \mathbf{y}^{r+1}\|^2 \text{ s.t. } A\mathbf{y} \leq \mathbf{b} \tag{14}$$

5: **end for**

C Additional Experiments

In this section, we include additional experiments on quadratic bilevel optimization problems and Adversarial training along with the implementation details. First, we evaluate the performance of the [D]SIGD and [S]SIGD on quadratic bilevel optimization problems.

C.1 Numerical Results

We consider two linearly constrained quadratic bilevel problems of the form (5) with the UL and the LL objectives defined as

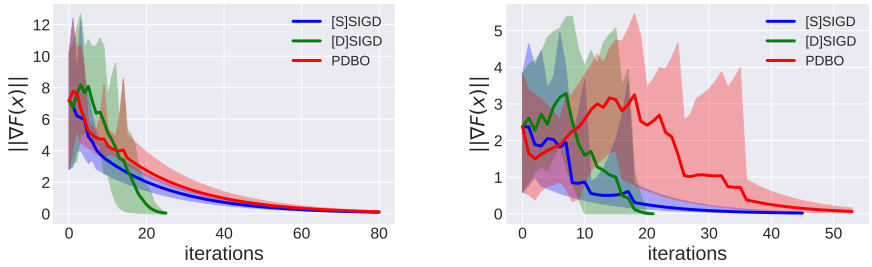
$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{4}\|\mathbf{x}\|^2 + 5\mathbf{x}^T\mathbf{y} - \frac{1}{4}\|\mathbf{y}\|^2 \ , \ h(\mathbf{x}, \mathbf{y}) = \frac{1}{4}\|\mathbf{x}\|^2 + \frac{1}{2}\mathbf{x}^T\mathbf{y} + \frac{1}{4}\|\mathbf{y}\|^2 \tag{15}$$

and

$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|^2 + 2\mathbf{x}^T\mathbf{y} - \frac{1}{2}\|\mathbf{y}\|^2 \ , \ h(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} + \frac{1}{2}\|\mathbf{y}\|^2. \tag{16}$$

For both the cases, we have $d_u = d_l = 2$, and the linear constraints in the LL are of the form $-1 \leq \mathbf{y}_i \leq 1, \ i \in \{1, 2\}$. We compare the performance of SIGD algorithms to recently proposed PDBO (Sow et al., 2022). In Figures 3a and 3b, we present the evolution of the stationarity gap $\|\nabla F(\mathbf{x})\|$ during the execution of the three algorithms, for the problems (15) and (16), respectively. The results are averaged over 10 random runs, and the variance of the results across these runs is reflected on the shaded region across the convergence curves. In our experiments, we choose the step-size using the backtracking line search for [D]SIGD as stated in Algorithm 1, while for [S]SIGD we choose a constant step-size. Note that since problems (15) and (16) are deterministic [S]SIGD utilizes a gradient estimator with zero variance.

In problem (15), we solve the LL problem using 10 steps of projected gradient descent with stepsize 0.1; in the case of [D]SIGD the stepsize is 1. For the stepsize of [S]SIGD, we choose $\beta = 0.1$, while in [D]SIGD we find the proper Armijo step-size by successively adapting (by increasing m) the quantity $a_r = (0.9)^m$ until condition (11) is met. In PDBO we select 0.1 for the stepsizes of both the primal and dual steps, and the number of inner iterations is set to 10. In problem (16), we solve the LL problem using 20 steps of projected gradient descent with stepsize 0.1; in the case of [D]SIGD the number of steps is 10 and the stepsize is 1. For the stepsize of [S]SIGD, we choose $\beta = 0.1$, while in [D]SIGD we find the proper Armijo step-size by successively adapting (by increasing m) the quantity $a_r = (0.95)^m$ until condition (11) is met. In PDBO we select 0.1 for the stepsizes of both the primal and dual steps, and the number of inner iterations is set to 20.



(a) Problem (15). (b) Problem (16).

Figure 3: Convergence curves w.r.t. number of iterations.

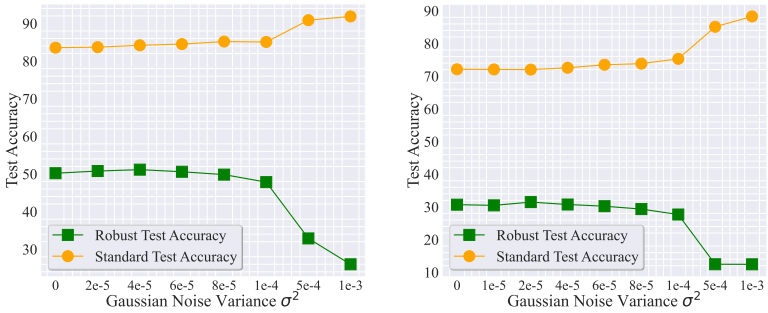
Table 2: Performance overview of different methods on CIFAR-10 (Krizhevsky et al., 2009) with ResNet-18 (He et al., 2016). The result $a_{\pm b}$ represents the mean value a with a standard deviation of b over 5 random trials.

CIFAR-10, $\epsilon = 8/255$							
Metrics	AT	TRADES	[S]SIGD (Gaussian variance σ^2)				
			2e−5	4e−5	6e−5	8e−5	1e−4
SA	$80.78_{\pm 0.23}$	$80.23_{\pm 0.23}$	$80.70_{\pm 0.14}$	$81.20_{\pm 0.22}$	$81.52_{\pm 0.21}$	$83.19_{\pm 0.24}$	$85.08_{\pm 0.44}$
RA	$50.71_{\pm 0.21}$	$51.17_{\pm 0.19}$	$50.78_{\pm 0.21}$	$51.15_{\pm 0.19}$	$50.59_{\pm 0.18}$	$49.83_{\pm 0.23}$	$47.83_{\pm 0.13}$
$\epsilon = 16/255$							
SA	$70.31_{\pm 0.11}$	$70.22_{\pm 0.29}$	$71.43_{\pm 0.14}$	$72.79_{\pm 0.24}$	$73.50_{\pm 0.09}$	$73.98_{\pm 0.35}$	$75.31_{\pm 0.33}$
RA	$32.12_{\pm 0.18}$	$33.35_{\pm 0.14}$	$32.72_{\pm 0.25}$	$31.73_{\pm 0.10}$	$29.97_{\pm 0.14}$	$29.39_{\pm 0.15}$	$27.67_{\pm 0.07}$

C.2 Adversarial Learning

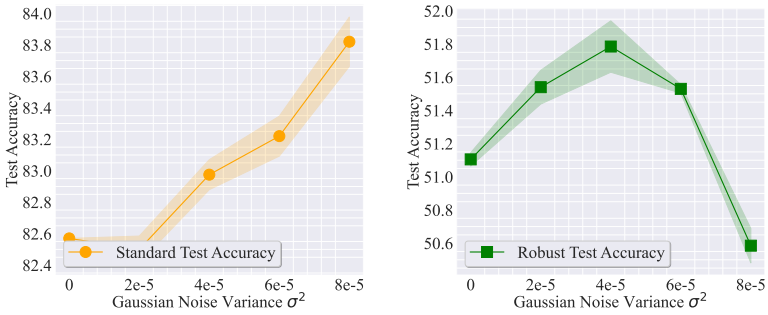
In this section, we present some additional results along with the implementation details for the adversarial learning problem. As noted earlier, we consider the adversarial learning problem of form (3). For learning the perturbation $\mathbf{y}^*(\mathbf{x})$, we focus on the ϵ-tolerant ℓ_∞-norm attack constraint, i.e., $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{d_\ell} \,|\, \|\mathbf{y}\|_\infty \leq \epsilon\}$. Note that this constraint can easily be expressed as a linear inequality constraint as in the LL problem in (3). In particular, we evaluate the performance of [S]SIGD on two widely used attack budget choices of $\epsilon \in \{8/255, 16/255\}$ (Madry et al., 2017; Zhang et al., 2019b; Wong et al., 2020; Andriushchenko & Flammarion, 2020; Zhang et al., 2019a). In the implementation of our [S]SIGD method, we adopt a perturbation generated by a Gaussian random vector $\mathbf{q}$ with variances from the following list $\sigma^2 \in \{2\text{e}{-5}, 4\text{e}{-5}, 6\text{e}{-5}, 8\text{e}{-5}, 1\text{e}{-4}, \}$, in order to study different levels of smoothness. We choose f_i to be cross-entropy loss and $h_i = -f_i + \lambda\|\mathbf{y}_i\|^2$ with $\lambda > 0$ as a hyper-parameter. For solving (3), in each iteration we select a fixed batch of samples for both the UL and LL problems. Also, note that the ReLU-based neural networks commonly lead to a piece-wise linear decision boundary w.r.t. the inputs (Moosavi-Dezfooli et al., 2019). This implies that the implicit gradient in (10) can be further approximated using a Hessian-free implementation, where the Hessian of the LL problem can be approximated by λI (Zhang et al., 2021, Eq. (25)). Note that these approximations are common in practice and do not lead to performance degradation compared to the case when full Hessian is used to compute the implicit-gradient (Zhang et al., 2021, Table 5). Next, we analyze the effect of adding different perturbations $\mathbf{q}$ in the LL problem on the performance of [S]SIGD. Specifically, we choose $\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and evaluate the performance of [S]SIGD with σ^2.

In Figure 4, we plot the robust accuracy (RA) and the standard accuracy (SA) with respect to the variance of the Gaussian perturbation vector used in the LL problem. As can be seen, the RA increases as the variance increases within a certain range. However, with stronger noise (i.e., $\sigma^2 > 10^{-4}$), the RA drops sharply, while the SA increases. This is reasonable, since high variance makes the true LL gradient noisy. For easier observation, in Figure 5 we zoom in the part of Figure 4 where $\sigma^2 \in [0, 8 \cdot 10^{-5}]$. It can be clearly seen that adding a small perturbation $\mathbf{q}$ helps in improving the RA.

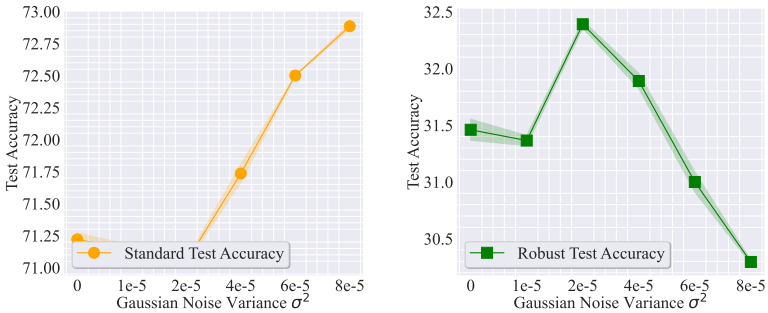


(a) $\epsilon = 8/255$ (b) $\epsilon = 16/255$

Figure 4: The influence of Gaussian variance on the RA and SA. The experiments are based on CIFAR-10 with ResNet-18 model.



(a) $\epsilon = 8/255$, CIFAR-10



(b) $\epsilon = 16/255$, CIFAR-10

Figure 5: The influence of Gaussian variance on the RA and SA.

Next, we compare the performance of [S]SIGD against two widely accepted adversarial learning methods as baselines, namely AT (Madry et al., 2017) and TRADES (Zhang et al., 2019b). Here, we present the results for CIFAR-10 dataset (Krizhevsky et al., 2009) and adopt the ResNet-18 (He et al., 2016). In Table 2, we compare the performance of [S]SIGD for different perturbation variances with classical AT (Madry et al., 2017) algorithm and TRADES (Zhang et al., 2019b). Note that for appropriate choice of perturbation variance [S]SIGD outperforms the classical AT algorithm while performs is only slightly worse compared to TRADES, especially, for higher attack budget of $\epsilon = 16/255$.

D PROOFS OF SECTION 2

D.1 PROOF OF PROPOSITION 1

Note that the goal of Proposition 1 is to establish the continuity of the mapping $\mathbf{y}^*(\mathbf{x})$ and the implicit function $F(\mathbf{x}) := f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$. In the following, we will show that under Assumption 1, $\mathbf{y}^*(\mathbf{x})$ is in fact continuous, which will then utilize to establish the continuity of $F(\mathbf{x})$. Before starting the proof we need a few definitions. Consider the LL problem (1b) and let us denote the set $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{d_\ell} | A\mathbf{y} \leq \mathbf{b}\}$. Note that in general the constraint set $\mathcal{Y}$ can depend on the UL variable $\mathbf{x} \in \mathcal{X}$. For such cases, $\mathcal{Y}(\mathbf{x})$ is a set valued map $\mathcal{Y} : \mathcal{X} \to \mathbb{R}^{d_\ell}$ and is referred to as a correspondence. However, for the bilevel problem in (1a) and (1b) the correspondence $\mathcal{Y}$ is independent of $\mathbf{x} \in \mathcal{X}$ and is a fixed set. Also, we define the upper-semi continuity (USC) and the lower-semi continuity (LSC) for the correspondence $\mathcal{Y}(\mathbf{x})$. To define these notions of continuity, we will utilize the notion of an ϵ-ball defined below.

Definition 2 (ϵ-Ball)**.** *For $\mathcal{Y} \subset \mathbb{R}^{d_\ell}$, and given $\epsilon > 0$, we define the open ball about $\mathcal{Y}$ as*

$$B_\epsilon(\mathcal{Y}) := \big\{\mathbf{y} \in \mathbb{R}^{d_\ell} \;\big|\; \|\mathbf{y} - \mathbf{y}'\| < \epsilon, \;\; \textit{for some } \mathbf{y}' \in \mathcal{Y}\big\},$$

where $\|\cdot\|$ is the standard Euclidean norm.

Using the ϵ-ball we define the Upper Semi-Continuity (USC) of the correspondence $\mathcal{Y}$.

Definition 3 (Upper Semi-Continuity (USC))**.** *The correspondence $\mathcal{Y} : \mathcal{X} \to \mathbb{R}^{d_\ell}$ is USC if for every $\mathbf{x} \in \mathcal{X}$ and $\epsilon > 0$, there exists a $\delta > 0$ such that $\mathcal{Y}(\mathbf{x}') \subset B_\epsilon(\mathcal{Y}(\mathbf{x})))$, if $\mathbf{x}' \in \mathcal{X}$ and $\|\mathbf{x} - \mathbf{x}'\| < \delta$.*

Next, we define the notion of Lower Semi-Continuity (LSC).

Definition 4 (Lower Semi-Continuity (LSC))**.** *The correspondence $\mathcal{Y} : \mathcal{X} \to \mathbb{R}^{d_\ell}$ is LSC if for any sequence $\mathbf{x}_n$ in $\mathcal{X}$ that converges to a point $\mathbf{x} \in \mathcal{X}$, and $\mathbf{y} \in \mathcal{Y}(\mathbf{x})$, there exists a sequence $\mathbf{y}_n$ such that $\mathbf{y}_n \in \mathcal{Y}(\mathbf{x}_n)$, for all $n \in \mathbb{N}$, and $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{y}$.*

Theorem 5 (Berge's Theorem of Maximum (Lecture, 2017))**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_u}$ be a non-empty set. Also, let $\mathcal{Y} : \mathcal{X} \to \mathbb{R}^{d_\ell}$ be a correspondence such that the set $\mathcal{Y}(\mathbf{x})$ is compact and non-empty for all $\mathbf{x} \in \mathcal{X}$, and $\mathcal{Y}$ is USC and LSC. Then, if $g : \mathcal{X} \times \mathbb{R}^{d_\ell} \to \mathbb{R}$ is a continuous function with $\mathbf{y}^*(\mathbf{x})$ defined as*

$$\mathbf{y}^*(\mathbf{x}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_\ell}} \Big\{g(\mathbf{x}, \mathbf{y}) \;\Big|\; \mathbf{y} \in \mathcal{Y}(\mathbf{x})\Big\},$$

the correspondence $\mathbf{y}^*(\mathbf{x})$ is non-empty for all $\mathbf{x} \in \mathcal{X}$, and USC.

Remark 2. *If $\mathbf{y}^*(\mathbf{x})$ is singleton, then USC implies the continuity of the map $\mathbf{y}^*(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$.* □

Next, we present the proof of Proposition 1.

Proof. The proof of proposition 1 follows from the application of Berge's theorem.

To begin with, note that for our problem the set $\mathcal{Y}$ is a fixed set independent of $\mathbf{x} \in \mathcal{X}$. We are going to verify the conditions of Theorem 5. First, note from Assumption 1 that the set $\mathcal{Y}$ is non-empty and compact. Then, it is easy to see that $\mathcal{Y} \subset B_\epsilon(\mathcal{Y})$, for every $\epsilon > 0$, and that implies the LSC of $\mathcal{Y}$. Moreover, since the set $\mathcal{Y}$ is independent of $\mathbf{x} \in \mathcal{X}$ and comapct, for every sequence $\mathbf{x}_n \to \mathbf{x}$ in $\mathcal{X}$, we can always find a sequence $\mathbf{y}_n \to \mathbf{y}$, such that $\mathbf{y}_n, \mathbf{y} \in \mathcal{Y}$. Therefore, $\mathcal{Y}$ is LSC. Finally, using Assumption 1 we see that the function $g(\mathbf{x}, \mathbf{y})$ is continuous. Then, Theorem 5 implies that the set $\mathbf{y}^*(\mathbf{x})$ is non-empty and the correspondence USC.

Using the strong-convexity of $g(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{y}$ (Assumption 1) we claim that $\mathbf{y}^*(\mathbf{x})$ will be a singleton, and thereby a continuous mapping. Then, the continuity of $\mathbf{y}^*(\mathbf{x})$ implies the continuity of $F(\mathbf{x}) := f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, since the composition of two continuous functions is continuous. The proof is now complete. □

D.2 PROOF OF PROPOSITION 2

Proof. To begin with, let us denote the global minimum of the perturbed LL problem $g(\mathbf{x}, \mathbf{y})$ with $\mathbf{y}^*(\mathbf{x})$, and the respective minimum of $h(\mathbf{x}, \mathbf{y})$ with $\widetilde{\mathbf{y}}^*(\mathbf{x})$. From Assumption 1 we know that $h(\mathbf{x}, \mathbf{y})$

(and thus $g(\mathbf{x}, \mathbf{y})$) is strongly convex in $\mathbf{y}$ with modulus $\mu_g = \mu_h$. As a result we have that

$$h(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \geq h(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})) + \langle \nabla_y h(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})), \mathbf{y}^*(\mathbf{x}) - \widetilde{\mathbf{y}}^*(\mathbf{x}) \rangle + \frac{\mu_g}{2} \|\mathbf{y}^*(\mathbf{x}) - \widetilde{\mathbf{y}}^*(\mathbf{x})\|^2 \quad (17)$$

$$g(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})) \geq g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \langle \nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})), \widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}) \rangle + \frac{\mu_g}{2} \|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2. \quad (18)$$

By definition $\widetilde{\mathbf{y}}^*(\mathbf{x})$ is the global minimum of the objective $h(\mathbf{x}, \mathbf{y})$, and so it holds that $\langle \nabla_y h(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})), \mathbf{y}^*(\mathbf{x}) - \widetilde{\mathbf{y}}^*(\mathbf{x}) \rangle \geq 0$. Similarly, $\mathbf{y}^*(\mathbf{x})$ is the global minimum of the objective $g(\mathbf{x}, \mathbf{y})$, and so it holds that $\langle \nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})), \widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}) \rangle \geq 0$. Then, using the above inequalities and adding (17) and (18), we get

$$\begin{aligned}
&h(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + g(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})) \geq h(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})) + g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \mu_g \|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2 \\
&\mu_g \|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2 \leq [h(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))] + [g(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x})) - h(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x}))] \\
&\mu_g \|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2 \leq -\mathbf{q}^T \mathbf{y}^*(\mathbf{x}) + \mathbf{q}^T \widetilde{\mathbf{y}}^*(\mathbf{x}) \\
&\|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2 \leq \frac{\mathbf{q}^T \left(\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}) \right)}{\mu_g} \\
&\|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|^2 \leq \frac{\|\mathbf{q}^T\| \|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\|}{\mu_g} \\
&\|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\| \leq \frac{\|\mathbf{q}\|}{\mu_g}.
\end{aligned}$$

Using the above bound and the fact that f is Lipschitz continuous (it follows from the bounded gradient assumption 2(a)) it is easy to see that

$$|F(\mathbf{x}) - G(\mathbf{x})| = |f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - f(\mathbf{x}, \widetilde{\mathbf{y}}^*(\mathbf{x}))| \leq \overline{L}_f \|\mathbf{y}^*(\mathbf{x}) - \widetilde{\mathbf{y}}^*(\mathbf{x})\| \leq \overline{L}_f \frac{\|\mathbf{q}\|}{\mu_g}.$$

Therefore, the proof is complete. □

E PROOFS OF SECTION 3

E.1 PROOF OF LEMMA 2

Proof. In this proof we follow a reasoning similar to (Parise & Ozdaglar, 2017, Thm. 1). However, differently from that work we consider bilevel problems rather than Nash games. To begin with, consider the Lagrangian of problem (5b), i.e.,

$$L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = g(\mathbf{x}, \mathbf{y}) + \boldsymbol{\lambda}^T \left(A\mathbf{y} - \mathbf{b} \right).$$

Then, for some fixed $\mathbf{x} \in \mathcal{X}$, consider a KKT point $(\mathbf{y}^*(\mathbf{x}), \boldsymbol{\lambda}^*(\mathbf{x}))$ of (5b), for which it holds that,

- $\nabla_y L(\mathbf{x}, \mathbf{y}^*(\mathbf{x}), \boldsymbol{\lambda}^*(\mathbf{x})) = \nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + A^T \boldsymbol{\lambda}^*(\mathbf{x}) = 0$
- $[\boldsymbol{\lambda}^*(\mathbf{x})]^T \left(A\mathbf{y}^*(\mathbf{x}) - \mathbf{b} \right) = 0$
- $\boldsymbol{\lambda}^*(\mathbf{x}) \geq 0$
- $A\mathbf{y}^*(\mathbf{x}) - \mathbf{b} \leq 0.$

Now, consider the active constraints at $(\mathbf{y}^*(\mathbf{x}), \boldsymbol{\lambda}^*(\mathbf{x}))$, and to simplify notation let us set $\overline{A} := \overline{A}(\mathbf{y}^*(\mathbf{x}))$. Using the notations defined in Section 2 and the SC property, the KKT conditions given above can be equivalently rewritten as

$$\nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \overline{A}^T \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = 0, \quad \overline{A}\mathbf{y}^*(\mathbf{x}) - \overline{\mathbf{b}} = 0, \quad \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) > 0, \quad (19)$$

where $\overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ is the subvector of $\boldsymbol{\lambda}^*(\mathbf{x})$ that contains only the elements whose indices correspond to the active constraints at $\mathbf{y} = \mathbf{y}^*(\mathbf{x})$. Moreover, notice that the point $(\mathbf{y}^*(\mathbf{x}), \boldsymbol{\lambda}^*(\mathbf{x}))$ is unique. The

uniqueness of $\mathbf{y}^*(\mathbf{x})$ follows from the strong convexity of $g(\mathbf{x}, \cdot)$; the uniqueness of $\boldsymbol{\lambda}^*(\mathbf{x})$ results from the fact that matrix $\overline{A}$ has full row rank (which guarantees regularity, e.g., see Bertsekas (1998)).

As mentioned in section 2, the SC condition (from Lemma 1) combined with Assumption 1 implies that the mapping $\mathbf{y}^*(\mathbf{x})$ is differentiable almost surely (Friesz & Bernstein, 2015, Theorem 2.22). As a result, at any given point $\mathbf{x}$, we can consider a sufficiently small neighborhood around it, such that the active constraints $\overline{A}$ remain unchanged. Then, we can compute the gradient of (19) using the implicit function theorem as follows

$$\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \nabla \mathbf{y}^*(\mathbf{x}) + \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = 0 \tag{20}$$

$$\overline{A} \nabla \mathbf{y}^*(\mathbf{x}) = 0. \tag{21}$$

Solving the (20) for $\nabla \mathbf{y}^*(\mathbf{x})$ yields

$$\nabla \mathbf{y}^*(\mathbf{x}) = \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \left[-\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right], \tag{22}$$

where we exploited the fact that the Hessian matrix $\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is positive definite and thus invertible. Substituting (22) into (21) gives

$$\begin{aligned} &\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \left[-\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] = 0 \\ \implies &\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = -\left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))^{-1}\right]\overline{A}^T\right]^{-1} \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]. \end{aligned}$$

Finally, note that the KKT point $\mathbf{y}^*(\mathbf{x})$ corresponds to the unique global minimum of (5b), due to the strong convexity of $g(\mathbf{x}, \cdot)$. The proof is now complete. □

E.2 The proof of Lemma 3

The proof of Lemma 3 requires several intermediate results which we provide below. Note that under Assumption 3(c) it holds that $\overline{A}(\mathbf{y}^*(\mathbf{x})) = \overline{A}(\widehat{\mathbf{y}}(\mathbf{x}))$; for simplicity we will denote these matrices as $\overline{A}$ in the derivations of this subsection. Moreover, for any given matrix A we will denote with L_A the maximum value of the quantity $\|\overline{A}\left(\widehat{\mathbf{y}}(\mathbf{x})\right)\|$, across all $\mathbf{x} \in \mathcal{X}$.

Lemma 5. *Suppose that Assumption 1, 2, and 3 hold. Then for any $\mathbf{x} \in \mathcal{X}$, we have:*

(a) $\left\|\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1}\right\| \leq \frac{1}{\mu_g}, \ \forall \mathbf{y} \in \mathbb{R}^{d_\ell}$.

(b) $\left\|\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1}\right\| \leq \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \delta$.

(c) $\left\|\left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T\right]^{-1}\right\| \leq \overline{L}_A, \ \forall \mathbf{y} \in \mathbb{R}^{d_\ell}$.

(d) $\left\|\left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} - \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1}\right\| \leq L_A^2 \overline{L}_A^2 \frac{1}{\mu_g^2} L_{g_{yy}} \delta$.

Proof. a) We know that $g(\mathbf{x}, \mathbf{y})$ is strongly convex in $\mathbf{y}$ with modulus μ_g. Therefore, for any $\mathbf{x} \in \mathcal{X}$ we have

$$\begin{aligned} &\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}) \succeq \mu_g I \succ 0, \forall \mathbf{y} \in \mathbb{R}^{d_\ell} \\ \implies &0 \prec \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \preceq \frac{1}{\mu_g} I, \forall \mathbf{y} \in \mathbb{R}^{d_\ell} \\ \implies &\left\|\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1}\right\| \leq \frac{1}{\mu_g}, \forall \mathbf{y} \in \mathbb{R}^{d_\ell}. \end{aligned} \tag{23}$$

b) To begin with, notice that for arbitrary square invertible matrices P, Q we have

$$\left\|P^{-1} - Q^{-1}\right\| = \left\|P^{-1}(Q - P)Q^{-1}\right\| \leq \left\|Q^{-1}(P - Q)\right\| \left\|P^{-1}\right\| \leq \left\|Q^{-1}\right\| \left\|P - Q\right\| \left\|P^{-1}\right\|. \tag{24}$$

Then, using the above inequality we get

$$\begin{aligned}
&\left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \\
&\leq \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \right\| \left\| \nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) \right\| \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \\
&\leq \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \left\| \mathbf{y}^*(\mathbf{x}) - \widehat{\mathbf{y}}(\mathbf{x}) \right\| \\
&\leq \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \delta,
\end{aligned}$$

where in the second inequality we used the result from Lemma 5(a) and the Lipschitz Hessian property of g in $\mathbf{yy}$ (Assumption 2(d)); in the third inequality we use the Assumption 3(a) for $\mathbf{y}(\mathbf{x})$.

c) In our problem we have that g strongly convex in $\mathbf{y}$ and Lipschitz gradient in $\mathbf{y}$. Thus, for any $\mathbf{x} \in \mathcal{X}$ we have

$$\begin{aligned}
L_y I \succeq \nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*) \succeq \mu_g I \succ 0 \\
\implies 0 \prec \frac{1}{L_y} I \preceq \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \preceq \frac{1}{\mu_g} I, \forall \mathbf{y} \in \mathbb{R}^{d_\ell}.
\end{aligned}$$

Also, for every $\mathbf{x} \in \mathcal{X}$, we have that

$$\left\| \overline{A}^T \mathbf{z} \right\|^2 = \mathbf{z}^T \overline{A}\overline{A}^T \mathbf{z} \geq \lambda_{min}(\overline{A}\overline{A}^T) \|\mathbf{z}\|^2, \forall \mathbf{z} \in \mathbb{R}^{d_\ell}.$$

Using the above two lower bound we get

$$\mathbf{z}^T \overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T \mathbf{z} \geq \frac{1}{L_y} \lambda_{min}(\overline{A}\overline{A}^T) \|\mathbf{z}\|^2 > 0, \forall \mathbf{z} \in \mathbb{R}^{d_\ell} \setminus \{0\},$$

where the last inequality follows from the fact that $\overline{A}$ is full row rank which implies that $\lambda_{min}(\overline{A}\overline{A}^T) > 0, \forall \mathbf{x} \in \mathcal{X}$.

Since the above inequality holds for every $\mathbf{z} \in \mathbb{R}^{d_\ell} \setminus \{0\}$, it $\forall \mathbf{x} \in \mathcal{X}$ we get that

$$\begin{aligned}
\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T &\succeq \frac{\lambda_{min}(\overline{A}\overline{A}^T)}{L_y} I \succ 0 \\
\left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T\right]^{-1} &\preceq \frac{L_y}{\lambda_{min}(\overline{A}\overline{A}^T)} I \\
\left\| \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T\right]^{-1} \right\| &\leq \frac{L_y}{\lambda_{min}(\overline{A}\overline{A}^T)}.
\end{aligned}$$

Finally, for the given matrix A, consider the submatrix $\overline{A} = \overline{A}(\widehat{\mathbf{y}}(\mathbf{x}))$ generated by considering only the subset of its rows corresponding to the active constraints at $\widehat{\mathbf{y}}(\mathbf{x})$. From Assumption 1(c) we know that $\overline{A}(\widehat{\mathbf{y}}(\mathbf{x}))$ is full row rank for every $\mathbf{x} \in \mathcal{X}$, and so we can ensure that $\lambda_{min}\left(\overline{A}\left(\widehat{\mathbf{y}}(\mathbf{x})\right)\overline{A}\left(\widehat{\mathbf{y}}(\mathbf{x})\right)^T\right) > 0, \forall \mathbf{x} \in \mathcal{X}$. Then, we denote with λ_{min} the minimum value of the quantity $\lambda_{min}\left(\overline{A}\left(\widehat{\mathbf{y}}(\mathbf{x})\right)\overline{A}\left(\widehat{\mathbf{y}}(\mathbf{x})\right)^T\right)$ across all $\mathbf{x} \in \mathcal{X}$. Therefore, we conclude that

$$\left\| \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y})\right]^{-1} \overline{A}^T\right]^{-1} \right\| \leq \frac{L_y}{\lambda_{min}} := \overline{L}_A.$$

d) Applying formula (24) with $P = \overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T, Q = \overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T$ we get

$$
\begin{aligned}
&\left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} - \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \\
&\leq \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \left\| \overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T - \overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \right\| \\
&\qquad \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \\
&\leq \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \left\| \overline{A}\left[\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{y}(\mathbf{x}))\right]^{-1}\right] \overline{A}^T \right\| \\
&\qquad \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \\
&\leq \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \|\overline{A}\| \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \left\| \overline{A}^T \right\| \\
&\qquad \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \\
&\leq L_A^2 \overline{L}_A^2 \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \delta,
\end{aligned}
$$

where in the final inequality we used the bounds derived in Lemma 5(b), 5(c), and the bound $\|\overline{A}\| \leq L_A$.

The proof is now complete. □

Now let us bound the norm of the gradients of the mappings $\boldsymbol{\lambda}^*(\mathbf{x})$ and $\mathbf{y}^*(\mathbf{x})$.

Lemma 6. *Under Assumptions 1, 2 and 3, the gradients of the mappings $\boldsymbol{\lambda}^*(\mathbf{x})$ and $\mathbf{y}^*(\mathbf{x})$ satisfy the following bounds for every $\mathbf{x} \in \mathcal{X}$,*

$$
\|\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\| \leq \overline{L}_{\boldsymbol{\lambda}^*}, \quad \left\|\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right\| \leq \overline{L}_{\boldsymbol{\lambda}^*}
$$

$$
\|\nabla \mathbf{y}^*(\mathbf{x})\| \leq \overline{L}_{\mathbf{y}^*}, \quad \|\widehat{\nabla} \mathbf{y}^*(\mathbf{x})\| \leq \overline{L}_{\mathbf{y}^*}
$$

where $\overline{L}_{\boldsymbol{\lambda}^*} = \frac{1}{\mu_g} \overline{L}_A L_A \overline{L}_{g_{xy}}$ and $\overline{L}_{\mathbf{y}^*} = \frac{1}{\mu_y}\left(\overline{L}_{g_{xy}} + L_A \overline{L}_{\boldsymbol{\lambda}^*}\right)$. Note that $\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ and $\widehat{\nabla} \mathbf{y}^*(\mathbf{x})$ are obtained by substituting the estimate $\widehat{\mathbf{y}}(\mathbf{x})$ in place of $\mathbf{y}^*(\mathbf{x})$ in the expressions $\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ and $\nabla \mathbf{y}^*(\mathbf{x})$, respectively (Please see Lemma 2).

Proof. From Lemma 2 we have

$$
\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = -\left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]
$$

Then, taking the norm of this quantity we get,

$$
\begin{aligned}
\left\|\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right\| &= \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right] \right\| \\
&\leq \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \right\| \|\overline{A}\| \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \right\| \left\|\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right\| \\
&\leq \overline{L}_A L_A \frac{1}{\mu_g} \overline{L}_{g_{xy}} := \overline{L}_{\boldsymbol{\lambda}^*},
\end{aligned}
$$

where in the last inequality we used Lemma 5(a), 5(c) and Assumption 2(f).

Similarly, for $\|\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\|$ we have that

$$
\begin{aligned}
\left\|\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right\| &= \left\| \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A}\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right] \right\| \\
&\leq \overline{L}_A L_A \frac{1}{\mu_g} \overline{L}_{g_{xy}} = \overline{L}_{\boldsymbol{\lambda}^*}.
\end{aligned}
$$

Moving to the bound of $\|\nabla \mathbf{y}^*(\mathbf{x})\|$, we know from Lemma 2 that the formula of the gradient of $\mathbf{y}^*(\mathbf{x})$ is

$$\nabla \mathbf{y}^*(\mathbf{x}) = \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \left[-\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right]. \tag{25}$$

Then, we have that

$$\begin{aligned}
\|\nabla \mathbf{y}^*(\mathbf{x})\| &= \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \left[-\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] \right\| \\
&\leq \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \right\| \left\| \left[-\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] \right\| \\
&\leq \frac{1}{\mu_g} \left(\|\nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| + \|\overline{A}\| \left\| \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \right) \\
&\leq \frac{1}{\mu_g} \left(\overline{L}_{g_{xy}} + L_A \overline{L}_{\boldsymbol{\lambda}^*} \right) := \overline{L}_{\mathbf{y}^*},
\end{aligned}$$

where in the second inequality we used we used Lemma 5(a); the third inequality follows from Assumption 2(f) and the bound for $\left\| \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\|$ we derived above.

Similarly, for $\left\| \widehat{\nabla} \mathbf{y}(\mathbf{x}) \right\|$ we can obtain the following bound

$$\begin{aligned}
\left\| \widehat{\nabla} \mathbf{y}(\mathbf{x}) \right\| &= \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \left[-\nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - \overline{A}^T \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] \right\| \\
&\leq \frac{1}{\mu_g} \left(\|\nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| + \|\overline{A}\| \left\| \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \right) \\
&\leq \frac{1}{\mu_g} \left(\overline{L}_{g_{xy}} + L_A \overline{L}_{\boldsymbol{\lambda}^*} \right) = \overline{L}_{\mathbf{y}^*}.
\end{aligned}$$

The proof is now complete. □

In the next two results we are going to present bounds for the difference of the exact and approximate gradients of the mappings $\overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ and $\nabla \mathbf{y}^*(\mathbf{x})$.

Lemma 7. *Suppose that Assumptions 1, 2 and 3 hold. Then, the following bound holds*

$$\|\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\| \leq L_{\boldsymbol{\lambda}^*} \delta,$$

where $L_{\boldsymbol{\lambda}^*} = \left(\frac{1}{\mu_g}\right)^3 \overline{L}_A^2 L_A^3 L_{g_{yy}} \overline{L}_{g_{xy}} + \frac{1}{\mu_g} \overline{L}_A L_A L_{g_{xy}} + \left(\frac{1}{\mu_g}\right)^2 \overline{L}_A L_A L_{g_{yy}} \overline{L}_{g_{xy}}$.

Proof. Using the derivation of $\nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ from Lemma 2, and its approximation $\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ where we substitute $\mathbf{y}^*(\mathbf{x})$ with $\widehat{\mathbf{y}}(\mathbf{x})$ in the formula of the former, that is,

$$\widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) = -\left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right],$$

we obtain

$$\begin{aligned}
\left\| \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| = \Bigg\| & \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right] \\
& - \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]^{-1} \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right] \Bigg\|.
\end{aligned}$$

Below, we use the following notation in order to simplify the derivations.

$H(\mathbf{x}) = \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T\right]$, $G(\mathbf{x}) = \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1}$, $M(\mathbf{x}) = \nabla^2_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$

$\widehat{H}(\mathbf{x}) = \left[\overline{A} \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T\right]$, $\widehat{G}(\mathbf{x}) = \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1}$, $\widehat{M}(\mathbf{x}) = \nabla^2_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))$

Then, we have that

$$
\begin{aligned}
\|\nabla\overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \widehat{\nabla}\overline{\boldsymbol{\lambda}}^*(\mathbf{x})\| &= \left\|H^{-1}(\mathbf{x})\overline{A}G(\mathbf{x})M(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\overline{A}\widehat{G}(\mathbf{x})\widehat{M}(\mathbf{x})\right\| \\
&\overset{(a)}{\leq} \left\|H^{-1}(\mathbf{x})\overline{A}G(\mathbf{x})M(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\overline{A}G(\mathbf{x})M(\mathbf{x})\right\| \\
&+ \left\|\widehat{H}^{-1}(\mathbf{x})\overline{A}G(\mathbf{x})M(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\overline{A}\widehat{G}(\mathbf{x})\widehat{M}(\mathbf{x})\right\| \\
&\leq \left\|H^{-1}(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \|G(\mathbf{x})\| \|M(\mathbf{x})\| \\
&+ \left\|\widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \left\|G(\mathbf{x})M(\mathbf{x}) - \widehat{G}(\mathbf{x})\widehat{M}(\mathbf{x})\right\| \\
&\overset{(b)}{\leq} \left\|H^{-1}(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \|G(\mathbf{x})\| \|M(\mathbf{x})\| \\
&+ \left\|\widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \left[\left\|G(\mathbf{x})M(\mathbf{x}) - G(\mathbf{x})\widehat{M}(\mathbf{x})\right\| + \left\|G(\mathbf{x})\widehat{M}(\mathbf{x}) - \widehat{G}(\mathbf{x})\widehat{M}(\mathbf{x})\right\|\right] \\
&\leq \left\|H^{-1}(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \|G(\mathbf{x})\| \|M(\mathbf{x})\| \\
&+ \left\|\widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \|G(\mathbf{x})\| \left\|M(\mathbf{x}) - \widehat{M}(\mathbf{x})\right\| + \left\|\widehat{H}^{-1}(\mathbf{x})\right\| \|\overline{A}\| \left\|G(\mathbf{x}) - \widehat{G}(\mathbf{x})\right\| \left\|\widehat{M}(\mathbf{x})\right\| \\
&\overset{(c)}{\leq} \overline{L}_A^2 L_A^2 \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}}\delta L_A \frac{1}{\mu_g}\overline{L}_{g_{xy}} + \overline{L}_A L_A \frac{1}{\mu_g} L_{g_{xy}}\delta + \overline{L}_A L_A \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}}\delta\overline{L}_{g_{xy}} \\
&= \left(\left(\frac{1}{\mu_g}\right)^3 \overline{L}_A^2 L_A^3 L_{g_{yy}}\overline{L}_{g_{xy}} + \frac{1}{\mu_g}\overline{L}_A L_A L_{g_{xy}} + \left(\frac{1}{\mu_g}\right)^2 \overline{L}_A L_A L_{g_{yy}}\overline{L}_{g_{xy}}\right)\delta.
\end{aligned}
$$

In (a) we add and subtract the term $\widehat{H}^{-1}(\mathbf{x})\overline{A}G(\mathbf{x})M(\mathbf{x})$ and apply the triangle inequality. In (b) we add and subtract the term $G(\mathbf{x})\widehat{M}(\mathbf{x})$ and apply the triangle inequality. In (c) we use Lemma 5(d) for $\|H^{-1}(\mathbf{x}) - \widehat{H}^{-1}(\mathbf{x})\|$, the bound $\|\overline{A}\| \leq L_A$, Lemma 5(a) for $\|G(\mathbf{x})\|$, Lemma 5(c) for $\|H^{-1}(\mathbf{x})\|$ and $\|\widehat{H}^{-1}(\mathbf{x})\|$, Assumption 2(f) for $\|M(\mathbf{x})\|$ and $\|\widehat{M}(\mathbf{x})\|$, Assumption 2(e) for $\|M(\mathbf{x}) - \widehat{M}(\mathbf{x})\|$, and finally Lemma 5(b) for $\|G(\mathbf{x}) - \widehat{G}(\mathbf{x})\|$.

The proof is now complete. □

Lemma 8. *Suppose that Assumptions 1, 2 and 3 hold. Then, the following bound holds*

$$\|\nabla\mathbf{y}^*(\mathbf{x}) - \widehat{\nabla}\mathbf{y}(\mathbf{x})\| \leq L_{\mathbf{y}^*}\delta,$$

where $L_{\mathbf{y}^*} = \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}}\overline{L}_{g_{xy}} + \frac{1}{\mu_g}L_{g_{xy}} + \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}}L_A\overline{L}_{\boldsymbol{\lambda}^*} + \frac{1}{\mu_g}L_A L_{\boldsymbol{\lambda}^*}$.

Proof. From Lemma 2 we have that

$$\nabla\mathbf{y}^*(\mathbf{x}) = \left[\nabla^2_{yy}g(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\right]^{-1}\left[-\nabla^2_{xy}g(\mathbf{x},\mathbf{y}^*(\mathbf{x})) - \overline{A}^T\nabla\overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right].$$

We can also get $\widehat{\nabla}\mathbf{y}(\mathbf{x})$ by substituting $\widehat{\mathbf{y}}(\mathbf{x})$ in place of $\mathbf{y}^*(\mathbf{x})$ in the above formula, i.e.,

$$\widehat{\nabla}\mathbf{y}(\mathbf{x}) = \left[\nabla^2_{yy}g(\mathbf{x},\widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1}\left[-\nabla^2_{xy}g(\mathbf{x},\widehat{\mathbf{y}}(\mathbf{x})) - \overline{A}^T\widehat{\nabla}\overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right].$$

Then, we have that

$$
\begin{aligned}
\left\|\nabla \mathbf{y}^*(\mathbf{x}) - \widehat{\nabla}\mathbf{y}(\mathbf{x})\right\| &= \Big\| \left[\nabla_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \left[-\nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] \\
&\quad - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \left[-\nabla_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - \overline{A}^T \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x})\right] \Big\| \\
&\overset{(a)}{\leq} \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \\
&\overset{(b)}{\leq} \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \left[\nabla_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \\
&\leq \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \left\| \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \left\| \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_{xy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right]^{-1} - \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \left\| \overline{A}^T \right\| \left\| \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \\
&+ \left\| \left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \right\| \left\| \overline{A}^T \right\| \left\| \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) - \widehat{\nabla} \overline{\boldsymbol{\lambda}}^*(\mathbf{x}) \right\| \\
&\overset{(c)}{\leq} \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \delta \overline{L}_{g_{xy}} + \frac{1}{\mu_g} L_{g_{xy}} \delta + \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \delta L_A \overline{L}_{\boldsymbol{\lambda}^*} + \frac{1}{\mu_g} L_A L_{\boldsymbol{\lambda}^*} \delta \\
&= \left(\left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} \overline{L}_{g_{xy}} + \frac{1}{\mu_g} L_{g_{xy}} + \left(\frac{1}{\mu_g}\right)^2 L_{g_{yy}} L_A \overline{L}_{\boldsymbol{\lambda}^*} + \frac{1}{\mu_g} L_A L_{\boldsymbol{\lambda}^*} \right) \delta.
\end{aligned}
$$

In (a) the triangle inequality was used. In (b) we add and subtract the expressions $\left[\nabla_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \nabla_{xy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ and $\left[\nabla^2_{yy} g(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right]^{-1} \overline{A}^T \nabla \overline{\boldsymbol{\lambda}}^*(\mathbf{x})$ in the first and second terms, respectively. In (c) we apply Lemma 5(a), 5(b), Assumption 2(e), 2(f), the bound $\|\overline{A}\| \leq L_A$, Lemmas 6 and 7.

The proof is now complete. □

Now we have all the results needed to prove Lemma 3.

Proof of Lemma 3. To begin with, the exact and approximate (due to the inexact solution of the LL problem) implicit gradients of the objective $F(\mathbf{x})$, are given below.

$$
\begin{aligned}
\nabla F(\mathbf{x}) &= \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + \left[\nabla \mathbf{y}^*(\mathbf{x})\right]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \\
\widehat{\nabla} F(\mathbf{x}) &= \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) + \left[\widehat{\nabla}\mathbf{y}(\mathbf{x})\right]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})).
\end{aligned}
$$

Then, we can compute the norm of their difference.

$$
\begin{aligned}
\|\widehat{\nabla} F(\mathbf{x}) - \nabla F(\mathbf{x})\| &= \|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) + [\nabla \widehat{\mathbf{y}}(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) \\
&\qquad - \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\overset{(a)}{\leq} \|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\qquad + \|[\widehat{\nabla} \mathbf{y}(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\overset{(b)}{\leq} \|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\qquad + \|[\widehat{\nabla} \mathbf{y}(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| \\
&\qquad + \| [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\overset{(c)}{\leq} L_f \|\widehat{\mathbf{y}}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}))\| + \|\widehat{\nabla} \mathbf{y}(\mathbf{x}) - \nabla \mathbf{y}^*(\mathbf{x}))\| \|\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| \\
&\qquad + \|\nabla \mathbf{y}^*(\mathbf{x}))\| \|\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) - \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\overset{(d)}{\leq} L_f \|\widehat{\mathbf{y}}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}))\| + \|\widehat{\nabla} \mathbf{y}(\mathbf{x}) - \nabla \mathbf{y}^*(\mathbf{x}))\| \|\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| \\
&\qquad + L_f \|\nabla \mathbf{y}^*(\mathbf{x}))\| \|\widehat{\mathbf{y}}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\| \\
&\overset{(e)}{\leq} L_f \delta + L_{\mathbf{y}^*} \delta \overline{L}_f + L_f \overline{L}_{\mathbf{y}^*} \delta \\
&= \left(L_f + L_{\mathbf{y}^*} \overline{L}_f + L_f \overline{L}_{\mathbf{y}^*}\right) \delta := L_F \delta, \qquad (26)
\end{aligned}
$$

where $L_F = L_f + L_{\mathbf{y}^*} \overline{L}_f + L_f \overline{L}_{\mathbf{y}^*}$. Also, in inequality (a) above we apply the triangle inequality; in (b) we add and subtract the term $[\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))$, and use triangle inequality; in (c) and (d) we use the Lipschitz gradient property of f (Assumption 3(b)); in (e) we apply Assumptions 2(a) and 3(a), and Lemmas 6 and 8.

Now consider the expression $\|\nabla F(\mathbf{x})\|$. We have that

$$
\begin{aligned}
\|\nabla F(\mathbf{x})\| &= \left\|\nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\right\| \\
&\leq |\nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| + \|\nabla \mathbf{y}^*(\mathbf{x})\| \, \|\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \\
&\leq \left(1 + \overline{L}_{\mathbf{y}^*}\right) \overline{L}_f := \overline{L}_F,
\end{aligned}
$$

where we applied Assumption 2(a) and Lemma 6. Similarly, we can see that

$$
\begin{aligned}
\left\|\widehat{\nabla} F(\mathbf{x})\right\| &= \left\|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) + \left[\widehat{\nabla} \mathbf{y}(\mathbf{x})\right]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\right\| \\
&\leq \|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| + \left\|\widehat{\nabla} \mathbf{y}(\mathbf{x})\right| \|\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\| \\
&\leq \left(1 + \overline{L}_{\mathbf{y}^*}\right) \overline{L}_f = \overline{L}_F.
\end{aligned}
$$

Therefore, the proof is completed. □

F Proofs of Section 4

F.1 Asymptotic Convergence Guarantees with the Line Search Method: Proof of Theorem 1

Lemma 9. *Under Assumption 1, 2 and 3, $\nabla F(\mathbf{x})$ is almost surely continuous at every $\mathbf{x} \in \mathcal{X}$.*

Proof. To begin with, we already established that F is almost surely differentiable at every $\mathbf{x} \in \mathcal{X}$. Therefore, for any $\mathbf{x} \in \mathcal{X}$ there exists (almost surely) a neighborhood around it such that the matrix $\overline{A}$ corresponding to the active constraints at $\mathbf{y}^*(\mathbf{x})$ remains unchanged, while the gradient of $\mathbf{y}^*(\mathbf{x})$ is defined in eq. (6), (7). Further, since $\overline{A}$ is locally (i.e., around any given $\mathbf{x}$) constant, and the formulas in (6), (7) can be seen as the results of a number of continuous operations over continuous functions, it is implied that $\nabla \mathbf{y}^*(\mathbf{x})$ is also a continuous function at $\mathbf{x}$ almost surely. As a

result, $\nabla F(\mathbf{x}) = \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) + [\nabla \mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is almost surely continuous at every $\mathbf{x} \in \mathcal{X}$. □

Proof of theorem1. Here we follow a reasoning similar to the proof of (Bertsekas, 1998, Prop. 1.2.1). However, there are a number of differences that make this proof more challenging. First, in our setting we are optimizing an inexact version of the objective $\widehat{F}(\mathbf{x}) = f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))$, using an approximate version of the gradient $\widehat{\nabla} F(\mathbf{x})$. Since the approximate gradient we are using is not the gradient of the objective $\widehat{F}(\mathbf{x})$ (the gradient of this function might not even exist), we consider a modification of the standard Armijo rule where an additional error term is present. Secondly, the statements below that involve the gradient of $F(\mathbf{x})$ (as well as the statements that result from them) are not deterministic, as in (Bertsekas, 1998, Prop. 1.2.1), but rather they hold in an almost sure sense; this stems from the fact that $F(\mathbf{x})$ is almost surely differentiable at any given point $\mathbf{x} \in \mathcal{X}$. Finally, in the proof below the (classical) mean value theorem (Bertsekas, 1998, Prop. 1.23) cannot be applied, since we cannot ensure that F is (surely) differentiable at any given interval over $\mathbf{x}$. As we are going to show below, we use an alternative mean value theorem that does not require such assumption.

To begin with, we know that for the exact implicit objective $F(\mathbf{x})$ and gradient $\nabla F(\mathbf{x})$ (quantities to which we do not have access to) we can find at each iteration r a step-size a^r such that the following condition holds

$$F(\mathbf{x}^r) - F(\mathbf{x}^r + a^r \mathbf{d}^r) \geq -\sigma a^r \left[\nabla F(\mathbf{x}^r)\right]^T \mathbf{d}^r, \tag{27}$$

where $\mathbf{d}^r = \widetilde{\mathbf{x}}^r - \mathbf{x}^r$ with $\widetilde{\mathbf{x}}^r = \text{proj}_{\mathcal{X}}(\mathbf{x}^r - \nabla F(\mathbf{x}^r))$.

Next, the difference between the (approximate) objective values of two successive iterates (for simplicity we will use the notation $\mathbf{x}^{r+1} = \mathbf{x}^r + a^r \mathbf{d}^r$, $\widehat{\mathbf{x}}^{r+1} = \mathbf{x}^r + a^r \widehat{\mathbf{d}}^r$; $\widehat{\mathbf{d}}^r$ is defined in Algorithm 1) is

$$\begin{aligned}
\widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) &= \widehat{F}(\mathbf{x}^r) - F(\mathbf{x}^r) + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + F(\mathbf{x}^{r+1}) - F(\widehat{\mathbf{x}}^{r+1}) + F(\widehat{\mathbf{x}}^{r+1}) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) \\
&= f(\mathbf{x}^r, \widehat{\mathbf{y}}(\mathbf{x}^r)) - f(\mathbf{x}^r, \mathbf{y}^*(\mathbf{x}^r)) + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + F(\mathbf{x}^{r+1}) - F(\widehat{\mathbf{x}}^{r+1}) \\
&\qquad + f(\widehat{\mathbf{x}}^{r+1}, \mathbf{y}^*(\widehat{\mathbf{x}}^{r+1})) - f(\widehat{\mathbf{x}}^{r+1}, \widehat{\mathbf{y}}(\widehat{\mathbf{x}}^{r+1})) \\
&\geq -L_f \|\mathbf{y}^*(\mathbf{x}^r) - \widehat{\mathbf{y}}(\mathbf{x}^r)\| + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) - \overline{L}_F \|\mathbf{x}^{r+1} - \widehat{\mathbf{x}}^{r+1}\| \\
&\qquad - L_f \|\mathbf{y}^*(\mathbf{x}^{r+1}) - \widehat{\mathbf{y}}(\mathbf{x}^{r+1})\| \\
&\geq -L_f \delta^r - \sigma a^r \left[\nabla F(\mathbf{x}^r)\right]^T \mathbf{d}^r - \overline{L}_F a^r \|\mathbf{d}^r - \widehat{\mathbf{d}}^r\| - L_f \delta^{r+1} \\
&\geq -L_f \delta^r - \sigma a^r \left[\nabla F(\mathbf{x}^r)\right]^T \mathbf{d}^r - \overline{L}_F L_F a^r \delta^r - L_f \delta^{r+1} \\
&= -\sigma a^r \left[\nabla F(\mathbf{x}^r)\right]^T \mathbf{d}^r - \epsilon_1(\delta; r),
\end{aligned} \tag{28}$$

where we set $\epsilon_1(\delta; r) = L_f \delta^r + \overline{L}_F L_F a^r \delta^r + L_f \delta^{r+1}$. In the first inequality, we used the Lipschitz continuity of f and F; in the second inequality Assumption 3(a) and condition (27) were applied; in the third inequality the non-expansive property of the projection operator was used.

Also, we have that

$$\begin{aligned}
\nabla^T F(\mathbf{x}^r) \mathbf{d}^r &= \left(\nabla F(\mathbf{x}^r) - \widehat{\nabla} F(\mathbf{x}^r) + \widehat{\nabla} F(\mathbf{x}^r)\right)^T \left(\mathbf{d}^r + \widehat{\mathbf{d}}^r - \widehat{\mathbf{d}}^r\right) \\
&= \left(\nabla F(\mathbf{x}^r) - \widehat{\nabla} F(\mathbf{x}^r)\right)^T \left(\mathbf{d}^r - \widehat{\mathbf{d}}^r\right) + \left(\nabla F(\mathbf{x}^r) - \widehat{\nabla} F(\mathbf{x}^r)\right)^T \widehat{\mathbf{d}}^r \\
&+ \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \left(\mathbf{d}^r - \widehat{\mathbf{d}}^r\right) + \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r \\
&\leq \|\nabla F(\mathbf{x}^r) - \widehat{\nabla} F(\mathbf{x}^r)\| \|\mathbf{d}^r - \widehat{\mathbf{d}}^r\| + \|\nabla F(\mathbf{x}^r) - \widehat{\nabla} F(\mathbf{x}^r)\| \|\widehat{\mathbf{d}}^r\| \\
&+ \|\widehat{\nabla} F(\mathbf{x}^r)\| \|\mathbf{d}^r - \widehat{\mathbf{d}}^r\| + \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r \\
&\leq L_F^2 \left(\delta^r\right)^2 + L_F \overline{L}_F \delta^r + L_F \overline{L}_F \delta^r + \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r \\
&= \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r + \epsilon_2(\delta; r),
\end{aligned} \tag{29}$$

where $\epsilon_2(\delta; r) = L_F^2 \left(\delta^r\right)^2 + 2L_F\overline{L}_F\delta^r$. Notice that the results in the second inequality follow from Lemma 3.

Then, combining (28) and (29) we get

$$\begin{aligned}\widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) &\geq -\sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r - \epsilon_1(\delta; r) - \sigma a^r \epsilon_2(\delta; r) \\ &= -\sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r - \epsilon(\delta; r),\end{aligned}$$

where $\epsilon(\delta; r) = \epsilon_1(\delta; r) + \sigma a^r \epsilon_2(\delta; r)$; notice that $\lim_{\delta \to 0} \epsilon(\delta) = 0$. In conclusion, we can follow this (inexact) Armijo-type rule in our inexact problem; the existence of the (Armijo) step-size is guaranteed by its existence for the exact problem (27).

Now let us move to the main part of the proof, which follows the reasoning used in (Bertsekas, 1998, Prop. 1.2.1). Let $\{\mathbf{x}^r\} \in \mathcal{X}$ be the iterate sequence of our algorithm, and let $\bar{\mathbf{x}} \in \mathcal{X}$ be a limit point; the existence of such point is guaranteed by the closedness of the set $\mathcal{X}$. Moreover, it is established in Proposition 1 that $F(\mathbf{x})$ is continuous, and as a result it holds that $\lim_{r \to +\infty} F(\mathbf{x}^r) = F(\bar{\mathbf{x}})$. The latter results combined with the fact that all convergent sequences are also Cauchy sequences, implies that $\lim_{r \to +\infty}(F(\mathbf{x}^r) - F(\mathbf{x}^{r+1})) = 0$.

We want to show that $\bar{\mathbf{x}}$ is a stationary point of $F(\mathbf{x})$. We are going to show that by assuming that the opposite holds, i.e., $\bar{\mathbf{x}}$ is not a stationary point of $F(\mathbf{x})$, and arriving at a contradiction. From the Armijo rule of our problem we have that

$$\widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) \geq -\sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r - \epsilon(\delta; r). \tag{30}$$

Then, consider the following

$$\begin{aligned}\widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) &= \widehat{F}(\mathbf{x}^r) - F(\mathbf{x}^r) + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + F(\mathbf{x}^{r+1}) - F(\widehat{\mathbf{x}}^{r+1}) + F(\widehat{\mathbf{x}}^{r+1}) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) \\ &= f(\mathbf{x}^r, \widehat{\mathbf{y}}(\mathbf{x}^r)) - f(\mathbf{x}^r, \mathbf{y}^*(\mathbf{x}^r)) + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + F(\mathbf{x}^{r+1}) - F(\widehat{\mathbf{x}}^{r+1}) \\ &+ f(\widehat{\mathbf{x}}^{r+1}, \mathbf{y}^*(\widehat{\mathbf{x}}^{r+1})) - f(\widehat{\mathbf{x}}^{r+1}, \widehat{\mathbf{y}}(\widehat{\mathbf{x}}^{r+1})) \\ &\leq L_f \|\mathbf{y}^*(\mathbf{x}^r) - \widehat{\mathbf{y}}(\mathbf{x}^r)\| + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + \overline{L}_F \|\mathbf{x}^{r+1} - \widehat{\mathbf{x}}^{r+1}\| \\ &\qquad + L_f \|\mathbf{y}^*(\mathbf{x}^{r+1}) - \widehat{\mathbf{y}}(\mathbf{x}^{r+1})\| \\ &\leq L_f \delta^r + F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + \overline{L}_F a^r \|\mathbf{d}^r - \widehat{\mathbf{d}}^r\| + L_f \delta^{r+1} \\ &\leq F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + L_f \delta^r + \overline{L}_F L_F a^r \delta^r + L_f \delta^{r+1} \\ &= F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + \epsilon_1(\delta; r).\end{aligned}$$

In the first inequality, we used the Lipschitz continuity of f and F; in the second inequality Assumption 3(a) and condition (27) were applied; in the third inequality the non-expansive property of the projection operator was used. Using the above derivation we can bound the left-hand side of inequality (30) as follows

$$F(\mathbf{x}^r) - F(\mathbf{x}^{r+1}) + \epsilon_1(\delta; r) \geq \widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) \geq -\sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r - \epsilon(\delta; r).$$

It is easy to see that the left-hand side in the above inequality tends to 0. Therefore, $\lim_{r \to +\infty} \left(-\sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r - \epsilon(\delta; r)\right) \leq 0$. In addition, we know that $\lim_{r \to +\infty} \epsilon(\delta; r) = 0$ and $-\sigma a^r \left[\widehat{\nabla} F(\mathbf{x})\right]^T \widehat{\mathbf{d}}^r \geq 0, \forall \mathbf{x} \in \mathcal{X}$. From the above statements we can conclude that

$$\lim_{r \to +\infty} \sigma a^r \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r = 0. \tag{31}$$

Moreover, from the gradient-related assumption we know that for a non-stationary point $\bar{\mathbf{x}}$ we have that

$$\limsup_{r \to \infty, r \in \mathcal{R}} \left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{d}}^r < 0, \tag{32}$$

where$\{\mathbf{x}^r\}_{\mathcal{R}}$ is subsequence with $\lim_{r\to\infty, r\in\mathcal{R}} \mathbf{x}^r = \bar{\mathbf{x}}$. Then, the conditions (31), (32) imply that

$$\lim_{r\to\infty, r\in\mathcal{R}} a^r = 0.$$

In the subsequence $\mathcal{R}$ we can find an index $\bar{r} \geq 0$ such that

$$\widehat{F}(\mathbf{x}^r) - \widehat{F}\left(\mathbf{x}^r + \left(\frac{a^r}{\beta}\right)\widehat{\mathbf{d}}^r\right) < -\sigma\left(\frac{a^r}{\beta}\right)\widehat{\nabla}^T F(\mathbf{x}^r)\widehat{\mathbf{d}}^r - \epsilon(\delta; r), \forall r \in \mathcal{R}, r \geq \bar{r}. \tag{33}$$

Similarly with the proof of (Bertsekas, 1998, Prop. 1.2.1) let us introduce the following sequences:

$$\widehat{\mathbf{p}}^r = \frac{\widehat{\mathbf{d}}^r}{\|\widehat{\mathbf{d}}^r\|}, \bar{a}^r = \frac{a^r\|\widehat{\mathbf{d}}^r\|}{\beta}$$

The first sequence $\{\widehat{\mathbf{p}}^r\}$ is bounded and so it admits a limit point $\bar{\mathbf{p}}$ with $\|\bar{\mathbf{p}}\| = 1$, that is $\lim_{r\to+\infty, r\in\overline{\mathcal{R}}} \mathbf{p}^r = \bar{\mathbf{p}}$, where $\overline{\mathcal{R}}$ denotes the indices of a subsequence of $\mathcal{R}$. In addition, taking into account the facts that $\lim_{r\to+\infty, r\in\mathcal{R}} a^r = 0$ and the fact that the sequence $\{\|\mathbf{d}^r\|\}_{\mathcal{R}}$ is bounded we can easily see that $\lim_{r\to+\infty, r\in\mathcal{R}} \bar{a}^r = 0$.

Dividing both sides of (33) by $\bar{a}^r$ and using the definitions of $\widehat{\mathbf{p}}^r$ and $\bar{a}_r$ from above we get

$$\frac{\widehat{F}(\mathbf{x}^r) - \widehat{F}\left(\mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r\right)}{\bar{a}^r} < -\sigma\left[\widehat{\nabla}F(\mathbf{x}^r)\right]^T \widehat{\mathbf{p}}^r - \epsilon(\delta; r), \forall r \in \overline{\mathcal{R}}, r > \bar{r}. \tag{34}$$

Then, we have that (for convenience we adopt the notation $\widehat{\mathbf{x}}^{r+1} = \mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r$)

$$\begin{aligned}
\widehat{F}(\mathbf{x}^r) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) &= \widehat{F}(\mathbf{x}^r) - F(\mathbf{x}^r) + F(\mathbf{x}^r) - F(\widehat{\mathbf{x}}^{r+1}) + F(\widehat{\mathbf{x}}^{r+1}) - \widehat{F}(\widehat{\mathbf{x}}^{r+1}) \\
&= f(\mathbf{x}^r, \widehat{\mathbf{y}}(\mathbf{x}^r)) - f(\mathbf{x}^r, \mathbf{y}^*(\mathbf{x}^r)) + F(\mathbf{x}^r) - F(\widehat{\mathbf{x}}^{r+1}) \\
&\qquad + f(\widehat{\mathbf{x}}^{r+1}, \mathbf{y}^*(\widehat{\mathbf{x}}^{r+1})) - f(\widehat{\mathbf{x}}^{r+1}, \widehat{\mathbf{y}}(\widehat{\mathbf{x}}^{r+1})) \\
&\geq -L_f\|\mathbf{y}^*(\mathbf{x}^r) - \widehat{\mathbf{y}}(\mathbf{x}^r)\| - L_f\|\mathbf{y}^*(\widehat{\mathbf{x}}^{r+1}) - \widehat{\mathbf{y}}(\widehat{\mathbf{x}}^{r+1})\| + F(\mathbf{x}^r) - F(\widehat{\mathbf{x}}^{r+1}) \\
&\geq F(\mathbf{x}^r) - F(\widehat{\mathbf{x}}^{r+1}) - L_f\delta^r - L_f\delta^{r+1},
\end{aligned} \tag{35}$$

where the first inequality above follows the Lipschitz continuity of f, and the second inequality is an application of Assumption 3(a). Incorporating inequality (35) into (34) results to

$$\frac{F(\mathbf{x}^r) - F(\mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r)}{\bar{a}_r} - L_f\frac{\delta^r + \delta^{r+1}}{\bar{a}_r} < -\sigma\left[\widehat{\nabla}F(\mathbf{x}^r)\right]^T \widehat{\mathbf{p}}^r - \epsilon(\delta; r), \forall r \in \overline{\mathcal{R}}, r > \bar{r}. \tag{36}$$

Lebourg's mean value theorem (Lebourg, 1979, Theorem 1.7) implies that

$$\frac{F(\mathbf{x}^r) - F(\mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r)}{\bar{a}_r} = \mathbf{u}^{\mathbf{T}}\widehat{\mathbf{p}}^{\mathbf{r}}$$

with $\mathbf{u} \in \vartheta\mathbf{F}(\mathbf{x}^{\mathbf{r}} + \widetilde{\mathbf{a}}^{\mathbf{r}}\widehat{\mathbf{p}}^{\mathbf{r}})$ and $\widetilde{a}^r \in [0, \bar{a}^r]$, where $\vartheta F(\cdot)$ is the Clarke subdifferntial of F. We know that F is almost surely continuously differentiable (Lemma 9) at every $\mathbf{x} \in \mathcal{X}$, and so the Clarke subdifferential at $\mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r$ becomes with probability 1 equal to $\nabla F(\mathbf{x}^r + \widetilde{a}^r\widehat{\mathbf{p}}^r)$. Note that the we cannot use here the (classical) mean value theorem (Bertsekas, 1998, Prop. 1.23), as in the proof of (Bertsekas, 1998, Prop. 1.2.1), because it requires that the function $F(\mathbf{x})$ is (surely) differentiable on the interval $[\mathbf{x}^r, \mathbf{x}^r + \bar{a}^r\widehat{\mathbf{p}}^r]$.

Then, we can rewrite the expression in (36) as follows

$$-L_f\beta\frac{\delta^r + \delta^{r+1}}{a_r\|\widehat{\mathbf{d}}^r\|} - \left[\nabla F(\mathbf{x}^r + \widetilde{a}^r\widehat{\mathbf{p}}^r)\right]^T \widehat{\mathbf{p}}^r < -\sigma\left[\widehat{\nabla}F(\mathbf{x}^r)\right]^T \widehat{\mathbf{p}}^r - \epsilon(\delta; r), \forall r \in \overline{\mathcal{R}}, r > \bar{r},$$

where $\widetilde{a}^r \in [0, \bar{a}^r]$.

Using the assumption that $0 \leq \frac{\delta^r}{a^r} \sim \mathcal{O}(c^r)$, where c^r is some sequence with $\lim_{r\to\infty, r\in\overline{\mathcal{R}}} c^r = 0$, and the fact that $\lim_{r\to\infty, r\in\overline{\mathcal{R}}} \|\widehat{\mathbf{d}}^r\| \neq 0$ (because of the assumption that the sequence $\mathbf{x}^r$ converges to a non-stationary point), we compute the limit in the above expression and get

$$\begin{aligned} -\left[\nabla F(\bar{\mathbf{x}})\right]^T \bar{\mathbf{p}} &< -\sigma \left[\nabla F(\bar{\mathbf{x}})\right]^T \bar{\mathbf{p}} \\ 0 &< (1-\sigma) \left[\nabla F(\bar{\mathbf{x}})\right]^T \bar{\mathbf{p}} \\ 0 &< \left[\nabla F(\bar{\mathbf{x}})\right]^T \bar{\mathbf{p}}. \end{aligned} \tag{37}$$

However, note that $\left[\widehat{\nabla} F(\mathbf{x}^r)\right]^T \widehat{\mathbf{p}}^r = \frac{\widehat{\nabla}^T F(\mathbf{x}^r)\widehat{\mathbf{d}}^r}{\|\widehat{\mathbf{d}}^r\|}$ and therefore if we take limits in both sides we obtain

$$\left[\nabla F(\bar{\mathbf{x}})\right]^T \bar{\mathbf{p}}^r \leq \frac{\limsup_{r\to\infty, r\in\overline{\mathcal{R}}} \widehat{\nabla}^T F(\mathbf{x}^r)\widehat{\mathbf{d}}^r}{\limsup_{r\to\infty, r\in\overline{\mathcal{R}}} \|\widehat{\mathbf{d}}^r\|} < 0, \tag{38}$$

due to the gradient-related assumption. We notice that expressions (37) and (38) lead to a contradiction. Therefore, $\bar{\mathbf{x}}$ is a stationary point of $F(\mathbf{x})$.

To conclude, note that in the above derivations we require the differentiability of $F(\mathbf{x})$ at a countably infinite number of points (because we require the existence of $\nabla F(\mathbf{x})$ at the sequence of the algorithm's iterates). This holds with probability 1 (i.e., almost surely), since an infinite but countable intersection of almost sure events is an almost sure event and the probability of any given point $\mathbf{x}$ to be differentiable is 1. As a result, the final conclusion, i.e., the limit point of the iterate sequence $\mathbf{x}^r$ is a stationary point, holds almost surely (i.e., with probability 1).

The proof is now complete. □

F.2 Proof of Lemma 4

From the definition of the stochastic gradient in (10) we have

$$\widehat{\nabla} F(\mathbf{x};\xi) = \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi).$$

Taking expectation on both sides and utilizing Assumption 4, we get

$$\begin{aligned} \mathbb{E}_\xi[\widehat{\nabla} F(\mathbf{x};\xi)] &= \mathbb{E}_\xi\big[\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi)\big] \\ &= \mathbb{E}_\xi\big[\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi)\big] + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \mathbb{E}_\xi\big[\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi)\big] \\ &= \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) \\ &= \widehat{\nabla} F(\mathbf{x}). \end{aligned}$$

Similarly, for the variance of the stochastic implicit gradient, we have

$$\begin{aligned} \mathbb{E}_\xi\|\widehat{\nabla} F(\mathbf{x};\xi) - \widehat{\nabla} F(\mathbf{x})\|^2 &= \mathbb{E}_\xi\big\|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) \\ &\qquad - \big[\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x})) + [\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]^T \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\big]\big\|^2 \\ &\overset{(a)}{\leq} 2\,\mathbb{E}_\xi\|\nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) - \nabla_x f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\|^2 \\ &\qquad + 2\,\|\widehat{\nabla}\mathbf{y}^*(\mathbf{x})]\|^2\,\mathbb{E}_\xi\|\nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x});\xi) - \nabla_y f(\mathbf{x}, \widehat{\mathbf{y}}(\mathbf{x}))\|^2 \\ &\overset{(b)}{\leq} 2\sigma_f^2 + 2\overline{L}_{\mathbf{y}^*}\sigma_f^2 := \sigma_F^2, \end{aligned}$$

where (a) follows from $\|\mathbf{x}+\mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ and (b) results from Assumption 4 and the application of Lemma 6.

Therefore, we have the proof.

F.3 WEAKLY-CONVEX OBJECTIVE: PROOF OF THEOREM 2

Proof. Define $\hat{\mathbf{x}}^r = \underset{\mathbf{z}\in\mathbb{R}^{d_u}}{\arg\min}\Big\{H(\mathbf{z}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \mathbf{z}\|^2\Big\}$. Using the definition of Moreau envelope, we have

$$
\begin{aligned}
\mathbb{E}[H_{1/\hat{\rho}}(\mathbf{x}^{r+1})] &\leq \mathbb{E}\Big[F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{r+1} - \hat{\mathbf{x}}^r\|^2\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\text{proj}_{\mathcal{X}}(\mathbf{x}^r - \beta\widehat{\nabla}F(\mathbf{x}^r;\xi^r)) - \text{proj}_{\mathcal{X}}(\hat{\mathbf{x}}^r)\|^2\Big] \\
&\overset{(b)}{\leq} \mathbb{E}\Big[F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \beta\widehat{\nabla}F(\mathbf{x}^r;\xi^r) - \hat{\mathbf{x}}^r\|^2\Big] \\
&= F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\mathbb{E}\Big[\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 - 2\langle \mathbf{x}^r - \hat{\mathbf{x}}^r, \beta\widehat{\nabla}F(\mathbf{x}^r;\xi^r)\rangle + \beta^2\|\widehat{\nabla}F(\mathbf{x}^r;\xi^r)\|^2\Big] \\
&\overset{(c)}{\leq} F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\mathbb{E}\Big[\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 - 2\langle \mathbf{x}^r - \hat{\mathbf{x}}^r, \beta\widehat{\nabla}F(\mathbf{x}^r;\xi^r)\rangle \\
&\qquad\qquad + 2\beta^2\|\widehat{\nabla}F(\mathbf{x}^r;\xi^r) - \widehat{\nabla}F(\mathbf{x}^r)\|^2 + 2\beta^2\|\widehat{\nabla}F(\mathbf{x}^r)\|^2\Big] \\
&\overset{(d)}{\leq} F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\Big[\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 - 2\beta\langle \mathbf{x}^r - \hat{\mathbf{x}}^r, \widehat{\nabla}F(\mathbf{x}^r)\rangle + 2\beta^2\big(\sigma_F^2 + \overline{L}_F^2\big)\Big] \\
&\leq F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\Big[\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 + 2\beta\underbrace{\langle \hat{\mathbf{x}}^r - \mathbf{x}^r, \nabla F(\mathbf{x}^r)\rangle}_{\text{Term I}} \\
&\qquad\qquad + 2\beta\underbrace{\langle \hat{\mathbf{x}}^r - \mathbf{x}^r, \widehat{\nabla}F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\rangle}_{\text{Term II}} + 2\beta^2\big(\sigma_F^2 + \overline{L}_F^2\big)\Big],
\end{aligned}
\tag{39}
$$

where (a) follows from the fact that $\mathbf{x}^{r+1} \in \mathcal{X}$ and $\hat{\mathbf{x}}^r \in \mathcal{X}$; (b) results from the non-expansiveness of the projection operator; (c) uses $\|\mathbf{a} - \mathbf{b}\|^2 = 2\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$; and (d) results from the application of Lemmas 3 and 4.

Next, considering Term I and Term II separately in (39) above. For Term I, we get using the weak convexity of $F(\cdot)$

$$
\text{Term I} = \langle \hat{\mathbf{x}}^r - \mathbf{x}^r, \nabla F(\mathbf{x}^r)\rangle \leq F(\hat{\mathbf{x}}^r) - F(\mathbf{x}^r) + \frac{\rho}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2
$$

We bound Term II using the Young's inequality as

$$
\begin{aligned}
\text{Term II} &= \langle \hat{\mathbf{x}}^r - \mathbf{x}^r, \widehat{\nabla}F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\rangle \\
&\leq \frac{\rho}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 + \frac{1}{2\rho}\|\widehat{\nabla}F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\|^2 \\
&\overset{(e)}{\leq} \frac{\rho}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 + \frac{1}{2\rho}L_F^2\delta^2
\end{aligned}
$$

where (e) follows from Lemma 3. Next, substituting the bounds of Term I and Term II in (39) and using the definition of $\hat{\mathbf{x}}^r$, we get

$$
\begin{aligned}
\mathbb{E}[H_{1/\hat{\rho}}(\mathbf{x}^{r+1})] &\leq F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 + \hat{\rho}\beta\Big[F(\hat{\mathbf{x}}^r) - F(\mathbf{x}^r) + \rho\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2\Big] \\
&\qquad\qquad + \frac{\hat{\rho}\beta}{2\rho}L_F^2\delta^2 + \beta^2\hat{\rho}\big(\sigma_F^2 + \overline{L}_F^2\big) \\
&\leq H_{1/\hat{\rho}}(\mathbf{x}^r) + \hat{\rho}\beta\underbrace{\Big[F(\hat{\mathbf{x}}^r) - F(\mathbf{x}^r) + \rho\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2\Big]}_{\text{Term III}} + \frac{\hat{\rho}\beta}{2\rho}L_F^2\delta^2 + \beta^2\hat{\rho}\big(\sigma_F^2 + \overline{L}_F^2\big).
\end{aligned}
\tag{40}
$$

Next, we bound Term III in (40) above.

$$\begin{aligned}\text{Term III} &= F(\hat{\mathbf{x}}^r) - F(\mathbf{x}^r) + \rho\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 \\ &= F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 - F(\mathbf{x}^r) + \frac{2\rho - \hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 \\ &\leq \frac{3\rho - 2\hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 \leq \frac{3\rho - 2\hat{\rho}}{2\hat{\rho}^2}\|\nabla H_{1/\hat{\rho}}(\mathbf{x}^r)\|^2,\end{aligned}$$

where the last equality follows from (12) and the first inequality follows from the fact that $F(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \mathbf{x}\|^2$ is $(\hat{\rho} - \rho)$-strongly convex. This implies the following

$$\begin{aligned}F(\hat{\mathbf{x}}^r) + \frac{\hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 - F(\mathbf{x}^r) &\leq -\langle \nabla F(\hat{\mathbf{x}}^r) + \hat{\rho}(\mathbf{x}^r - \hat{\mathbf{x}}^r), x^r - \hat{x}^r\rangle - \frac{\hat{\rho} - \rho}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2 \\ &\leq \frac{\rho - \hat{\rho}}{2}\|\mathbf{x}^r - \hat{\mathbf{x}}^r\|^2,\end{aligned}$$

where the second inequality results from the definition of $\hat{\mathbf{x}}^r$.

Finally, substituting Term III in (40) and rearranging the terms we get:

$$\beta\left[\frac{2\hat{\rho} - 3\rho}{2\hat{\rho}}\right]\|\nabla H_{1/\hat{\rho}}(\mathbf{x}^r)\|^2 \leq \mathbb{E}\big[H_{1/\hat{\rho}}(\mathbf{x}^r) - H_{1/\hat{\rho}}(\mathbf{x}^{r+1})\big] + \beta^2\hat{\rho}\big(\sigma_F^2 + \overline{L}_F^2\big) + \frac{\hat{\rho}\beta}{2\rho}L_F^2\delta^2.$$

Summing over all $r \in \{0, 1, \ldots, T-1\}$ and dividing by T, we get

$$\frac{1}{T}\sum_{r=0}^{T-1}\|\nabla H_{1/\hat{\rho}}(\mathbf{x}^r)\|^2 \leq \left[\frac{2\hat{\rho}}{2\hat{\rho} - 3\rho}\right]\left[\frac{H_{1/\hat{\rho}}(\mathbf{x}^0) - H^*}{\beta T} + \beta\hat{\rho}\big(\sigma_F^2 + \overline{L}_F^2\big) + \frac{\hat{\rho}}{2\rho}L_F^2\delta^2\right].$$

Therefore, we have the result. □

F.4 Strongly-Convex Objective: Proof of Theorem 3

Proof. Using the update rule of the Algorithm 2, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 &= \mathbb{E}\|\text{proj}_X(\mathbf{x}^r - \beta^r\widehat{\nabla}F(\mathbf{x}^r;\xi^r)) - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\|\text{proj}_X(\mathbf{x}^r - \beta^r\widehat{\nabla}F(\mathbf{x}^r;\xi^r)) - \text{proj}_X(\mathbf{x}^*)\|^2 \\ &\overset{(a)}{\leq} \mathbb{E}\|\mathbf{x}^r - \beta^r\widehat{\nabla}F(\mathbf{x}^r;\xi^r) - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\Big[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + (\beta^r)^2\|\widehat{\nabla}F(\mathbf{x}^r;\xi^r)\|^2 - 2\beta^r\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla}F(\mathbf{x}^r;\xi^r)\rangle\Big] \\ &\overset{(b)}{\leq} \mathbb{E}\Big[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2(\beta^r)^2\|\widehat{\nabla}F(\mathbf{x}^r;\xi^r) - \widehat{\nabla}F(\mathbf{x}^r)\|^2 \\ &\qquad + 2(\beta^r)^2\|\widehat{\nabla}F(\mathbf{x}^r)\|^2 - 2\beta^r\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla}F(\mathbf{x}^r)\rangle\Big] \\ &\overset{(c)}{\leq} \mathbb{E}\Big[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2(\beta^r)^2(\sigma_F^2 + B_F^2) - 2\beta^r\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla}F(\mathbf{x}^r)\rangle\Big] \\ &\leq \mathbb{E}\Big[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2(\beta^r)^2(\sigma_F^2 + B_F^2) - 2\beta^r\langle \mathbf{x}^r - \mathbf{x}^*, \nabla F(\mathbf{x}^r)\rangle \\ &\qquad - 2\beta^r\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla}F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\rangle\Big] \\ &\overset{(d)}{\leq} \mathbb{E}\Big[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2(\beta^r)^2(\sigma_F^2 + B_F^2) - 2\beta^r[F(\mathbf{x}^r) - F^*] - \mu_F\beta^r\|\mathbf{x}^r - \mathbf{x}^*\|^2 \\ &\qquad + 2\beta^r\|\mathbf{x}^r - \mathbf{x}^*\|\|\widehat{\nabla}F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\|\Big] \\ &\overset{(e)}{\leq} \mathbb{E}\Big[(1 - \mu_F\beta^r)\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2(\beta^r)^2(\sigma_F^2 + B_F^2) - 2\beta^r[F(\mathbf{x}^r) - F^*] + 2\beta^r D_{\mathcal{X}}L_F\delta\Big]\end{aligned}$$

where in (a) the non-expansive property of the projection operator is applied, in (b) we add and subtract the term $\widehat{\nabla}F(\mathbf{x}^r)$, use the well-known inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, and in

the last term we utilize Lemma 4; (c) results from the application of Lemmas 3 and 4; (d) uses strong-convexity (Assumption 6) of $F(\cdot)$ and Cauchy-Schwartz inequality; finally, (e) results from the application of Lemma 3 and Assumption 1(b).

Rearranging the terms, we get

$$2\beta^r \mathbb{E}[F(\mathbf{x}^r) - F^*] \leq \mathbb{E}\Big[(1-\mu_F\beta^r)\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 + 2(\beta^r)^2(\sigma_F^2 + B_F^2) + 2\beta^r D_{\mathcal{X}} L_F \delta\Big]$$

$$\mathbb{E}[F(\mathbf{x}^r) - F^*] \leq \mathbb{E}\Bigg[\left(\frac{1}{2\beta^r} - \frac{\mu_F}{2}\right)\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \frac{1}{2\beta^r}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 + \beta^r(\sigma_F^2 + B_F^2) + D_{\mathcal{X}} L_F \delta\Bigg].$$

Summing over $r \in \{0, \ldots, T-1\}$, multiplying by $1/T$ and using the fact that $\beta^r = \frac{1}{\mu_F(r+1)}$, we get

$$\frac{1}{T}\sum_{r=0}^{T-1} \mathbb{E}[F(\mathbf{x}^r) - F^*] \leq \frac{\mu_F}{2T}\sum_{r=0}^{T-1} \mathbb{E}\Big[r\|\mathbf{x}^r - \mathbf{x}^*\|^2 - (r+1)\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2\Big] + \frac{1}{T}\sum_{r=0}^{T-1} \beta^r(\sigma_F^2 + B_F^2) + D_{\mathcal{X}} L_F \delta.$$

Telescoping the sum we get the following

$$\begin{aligned}\frac{1}{T}\sum_{r=0}^{T-1} \mathbb{E}[F(\mathbf{x}^r) - F^*] &\leq -\frac{\mu_F}{2}\|\mathbf{x}^T - \mathbf{x}^*\|^2 + \frac{(\sigma_F^2 + B_F^2)}{\mu_F}\frac{\log(T)}{T} + D_{\mathcal{X}} L_F \delta \\ &\leq \frac{(\sigma_F^2 + B_F^2)}{\mu_F}\frac{\log(T)}{T} + D_{\mathcal{X}} L_F \delta.\end{aligned}$$

Therefore, we have the proof. □

F.5 Convex Objective: Proof of Theorem 4

Proof. From the update rule of Algorithm 2, we have for $\beta^r = \beta$ for all $r \in \{0, 1, \ldots, T-1\}$

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2] &= \mathbb{E}[\|\text{proj}_{\mathcal{X}}(\mathbf{x}^r - \beta\widehat{\nabla} F(\mathbf{x}^r;\xi^r)) - \mathbf{x}^*\|^2] \\ &\overset{(a)}{=} \mathbb{E}[\|\text{proj}_{\mathcal{X}}(\mathbf{x}^r - \beta\widehat{\nabla} F(\mathbf{x}^r);\xi^r) - \text{proj}_{\mathcal{X}}(\mathbf{x}^*)\|^2] \\ &\overset{(b)}{\leq} \mathbb{E}[\|\mathbf{x}^r - \beta\widehat{\nabla} F(\mathbf{x}^r;\xi^r) - \mathbf{x}^*\|^2] \\ &\overset{(c)}{=} \mathbb{E}[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + \beta^2\|\widehat{\nabla} F(\mathbf{x}^r;\xi^r)\|^2 - 2\beta\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla} F(\mathbf{x}^r;\xi^r)\rangle] \\ &\overset{(d)}{\leq} \mathbb{E}[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2\beta^2\|\widehat{\nabla} F(\mathbf{x}^r;\xi^r) - \widehat{\nabla} F(\mathbf{x}^r)\|^2 + 2\beta^2\|\widehat{\nabla} F(\mathbf{x}^r)\|^2 \\ &\qquad - 2\beta\langle \mathbf{x}^r - \mathbf{x}^*, \nabla F(\mathbf{x}^r)\rangle - 2\beta\langle \mathbf{x}^r - \mathbf{x}^*, \widehat{\nabla} F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\rangle] \\ &\overset{(e)}{\leq} \mathbb{E}[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2\beta^2\sigma_F^2 + 2\beta^2\overline{L}_F^2 - 2\beta(F(\mathbf{x}^r) - F^*) \\ &\qquad + 2\beta\|\mathbf{x}^r - \mathbf{x}^*\|\,\|\widehat{\nabla} F(\mathbf{x}^r) - \nabla F(\mathbf{x}^r)\|] \\ &\overset{(f)}{\leq} \mathbb{E}[\|\mathbf{x}^r - \mathbf{x}^*\|^2 + 2\beta^2(\sigma_F^2 + \overline{L}_F^2) - 2\beta(F(\mathbf{x}^r) - F^*) + 2\beta D_{\mathcal{X}} L_F \delta]\end{aligned}$$

where (a) follows from the fact that $\mathbf{x}^* \in \mathcal{X}$; (b) results from the non-expansiveness of the projection operator; (c) uses $\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\langle \mathbf{a}, \mathbf{b}\rangle$; (d) utilizes the fact that $\|\mathbf{a} - \mathbf{b}\|^2 = 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ and unbiased gradient from Lemma 4; (e) results from Lemmas 3 and 4, the convexity assumption of the implicit function (Assumption 6 with $\mu_F = 0$) and the Cauchy-Schwartz inequality; and (f) results from Assumption 1(b) and Lemma 3.

Summing over $r \in \{0, 1 \ldots, T-1\}$, multiplying by $1/T$ and rearranging the terms, we get

$$\frac{1}{T}\sum_{r=0}^{T-1} \mathbb{E}[F(\mathbf{x}^r) - F^*] \leq \frac{\|\mathbf{x}^1 - \mathbf{x}^*\|^2}{2\beta T} + \beta(\sigma_F^2 + \overline{L}_F^2) + D_{\mathcal{X}} L_F \delta.$$

Using Jensen's inequality and denoting $\underline{\mathbf{x}} = \frac{1}{T}\sum_{r=0}^{T-1}\mathbf{x}^r$, we get

$$\mathbb{E}[F(\underline{\mathbf{x}}) - F^*] \leq \frac{\|\mathbf{x}^1 - \mathbf{x}^*\|^2}{2\beta T} + \beta(\sigma_F^2 + \overline{L}_F^2) + D_{\mathcal{X}} L_F \delta.$$

Therefore, we have the proof. □