

MINIMAX PROBLEMS WITH COUPLED LINEAR CONSTRAINTS: COMPUTATIONAL COMPLEXITY AND DUALITY*

IOANNIS TSAKNAKIS[†], MINGYI HONG[‡], AND SHUZHONG ZHANG[‡]

Abstract. In this work we study a special minimax problem where there are linear constraints that couple both the minimization and maximization decision variables. The problem is a generalization of the traditional saddle point problem (which does not have the coupling constraint), and it finds applications in wireless communication, game theory, transportation, just to name a few. We show that the considered problem is challenging, in the sense that it violates the classical max-min inequality, and that it is NP-hard even under very strong assumptions (e.g., when the objective is strongly-convex–strongly-concave). We then develop a duality theory for it, and analyze conditions under which the duality gap becomes zero. Finally, we study a class of stationary solutions defined based on the dual problem, and evaluate their practical performance in an application on adversarial attacks on network flow problems.

Key words. min-max problem, coupled constraints, duality theory, computational complexity, max-min inequality

MSC codes. 49K35, 65K05, 90C47

DOI. 10.1137/21M1462428

1. Introduction. The *minimax optimization problem*, given below,

$$(1.1) \quad \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}),$$

finds applications in areas such as machine learning, game theory, signal processing, and it has been extensively studied in recent years [37, 30, 26, 27, 51]. Its specific applications include adversarial learning [32], reinforcement learning [41], resource allocation in wireless communication [30], and generative adversarial networks [19], among others.

In this work¹ we consider a class of more general minimax problems, where the constraint set of the inner problem linearly depends on *both* optimization variables:

$$(mM-I) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right).$$

In the above expression, $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$, $A \in \mathbb{R}^{k \times n}$, $B \in \mathbb{R}^{k \times m}$, $\mathbf{c} \in \mathbb{R}^k$. Let us use $[A\mathbf{x} + B\mathbf{y}]_i \leq \mathbf{c}_i$ to denote the i th constraint, and define $\mathcal{K} := \{1, \dots, k\}$ as the index set of the constraints. As a naming convention, in this work we will use “M” and “m” to denote the maximization and minimization problems, respectively. Therefore, we refer to the above problem as a *minimax with*

*Received by the editors November 29, 2021; accepted for publication (in revised form) April 25, 2023; published electronically October 13, 2023.

<https://doi.org/10.1137/21M1462428>

Funding: The first and second authors are supported in part by NSF grants CIF-1910385, CMMI-1727757.

[†]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (tsakn001@umn.edu, mhong@umn.edu).

[‡]Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455 USA (zhangs@umn.edu).

¹Due to space limitations we were unable to include all the results of our work in this manuscript. The interested readers can refer to an extended version of this paper which is available online [46].

inner-level coupling (mM-I) problem. Further, the maximization variable is always $\mathbf{y}$, while the minimization variable is always $\mathbf{x}$.

Problems (1.1) and (mM-I) appear to be closely related, so one might think that they have similar properties. For example, the original problem (1.1) is easily solvable when $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, so this may lead to the belief that the linearly constrained version is still relatively easy. However, we will show that problem (mM-I) is *NP-hard* in general, even when $f(\mathbf{x}, \mathbf{y})$ is strongly convex and strongly concave. Further, the classical max-min inequality [35, Theorem 1.3.1] does not hold for such a class of problems. Finally, existing algorithms developed for minimax problems cannot be applied directly because they can get stuck at some uninteresting solution points. On the other hand, compared with (1.1), these problems can be used to model a wider class of applications, a few of which are presented below.

1.1. Representative applications.

1.1.1. Adversarial attacks in resource allocation problems. Consider a setting where a player (the “user”) aims to optimally allocate a resource of fixed total amount $c_0 \in \mathbb{R}$, across n different tasks. The goal is to maximize a certain utility which is a function of the allocation. Also, there is an adversary that tries to reduce the utility of the user, by designing an allocation strategy across the n tasks that “forces” the user (who utilizes the remaining resource) to select an allocation that lowers its utility. As a concrete example, let us consider a transmission rate maximization problem under the presence of a jammer, which arises in wireless communications.

Specifically, we consider a problem where a single user (i.e., a transmitter-receiver pair) transmits messages over n channels. The goal of this user is to allocate power (the “resource”) across the n channels (the “tasks”) such that its transmission rate (the “utility”) is maximized. Suppose that there is a jammer (the “adversary”) in the system, whose goal is to minimize the user’s rate (e.g., [17]). In this problem we assume that the channel is Gaussian, and denote with $\mathbf{x}, \mathbf{y}$ the power allocations of the user and the jammer, respectively. Then, the user’s system rate is given by [17, 30]

$$R(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \log \left(1 + \frac{g_i^u x_i}{\sigma^2 + g_i^j y_i} \right),$$

where g_i^u, g_i^j are the gains of the user and the jammer on channel i , respectively, and σ^2 is the noise power. Note that both players’ power allocations are subject to constraints $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^n x_i \leq \bar{x}\}$ and $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} \geq \mathbf{0}, \sum_{i=1}^n y_i \leq \bar{y}\}$, where $\bar{x}, \bar{y}$ are the total power budgets for the user and the adversary, respectively. Moreover, in certain communication systems such as a cognitive radio network [52], it is required that the total transmission power on each channel is upper bounded by the so-called *interference temperature* (denoted by a constant $c > 0$), in order to limit the *total* interference caused to other (perhaps more important) cochannel users in the system. Consequently, the problem of how to *optimally* attack the user in such a multichannel wireless system can be formulated as the following linearly constrained minimax problem, which is a special case of (mM-I):

$$\min_{\mathbf{y} \in \mathcal{Y}} \left(\max_{\mathbf{x} \in \mathcal{X}, \mathbf{x} + \mathbf{y} \leq \mathbf{c}} R(\mathbf{x}, \mathbf{y}) \right).$$

1.1.2. Adversarial attacks in network flow problems. Consider a *flow network* represented by a directed graph $G = (V, E)$, where V is the set of vertices and E

is the set of edges. Let x_e and p_e denote the flow and capacity on each edge $e \in E$, respectively; let $\mathbf{x} := \{x_e\}_{e \in E} \in \mathbb{R}^{|E|}$ and $\mathbf{p} := \{p_e\}_{e \in E} \in \mathbb{R}^{|E|}$ denote the vectors of edge flows and the edge capacities, respectively. Let $q_e(x_e)$ denote the cost of moving one unit of flow across edge e , and s and t the source and sink nodes, respectively. Suppose F is a set that collects all the edges used by $\mathbf{x}$ to deliver a total of r_t units of flow from s to t . Then the total transportation cost is defined as $q_{tot}(\mathbf{x}) = \sum_{e \in F} q_e(x_e)x_e$.

A minimum cost network flow problem can be defined as finding the paths with the minimum cost from source to sink, so that we can successfully transport a certain amount of flow [8]. Let us consider an extension to such a problem, where in addition to the regular network user, there is an adversary who will inject flows onto the network to force the regular user to use more expensive paths. Let $\mathbf{y} := \{y_e\}_{e \in E}$ denote the set of flows controlled by the adversary, and $b > 0$ be its total budget. Then, the problem of an adversary can be formulated as follows,

$$(1.2) \quad \begin{aligned} & \max_{\substack{0 \leq \mathbf{y} \leq \mathbf{p} \\ \sum_{(i,j) \in E} y_{ij} = b}} \min_{\substack{0 \leq \mathbf{x} \leq \mathbf{p} \\ \sum_{(i,t) \in E} x_{it} = r_t}} \sum_{(i,j) \in E} \bar{q}_{ij}(x_{ij}, y_{ij}) x_{ij} \\ & \text{s.t. } \mathbf{x} + \mathbf{y} \leq \mathbf{p}, \\ & \sum_{(i,j) \in E} x_{ij} - \sum_{(j,k) \in E} x_{jk} = 0 \quad \forall j \in V \setminus \{s, t\}, \end{aligned}$$

where $\bar{q}_{ij}(x_{ij}, y_{ij})$ is related to the cost of routing both x_{ij} and y_{ij} onto the network; a simple choice is $\bar{q}_{ij}(x_{ij}, y_{ij}) = q_{ij}(x_{ij} + y_{ij})$. We note that such a kind of attack falls under the scope of network interdiction problems [45] and there are several works in literature that study variants of this application [15, 43, 44]. We note that problem (1.2) can be used to model attacks for real systems such as communication networks [15] or power networks [43], in which the adversary aims to maximize the cost of the network owner by injecting spurious traffic or disabling network components.

1.2. Contributions. In this work, we study the minimax problem with coupled linear constraints (mM-I). Our main contributions are listed below:

- We first identify relationships between (mM-I) and a number of its variants, and show that even when $f(\mathbf{x}, \mathbf{y})$ is strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$, the well-known max-min inequality does not hold true. Additionally, we show that (mM-I) and its variants are NP-hard in general.
- We develop a duality theory under the assumption that the objective f is strongly concave w.r.t $\mathbf{y}$. Specifically, we define three different dual problems, identify the conditions under which strong duality holds, and establish the equivalence of their solutions to that of the original primal problem.
- Based on our developed dual problems, we identify a new stationary solution concept, and develop a first-order algorithm to evaluate the practical performance of such a solution concept on an adversarial attack problem for network flows.

We emphasize that in order to evaluate the quality of the proposed stationary solution concept, we need to develop an efficient numerical algorithm. The algorithm itself is new (to our knowledge), but its analysis is relatively standard, and we do not have the space to include the details here; the algorithmic development is indeed not the focus of this paper.

1.3. Related works. The class of problems (mM-I) is closely related to problems in optimization and game theory. Below, we review the related literature.

Minimax problems. These problems have been extensively studied under different assumptions for the objective f . For instance, there are works for

(strongly-)convex (strongly-)concave [33, 34, 50], nonconvex (strongly-)concave [26, 27, 30, 37], and nonconvex nonconcave [28, 37, 51] problems. We refer the interested readers to a recent survey about detailed developments [42]. Specifically, in [26] the “classical” gradient descent-ascent (GDA) algorithm is analyzed, and convergence is established to stationary solutions, under the assumption that the objective is nonconvex (strongly-) concave. Moreover, in [34] two extensions of GDA are studied, namely, the optimistic GDA (OGDA), and the extragradient algorithm, albeit in a convex-concave setting. In another extension, the inner-level problem is solved using multiple steps of a first-order algorithm; the convergence of this scheme is analyzed in [37] assuming that the objective is nonconvex in $\mathbf{x}$, and satisfies either the Polyak–Łojasiewicz condition or is concave in $\mathbf{y}$.

The above works focus on problems *without* any constraints that couple the inner and outer variables. In these problems, the optimality conditions that are typically used are (a) the stationarity conditions of objective f w.r.t $\mathbf{x}$ and $\mathbf{y}$, and (b) the stationarity condition of the upper-level objective $\rho(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$, i.e., $0 \in \partial \rho(\mathbf{x})$. However, these conditions were defined for problems with no coupled constraints, and thus they cannot be applied in our case. Therefore, it is no longer clear how algorithms such as OGDA can be used. Consequently, the analysis of (mM-I) requires different approaches.

There are a couple of very recent works that study min-max problems with coupled constraints similar to (mM-I). In [9] the authors extend the notion of local minimax points (e.g., [23]) for min-max problems with coupled constraints, and derive the respective necessary and sufficient optimality conditions. However, no duality theory or algorithms are developed. Moreover, in [18] the authors study a min-max problem where the objective is convex-concave, the coupled constraints are of the form $g(\mathbf{x}, \mathbf{y}) \geq 0$, for some function $g(\mathbf{x}, \mathbf{y})$ that is concave in $\mathbf{y}$, and the solution concept they consider is the same as in this work. Most importantly they also assume that the inner function $\phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}, g(\mathbf{x}, \mathbf{y}) \geq 0} f(\mathbf{x}, \mathbf{y})$ is convex, which makes the problem tractable. On the contrary, we make no such assumption, and develop our theory and algorithm for the case where the problem is still difficult (i.e., NP-hard).

Generalized Nash equilibrium games. Our problem formulation is also related to the *generalized Nash equilibrium* (GNE) game [12, 11, 14], which is an extension of Nash games where the action space of each player depends on the actions of the other players. Specifically, a two player GNE game can be formulated as follows [12]:

$$(1.3) \quad \min_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} f_1(\mathbf{x}, \mathbf{y}), \quad \min_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} f_2(\mathbf{x}, \mathbf{y}),$$

where f_1, f_2 are the utilities of two players. Problem (mM-I) is related to GNE games in the zero-sum case where $f_1 = -f_2$. However, in GNE games the goal is to attain a Nash equilibrium, and not a minimax/maximin point. In other words, unlike problem (mM-I) the two problems do not have an order. In addition, note that for GNE games, the equilibrium solutions do not always exist, and certain (strong) assumptions are required; for instance, the utilities f_1, f_2 need to be quasi-convex [22].

One technique that was developed for GNEs [47] utilizes the Nikaidô–Isoda function [36] to reformulate the GNE-finding problem to a constraint optimization task. However, the main results in this case have been obtained under the assumption that the utilities are convex. Also, a number of penalty approaches [39, 13] have been proposed; nonetheless, these are considered difficult to solve in practice. Another approach reduces the problem of finding the KKT points of a GNE game to a variational

inequality problem [2]. Overall, the above approaches cannot be applied in problem (mM-I) because they were developed for problems of the form (1.3), where each player has its own utility function, and the order of play does not matter.

Stackelberg games. Another class of game-theoretic problems that are closely related to problem (mM-I) are *Stackelberg games* [48, 38, 49], in which there are two players, the leader and the follower. The leader can anticipate the follower's actions. In this case, the order in which the two players act matters. Then a (general-sum) Stackelberg game can be formulated as the following bilevel program:

$$\min_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad \mathbf{y} \in \arg \min_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}).$$

Therefore, these games are in some sense more general than problem (mM-I), which can only model zero-sum Stackelberg games. On the other hand, typically in Stackelberg games the action space of the follower (resp., the leader) is independent of the action of the leader (resp., the follower); see, e.g., [54, section 2.6.1]. As a result, problem (mM-I) extends the zero-sum Stackelberg games by taking into consideration the interactions between the players' actions.

Bilevel optimization. Finally, we would like to discuss the relationship between problem (mM-I) and bilevel optimization problems of the following form:

$$\min_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad \mathbf{y} \in \arg \min_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} f(\mathbf{x}, \mathbf{y}).$$

Bilevel optimization was formally introduced in [5], and it is also related to the broader class of problems of mathematical programming with equilibrium constraints [31]. A generic instance of bi-level problems includes all minimax problems as special cases, with or without coupled constraint; see [7, 10, 29] for a number of survey papers. However, bilevel optimization problems are in general very challenging to solve. More precisely [29] (1) even linear bilevel problems are NP-hard, (2) the lower-level problem might have multiple solutions, (3) the feasible region defined by the lower-level problem can be a nonconvex set, and (4) they are nonsmooth in general.

Several techniques have been developed in the literature for solving these problems. One such technique uses the value function $V(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} f(\mathbf{x}, \mathbf{y})$ to express the lower-level problem as an inequality constraint, and transforms the bilevel problem into a single-level optimization task [53, 25]. However, not only does this reformulated problem not satisfy any of the known constraint qualifications (CQ), but also the nonsmoothness of $V(\mathbf{x})$ makes the problem difficult. Alternative CQs such as calmness conditions have been proposed, which are trivially satisfied in minimax problems [53]. Based on the calmness condition, a penalty method has been developed in [25], but it only applies to bilevel problems in which the constraints set of the lower-level problem is independent of the variable of the upper one. Overall, this generic theory does not provide much understanding about the structure of our problem.

There is a recent line of work in (stochastic) single and two-time scale gradient descent methods [16, 21, 24, 6]. Under the assumption that the lower-level problem is strongly convex (and thus admits a unique solution) and unconstrained, the authors attempt to minimize the objective $\phi(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$. In this case the implicit function theorem provides access to the gradient $\nabla \phi(\mathbf{x})$. Unfortunately, in our case the constraints on the lower-level problem prevent us from using these type of methods.

All the above approaches were developed for general bilevel problems. As a result, they were designed to be suitable for a wide range of problems. Moreover, these approaches are either not applicable to our problem, or even if they are, they are not

very attractive (e.g., not very efficient, do not have iteration complexity guarantees, difficult to analyze due to nonsmoothness). On the other hand, problem (mM-I) has a special minimax structure, that is the lower-level objective f is the negative of the upper level one F , and the coupling constraints are linear in the decision variables. Therefore, in contrast to the general methods presented here, our goal is to leverage the special structure of problem (mM-I) to design methods tailored for it.

2. Minimax problems with coupled linear constraints. In this section, we derive a few unique properties about problem (mM-I), and outline the challenges in designing efficient algorithms for it. To begin with our analysis, let us first introduce a number of variants of problem (mM-I), and understand their relations.

DEFINITION 2.1. *We define the following problems:*

A. *Inner-level coupling (I)*

1. *miniMax with inner-level coupling (mM-I):*

$$(mM-I) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right).$$

2. *Maximin with inner-level coupling (Mm-I):*

$$(Mm-I) \quad \max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right).$$

B. *Outer-level coupling (O)*

1. *miniMax with outer-level coupling (mM-O):*

$$(mM-O) \quad \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right), \quad \mathbf{y}^*(\mathbf{x}) \in \arg \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}).$$

2. *Maximin with outer-level coupling (Mm-O):*

$$(Mm-O) \quad \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right), \quad \mathbf{x}^*(\mathbf{y}) \in \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}).$$

We note that for the two outer-level coupled problems, the constraint set of the outer problem depends on the solutions $\mathbf{y}^*(\mathbf{x})$ of the inner one. Since $\mathbf{y}^*(\mathbf{x})$ is typically a nonlinear function, technically such a constraint is not a linear one any more. Further, at this point these problems are underspecified, as it is not clear which of the solutions of $\mathbf{y}^*(\mathbf{x})$ and $\mathbf{x}^*(\mathbf{y})$ (e.g., one, some, or all) are expected to satisfy the constraints $A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}$ and $A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}$, respectively. Note that these two outer-level coupling problems are mainly introduced to better understand the two inner-level coupling problems. They have their own special structures, therefore require a separate treatment (which is beyond the scope of the current paper).

To proceed, let us define the global minimax (or maxmin) solutions for these problems. For problem (mM-I), a point $(\mathbf{x}^*, \mathbf{y}^*)$ is the solution, if the following holds:

$$(2.1a) \quad \mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}), \quad \text{where} \quad \phi(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}),$$

$$(2.1b) \quad \mathbf{y}^* \in \arg \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^* + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}^*, \mathbf{y}).$$

For problem (Mm-I), the global maxmin solution can be defined similarly. Moreover, a solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (mM-O) is defined as

$$(2.2a) \quad \mathbf{y}^* \in \arg \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^*, \mathbf{y}),$$

$$(2.2b) \quad \mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^* \leq \mathbf{c}} \tilde{\phi}(\mathbf{x}), \quad \text{where} \quad \tilde{\phi}(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}).$$

The solution of problem (Mm-O) can be defined similarly. Next, we provide necessary conditions for the existence of those solutions.

Assumption 2.2. Suppose the following conditions hold true.

- The function $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is continuous w.r.t. both $\mathbf{x}$ and $\mathbf{y}$.
- The sets $\mathcal{X}$, $\mathcal{Y}$ are convex and compact, with $\|\mathbf{x}\| \leq D \ \forall \mathbf{x} \in \mathcal{X}$, $\|\mathbf{y}\| \leq D \ \forall \mathbf{y} \in \mathcal{Y}$, for some constant $D > 0$.
- (feasibility) The following feasibility conditions hold:
 - In problem (mM-I), $\forall \mathbf{x} \in \mathcal{X}, \exists \mathbf{y} \in \mathcal{Y}$, such that $A\mathbf{x} + B\mathbf{y} - \mathbf{c} \leq \mathbf{0}$.
 - In problem (Mm-I), $\forall \mathbf{y} \in \mathcal{Y}, \exists \mathbf{x} \in \mathcal{X}$, such that $A\mathbf{x} + B\mathbf{y} - \mathbf{c} \leq \mathbf{0}$.
 - In problem (mM-O), $\exists \mathbf{x} \in \mathcal{X}$ such that $A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) - \mathbf{c} \leq \mathbf{0}$ holds for at least one solution $\mathbf{y}^*(\mathbf{x})$.
 - In problem (Mm-O), $\exists \mathbf{y} \in \mathcal{Y}$ such that $A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} - \mathbf{c} \leq \mathbf{0}$ holds for at least one solution $\mathbf{x}^*(\mathbf{y})$.

Remark 2.3. Under Assumption 2.2, the inner-level tasks introduced in Definition 2.1 are well-defined. Specifically, $\forall \mathbf{x} \in \mathcal{X}$ in problems (mM-I), (mM-O), the optimal solution of the inner problem exists, (or $\forall \mathbf{y} \in \mathcal{Y}$ in problems (Mm-I), (Mm-O), the same holds). This is a result of Assumption 2.2.c, which ensure the feasibility of the inner problem, and Assumptions 2.2.a, 2.2.b, which guarantee the existence of a global solution, through Weierstrass' theorem. In addition, Berge's theorem of the maximum [1, Theorem 17.31] implies that $\phi(\mathbf{x})$ and $\tilde{\phi}(\mathbf{x})$ are continuous functions. This continuity property combined the compactness of the constraint sets, and ensured the existence of global solutions in problems (2.1b) and (2.2b), through Weierstrass' theorem.

Under Assumption 2.2, it is easy to argue that f is lower/upper bounded over $\mathcal{X} \times \mathcal{Y}$. That is, there exist finite constants $\underline{f}, \bar{f}$ such that

$$(2.3) \quad \underline{f} \leq f(\mathbf{x}, \mathbf{y}) \leq \bar{f} \ \forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}.$$

Next, we characterize the relationships between the four problems in Definition 2.1. As a convention, we will denote the *optimal* objective value of a problem P as $v(P)$.

PROPOSITION 2.4 (relations between problems). *For the values of the problems (mM-I), (Mm-I), (mM-O), and (Mm-O), the following relationships hold:*

- $v(\text{mM-I}) \leq v(\text{mM-O});$
- $v(\text{Mm-I}) \geq v(\text{Mm-O});$
- $v(\text{mM-O}) \geq v(\text{Mm-O}).$

Moreover, there is no definitive relation between the rest of the pairs, namely, the pairs $\{v(\text{mM-I}), v(\text{Mm-I})\}$, $\{v(\text{mM-I}), v(\text{Mm-O})\}$, $\{v(\text{mM-O}), v(\text{Mm-I})\}$. That is, any relation of $<$, $=$, or $>$ can hold between them.

Proof. See Appendix A.1. □

The most remarkable fact concerns the relation between (mM-I) and (Mm-O) [or between (mM-O) and (Mm-I)]. Note that after exchanging the min and max operators, (Mm-O) becomes (mM-I). Proposition 2.4 shows that there is no

definite relation between (mM-I) and (mM-O). On the contrary, in standard minimax problems (1.1), the max-min inequality [35, Theorem 1.3.1] holds, that is,

$$(2.4) \quad \max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right) \leq \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right)$$

for any function f and nonempty, closed sets $\mathcal{X}, \mathcal{Y}$. In addition, if f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, and $\mathcal{X}, \mathcal{Y}$ are compact, then the values of these problems are equal.

The results shown in proposition (2.4) indicate that the four problems we studied above should be harder than the standard minimax problems. Indeed, as we show below, problems (mM-I)–(mM-O) are all NP-hard in general, even when the objective function f is strongly-convex–strongly-concave, and the sets $\mathcal{X}$ and $\mathcal{Y}$ are compact.

PROPOSITION 2.5. *Consider the (mM-I) problem with a strongly-convex–strongly-concave objective $f(\mathbf{x}, \mathbf{y})$. This problem is NP-hard.*

Proof. Consider the problem

$$(2.5) \quad \min_{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}} \left(\max_{\substack{-\mathbf{1} \leq \mathbf{y} \leq \mathbf{1} \\ \mathbf{x} - 2\mathbf{y} = \mathbf{0}}} \|\mathbf{x}\|^2 + \frac{1}{2} \mathbf{x}^T Q \mathbf{y} - 4 \|\mathbf{y}\|^2 + \mathbf{d}^T \mathbf{x} \right),$$

where $Q \in \mathbb{R}^{n \times n}$ is a symmetric matrix with $Q \preceq 0$, $\mathbf{d} \in \mathbb{R}^n$, and $\mathbf{1}$ is a vector of ones. We can easily verify that problem (2.5) is of the form (mM-I), and satisfies Assumption 2.2. Then, using the coupled equality constraint $\mathbf{x} - 2\mathbf{y} = \mathbf{0}$ (which can be also expressed with two inequality constraints, i.e., $\mathbf{x} - 2\mathbf{y} \leq \mathbf{0}$ and $-\mathbf{x} + 2\mathbf{y} \leq \mathbf{0}$), problem (2.5) can be written as

$$\min_{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}} \frac{1}{4} \mathbf{x}^T Q \mathbf{x} + \mathbf{d}^T \mathbf{x}.$$

Since $Q \preceq 0$ the above is a difficult constrained concave minimization problem. To be more precise, it is established in [40] that this box-constrained nonconvex quadratic problem is NP-hard. The proof is now completed. $\square$

PROPOSITION 2.6. *Consider the (mM-O) problem with a strongly convex concave objective $f(\mathbf{x}, \mathbf{y})$. This problem is NP-hard.*

Proof. Consider the problem

$$(2.6) \quad \min_{\|\mathbf{x}\| \leq 1, \mathbf{x} - \mathbf{y}^*(\mathbf{x}) = \mathbf{0}} \left(\max_{\|\mathbf{y}\| \leq 1} \left\{ M \mathbf{x}^T \mathbf{x} + \sum_{i,j}^n c_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2 + \mathbf{x}^T \mathbf{y} \right\} \right),$$

where $c_{i,j}$ are constants, and $M > 0$ is selected such that the function $g(\mathbf{x}) = M \mathbf{x}^T \mathbf{x} + \sum_{i,j}^n c_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2$ is strongly convex in a compact domain (which is a superset of the set $\|\mathbf{x}\| \leq 1$). Therefore, the objective of problem (2.6) is strongly convex in $\mathbf{x}$, and concave (linear) in $\mathbf{y}$. Then, (2.6) can be equivalently written as

$$(2.7) \quad \min_{\|\mathbf{x}\| \leq 1, \mathbf{x} - \frac{\mathbf{x}}{\|\mathbf{x}\|} = \mathbf{0}} \left\{ M \mathbf{x}^T \mathbf{x} + \sum_{i,j}^n c_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2 + \|\mathbf{x}\| \right\} \Leftrightarrow \min_{\|\mathbf{x}\|=1} \left\{ \sum_{i,j}^n c_{ij} \mathbf{x}_i^2 \mathbf{x}_j^2 \right\} + M + 1.$$

It is a known fact that the problem in the right-hand side (RHS) of (2.7) is in general NP-hard, which establishes the NP-hardness of problem (2.6). The proof is now completed. $\square$

We would like to stress here that the classical min-max problem of the form (1.1), where the objective $f(\mathbf{x}, \mathbf{y})$ is differentiable, strongly convex in $\mathbf{x}$, and strongly concave in $\mathbf{y}$, is not NP-hard. To see this, one needs only to observe that for any $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, the direction $(\nabla_x f(\mathbf{x}, \mathbf{y})^T, -\nabla_y f(\mathbf{x}, \mathbf{y})^T)^T$ defines a separating hyperplane regarding any stationary solution. In particular, for any given stationary solution $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ we have

$$\begin{aligned} & (\nabla_x f(\mathbf{x}, \mathbf{y}), -\nabla_y f(\mathbf{x}, \mathbf{y}))^T (\mathbf{x} - \mathbf{x}^*, \mathbf{y} - \mathbf{y}^*) \\ &= (\nabla_x f(\mathbf{x}, \mathbf{y})^T (\mathbf{x} - \mathbf{x}^*) - (\nabla_y f(\mathbf{x}, \mathbf{y}))^T (\mathbf{y} - \mathbf{y}^*)) \\ &\leq f(\mathbf{x}^*, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - [f(\mathbf{x}, \mathbf{y}^*) - f(\mathbf{x}, \mathbf{y})] \\ &= f(\mathbf{x}^*, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}^*) \\ &\leq 0, \end{aligned}$$

where the first inequality in the above derivation is due to the gradient inequality of the convex-concave property, and the second inequality is due to the stationary property,

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*)$$

for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. Using the ellipsoid method (see, e.g., [20]), this implies that finding a stationary solution can be done in polynomial time. In terms of finding stationary solutions effectively in practice without resorting to the ellipsoid method, there are several works in the literature in which convergence is shown for the above problem to approximate global solutions (i.e., saddle points) in a finite number of iterations for any given precision; see, for instance, [34, 50]. Therefore, it is the introduction of the linear coupling between the inner and the outer variables in the constraints (of one of the two problems) that renders the problem intractable, even under strong assumptions. To illustrate this point more clearly consider the following example.

Example 2.7. Consider the following simple linearly constrained min-max problem

$$\min_{x \in [0, 1]} \left(\max_{\substack{y \in [-1, 1] \\ x - 2y = 0}} f(x, y) := x^2 - \frac{1}{2}xy - 4y^2 \right).$$

Notice that the objective $f(x, y)$ is strongly convex in x and strongly concave in y . Then, it is easy to see that inner function takes the form

$$\phi(x) = \max_{y \in [-1, 1] : x - 2y = 0} x^2 - \frac{1}{2}xy - 4y^2 = x^2 - \frac{1}{2}x \frac{1}{2}x - 4 \frac{1}{4}x^2 = -\frac{1}{4}x^2.$$

Clearly, $\phi(x)$ is nonconvex; in fact it is strongly concave. Therefore we see that while the objective $f(x, y)$ is a simple strongly-convex-strongly-concave function, the resulting value function $\phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y})$ is nonconvex.

Despite the fact that the above results show that the linearly constrained minimax problems are NP-hard in general, it is still desirable to design efficient algorithms for

them (for computing certain stationary solutions). As we have noted in section 1.3, standard algorithms for minimax problems do not apply, and generic methods for bilevel optimization can be inefficient. Our idea is to explore the structure of these problems from their dual perspective, and identify equivalent problems that are much easier to optimize. Since our focus is mainly given to the inner-level coupled problems, in the next section we develop a duality theory for problem (mM-I).

3. Duality theory for (mM-I). To begin with, let us define the dual problems.

DEFINITION 3.1 (dual problems). *Consider the Lagrangian of problem (mM-I),*

$$(3.1) \quad L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}, \mathbf{y}) - \boldsymbol{\lambda}^T (A\mathbf{x} + B\mathbf{y} - \mathbf{c}),$$

where $\boldsymbol{\lambda} \geq \mathbf{0}$ is the vector of Lagrangian multipliers. Then, define the following three dual problems:

$$(D1) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right),$$

$$(D2) \quad \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \left(\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right),$$

$$(D3) \quad \min_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right).$$

Note that due to the special structure of our problem, we can potentially have a few more possible “dual” problems. However, our analysis will focus on three of them listed above, since they will be instrumental in our subsequent analysis and development of stationary solution concept. Below, we provide the weak duality theorem of problem (mM-I).

THEOREM 3.2 (weak duality). *Under Assumption 2.2, we have:*

$$v(\text{mM-I}) \leq v(D1) = v(D2) = v(D3).$$

Proof. First, notice that we can exchange the order of the min operators:

$$\min_{\mathbf{x} \in \mathcal{X}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) = \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \left(\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) = \min_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right).$$

So $v(D1) = v(D2) = v(D3)$. Moreover, using the max-min inequality we have that

$$\max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \leq \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \quad \forall \mathbf{x} \in \mathcal{X}.$$

Minimizing the problems on both sides over the set $\mathcal{X}$, we can obtain

$$(3.2) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) \leq \min_{\mathbf{x} \in \mathcal{X}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right).$$

Moreover, consider an arbitrary $\mathbf{x} \in \mathcal{X}$, and observe that if there exists an index i such that $[A\mathbf{x} + B\mathbf{y} - \mathbf{c}]_i > 0$, then we have

$$(3.3) \quad \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}, \mathbf{y}) + \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \left\{ -\boldsymbol{\lambda}^T (A\mathbf{x} + B\mathbf{y} - \mathbf{c}) \right\} = -\infty.$$

On the other hand, if it holds that $A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}$, then

$$(3.4) \quad \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}, \mathbf{y}) > \underline{f} \quad \forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}.$$

As a result, (3.3) and (3.4) imply that

$$\max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) = \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) = \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}).$$

Minimizing both sides of the above equality over $\mathcal{X}$ we get

$$(3.5) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right) = \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right).$$

Finally, combining (3.2) and (3.5), we conclude that

$$\min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right) \leq \min_{\mathbf{x} \in \mathcal{X}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right).$$

The proof is now completed. $\square$

Next, we develop the strong duality of problem (mM-I). That is, we identify the conditions under which strong duality holds, and establish the equivalence of the solutions of the primal problem with those of the dual problems. To begin with, we impose the following assumptions.

Assumption 3.3. Assume that the following hold:

1. $f(\mathbf{x}, \mathbf{y})$ is strongly concave in $\mathbf{y}$ for every $\mathbf{x} \in \mathcal{X}$ with modulus μ_y .
2. For every $\mathbf{x} \in \mathcal{X}$ there exists $\mathbf{y} \in \text{relint}(\mathcal{Y})$ such that $(A\mathbf{x} + B\mathbf{y} - \mathbf{c}) \leq \mathbf{0}$.

Based on the above assumptions, we have the following duality theorem.

THEOREM 3.4 (duality theorem). *Under Assumptions 2.2 and 3.3, strong duality holds, that is, we have the following relations:*

$$v(\text{mM-I}) = v(\text{D1}) = v(\text{D2}) = v(\text{D3}).$$

Also, for the solutions of the dual problems we have that the following:

1. There exists a $\boldsymbol{\lambda}^* \geq \mathbf{0}$ such that $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a solution of (D2) if and only if $(\mathbf{x}^*, \mathbf{y}^*)$ is a solution of (mM-I).
That is, the tuple $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ satisfies the following conditions: (1) $(\mathbf{x}^*, \mathbf{y}^*)$ is a solution (i.e., a minimax point) of the inner min-max problem with $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, i.e., $(\mathbf{x}^*, \mathbf{y}^*) \in \arg \min_{\mathbf{x} \in \mathcal{X}} \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}^*)$; (2) $\boldsymbol{\lambda}^*$ is a global minimizer of the function $G(\boldsymbol{\lambda}) := \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$, $\boldsymbol{\lambda}^* \in \arg \min_{\boldsymbol{\lambda} \geq \mathbf{0}} G(\boldsymbol{\lambda})$.
2. There exists a $\boldsymbol{\lambda}^* \geq \mathbf{0}$ such that $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a solution of (D3) if and only if $(\mathbf{x}^*, \mathbf{y}^*)$ is a solution of (mM-I). That is $((\mathbf{x}^*, \boldsymbol{\lambda}^*), \mathbf{y}^*) \in \arg \min_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{x} \in \mathcal{X}} \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$.

Proof. We divide the proof into two parts.

Objective equivalence. We already established in Theorem 3.2 that $v(\text{D1}) = v(\text{D2}) = v(\text{D3})$. Note that strong duality holds for problem $\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y})$ for any fixed $\mathbf{x}$, as a consequence of Assumption 3.3 (see, e.g., [4, Proposition 5.3.1]). Then,

$$\begin{aligned} \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) &= \max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \quad \forall \mathbf{x} \in \mathcal{X}, \\ \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right) &= \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) = \min_{\mathbf{x} \in \mathcal{X}} \left(\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right). \end{aligned}$$

That is, we have $v(\text{mM-I}) = v(\text{D1}) = v(\text{D2}) = v(\text{D3})$.

Solution equivalence. Next we show the second part of the claim.

First, we show that the solutions of (mM-I) and (D1) are equivalent. Consider an arbitrary $\mathbf{x} \in \mathcal{X}$, and let us define the following problems:

$$\begin{aligned} \text{(A): } g_A(\mathbf{x}) &:= \max_{\mathbf{y} \in \mathcal{Y}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}), \\ \text{(B): } g_B(\mathbf{x}) &:= \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}), \\ \text{(C): } g_C(\mathbf{x}) &:= \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Let us denote $(\mathbf{y}_A^*, \boldsymbol{\lambda}_A^*)$ as a solution of (A) and $(\mathbf{y}_B^*, \boldsymbol{\lambda}_B^*)$ a solution of (B). We will show below that if $(\mathbf{y}_B^*, \boldsymbol{\lambda}_B^*)$ is a solution of (B), then $\mathbf{y}_B^*$ is a solution of (C). Conversely, we will show that if $\mathbf{y}^*$ is a solution of (C), then there exists a $\boldsymbol{\lambda}^*$ such that $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a solution of (B).

To begin with, using [4, Proposition 5.3.1], we see that Assumption 3.3 implies that strong duality holds for problem (C). From [4, Proposition 5.3.2] we know that when strong duality holds, $(\mathbf{y}_A^*, \boldsymbol{\lambda}_B^*)$ satisfies the following conditions: $\boldsymbol{\lambda}_B^* \geq \mathbf{0}$, $\mathbf{y}_A^* \in \mathcal{Y}$, and $A\mathbf{x} + B\mathbf{y}_A^* - \mathbf{c} \leq \mathbf{0}$, and

$$(3.6) \quad \mathbf{y}_A^* = \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}_B^*).$$

Moreover, observe that $(\mathbf{y}_B^*, \boldsymbol{\lambda}_B^*)$ is a solution of (B), and thus

$$(3.7) \quad \mathbf{y}_B^* = \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}_B^*) = \arg \max_{\mathbf{y} \in \mathcal{Y}} \{f(\mathbf{x}, \mathbf{y}) - (\boldsymbol{\lambda}_B^*)^T (A\mathbf{x} + B\mathbf{y} - \mathbf{c})\}.$$

Then, combining (3.6) and (3.7), and considering the strong concavity of $L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}_B^*)$ w.r.t $\mathbf{y}$, we can infer that $\mathbf{y}_A^* = \mathbf{y}_B^*$. Therefore, $\mathbf{y}_B^*$ is a solution of problem (A), and as a result a solution of (C).

Conversely, we will show that if $\mathbf{y}^*$ is a solution of (C) then there exists a $\boldsymbol{\lambda}^*$ such that $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a solution of (B). Indeed, under strong duality, for the unique (due to strong concavity) solution $\mathbf{y}^*$ of (C) there exist Lagrange multipliers $\boldsymbol{\lambda}^* \geq \mathbf{0}$ such that $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a saddle point of the Lagrangian $L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ [4, Proposition 5.3.2]. We know that the set of saddle points of $L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ is the intersection of the set of minimax points $(\mathbf{y}_A^*, \boldsymbol{\lambda}_A^*)$ with the set of maximin points $(\mathbf{y}_B^*, \boldsymbol{\lambda}_B^*)$, and thus it is implied that $(\boldsymbol{\lambda}^*, \mathbf{y}^*)$ is a solution of (B).

We have shown for some arbitrary $\mathbf{x} \in \mathcal{X}$ that $\mathbf{y}_{\mathbf{x}}^*$ is a solution of (C) if and only if $(\mathbf{y}_{\mathbf{x}}^*, \boldsymbol{\lambda}_{\mathbf{x}}^*)$ is a solution of (B) for some $\boldsymbol{\lambda}_{\mathbf{x}}^* \geq \mathbf{0}$; here we used the notation $\mathbf{y}_{\mathbf{x}}^*, \boldsymbol{\lambda}_{\mathbf{x}}^*$ to make the dependence to that arbitrary $\mathbf{x}$ clear. Finally, note that the strong duality holds for problem (C) for any fixed $\mathbf{x} \in \mathcal{X}$. This implies that the values of the problems (A), (B), and (C) are the same at every $\mathbf{x} \in \mathcal{X}$, that is

$$g_C(\mathbf{x}) = g_A(\mathbf{x}) = g_B(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X}.$$

Therefore, g_C, g_A, g_B have the same set of global minima $\mathcal{X}^*$. Then, at any given $\mathbf{x}^* \in \mathcal{X}^*$, $\mathbf{y}_{\mathbf{x}^*}^*$ is a solution of (C) if and only if $(\mathbf{y}_{\mathbf{x}^*}^*, \boldsymbol{\lambda}_{\mathbf{x}^*}^*)$ is a solution of (B) for some $\boldsymbol{\lambda}_{\mathbf{x}^*}^* \geq \mathbf{0}$. In conclusion, a point $(\mathbf{x}^*, \mathbf{y}_{\mathbf{x}^*}^*)$ is a solution of (mM-I) if and only if $(\mathbf{x}^*, \mathbf{y}_{\mathbf{x}^*}^*, \boldsymbol{\lambda}_{\mathbf{x}^*}^*)$ is a solution of (D1). The claim is proved.

Next, we show that the solutions of (mM-I) and (D2)/(D3) are equivalent. Note that the solution sets of the problems (D1), (D2), and (D3) are equivalent, since the order of the min operators can be exchanged. That is, a solution $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ of (D2) is also a solution of (D1), and vice versa; a similar result holds true between (D3) and (D1). In addition, using the (solution) equivalence between problems (mM-I) and (D1) established above, we can conclude that a point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a solution of (D2) if and only if $(\mathbf{x}^*, \mathbf{y}^*)$ is a solution of (mM-I). The same result also holds for (D3).

The proof is now completed. $\square$

In the above proof, note that despite Assumption 3.3, which ensures strong duality, the solution sets of the problems (A) and (B) (denoted with S_A, S_B , respectively) are not the same. Nonetheless, the original problem (mM-I) and the dual one (D2) are equivalent (in the sense mentioned in Theorem 3.4). This is due to the fact that $S_B \subseteq S_A$, which implies that every solution of B is a saddle point (i.e., the points in the set $S_{\text{saddle}} = S_B \cap S_A$) of the Lagrangian L and vice versa. Below we provide a relevant example.

Example 3.5. Consider a problem of the form (mM-I), where the inner task takes the form $\max_{y \in [0,2], y \leq 1} -y^2 + 2y$; the exact form of the outer task is irrelevant for the context of this example. Then, the primal (A) and dual (B) problems (of the inner task) are

$$p^* = \max_{y \in [0,2]} \min_{\lambda \geq 0} \{-y^2 + 2y - \lambda(y - 1)\}$$

$$d^* = \min_{\lambda \geq 0} \max_{y \in [0,2]} \{-y^2 + 2y - \lambda(y - 1)\}.$$

Note that Assumption 3.3 is satisfied in this example, and thus $p^* = d^*$ (i.e., strong duality). However, it can be easily shown that the solutions set of the primal problem is $S_A = \{(1, \lambda) : \lambda \geq 0\}$, while the respective set of the dual problem is different and consists only of one point, namely, $S_B = \{(1, 0)\} \subseteq S_A$. As a result, there is a unique saddle point, i.e., $S_{\text{saddle}} = S_B \cap S_A = S_B = \{(1, 0)\}$.

4. First-order stationarity conditions. In section 2 we established that finding the globally optimal solutions of problem (mM-I) is NP-hard in general. It is then useful to identify some high-quality solution concepts that can be computed efficiently. In this section, we propose to leverage the duality theory developed in the previous section to identify such a class of first-order stationary solutions. More specifically, we will define first-order stationary solutions based on the dual problem (D2). Note that one can directly work with the problem $\phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y})$, but this is challenging since $\phi(\mathbf{x})$ is neither smooth nor convex.

Before we proceed let us make a few additional assumptions.

Assumption 4.1. We impose the following assumptions:

1. The function $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable w.r.t. both $\mathbf{x}$ and $\mathbf{y}$.
2. The function $f(\mathbf{x}, \mathbf{y})$ is strongly convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{Y}$ with modulus μ_x .
3. The function f has Lipschitz continuous gradients, i.e.,

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x}_2, \mathbf{y}_2)\| \leq L_x \left\| \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{bmatrix} - \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix} \right\| \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}.$$

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)\| \leq L_y \left\| \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{bmatrix} - \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix} \right\| \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}.$$

Next, let us define the following quantities:

$$(4.1a) \quad H(\mathbf{x}, \boldsymbol{\lambda}) := \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}),$$

$$(4.1b) \quad G(\boldsymbol{\lambda}) := \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}),$$

$$(4.1c) \quad \bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda}) := \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}), \quad \bar{\mathbf{x}}(\boldsymbol{\lambda}) := \arg \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \boldsymbol{\lambda}).$$

Using these new notations, problem (D2) can be decomposed into the following two problems:

$$(4.2) \quad \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \boldsymbol{\lambda}) := \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \right) \quad (\text{inner-level problem}),$$

$$(4.3) \quad \min_{\boldsymbol{\lambda} \geq \mathbf{0}} G(\boldsymbol{\lambda}) := \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \left(\min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \boldsymbol{\lambda}) \right) \quad (\text{outer-level problem}).$$

Note that the assumption that f is strongly convex in $\mathbf{x}$ provides us with a number of useful properties which facilitate algorithm design. Specifically, it implies that the function $G(\boldsymbol{\lambda})$ defined above, which is the objective of the dual problem (D2), is differentiable. However, notice that $G(\boldsymbol{\lambda})$ is not necessarily convex and therefore any reasonable notion of stationary conditions for (D2) will involve the gradient and the stationary points of $G(\boldsymbol{\lambda})$. Therefore, below we derive the formula of the gradient of $G(\boldsymbol{\lambda})$.

LEMMA 4.2. *Suppose that Assumptions 2.2, 3.3, and 4.1 hold. Then $G(\boldsymbol{\lambda})$ is differentiable, and its gradient is given by*

$$\nabla G(\boldsymbol{\lambda}) = -A\bar{\mathbf{x}}(\boldsymbol{\lambda}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) + \mathbf{c},$$

where $\bar{\mathbf{x}}(\boldsymbol{\lambda}), \bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda})$ are defined in (4.1c).

Proof. See Appendix A.2. □

Now we are ready to provide definitions of (exact and approximate) stationary solutions, based on the dual problem (D2).

DEFINITION 4.3 (stationary solutions). *A solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\boldsymbol{\lambda}})$ is an exact stationary solution of (D2) if the following hold:*

- $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\mathbf{y}}(\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}}))$;
- $\bar{\boldsymbol{\lambda}}$ is a stationary point of $G(\boldsymbol{\lambda})$, that is, $\langle \nabla G(\bar{\boldsymbol{\lambda}}), \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \geq 0 \quad \forall \boldsymbol{\lambda} \geq \mathbf{0}$.

We will later show that the stationarity conditions for $G(\boldsymbol{\lambda})$ imply the complementarity condition $\bar{\boldsymbol{\lambda}}^T \nabla G(\bar{\boldsymbol{\lambda}}) = 0$ with $\boldsymbol{\lambda} \geq \mathbf{0}$. Next, we provide the definition of the approximate stationary solutions.

DEFINITION 4.4 ((ϵ, δ) -approximate stationary solution). *Let us define the following quantities:*

$$Q(\boldsymbol{\lambda}) := \frac{1}{\alpha} \left(\boldsymbol{\lambda} - \text{proj}_{\mathbb{R}_+^k}(\boldsymbol{\lambda} - \alpha \nabla G(\boldsymbol{\lambda})) \right),$$

$$d(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) := \|\bar{\mathbf{x}}(\boldsymbol{\lambda}) - \mathbf{x}\|^2 + \|\bar{\mathbf{y}}(\bar{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) - \mathbf{y}\|^2,$$

where $\alpha > 0$ is a constant independent of ϵ and δ . Then, a point $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}}) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}_+^k$ is called an (ϵ, δ) -approximate stationary solution if it holds that

$$\|Q(\tilde{\boldsymbol{\lambda}})\| \leq \epsilon, \quad d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}}) \leq \delta^2.$$

In words, at an exact stationary solution of (D2), the inner-level minimax problem is solved to global optimality, while the outer-level one reaches an exact stationary solution. However, finding such points in practice is unrealistic. The notion of approximate stationary solutions is then introduced as a relaxation. At an approximate stationary solution $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}})$ the inner-level problem is solved inexactly, in the sense

that the distance of $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ from the exact solution (of the inner-level problem (4.2) with $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$) is small. In addition, the function $Q(\tilde{\boldsymbol{\lambda}})$ is the stationarity gap of the outer-level problem at $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$.

Next, we analyze the implications of a point being an (exact or approximate) stationary solution. In particular, we will show that at an exact stationary solution, the coupled linear constraints as well as the complementary slackness condition will be satisfied. Further, at an approximate stationary solution, the choice of $\tilde{\boldsymbol{\lambda}}$ ensures that the constraint violation is small.

PROPOSITION 4.5. *Suppose that Assumptions 2.2, 3.3, 4.1 hold, and let $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\boldsymbol{\lambda}})$ be a stationary solution of (D2). Then the following hold:*

$$(4.4a) \quad \bar{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\bar{\mathbf{x}}, \mathbf{y}, \bar{\boldsymbol{\lambda}}), \quad A\bar{\mathbf{x}} + B\bar{\mathbf{y}} - \mathbf{c} \leq \mathbf{0}, \quad \bar{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{H(\mathbf{x}, \bar{\boldsymbol{\lambda}})\},$$

$$(4.4b) \quad \mathbf{0} = \bar{\boldsymbol{\lambda}}^T (-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}), \quad \bar{\boldsymbol{\lambda}} \geq \mathbf{0}.$$

Proof. See Appendix A.2. □

PROPOSITION 4.6. *Suppose that Assumptions 2.2, 3.3, 4.1 hold, and let us denote with $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}}) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}_+^k$ an (ϵ, δ) -approximate stationary solution. Then, $\forall i \in \mathcal{K}$ it holds that*

$$\begin{aligned} \max\{0, [A\tilde{\mathbf{x}} + B\tilde{\mathbf{y}} - \mathbf{c}]_i\} &\leq 2\sigma_{\max} \cdot \delta + \epsilon, \\ \|\bar{\mathbf{y}} - \tilde{\mathbf{y}}\| &\leq \delta, \quad \|\bar{\mathbf{x}} - \tilde{\mathbf{x}}\| \leq \delta, \end{aligned}$$

where $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \bar{\mathbf{y}}(\bar{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}))$, and the two functions $\bar{\mathbf{x}}(\cdot), \bar{\mathbf{y}}(\cdot)$ are defined in (4.1c); $\sigma_{\max} := \max\{\|A\|, \|B\|\}$.

Proof. See Appendix A.2. □

So far, we have defined a suitable (dual) reformulation of the original problem (mM-I), and introduced a set of first-order stationarity conditions for it. To evaluate the quality of these solutions, we introduce an algorithm named multiplier gradient descent (MGD) that can efficiently compute such solutions. Next, we provide a description of the proposed algorithm.

First, we obtain an approximate solution of the inner-level problem (4.2) by using any reasonable iterative subroutine that can solve a strongly-convex-strongly-concave minimax problem. In Algorithm 4.1, such a subroutine is referred to as “Alg(·),” which takes the current iterates $\mathbf{x}^r, \mathbf{y}^r, \boldsymbol{\lambda}^r$ as well as the desired number of inner iterations as input. Such an algorithm can be, for instance, the gradient descent-(multistep) ascent [37] or the extragradient method [34]. Afterwards, one iteration of gradient descent is performed on the outer problem (4.3). In the subsequent discussion, we refer to r as the *outer-iteration* index.

We should note here that the MGD algorithm resembles the classical dual-ascent algorithm (for constrained minimization problems), and its analysis also follows similar ideas. The interested readers are referred to the online version of this paper for the formal convergence claims and the proof [46]. We emphasize that algorithm design and analysis are *not* the central focus of this work. The MGD algorithm just gives us a way to evaluate the quality of the proposed stationary solution. One may certainly be able to design more efficient algorithms (for example, based on extra-gradient, or momentum acceleration techniques), but this is beyond the scope of this work.

Below we briefly comment on the convergence properties of MGD.

Algorithm 4.1. MGD.

Input: $\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0, T, K, \alpha, A, B, \mathbf{c}$
for $r = 0$ **to** $T - 1$ **do**
 # Solve inner minimax problem by running Alg using K iterations;
 # e.g., Alg: gradient descent-(multistep) ascent;
 $(\mathbf{x}^{r+1}, \mathbf{y}^{r+1}) \leftarrow \text{Alg}(\mathbf{x}^r, \mathbf{y}^r, \boldsymbol{\lambda}^r; K)$
 # Solve outer min problem
 $\boldsymbol{\lambda}^{r+1} = \text{proj}_{\mathbb{R}_+^k} [\boldsymbol{\lambda}^r - \alpha (-A\mathbf{x}^{r+1} - B\mathbf{y}^{r+1} + \mathbf{c})]$
end for

Remark 4.7. First of all, under the Assumptions 2.2, 3.3, and 4.1 we can show that the outer-level objective $G(\boldsymbol{\lambda})$ has Lipschitz continuous gradients. Based on this result, we can show (under the same assumptions as above) that the MGD algorithm can asymptotically compute an *exact stationary solution*, i.e.,

$$\lim_{r \rightarrow \infty} \|Q(\boldsymbol{\lambda}^r)\| = 0, \quad \lim_{r \rightarrow \infty} d(\mathbf{x}^r, \mathbf{y}^r, \boldsymbol{\lambda}^r) = 0.$$

This result implies that the properties of Proposition 4.5 are also satisfied asymptotically, in particular, the complementary slackness condition holds, i.e.,

$$\lim_{r \rightarrow \infty} \langle \boldsymbol{\lambda}^r, A\bar{\mathbf{x}}(\boldsymbol{\lambda}^r) + B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\boldsymbol{\lambda}^r), \boldsymbol{\lambda}^r) - \mathbf{c} \rangle = 0.$$

Moreover, we can prove that MGD can reach an (ϵ, δ) -approximate stationary solution using at most $\mathcal{O}(\frac{1}{\epsilon^2} + \frac{1}{\sqrt{\delta}})$ outer iterations. However, in this case in order to ensure that the complementary slackness violation also goes to zero (as a function of ϵ and δ), we need to impose the following regularity assumption: for every $\mathbf{x} \in \mathcal{X}$, 0 is in the interior of the set $\{A\mathbf{x} + B\mathbf{y} - \mathbf{c} \mid \mathbf{y} \in \mathcal{Y}\}$; note that this assumption is only slightly more restrictive than the one made in Assumption 3.3.2. Then, we can show that at an (ϵ, δ) -approximate stationary solution $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}})$ obtained by the MGD algorithm, the complementary slackness violation $\langle \tilde{\boldsymbol{\lambda}}, A\tilde{\mathbf{x}} + B\tilde{\mathbf{y}} - \mathbf{c} \rangle$ is upper and lower bounded by functions of ϵ and δ , and these functions converge to zero as $\epsilon \rightarrow 0, \delta \rightarrow 0$.

5. Experiments. In this section we perform a number of experiments that develop attacks for a minimum cost network flow problem, using the proposed formulation (1.2). Note that our main goal is to evaluate the quality of the formulation and the proposed stationary solutions. To begin with, in our experiments we adopt the setting described in section 1.1.2, and generate networks of n nodes at random using the Erdős–Rényi model with parameter p (i.e., the probability that an edge appears on the graph is p). Moreover, the capacity p_{ij} and the cost coefficients w_{ij} of the edges are generated uniformly at random in the interval $[1, 2]$, while we set the demand (on the sink) as $d\%$ of the sum of capacities of the edges exiting the source, and d is a parameter to be chosen.

The adversary will generate the attack by solving problem (1.2). We choose $q_e(x_e) = w_e x_e$, that is, the cost per unit flow is a function of the amount of flow. Further, we add a small regularizer to the adversarial's problem to make it strongly concave. The specific problem to be solved is listed below ($\eta > 0$ is a small constant):

$$\begin{aligned}
(5.1) \quad & \max_{\substack{0 \leq \mathbf{y} \leq \mathbf{p} \\ \sum_{(i,j) \in E} y_{ij} = b}} \min_{\substack{0 \leq \mathbf{x} \leq \mathbf{p} \\ \sum_{(i,t) \in E} x_{it} = r_t}} \sum_{(i,j) \in E} w_{ij} \cdot (x_{ij} + y_{ij}) \cdot x_{ij} - \frac{\eta}{2} \|\mathbf{y}\|^2 \\
& \text{s.t. } \mathbf{x} + \mathbf{y} \leq \mathbf{p}, \\
& \sum_{(i,j) \in E} x_{ij} - \sum_{(j,k) \in E} x_{jk} = 0 \quad \forall j \in V \setminus \{s, t\}.
\end{aligned}$$

The above problem is then solved using the following two different approaches, while the algorithm parameters are chosen so that the performances are optimized:

1. Apply the MGD algorithm (Algorithm 4.1) to solve problem (D2). Choose $T = 100$, and use $K = 5$ steps of the gradient descent-ascent method [26] to solve the inner-problem. The step sizes for all variables are set to 0.5.
2. Apply the gradient descent-ascent algorithm (with multiple descent steps) [37] to solve problem (D3). Choose $T = 100$, and run 5 iterations for the inner minimization problem. The step sizes for all variables are set to 0.5. We refer to this approach as GDA.

To evaluate the performance of the proposed attack we test the following baselines:

- *Random attack*: An attack flow vector $\mathbf{x}$ is generated at random, under the given budget constraints b .
- *Max capacity attack*: The attack removes the edge with the largest capacity.
- *Greedy attack*: The edges are sorted in an ascending order w.r.t. the cost coefficients, and they are removed following this order until the capacity budget b is met. If the remaining budget does not suffice to remove an edge (i.e., zero its capacity), then the capacity of the edge is reduced by this amount.
- *NI attack*: We generate an attack following a method that is used in (generalized) Nash equilibria games. Specifically, we minimize a Nikaido–Isoda (NI)-based reformulation [12, 47] of our problem, that is

$$(5.2a) \quad V(\mathbf{x}, \mathbf{y}) = -V_1(\mathbf{y}) - V_2(\mathbf{x}),$$

$$(5.2b) \quad \text{where } V_1(\mathbf{y}) := \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}), \quad V_2(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}).$$

We approximately solve the (strongly-convex or strongly-concave) problems in (5.2b) using 25 steps of gradient descent and gradient ascent, respectively.

Then, the objective V is minimized by using 100 steps of gradient descent.

The performance of the experiments are evaluated according to the following procedure. First, an attack is generated, and the corresponding link capacities are reduced according to the attack pattern. Then, the minimum cost flow is computed based on the available link capacities. Let $\mathbf{x}_{\text{att}}, \mathbf{x}_{\text{cl}}$ denote the minimum cost flow assignment after and before the attack, respectively. We define the following *relative* increase of the cost as a performance measure: $\rho := \frac{q_{\text{tot}}(\mathbf{x}_{\text{att}}) - q_{\text{tot}}(\mathbf{x}_{\text{cl}})}{q_{\text{tot}}(\mathbf{x}_{\text{cl}})}$. The higher the increase the more successful/powerful the attack. Furthermore, in our experiments we compute the above measure as a function of the adversary's budget, and for each budget level we perform the experiment 15 times (each time a new graph is created, with different capacity and cost vectors) and average the results.

The results are illustrated in Figure 1. Observe that the proposed minimax based attack is more powerful compared to the three baselines, as it leads to the largest minimum cost flow after the attack in both networks (regardless of which algorithm we use). This also implies that the stationary solutions defined in the previous section

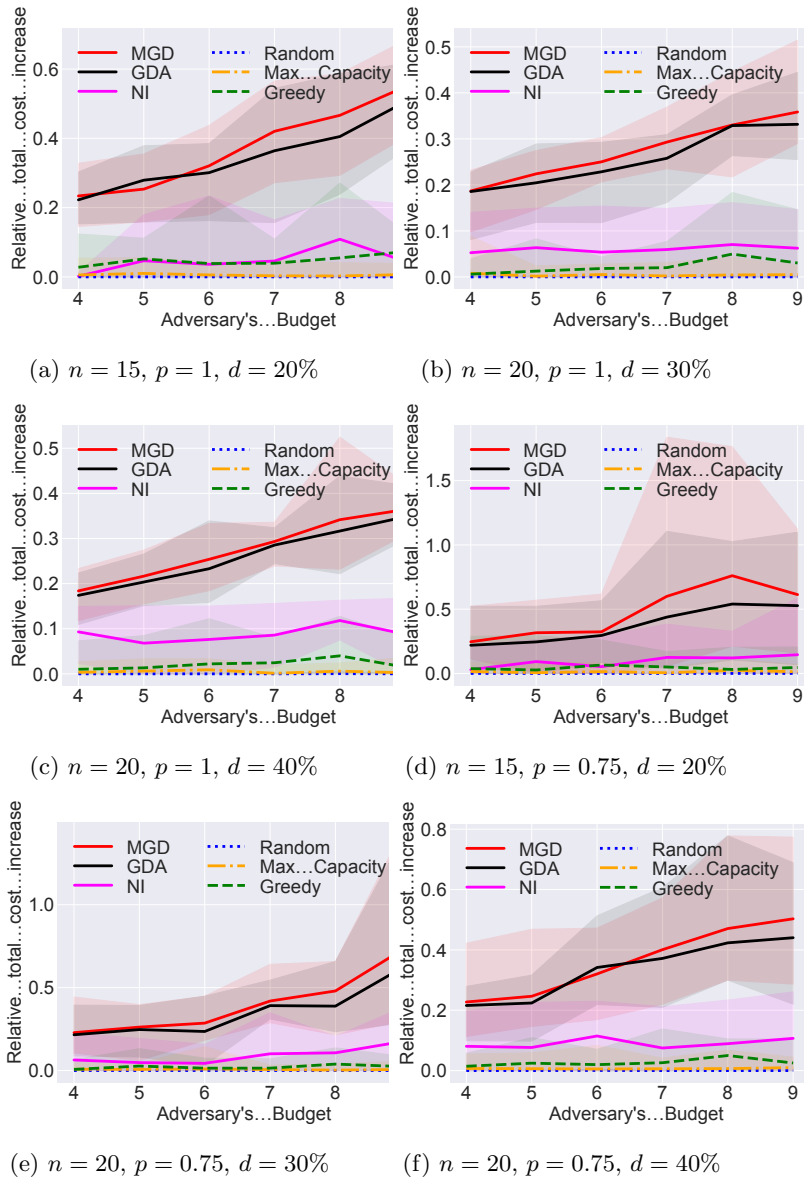


FIG. 1. The evolution of the relative increase (due to the attack) of the total cost as a function of the budget of the adversary, for different settings. The range of the results (between minimum and maximum values across all runs) is also depicted using a shaded region around the average cost curve.

correspond to relatively “strong” attacks. We also notice that MGD is slightly more advantageous compared to the GDA. It is worth mentioning that the network we generated makes the attack problem *nontrivial*, in the sense that a random attack does not work (the relative cost increase is imperceptible). Further, we would like to point out that the relatively high variance exhibited in the experiments is not unexpected since at every run we generate a different graph. In particular, in the experiments with $p = 0.75$, every new graph differs significantly from the previous one, leading to higher variance in this case.

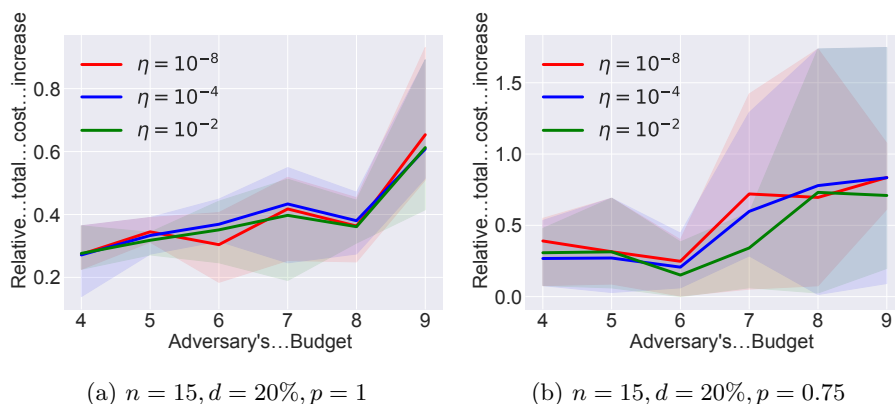


FIG. 2. The evolution of the relative increase (due to the attack) of the total cost as a function of the budget of the adversary, for different values of the regularization parameter η . The range of the results (between the minimum and maximum values across all runs) is also depicted using a shaded region around the average cost curve.

5.1. Effect of the regularization parameter. We perform an additional set of experiments, in which we study the effect of the regularization parameter η on the problem solution by examining the performance of the MGD algorithm. We would like to stress here that we select relatively small values of η , because the regularizer was introduced to impose the strong concavity assumption in $\mathbf{y}$ without significantly altering the original problem. In Figure 2, we plot the evolution of the relative increase of the total cost as a function of the adversary's budget for $\eta = 10^{-8}, 10^{-4}, 10^{-2}$. We consider two different settings by changing the parameter p , while we only consider graphs with $n = 15$ nodes; we also use the same step size and the same number of inner iterations as in the previous experiments.

In Figure 2 we see that it is not clear which value of η leads to the best performance. Therefore, we cannot reach a clear conclusion about the effect of the size of the regularization parameter. However, we also notice that the performance (such as the relative total cost increase) attained does not differ significantly among the different values of η , as long as the regularization parameter η is small enough (such as the ones we test in our experiments, i.e., $\eta \leq 10^{-2}$). This behavior is desirable, since our goal is only to make the objective strongly concave in $\mathbf{y}$ in a way that the original problem (and thus its solutions) is not altered significantly.

6. Concluding remarks. In this paper we study a minimax problem, where the constraints of the inner maximization problem depends linearly on the variable of the outer one. This special structure makes the problem NP-hard, even when the objective is strongly-convex-strongly-concave. We then develop the duality theory of the problem, and establish conditions under which strong duality holds. These results allow us to introduce duality-based reformulations of the original problem, and propose a set of first-order stationarity conditions. Moreover, we demonstrate that the proposed formulation can model adversarial attacks on network flow problems. In the future, we plan to study more applications of this formulation, as well as to related problems where the linear constraints are in the outer problem.

It is possible to extend the main results (e.g., the duality theory, the definition and properties of stationary points) to problems with general, not necessarily linear, constraints of the form $g(\mathbf{x}, \mathbf{y}) \leq 0$, where $g: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^k$. This is possible with addition of the following assumptions (below we provide only the key assumptions):

1. $g_i(\cdot, \mathbf{y})$ is concave $\forall \mathbf{y} \in \mathcal{Y}$ and $g_i(\mathbf{x}, \cdot)$ is convex $\forall \mathbf{x} \in \mathcal{X} \forall i \in \{1, \dots, k\}$.
2. For every $\mathbf{x} \in \mathcal{X}$ there exists $\mathbf{y} \in \text{relint}(\mathcal{Y})$ such that $g(\mathbf{x}, \mathbf{y}) < 0$.
3. $|g_i(\mathbf{x}_1, \mathbf{y}_1) - g_i(\mathbf{x}_2, \mathbf{y}_2)| \leq L[\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|] \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ for some $L > 0 \forall i \in \{1, \dots, k\}$.

Then, the duality theory developed in Theorems 3.2 and 3.4 will hold. Also, the definition of the (exact and approximate) stationary points in Definitions 4.3 and 4.4, as well as their respective properties in Propositions 4.5, 4.6, will apply; we just need to substitute the linear constraints $A\mathbf{x} + B\mathbf{y} + \mathbf{c} \leq 0$ with the more general ones $g(\mathbf{x}, \mathbf{y}) \leq 0$. Finally, let us note that the NP-hardness results (Propositions 2.5 and 2.6) and the relations between the coupled constrained problems (Proposition (2.4)) automatically apply in the case of the general constraints $g(\mathbf{x}, \mathbf{y}) \leq 0$. Note that despite the ability to generalize our results, in this work we focus on linear constraints, since the changes required for this generalization are mainly technical and do not lead to new insights to the problem itself.

Appendix A. Proofs.

A.1. Minimax problems with coupled linear constraints.

Proof of Proposition 2.4. Below we provide the proof for the relations described in Proposition 2.4.

$v(\text{mM-I}) \leq v(\text{mM-O})$:

For the function $\phi(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y})$ defined in (2.1b), it holds that

$$\phi(\mathbf{x}) \leq \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{x} \in \mathcal{X}.$$

This is because the RHS has a larger constraints set. Then, minimizing both sides of the above inequality, we obtain

$$(A.1) \quad \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \phi(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right),$$

where $\mathbf{y}^*(\mathbf{x})$ is defined in (mM-O).

Similarly, observe that the following holds:

$$(A.2) \quad \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \phi(\mathbf{x}).$$

Combining (A.1) and (A.2) implies that:

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}) &\leq \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right) \\ &\stackrel{(2.1b)}{\Rightarrow} \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right) \leq \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right). \end{aligned}$$

That is, $v(\text{mM-I}) \leq v(\text{mM-O})$.

To show that the strict inequality can also hold, consider the following problems, in which it is shown that $v(\text{mM-I}) = -1$ and $v(\text{mM-O}) = 1$:

$$\begin{aligned} \min_{x \in [0, 1]} \left(\max_{y \in [0, 1], x+y=1} x^2 - y^2 \right) &= \min_{x \in [0, 1]} x^2 - (1-x)^2 = \min_{x \in [0, 1]} 2x - 1 = -1, \\ \min_{x \in [0, 1], x+y^*(x)=1} \left(\max_{y \in [0, 1]} x^2 - y^2 \right) &= \min_{x \in [0, 1], x=1} x^2 = 1. \end{aligned}$$

Therefore, in the above example it holds that $v(\text{mM-I}) < v(\text{mM-O})$. This allows us to exclude the possibility that equality holds between (mM-I) and (mM-O) for all possible problem instances.

$v(\text{Mm-I}) \geq v(\text{Mm-O})$:

Let us define $\psi(\mathbf{y}) := \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y})$; then it holds that

$$\psi(\mathbf{y}) \geq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{Y},$$

since the RHS has a larger constraints set. Then, we can maximize both sides of the above inequality, and obtain:

$$(A.3) \quad \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \psi(\mathbf{y}) \geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right),$$

where $\mathbf{x}^*(\mathbf{y})$ is defined in (Mm-O). In addition, we can easily see that

$$(A.4) \quad \max_{\mathbf{y} \in \mathcal{Y}} \psi(\mathbf{y}) \geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \psi(\mathbf{y}).$$

Combining (A.3) and (A.4) we obtain

$$\begin{aligned} \max_{\mathbf{y} \in \mathcal{Y}} \psi(\mathbf{y}) &\geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right) \\ \Rightarrow \max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y} \leq \mathbf{c}} f(\mathbf{x}, \mathbf{y}) \right) &\geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right). \end{aligned}$$

That is, $v(\text{Mm-I}) \geq v(\text{Mm-O})$.

To show that the strict inequality can also hold, consider the following problems, in which it is shown that $v(\text{Mm-I}) = 1$, and $v(\text{Mm-O}) = -1$:

$$\begin{aligned} \max_{y \in [0,1]} \left(\min_{x \in [0,1], x+y=1} x^2 - y^2 \right) &= \max_{y \in [0,1]} (1-y)^2 - y^2 = \max_{y \in [0,1]} -2y + 1 = 1, \\ \max_{y \in [0,1], x^*(y)+y=1} \left(\min_{x \in [0,1]} x^2 - y^2 \right) &= \max_{y \in [0,1], y=1} -y^2 = -1. \end{aligned}$$

In the above example it holds that $v(\text{Mm-I}) > v(\text{Mm-O})$. This allows us to exclude the possibility that equality can hold between (Mm-I) and (Mm-O) for all possible problem instances.

$v(\text{mM-O}) \geq v(\text{Mm-O})$:

To begin with, observe that

$$(A.5) \quad \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right) \geq \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right),$$

since the outer problem in the RHS has a larger constraints set. Further, the max-min inequality implies that:

$$(A.6) \quad \min_{\mathbf{x} \in \mathcal{X}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right) \geq \max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right).$$

Moreover, we have that:

$$(A.7) \quad \max_{\mathbf{y} \in \mathcal{Y}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right) \geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right).$$

Combining equations (A.5), (A.6), (A.7), we obtain the following relation:

$$(A.8) \quad \min_{\mathbf{x} \in \mathcal{X}, A\mathbf{x} + B\mathbf{y}^*(\mathbf{x}) \leq \mathbf{c}} \left(\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \right) \geq \max_{\mathbf{y} \in \mathcal{Y}, A\mathbf{x}^*(\mathbf{y}) + B\mathbf{y} \leq \mathbf{c}} \left(\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}) \right).$$

That is, $v(\text{mM-O}) \geq v(\text{Mm-O})$.

$v(\text{Mm-I})$, $v(\text{Mm-O})$:

To begin with, for $A = B = c = 0$ problems (mM-I) and (Mm-O) reduce to classical (without coupled constraints) minimax problems. The max-min inequality implies that $v(\text{mM-I}) \geq v(\text{Mm-O})$.

Moreover, consider the following problems, in which it is shown that $v(\text{mM-I}) = -3$ and $v(\text{Mm-O}) = -1$:

$$\begin{aligned} \min_{x \in [-1, 1]} \left(\max_{y \in [0, 2], x+y=1} x^2 - y^2 \right) &= \min_{x \in [-1, 1]} x^2 - (1-x)^2 = \min_{x \in [-1, 1]} 2x - 1 = -3, \\ \max_{y \in [0, 2], x^*(y) + y = 1} \left(\min_{x \in [-1, 1]} x^2 - y^2 \right) &= \max_{y \in [0, 2], y=1} -y^2 = -1. \end{aligned}$$

The above examples show that there are problem instances where the relationship $v(\text{mM-I}) < v(\text{Mm-O})$ holds between the two problems.

$v(\text{Mm-I})$, $v(\text{mM-O})$:

To begin with, for $A = B = c = 0$ problems (Mm-I) and (mM-O) reduce to the classical (without coupled constraints) minimax problems. The max-min inequality implies that $v(\text{Mm-I}) \leq v(\text{mM-O})$.

Moreover, consider the following problems, in which it is shown that $v(\text{Mm-I}) = 3$ and $v(\text{mM-O}) = 1$:

$$\begin{aligned} \max_{y \in [-1, 0]} \left(\min_{x \in [1, 2], x+y=1} x^2 - y^2 \right) &= \max_{y \in [-1, 0]} (1-y)^2 - y^2 = \max_{y \in [-1, 0]} -2y + 1 = 3, \\ \min_{x \in [1, 2], x+y^*(x) = 1} \left(\max_{y \in [-1, 0]} x^2 - y^2 \right) &= \min_{x \in [1, 2], x=1} x^2 = 1. \end{aligned}$$

The above examples show that there are problem instances where the relationship $v(\text{Mm-I}) > v(\text{mM-O})$ is realized between the two problems.

$v(\text{Mm-I})$, $v(\text{Mm-I})$:

To begin with, for $A = B = c = 0$ problems (mM-I) and (Mm-I) reduce to the classical (without coupled constraints) minimax problems. The max-min inequality holds, that is $v(\text{Mm-I}) \geq v(\text{Mm-I})$.

Moreover, consider the following problems, in which it is shown that $v(\text{Mm-I}) = -1$ and $v(\text{Mm-I}) = 1$:

$$(A.9) \quad \min_{x \in [0, 1]} \left(\max_{y \in [0, 1], x+y=1} x^2 - y^2 \right) = \min_{x \in [0, 1]} x^2 - (1-x)^2 = \min_{x \in [0, 1]} -1 + 2x = -1,$$

$$(A.10) \quad \max_{y \in [0, 1]} \left(\min_{x \in [0, 1], x+y=1} x^2 - y^2 \right) = \max_{y \in [0, 1]} (1-y)^2 - y^2 = \max_{y \in [0, 1]} 1 - 2y = 1.$$

The above examples show that there are problem instances where the relationship $v(\text{mM-I}) < v(\text{Mm-O})$ is realized between the two problems. $\square$

A.2. First-order stationarity conditions. In the proof of Lemma 4.2 we make use of Danskin's theorem [3, Corollary, page 1167]. Below we provide its formal statement.

LEMMA A.1 (Danskin's theorem). Consider the problem $\phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$, where $f: \mathbb{R}^n \times Y \rightarrow \mathbb{R}$ is differentiable and strongly concave in $\mathbf{y}$, and $\mathcal{Y} \subseteq \mathbb{R}^m$ is a nonempty compact set. Also, consider the set $\mathcal{Y}^*(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ which is a singleton for any $\mathbf{x} \in \mathbb{R}^n$. Then, $\phi(\mathbf{x})$ is differentiable and it holds that $\nabla \phi(\mathbf{x}) = \nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, where $\mathbf{y}^*(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$.

Proof of Lemma 4.2. To begin with, we will compute the gradient of function $H(\mathbf{x}, \boldsymbol{\lambda}) = \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$, defined in (4.1a). Indeed, notice that $\mathcal{Y}$ is a compact set, L is differentiable and has a unique maximum due to its strong concavity in $\mathbf{y}$. Then, from Danskin's theorem [3] we can infer that $H(\mathbf{x}, \boldsymbol{\lambda})$ is differentiable with

$$(A.11) \quad \nabla H(\mathbf{x}, \boldsymbol{\lambda}) = \nabla L(\mathbf{x}, \bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda}), \boldsymbol{\lambda}) = \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda})) - A^T \boldsymbol{\lambda} \\ -A\mathbf{x} - B\bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda}) + \mathbf{c} \end{bmatrix},$$

where $\bar{\mathbf{y}}(\mathbf{x}, \boldsymbol{\lambda}) := \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$.

Next we will show that $H(\mathbf{x}, \boldsymbol{\lambda})$ is strongly convex in $\mathbf{x}$, with modulus μ_x . Indeed for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $\rho \in [0, 1]$ we have the following series of relations:

$$\begin{aligned} H(\rho \mathbf{x}_1 + (1 - \rho) \mathbf{x}_2, \boldsymbol{\lambda}) &= \max_{\mathbf{y} \in \mathcal{Y}} \{L(\rho \mathbf{x}_1 + (1 - \rho) \mathbf{x}_2, \mathbf{y}, \boldsymbol{\lambda})\} \\ &\stackrel{(a)}{=} \max_{\mathbf{y} \in \mathcal{Y}} \left\{ f(\rho \mathbf{x}_1 + (1 - \rho) \mathbf{x}_2, \mathbf{y}) - \boldsymbol{\lambda}^T (A(\rho \mathbf{x}_1 + (1 - \rho) \mathbf{x}_2) + B\mathbf{y} - \mathbf{c}) \right\} \\ &\stackrel{(b)}{\leq} \max_{\mathbf{y} \in \mathcal{Y}} \left\{ \rho f(\mathbf{x}_1, \mathbf{y}) + (1 - \rho) f(\mathbf{x}_2, \mathbf{y}) - \frac{\mu_x}{2} \rho(1 - \rho) \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \right. \\ &\quad \left. - \rho \boldsymbol{\lambda}^T A \mathbf{x}_1 - (1 - \rho) \boldsymbol{\lambda}^T A \mathbf{x}_2 - \boldsymbol{\lambda}^T B \mathbf{y} + \boldsymbol{\lambda}^T \mathbf{c} \right\} \\ &\stackrel{(c)}{\leq} \max_{\mathbf{y} \in \mathcal{Y}} \left\{ \rho f(\mathbf{x}_1, \mathbf{y}) - \rho \boldsymbol{\lambda}^T (A \mathbf{x}_1 + B \mathbf{y} - \mathbf{c}) \right\} - \frac{\mu_x}{2} \rho(1 - \rho) \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ &\quad + \max_{\mathbf{y} \in \mathcal{Y}} \left\{ (1 - \rho) f(\mathbf{x}_2, \mathbf{y}) - (1 - \rho) \boldsymbol{\lambda}^T (A \mathbf{x}_2 + B \mathbf{y} - \mathbf{c}) \right\} \\ &\leq \rho H(\mathbf{x}_1, \boldsymbol{\lambda}) + (1 - \rho) H(\mathbf{x}_2, \boldsymbol{\lambda}) - \frac{\mu_x}{2} \rho(1 - \rho) \|\mathbf{x}_1 - \mathbf{x}_2\|^2, \end{aligned}$$

where in (a) we use the definition of the Lagrangian from (3.1), in (b) the inequality follows from the strong convexity of f w.r.t. $\mathbf{x}$ (with modulus μ_x), and in (c) the triangle inequality of the max operator was used.

Moreover, note that

$$(A.12) \quad G(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \boldsymbol{\lambda}) = -\max_{\mathbf{x} \in \mathcal{X}} \{-H(\mathbf{x}, \boldsymbol{\lambda})\}.$$

Using the fact that $\mathcal{X}$ is a compact set, $-H$ is differentiable and strongly concave in $\mathbf{x}$, we can apply the Danskin's theorem [3] and obtain

$$\nabla G(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} H(\bar{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = -A\bar{\mathbf{x}}(\boldsymbol{\lambda}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) + \mathbf{c},$$

where in the last equality we used the gradient of function $H(\mathbf{x}, \boldsymbol{\lambda})$ from (A.11), and the definition of $\bar{\mathbf{x}}(\boldsymbol{\lambda}) = \arg \min_{\mathbf{x} \in \mathcal{X}} H(\mathbf{x}, \boldsymbol{\lambda})$. The proof is now completed. $\square$

Proof of Proposition 4.5. Let $\bar{\boldsymbol{\lambda}} \geq \mathbf{0}$ be a stationary point of $G(\boldsymbol{\lambda})$, that is

$$\langle \nabla G(\bar{\boldsymbol{\lambda}}), \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \geq 0 \quad \forall \boldsymbol{\lambda} \geq \mathbf{0}.$$

Substituting the formula of the gradient of $G(\boldsymbol{\lambda})$ from Lemma 4.2 gives

$$(A.13) \quad \langle -A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}, \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \geq 0 \quad \forall \boldsymbol{\lambda} \geq \mathbf{0}.$$

We will first show that these conditions imply that $A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) + B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) \leq \mathbf{c}$. Suppose that there exists index $i \in \mathcal{K}$ such that $[-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i < 0$. Then, if we select $\lambda_i = \bar{\lambda}_i + \epsilon$ for some $\epsilon > 0$, and $\lambda_j = \bar{\lambda}_j \forall j \neq i$, (A.13) becomes

$$\begin{aligned} & \langle -A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}, \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \\ &= [-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i (\bar{\lambda}_i + \epsilon - \bar{\lambda}_i) \\ &= [-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i \epsilon < 0. \end{aligned}$$

This is a contradiction. As a result, $[-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i \geq 0 \forall i \in \mathcal{K}$.

Next, we show that the following holds:

$$(A.14) \quad \bar{\lambda}_i [-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i = 0 \quad \forall i \in \mathcal{K}.$$

If $[-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i = 0$, then the above equality holds trivially. Suppose that there exists $[-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i > 0$ such that $\bar{\lambda}_i > 0$. Then, notice that we can select $\lambda_i = \bar{\lambda}_i - \epsilon$ (for some sufficiently small $\epsilon > 0$) and $\lambda_j = \bar{\lambda}_j \forall j \neq i$, substitute them into condition (A.13), and obtain

$$\begin{aligned} & \langle -A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}, \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \\ &= [-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i (\bar{\lambda}_i - \epsilon - \bar{\lambda}_i) \\ &= [-A\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}) + \mathbf{c}]_i (-\epsilon) < 0, \end{aligned}$$

which is a contradiction to (A.13). This completes the proof of (4.4b).

From Definition 4.3 we know that at a stationary point $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\boldsymbol{\lambda}})$ it holds that $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}), \bar{\boldsymbol{\lambda}}))$. Then, the optimality condition of $\bar{\mathbf{y}}$ follows directly from the definition of $\bar{\mathbf{y}}(\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}})$ [see (4.1c)], i.e.,

$$(A.15) \quad \bar{\mathbf{y}} = \bar{\mathbf{y}}(\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}}) = \arg \max_{\mathbf{y} \in \mathcal{Y}} L(\bar{\mathbf{x}}, \mathbf{y}, \bar{\boldsymbol{\lambda}}).$$

Similarly, from (4.1c), we have $\bar{\mathbf{x}} = \bar{\mathbf{x}}(\bar{\boldsymbol{\lambda}}) = \arg \min_{\mathbf{x} \in \mathcal{X}} \{H(\mathbf{x}, \bar{\boldsymbol{\lambda}})\}$. This concludes the proof. $\square$

Proof of Proposition 4.6. Let $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}})$ be an (ϵ, δ) -approximate stationary solution. In order to compute a bound for the (magnitude of the) violation of the constraint at $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}})$, we consider the following quantity, for an arbitrary $i \in \mathcal{K}$:

$$\begin{aligned} \max \{0, [A\tilde{\mathbf{x}} + B\tilde{\mathbf{y}} - \mathbf{c}]_i\} &= \max \left\{ 0, \left[A \left(\tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right) + B \left(\tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right) \right]_i \right. \\ &\quad \left. + \left[A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) + B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \mathbf{c} \right]_i \right\} \\ &\leq \max \left\{ 0, \left[A \left(\tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right) + B \left(\tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right) \right]_i \right\} \\ (A.16) \quad &\quad + \max \left\{ 0, \left[A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) + B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \mathbf{c} \right]_i \right\}. \end{aligned}$$

We can bound the first term in the above inequality as follows:

$$\begin{aligned}
 & \max \left\{ 0, \left[A \left(\tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right) + B \left(\tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right) \right]_i \right\} \\
 & \leq \left| \left[A \left(\tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right) + B \left(\tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right) \right]_i \right| \\
 & \stackrel{(a)}{\leq} \left\| A \left(\tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right) + B \left(\tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right) \right\| \\
 & \stackrel{(b)}{\leq} \|A\| \left\| \tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right\| + \|B\| \left\| \tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right\| \\
 & \stackrel{(c)}{\leq} \sigma_{\max} \cdot \left(\left\| \tilde{\mathbf{x}} - \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) \right\| + \left\| \tilde{\mathbf{y}} - \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \right\| \right) \stackrel{(d)}{\leq} 2\sigma_{\max} \cdot \delta,
 \end{aligned} \tag{A.17}$$

where in (a) we used the fact that for a vector $\mathbf{x}$ we have $|x_i| \leq \sqrt{\sum_{i=1}^n x_i^2} = \|\mathbf{x}\|$; in (b) we used the triangle inequality and the following property of norms: $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$; in (c) we used the definition $\sigma_{\max} := \max\{\|A\|, \|B\|\}$; and in (d) we used the condition $d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}}) = \|\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{x}}\|^2 + \|\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{y}}\|^2 \leq \delta^2$ from Definition 4.4, which implies that $\|\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{x}}\| \leq \delta$ and $\|\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{y}}\| \leq \delta$.

To derive a bound for the second term of (A.16), recall that the definition of approximate stationarity requires that $\|Q(\tilde{\boldsymbol{\lambda}})\| \leq \epsilon$. This implies that

$$\begin{aligned}
 \epsilon & \geq \left| \left[\frac{1}{\alpha} \left(\tilde{\boldsymbol{\lambda}} - \text{proj}_{\mathbb{R}_+^k} \left(\tilde{\boldsymbol{\lambda}} - \alpha \nabla G(\tilde{\boldsymbol{\lambda}}) \right) \right) \right]_i \right| \\
 & \stackrel{(i)}{=} \left| \frac{1}{\alpha} \left(\tilde{\lambda}_i - \max \left\{ 0, \tilde{\lambda}_i - \alpha \nabla_i G(\tilde{\boldsymbol{\lambda}}) \right\} \right) \right| \\
 & \stackrel{(ii)}{=} \left| \frac{1}{\alpha} \left(\tilde{\lambda}_i - \max \left\{ 0, \tilde{\lambda}_i - \alpha \left[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c} \right]_i \right\} \right) \right|,
 \end{aligned} \tag{A.18}$$

where (i) uses the fact that the projection is performed w.r.t. the set $\mathbb{R}_+^k$, thus it can be applied component-wise; (ii) uses the expression for $\nabla G(\tilde{\boldsymbol{\lambda}})$ derived in Lemma 4.2.

Suppose that the i th constraint is violated, i.e., $[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c}]_i < 0$. Then condition (A.18) takes the following form:

$$\begin{aligned}
 & \left| \frac{1}{\alpha} \left(\tilde{\lambda}_i - \tilde{\lambda}_i + \alpha \left[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c} \right]_i \right) \right| \leq \epsilon \\
 & \Rightarrow \left| \left[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c} \right]_i \right| \leq \epsilon \\
 & \Rightarrow \left[A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) + B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \mathbf{c} \right]_i \leq \epsilon,
 \end{aligned} \tag{A.19}$$

where the first relation uses the fact that $\tilde{\lambda}_i \geq 0$, so the argument inside the max function is positive.

Overall, in the case where $[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c}]_i \geq 0$ holds, then (A.16) and (A.17) together imply that:

$$\max \{0, [A\tilde{\mathbf{x}} + B\tilde{\mathbf{y}} - \mathbf{c}]_i\} \leq 2\sigma_{\max}\delta + 0 \leq 2\sigma_{\max}\delta.$$

On the other hand, if $[-A\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - B\bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) + \mathbf{c}]_i < 0$, then combining (A.16), (A.17), and (A.19) yields $\max \{0, [A\tilde{\mathbf{x}} + B\tilde{\mathbf{y}} - \mathbf{c}]_i\} \leq 2\sigma_{\max}\delta + \epsilon$.

Finally, we know from Definition 4.4 that the following condition holds at the approximate stationary point $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}})$:

$$d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\lambda}}) = \left\| \bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{x}} \right\|^2 + \left\| \bar{\mathbf{y}}(\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) - \tilde{\mathbf{y}} \right\|^2 \leq \delta^2.$$

Therefore, it follows directly that $\|\bar{\mathbf{y}} - \tilde{\mathbf{y}}\| \leq \delta$, $\|\bar{\mathbf{x}} - \tilde{\mathbf{x}}\| \leq \delta$, where $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\bar{\mathbf{x}}(\tilde{\boldsymbol{\lambda}}), \bar{\mathbf{y}}(\bar{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}))$. This completes the proof. $\square$

REFERENCES

- [1] C. ALIPRANTIS, K. BORDER, AND K. BORDER, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Studies in Economic Theory, Springer, Berlin, 1999.
- [2] D. AUSSEL AND J. DUTTA, *Generalized Nash equilibrium problem, variational inequality and quasiconvexity*, Oper. Res. Lett., 36 (2008), pp. 461–464.
- [3] P. BERNHARD AND A. RAPAPORT, *On a theorem of Danskin with an application to a theorem of von Neumann–Sion*, Nonlinear Anal., 24 (1995), pp. 1163–1181.
- [4] D. P. BERTSEKAS, *Convex Optimization Theory*, Athena Scientific, Belmont, MA, 2009.
- [5] J. BRACKEN AND J. T. MCGILL, *Mathematical programs with optimization problems in the constraints*, Oper. Res., 21 (1973), pp. 37–44.
- [6] T. CHEN, Y. SUN, AND W. YIN, *A single-timescale method for stochastic bilevel optimization*, Proc. Mach. Learn. Res. (PMLR), 151 (2022), pp. 2466–2488.
- [7] B. COLSON, P. MARCOTTE, AND G. SAVARD, *Bilevel programming: A survey*, 4OR, 3 (2005), pp. 87–107.
- [8] T. H. CORMEN, C. E. LEISERSON, R. L. RIVEST, AND C. STEIN, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2009.
- [9] Y.-H. DAI AND L. ZHANG, *Optimality conditions for constrained minimax optimization*, CSIAM Trans. Appl. Math., 1 (2020), pp. 296–315.
- [10] S. DEMPE AND A. ZEMKOHO, *Bilevel Optimization*, Springer, Cham, Switzerland, 2020.
- [11] A. DREVES, F. FACCHINEI, C. KANZOW, AND S. SAGRATELLA, *On the solution of the KKT conditions of generalized Nash equilibrium problems*, SIAM J. Optim., 21 (2011), pp. 1082–1108.
- [12] F. FACCHINEI AND C. KANZOW, *Generalized Nash equilibrium problems*, Ann. Oper. Res., 175 (2010), pp. 177–211.
- [13] F. FACCHINEI AND C. KANZOW, *Penalty methods for the solution of generalized Nash equilibrium problems*, SIAM J. Optim., 20 (2010), pp. 2228–2253.
- [14] A. FISCHER, M. HERRICH, AND K. SCHÖNEFELD, *Generalized Nash equilibrium problems-recent advances and challenges*, Pesquisa Oper., 34 (2014), pp. 521–558.
- [15] X. FU AND E. MODIANO, *Network interdiction using adversarial traffic flows*, in IEEE INFOCOM 2019-IEEE Conference on Computer Communications, IEEE, Piscataway, NJ, 2019, pp. 1765–1773.
- [16] S. GHADIMI AND M. WANG, *Approximation Methods for Bilevel Programming*, preprint, arXiv:1802.02246, 2018.
- [17] R. H. GOHARY, Y. HUANG, Z.-Q. LUO, AND J.-S. PANG, *A generalized iterative water-filling algorithm for distributed power control in the presence of a jammer*, IEEE Trans. Signal Process., 57 (2009), pp. 2660–2674.
- [18] D. GOKTAS AND A. GREENWALD, *Convex-concave min-max Stackelberg games*, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Red Hook, NY, 2021, pp. 2991–3003.
- [19] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAI, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, Vol. 3, Curran Associates, Red Hook, NY, 2014, pp. 2672–2680.
- [20] M. GRÖTSCHEL, L. LOVÁSZ, AND A. SCHRIJVER, *Geometric Algorithms and Combinatorial Optimization*, Vol. 2, Springer, Berlin, 1988.
- [21] M. HONG, H.-T. WAI, Z. WANG, AND Z. YANG, *A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic*, SIAM J. Optim., 33 (2023), pp. 147–180.
- [22] T. ICHISHI AND K. SHELL, *Game Theory for Economic Analysis*, Economic Theory, Econometrics, and Mathematical Economics, Elsevier Science, St. Louis, MO, 2014.
- [23] C. JIN, P. NETRAPALLI, AND M. I. JORDAN, *What is local optimality in nonconvex-nonconcave minimax optimization?*, Proc. Mach. Learn. Res. (PMLR), 119, 2020, pp. 4880–4889.

- [24] P. KHANDURI, S. ZENG, M. HONG, H.-T. WAI, Z. WANG, AND Z. YANG, *A near-optimal algorithm for stochastic bilevel optimization via double-momentum*, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Red Hook, NY, 2021, pp. 30271–30283.
- [25] G.-H. LIN, M. XU, AND J. J. YE, *On solving simple bilevel programs with a nonconvex lower level program*, Math. Program., 144 (2014), pp. 277–305.
- [26] T. LIN, C. JIN, AND M. JORDAN, *On gradient descent ascent for nonconvex-concave minimax problems*, in Proc. Mach. Learn. Res. (PMLR), 119, 2020, pp. 6083–6093.
- [27] T. LIN, C. JIN, AND M. I. JORDAN, *Near-optimal algorithms for minimax optimization*, in Conference on Learning Theory, PMLR, 119, 2020, pp. 2738–2779.
- [28] M. LIU, Y. MROUEH, J. ROSS, W. ZHANG, X. CUI, P. DAS, AND T. YANG, *Towards better understanding of adaptive gradient algorithms in generative adversarial nets*, in International Conference on Learning Representations, ICML, San Diego, CA, 2019.
- [29] R. LIU, J. GAO, J. ZHANG, D. MENG, AND Z. LIN, *Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond*, IEEE Trans. Pattern Anal. Mach. Intell., 44 (2021), pp. 10045–10067.
- [30] S. LU, I. TSAKNAKIS, M. HONG, AND Y. CHEN, *Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications*, IEEE Trans. Signal Process., 68 (2020), pp. 3676–3691.
- [31] Z.-Q. LUO, J.-S. PANG, AND D. RALPH, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, 1996.
- [32] A. MADRY, A. MAKELOV, L. SCHMIDT, D. TSIPRAS, AND A. VLADU, *Towards deep learning models resistant to adversarial attacks*, in International Conference on Learning Representations, ICML, San Diego, CA, 2018.
- [33] A. MOKHTARI, A. OZDAGLAR, AND S. PATTATHIL, *Convergence rate of $O(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems*, SIAM J. Optim., 30 (2020), pp. 3230–3251.
- [34] A. MOKHTARI, A. OZDAGLAR, AND S. PATTATHIL, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, in Proc. Mach. Learn. Res. (PMLR), 108, 2020, pp. 1497–1507.
- [35] Y. NESTEROV, *Lectures on Convex Optimization*, Springer Optim. Appl. 137, Springer, Cham, Switzerland, 2018.
- [36] H. NIKAIDŌ AND K. ISODA, *Note on non-cooperative convex games*, Pacific J. Math., 5 (1955), pp. 807–815.
- [37] M. NOUIEHED, M. SANJABI, T. HUANG, J. D. LEE, AND M. RAZAVIYAYN, *Solving a class of non-convex min-max games using iterative first order methods*, in Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Red Hook, NY, 2019, pp. 14934–14942.
- [38] M. OSBORNE AND A. RUBINSTEIN, *A Course in Game Theory*, MIT Press, Cambridge, MA, 1994.
- [39] J.-S. PANG AND M. FUKUSHIMA, *Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games*, Comput. Manag. Sci., 2 (2005), pp. 21–56.
- [40] P. M. PARDALOS AND S. A. VAVASIS, *Quadratic programming with one negative eigenvalue is NP-hard*, J. Global Optim., 1 (1991), pp. 15–22.
- [41] S. QIU, Z. YANG, X. WEI, J. YE, AND Z. WANG, *Single-Timescale Stochastic Nonconvex-Concave Optimization for Smooth Nonlinear TD Learning*, preprint, arXiv:2008.10103, 2020.
- [42] M. RAZAVIYAYN, T. HUANG, S. LU, M. NOUIEHED, M. SANJABI, AND M. HONG, *Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances*, IEEE Signal Proc. Mag., 37 (2020), pp. 55–66.
- [43] J. SALMERON, K. WOOD, AND R. BALDICK, *Analysis of electric grid security under terrorist threat*, IEEE Trans. Power Syst., 19 (2004), pp. 905–912.
- [44] J. SALMERON, K. WOOD, AND R. BALDICK, *Worst-case interdiction analysis of large-scale electric power grids*, IEEE Trans. Power Syst., 24 (2009), pp. 96–104.
- [45] J. C. SMITH, M. PRINCE, AND J. GEUNES, *Modern network interdiction problems and algorithms*, in Handbook of Combinatorial Optimization, Springer, New York, 2013, pp. 1949–1987.
- [46] I. TSAKNAKIS, M. HONG, AND S. ZHANG, *Minimax Problems with Coupled Linear Constraints: Computational Complexity, Duality and Solution Methods*, preprint, arXiv:2110.11210v1, 2021.
- [47] A. VON HEUSINGER AND C. KANZOW, *Optimization reformulations of the generalized Nash equilibrium problem using Nikaido–Isoda-type functions*, Comput. Optim. Appl., 43 (2009), pp. 353–377.

- [48] H. VON STACKELBERG, *The Theory of the Market Economy*, Oxford University Press, London, 1952.
- [49] Y. WANG, G. ZHANG, AND J. BA, *On solving minimax optimization locally: A follow-the-ridge approach*, in International Conference on Learning Representations, ICML, San Diego, CA, 2020.
- [50] Y. YAN, Y. XU, Q. LIN, W. LIU, AND T. YANG, *Optimal epoch stochastic gradient descent ascent methods for min-max optimization*, in Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Red Hook, NY, 2020, pp. 5789–5800.
- [51] J. YANG, N. KİYAVASH, AND N. HE, *Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems*, in Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Red Hook, NY, 2020, pp. 1153–1165.
- [52] Y. YANG, G. SCUTARI, P. SONG, AND D. P. PALOMAR, *Robust MIMO cognitive radio systems under interference temperature constraints*, IEEE J. Sel. Areas Commun., 31 (2013), pp. 2465–2482.
- [53] J. YE AND D. ZHU, *Optimality conditions for bilevel programming problems*, Optimization, 33 (1995), pp. 9–27.
- [54] G. ZHANG, J. LU, AND Y. GAO, *Multi-level Decision Making: Models, Methods and Applications*, Springer, Berlin, 2015.