Noname manuscript No.
(will be inserted by the editor)

First-order algorithms for robust optimization problems via convex-concave saddle-point Lagrangian reformulation

Krzysztof Postek · Shimrit Shtern

Abstract Robust optimization (RO) is one of the key paradigms for solving optimization problems affected by uncertainty. Two principal approaches for RO, the robust counterpart method and the adversarial approach, potentially lead to excessively large optimization problems. For that reason, first order approaches, based on online-convex-optimization, have been proposed [6,13] as alternatives for the case of large-scale problems. However, these methods are either stochastic in nature or involve a binary search for the optimal value. We propose deterministic first-order algorithms based on a saddle-point Lagrangian reformulation that avoids both of these issues. Our approach recovers the other approaches' $\mathcal{O}(1/\epsilon^2)$ convergence rate in the general case, and offers an improved $\mathcal{O}(1/\epsilon)$ rate for problems with constraints which are affine both in the decision and in the uncertainty. Experiment involving robust quadratic optimization demonstrates the numerical benefits of our approach.

1 Introduction

When solving optimization problems, one often has to deal with uncertainty present in the parameters of the objective and constraint functions. This uncertainty may stem from measurement, implementation or prediction errors. A common paradigm used to ensure that the solution remains feasible under uncertainty is *robust optimization (RO)* [5]. In RO, the uncertainty is assumed to be adversarial to the decision maker and to lie in a predefined *uncertainty set*, and the decision maker finds the best solution which remains feasible for all parameter values within this set.

Any RO algorithm involves simultaneous solving of two problems: the decision maker's problem of finding the best-possible decision which is feasible for the given uncertainty set, and nature's implicit adversarial problem of selecting the worst possible realization of the parameters.

Krzysztof Postek
Faculty of Electrical Enginnering, Mathematics and Computer Science
Delft University of Technology, Delft, The Netherlands
E-mail: k.s.postek@tudelft.nl

Shimrit Shtern
Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel
E-mail: shimrits@technion.ac.il

Denoting the decision maker's decision as $\mathbf{x}$, taken from a predefined set X, and the uncertain parameter for constraint i as $\mathbf{z}_i$, taken from a predefined uncertainty set Z^i, a general RO problem is given by:

$$\begin{aligned} &\min_{\mathbf{x}\in X} \mathbf{c}^\top \mathbf{x} && (1)\\ &\text{s.t.} \sup_{\mathbf{z}_i\in Z^i} g_i(\mathbf{x},\mathbf{z}_i) \leq 0, && i\in[m]. \end{aligned}$$

In this problem, the decision maker minimizes over $\mathbf{x} \in X$, while the nature selects $\mathbf{z} \in Z = Z^1 \times Z^2 \times \ldots \times Z^m$ such that the left-hand sides of the inequality constraints are maximized, in an attempt to violate the constraints.

There are two key methods for handling the opposite optimization problems simultaneously: (1) the *robust counterpart* (RC) reformulation method [4,5], (2) *adversarial approach*, also known as cutting planes [8,15]. In the RC approach, nature's maximization problem for each constraint is dualized, and the objective of the dual problem is required to be nonpositive. In this way, a constraint is substituted by a system of inequalities that are satisfied for a given $\mathbf{x}$ if and only if the original left-hand sides are nonpositive. This method's key advantage is that it requires to solve only one optimization problem and it ensures by-design the robustness of the solution. Its disadvantage is the increase of the problem size due to the added dual variables and constraints generated by dualizing nature's problems. Moreover, for strong duality to hold for nature's problem in each constraint, it has to be convex, or equivalent to a convex problem.

In the adversarial approach, a finite subset $\underline{Z} \subset Z$ of scenarios is iteratively built up, until it contains enough points to ensure that $\mathbf{x}$ is feasible for all realizations in the subset if and only if it is (almost) feasible for all realizations in Z. The set $\underline{Z}$ is intialized using an arbitrarily chosen realization $\mathbf{z} \in Z$. Then two steps are repeated alternatingly: an optimization step, in which a solution $\mathbf{x}$ which minimizes the objective function and is feasible for $\mathbf{z} \in \underline{Z}$ is found; a *pessimization step*, in which a new realization $\mathbf{z}$ violating at least one constraint is found and added to $\underline{Z}$. This iterative procedure continues until no violating scenario is found. While this method is simple and enables solving problems in which nature's problem is not necessarily convex, the size of $\underline{Z}$ increases at each iteration, which may result in extremely large optimization problems in $\mathbf{x}$ during the optimization step.

Both of the above methods potentially lead to excessively large optimization problems which creates space for approaches in which the decision maker's and the adversary's problems are simultaneously solved in a lightweight fashion. A recently suggested idea [6,13] is to use online convex optimization to solve problem (1), proving that the number of iterations needed to obtain an ϵ-feasible solution is $O(1/\epsilon^2)$. Through the online optimization lens, the robust problem is seen as a problem of minimizing a partly unknown objective with likewise constraints, whose shape is learned throughout the algorithm via samples.

The approach of [6] consists in iteratively solving a nominal version of (1) in which the set Z is replaced by a fixed realization $\mathbf{z}$. The value of $\mathbf{z}$ is updated at each iteration through first-order updates/pessimization and randomization. The approach of [13] is to use binary search to determine the minimal τ for which the feasibility problem

$$\{\mathbf{c}^\top \mathbf{x} \leq \tau,\ \sup_{\mathbf{z}\in Z} g_i(\mathbf{x},\mathbf{z}) \leq 0, \quad i\in[m]\}$$

has a solution up to a given accuracy. This requires running an online first-order algorithm for each tested τ. Each of the feasibility problems is solved by a first order iteration on $\mathbf{x}$, simultaneously with a pessimization/first order steps on the dual parameter $\mathbf{z}$.

Thanks to the online optimization framework, both [6,13] work directly with the functions $\{g_i(\cdot,\cdot)\}_{i\in[m]}$, without the need to build an ever-increasing list of scenarios, or to dualize the constraints. This has a price: the functions g_i have to be convex-concave, the set X needs to be bounded, and the maximum achievable convergence rate is $\mathcal{O}(1/\epsilon^2)$ to obtain an ϵ-feasible and ϵ-optimal solution. Also, [6] requires to solve multiple nominal problems, while in [13] one needs to run binary search for the optimal objective value.

We address these issues by proposing to solve (1) with single-run first-order algorithms. We leverage a natural convex-concave saddle-point reformulation of (1), based on its Lagrangian. Although the existence of such a formulation was noted and discussed in [13, Appendix A], it has been claimed that the problem loses a lot of convenient structure due to the lifting. In this paper, we demonstrate that if the problem satisfies the Slater-type condition, it can be solved using a subgradient saddle-point algorithm (SGSP) of [16] which requires $\mathcal{O}(1/\epsilon^2)$ iterations to obtain an ϵ-feasible and ϵ-optimal solution. For the special case where $g_i(\cdot,\cdot)$ are biaffine, we show that the problem can be solved using the proximal alternating predictor-corrector (PAPC) algorithm of [10], which takes $\mathcal{O}(1/\epsilon)$ iterations. Moreover, we show that when Z^i are projection- or proximal-friendly one may also find the projection onto the lifted space relatively easily, either analytically or by using bi-section. Finally, we show that the saddle-point formulation actually allows for more flexibility and enables to tackle problems where either Z^i or X are more complicated by using splitting techniques, where these problems prove to be more challenging for the previously suggested methods.

The remainder of the paper is structured as follows. In Section 2, we introduce the problem we solve along with the corresponding assumptions and the Lagrangian saddle-point reformulation. In Section 3, we introduce the two algorithms for the case of simple uncertainty sets and state the corresponding convergence results. In Section 4, we present the convergence analysis of the algorithms for a generalized problem form through a unified framework of showing that both are ergodically bounded (EB) algorithms. In Section 5, we compare the performance of our SGSP approach to the online first-order approaches of [13] and the adversarial approach on randomly sampled robust quadratic optimization problems with and without constraints. Section 6 concludes the paper. All proofs not given in the body are given in Appendix A.

1.1 Notation

Throughout the paper we use bolded small letters $\mathbf{x}$, $\mathbf{z}$ for vectors, and bolded capital letters $\mathbf{P}$ for matrices, and capital letters for sets. For any $k \in \mathbb{N}$, we use shorthand notation $[k]$ to denote the set of indexes $\{1, 2, \ldots, k\}$. Unless specified otherwise, $\|\cdot\|$ refers to the Euclidean norm.

2 Problem setting and assumptions

2.1 Introduction

In this paper, we consider the following general RO problem.

$$\begin{aligned} \min_{\mathbf{x}\in X}\ & \mathbf{c}^\top \mathbf{x} && && (2)\\ \text{s.t. } & f_i(\mathbf{x}) \coloneqq \max_{\mathbf{z}_i \in Z^i} g_i(\mathbf{x}, \mathbf{z}_i) \le 0 && i \in [m] \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

where $X \subseteq \mathbb{R}^n$ is a closed and convex set, for all $i \in [m]$ the set $Z^i \subset \mathbb{R}^{d_i}$ is a convex and compact set and without loss of generality $\mathbf{0} \in Z^i$, $g_i(\cdot, \mathbf{z}_i) : X \to \mathbb{R}$ is convex for any fixed $\mathbf{z}_i \in Z^i$, $g_i(\mathbf{x}, \cdot) : Z^i \to \mathbb{R}$ is concave for any fixed $\mathbf{x} \in X$, and $\mathbf{A} \in \mathbb{R}^{r \times n}$, $\mathbf{b} \in \mathbb{R}^r$. Note that, contrary to Problem 1, here we allow to separate some of the affine constraints that may be involved in the definition of the domain of $\mathbf{x}$ from the set X.

Note that formulation (2) also encompasses robust problems involving uncertainty in the objective function, since such problems can be transformed to form (2) using the epigraph formulation of the objective. Thus, formulation (2) is general and includes many useful problems (*c.f.*, [7,11]).

Functions $f_i(\mathbf{x})$ are known as the robust counterpart formulation of the robust constraint

$$g_i(\mathbf{x}, \mathbf{z}_i) \leq 0, \ \forall \mathbf{z}_i \in Z^i.$$

Note that while $f_i(\mathbf{x})$ are convex functions of $\mathbf{x}$ (as a maximum of convex functions), they are not necessarily easily representable due to their implicit formulation as maxima.

Our aim is to solve (2) through its saddle point Lagrangian formulation. For this formulation to be well-defined, we make three standard assumptions.

Assumption 1 *Problem* (2) *has an optimal solution.*

Assumption 2 *There exists* $\hat{\mathbf{x}} \in int(X)$ *such that* $\mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$ *and there exists a* $\epsilon_{\hat{\mathbf{x}}} > 0$ *such that* $\hat{\mathbf{x}} + \mathbf{y} \in X$ *and* $f_i(\hat{\mathbf{x}} + \mathbf{y}) < 0$ *for all* $\|\mathbf{y}\| \leq \epsilon_{\hat{\mathbf{x}}}$.

Assumption 3 *Matrix* $\mathbf{A}$ *has full row rank.*

With respect to Assumption 1, if the problem does not have an optimal solution there are three options: (i) it is infeasible, in which case the uncertainty sets defined for the problem may be too large, (ii) it is unbounded, *i.e.*, , it might not be constrained enough, or (iii) it is bounded but the optimal solution is not attained, in which case we can restrict the set X to a subset containing ϵ-optimal solutions of the original problem. Assumption 2, known as the Slater condition, ensures that the problem is stable, *i.e.*, slight perturbations in the feasible set do make the problem infeasible. Assumption 3 states that there are no redundant equality constraints.

In order to solve problem (2) we consider its Lagrangian:

$$L(\mathbf{x}, (\boldsymbol{\lambda}, \mathbf{w})) \equiv \mathbf{c}^\top \mathbf{x} + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \mathbf{w}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{3}$$

Specifically, we are interested in saddle points of function $L(\cdot, \cdot)$, *i.e.*, points $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ which satisfy

$$L(\mathbf{x}^*, (\boldsymbol{\lambda}, \mathbf{w})) \leq L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \leq L(\mathbf{x}, (\boldsymbol{\lambda}^*, \mathbf{w}^*)), \quad \forall \mathbf{x} \in X, \lambda \in \mathbb{R}^m_+, \mathbf{w} \in \mathbb{R}^r.$$

Under Assumptions 1 and 2, the Lagrangian function $L(\cdot, \cdot)$ has a saddle point, and $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ is a saddle point of $L(\cdot, \cdot)$ if and only if $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \mathbf{w}^*)$ are optimal solutions to the primal and dual problems respectively. Thus, instead of solving problem (2), we want to find a solution to

$$\inf_{\mathbf{x} \in X} \sup_{\boldsymbol{\lambda} \in \mathbb{R}^m_+, \mathbf{w} \in \mathbb{R}^r} L(\mathbf{x}, (\boldsymbol{\lambda}, \mathbf{w})). \tag{4}$$

This reformulation eliminates the constraints and renders the problem as a saddle point one, enabling the use of various first-order methods. However, such methods require computing at each iteration not only the functions $f_i(\mathbf{x})$, but also their sub-gradients or proximal operators, which may be challenging with the implicitly-defined f_i functions. We will therefore consider an alternative formulation where our goal is to work with functions g_i, and show how to obtain an ϵ-optimal and ϵ-feasible solution of problem (2).

2.2 Conversion to a convex-concave saddle point problem

We begin our transformation of problem (4) plugging in the explicit definitions of of functions f_i, using functions g_i, as follows:

$$\begin{aligned}\sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\mathbf{w}\in\mathbb{R}^r}\inf_{\mathbf{x}\in X} L(\mathbf{x},(\boldsymbol{\lambda},\mathbf{w})) &= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\mathbf{w}\in\mathbb{R}^r}\inf_{\mathbf{x}\in X}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i\sup_{\mathbf{z}_i\in Z^i} g_i(\mathbf{x},\mathbf{z}_i)+\mathbf{w}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})\\ &= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\mathbf{w}\in\mathbb{R}^r}\inf_{\mathbf{x}\in X}\max_{\mathbf{z}_i\in Z^i,\ i\in[m]}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i g_i(\mathbf{x},\mathbf{z}_i)+\mathbf{w}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})\\ &= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\mathbf{w}\in\mathbb{R}^r}\sup_{\mathbf{z}_i\in Z^i,\ i\in[m]}\inf_{\mathbf{x}\in X}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i g_i(\mathbf{x},\mathbf{z}_i)+\mathbf{w}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \quad (5)\end{aligned}$$

where the equalities follow from Sion's Theorem, the fact Z^i are convex and compact, X is convex, and g_i are convex-concave. However, the resulting saddle point problem (5) is over a function which is convex in $\mathbf{x}$ but is not jointly concave in $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_m)$ and $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_m)$. Since convergence results for saddle point algorithms typically require a convex-concave structure, we need to reformulate the problem to achieve such a structure. For this, we will use a change of variables $\tilde{\mathbf{z}}_i=\lambda_i\mathbf{z}_i$, and inversely

$$\mathbf{z}_i=\begin{cases}\frac{\tilde{\mathbf{z}}_i}{\lambda_i}, & \lambda_i>0,\\ \mathbf{0}, & \text{otherwise.}\end{cases}$$

Using this definition we have $\lambda_i g_i(\mathbf{x},\mathbf{z}_i)=\lambda_i g_i(\mathbf{x},\tilde{\mathbf{z}}_i/\lambda_i)$, and since $-g_i(\mathbf{x},\cdot)$ is convex for every $\mathbf{x}$, $-\lambda_i g_i(\mathbf{x},\tilde{\mathbf{z}}_i/\lambda_i)$ is jointly convex in $\mathbf{u}_i=(\tilde{\mathbf{z}}_i,\lambda_i)$ for every $\mathbf{x}$, as a *perspective* of a convex function [2, Proposition 8.23]. Moreover, $\lambda_i g_i(\mathbf{x},\tilde{\mathbf{z}}_i/\lambda_i)$ is continuous for all $\mathbf{u}_i\in U^i$, where $U^i=\{\mathbf{u}_i=(\tilde{\mathbf{z}}_i,\lambda_i):\tilde{\mathbf{z}}_i\in\lambda_i Z^i,\lambda_i\geq 0\}$, obtaining a value of 0 whenever $\lambda_i=0$. Defining $\mathbf{u}=(\mathbf{u}_1,\ldots,\mathbf{u}_m)$ and the set $U=U^1\times\ldots\times U^m$, we have

$$\sup_{\mathbf{u}\in U,\mathbf{w}\in\mathbb{R}^r}\inf_{\mathbf{x}\in X}\bar{L}(\mathbf{x},(\mathbf{u},\mathbf{w}))\coloneqq\sup_{\mathbf{u}\in U,\mathbf{w}\in\mathbb{R}^r}\inf_{\mathbf{x}\in X}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i g_i\left(\mathbf{x},\frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right)+\mathbf{w}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}). \quad (6)$$

For ease of notation, in henceforth we denote the perspective version of g_i as $\tilde{g}_i(\mathbf{x},\mathbf{u}_i)\equiv\lambda_i g_i(\mathbf{x},\tilde{\mathbf{z}}_i/\lambda_i)$ where $\mathbf{u}_i=(\tilde{\mathbf{z}}_i,\lambda_i)$.

The following result shows that solving (6) is sufficient for solving (4), *i.e.*, , the saddle points of $\bar{L}$ can be reduced to those of L.

Proposition 1 *Let $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))\in X\times U\times\mathbb{R}^r$, where $\mathbf{u}^*=(\mathbf{u}_1^*,\ldots,\mathbf{u}_m^*)$ and $\mathbf{u}_i^*=(\tilde{\mathbf{z}}_i^*,\lambda_i^*)$ for $i\in[m]$. Then $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))\in X\times U$ is a saddle point of $\bar{L}$ over $X\times U\times\mathbb{R}^r$ if and only if $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ is a saddle point of L over $X\times\mathbb{R}^m_+\times\mathbb{R}^r$.*

3 Algorithms for solving saddle point formulation

With the robust problem (2) in the desired convex-concave structure (6), we move towards introducing two algorithms for solving it. Before presenting them, we show a property of the optimal dual solution $(\boldsymbol{\lambda}^*,\mathbf{w}^*)$ used for both algorithms (either for running them or for obtaining their convergence guarantees). Namely, thanks to Assumptions 1-3, we can bound (i) the optimal value of $\boldsymbol{\lambda}$, which implies that we can bound the sets U^i, (ii) the optimal value of $\mathbf{w}$.

Proposition 2 *Let Assumption 1 hold, let $\hat{\mathbf{x}}$ be the point satisfying Assumption 2, let $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*))$ be a saddle point of $\bar{L}$ on the set $X \times U \times \mathbb{R}^r$, and let $\underline{v}$ be a strict lower bound on the optimal value of problem (2). Then, $\mathbf{u}^* = (\mathbf{u}_1^*, \ldots, \mathbf{u}_m^*)$, where $\mathbf{u}_i^* = (\tilde{\mathbf{z}}_i^*, \lambda_i^*)$, and $\mathbf{w}^*$ satisfy*

$$\begin{aligned}
\lambda_i^* &\leq \bar{\lambda} := \frac{\mathbf{c}^\top \hat{\mathbf{x}} - \underline{v}}{-\min_{i\in[m]} f_i(\hat{\mathbf{x}})} \\
\|\mathbf{u}_i^*\| &\leq \bar{\lambda}\sqrt{1+R_i^2} \\
\|\mathbf{w}^*\| &\leq R_{\mathbf{w}} := \frac{1}{\sigma_{\min}(\mathbf{A})}\left(\frac{\mathbf{c}^\top \hat{\mathbf{x}} - \underline{v}}{\epsilon_{\hat{\mathbf{x}}}} + \|\mathbf{c}\|\right),
\end{aligned}$$

where $R_i := \max_{\mathbf{z}\in Z^i} \|\mathbf{z}\|$ and $\sigma_{\min}(\mathbf{A}) > 0$ is the smallest singular value of $\mathbf{A}$.

Proof **Boundedness of λ_i^*.** Let $\mathbf{x}^*$ be an optimal solution to (2). Let $\hat{\mathbf{x}} \in X$ be a Slater point. We have

$$\begin{aligned}
\mathbf{c}^\top \mathbf{x}^* &\leq \mathbf{c}^\top \hat{\mathbf{x}} + \sum_{i=1}^m \lambda_i^* f_i(\hat{\mathbf{x}}) + (\mathbf{w}^*)^\top (\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}) \\
&= \mathbf{c}^\top \hat{\mathbf{x}} + \sum_{i=1}^m \lambda_i^* f_i(\hat{\mathbf{x}}) \\
&\leq \mathbf{c}^\top \hat{\mathbf{x}} + \|\boldsymbol{\lambda}^*\|_1 \max_{i\in[m]} f_i(\hat{\mathbf{x}})
\end{aligned}$$

whence

$$\|\boldsymbol{\lambda}^*\|_1 \leq \frac{\mathbf{c}^\top(\hat{\mathbf{x}} - \mathbf{x}^*)}{-\min_{i\in[m]} f_i(\hat{\mathbf{x}})} = \bar{\lambda}.$$

Thus, we have that $\lambda_i^* \leq \|\boldsymbol{\lambda}^*\|_1 \leq \bar{\lambda}$ for all $i \in [m]$.

Boundedness of $\mathbf{u}_i^*$. Since $\mathbf{u}_i^* = (\tilde{\mathbf{z}}_i^*, \lambda_i^*) \in U^i$ we have that $\tilde{\mathbf{z}}_i^* \in \lambda_i^* Z^i$, by the Cauchy-Schwartz inequality we have that $\|\tilde{\mathbf{z}}_i^*\| \leq |\lambda_i^*| R_i \leq \bar{\lambda} R_i$. Thus, $\|\mathbf{u}_i^*\|^* = \|\tilde{\mathbf{z}}_i^*\|^2 + |\lambda_i^*|^2 \leq \bar{\lambda}^2(R_i^2 + 1)$.

Boundedness of $\mathbf{w}^*$ Consider the saddle point formulation:

$$\inf_{\mathbf{x}\in X} \sup_{\mathbf{u}_i\in U^i, \mathbf{w}} \mathbf{c}^\top \mathbf{x} + \mathbf{w}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_i) = \inf_{\mathbf{x}} \sup_{\mathbf{u}_i\in U^i, \mathbf{w}} \mathbf{c}^\top \mathbf{x} + \mathbf{w}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_i) + \delta_X(\mathbf{x})$$

Assume $\mathbf{x}^*$, $\mathbf{w}^*$, $\mathbf{u}_i^*$ are the saddle point of this problem. The optimality conditions of the problem are:

$$\begin{aligned}
\mathbf{A}\mathbf{x}^* &= \mathbf{b} \\
\mathbf{c} + \mathbf{A}^\top \mathbf{w}^* + \sum_{i=1}^m \boldsymbol{\alpha}_i + \boldsymbol{\beta} &= \mathbf{0}
\end{aligned}$$

where $\boldsymbol{\alpha}_i \in \partial_{\mathbf{x}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*)$ and $\boldsymbol{\beta} \in \partial_{\mathbf{x}} \delta_X(\mathbf{x}^*)$. Let $\hat{\mathbf{x}} \in \operatorname{int}(X)$ be the Slater point so we have

$$\hat{\mathbf{x}} + \boldsymbol{\kappa} \in X, \quad \tilde{g}_i(\hat{\mathbf{x}} + \boldsymbol{\kappa}, \mathbf{u}_i^*) \leq 0 \qquad \forall \boldsymbol{\kappa} : \|\boldsymbol{\kappa}\| \leq \epsilon_{\hat{\mathbf{x}}} \tag{7}$$

Since $\boldsymbol{\alpha}_i \in \partial_{\mathbf{x}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*)$, we have that

$$\tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*) + \boldsymbol{\alpha}_i^\top (\hat{\mathbf{x}} + \boldsymbol{\kappa} - \mathbf{x}^*) \leq \tilde{g}_i(\hat{\mathbf{x}} + \boldsymbol{\kappa}, \mathbf{u}_i^*)$$

Algorithm	SGSP	PAPC
Domain X	Bounded	No restriction
Structure $g_i(\cdot)$	Any	Biaffine or reducible to biaffine
Optimality convergence rate	$\mathcal{O}(1/\sqrt{N})$	$\mathcal{O}(1/N)$
Feasibility convergence rate	$\mathcal{O}(1/\sqrt{N})$	$\mathcal{O}(1/N)$
Slater point needed to compute the stepsize	Yes	No

Table 1 Comparison of the two saddle point algorithms.

so that

$$\underbrace{\sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*)}_{=0} + \sum_{i\in[m]} \boldsymbol{\alpha}_i^\top (\hat{\mathbf{x}} + \boldsymbol{\kappa} - \mathbf{x}^*) \leq \underbrace{\sum_{i\in[m]} \tilde{g}_i(\hat{\mathbf{x}} + \boldsymbol{\kappa}, \mathbf{u}_i^*)}_{\leq 0}$$

where the first sum is equal to zero due to complementary slackness and the second sum is nonpositive due to (7). Since $\boldsymbol{\beta} \in \partial_{\mathbf{x}} \delta_X(\mathbf{x}^*)$, we also have that

$$\delta_X(\mathbf{x}^*) + \boldsymbol{\beta}^\top (\hat{\mathbf{x}} + \boldsymbol{\kappa} - \mathbf{x}^*) \leq \delta_X(\hat{\mathbf{x}} + \boldsymbol{\kappa}) = 0,$$

which, combined with the optimality conditions, implies

$$0 \leq \left(\mathbf{c} + \mathbf{A}^\top \mathbf{w}^* + \sum_{i=1}^m \boldsymbol{\alpha}_i\right)^\top (\hat{\mathbf{x}} + \boldsymbol{\kappa} - \mathbf{x}^*) \leq \left(\mathbf{c} + \mathbf{A}^\top \mathbf{w}^*\right)^\top (\hat{\mathbf{x}} + \boldsymbol{\kappa} - \mathbf{x}^*), \quad \forall \|\boldsymbol{\kappa}\| \leq \epsilon_{\hat{\mathbf{x}}}.$$

This gives us the property

$$\epsilon_{\hat{\mathbf{x}}} \left\|\mathbf{c} + \mathbf{A}^\top \mathbf{w}^*\right\| \leq \left(\mathbf{c} + \mathbf{A}^\top \mathbf{w}^*\right)^\top (\hat{\mathbf{x}} - \mathbf{x}^*)$$

Because $\mathbf{A}\mathbf{x}^* = \mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$, we obtain $\epsilon_{\hat{\mathbf{x}}} \left\|\mathbf{c} + \mathbf{A}^\top \mathbf{w}^*\right\| \leq \mathbf{c}^\top (\hat{\mathbf{x}} - \mathbf{x}^*)$. Since by Assumption 3, $\mathbf{A}$ has full row rank, by the reverse triangle inequality we can bound $\mathbf{w}^*$ as follows:

$$\epsilon_{\hat{\mathbf{x}}} \left(\sigma_{\min}(\mathbf{A}^\top) \|\mathbf{w}^*\| - \|\mathbf{c}\|\right) \leq \epsilon_{\hat{\mathbf{x}}} \left(\left\|\mathbf{A}^\top \mathbf{w}^*\right\| - \|\mathbf{c}\|\right) \leq \epsilon_{\hat{\mathbf{x}}} \left\|\mathbf{c} + \mathbf{A}^\top \mathbf{w}^*\right\| \leq \mathbf{c}^\top (\hat{\mathbf{x}} - \mathbf{x}^*).$$

□

We now move to presenting the two algorithms, summarized in Table 1. We will first state both for problem (6) and give their convergence without proofs. In Section 4, we shall prove the convergence of a generalized problem in a unified framework from which the 'simple cases' will follow as straightforward corollaries.

The first algorithm, presented in Section 3.1, applies to the case with the additional assumptions that X is bounded and the functions g_i have bounded subgradients over $X \times Z^i$. For this setting, convergence rate of $O(1/\epsilon^2)$ is attained, similar to the one obtained by [13] under almost identical assumptions. In Section 3.2, we consider the case where $g_i(\mathbf{x}, \mathbf{z}_i)$ are biaffine functions (or can be transformed to this form), and show that in this setting we can obtain a superior convergence rate of $O(1/\epsilon)$.

3.1 Subgradient Saddle Point algorithm

In this section, we show how problem (6) can be solved using the SGSP suggested in [16]. This algorithm requires that the both primal and dual variables be contained in compact sets. Thus, we first need to replace each set U^i by its compact counterpart $\tilde{U}^i = \{(\tilde{\mathbf{z}}_i, \lambda_i) \in U^i : \lambda_i \leq \bar{\lambda}\}$. Indeed, by the property of λ^* presented in Proposition 2, this restriction of U^i would not change the set of saddle points of problem (6). Similarly, we can restrict $\mathbf{w}$ to reside in a set $W = \{\mathbf{w} \in \mathbb{R}^r : \ \|\mathbf{w}\| \leq R_{\mathbf{w}}\}$. Using these new sets, the algorithm is as follows.

SGSP: SubGradient Algorithm for Saddle Point
Input: $\tau > 0, \theta_i > 0, \theta_{\mathbf{w}} > 0$ for $i \in [m]$, and $N \in \mathbb{N}$
Initialization. Initialize $\mathbf{x}^0 \in X$ and $\mathbf{u}_i^0 \in \tilde{U}^i$, for $i \in [m]$, $\mathbf{w}^0 \in W$.
General step: For $k \in [N]$
Compute subgradients $\mathbf{v}_x^k \in \partial_{\mathbf{x}} \bar{L}(\mathbf{x}^k, (\mathbf{u}^k, \mathbf{w}^k))$, $\mathbf{v}_i^k \in \partial_{\mathbf{u}_i} \left(-\bar{L}(\mathbf{x}^k, (\mathbf{u}^k, \mathbf{w}^k))\right)$, for all $i \in [m]$

$$\begin{aligned} \mathbf{x}^{k+1} &= P_X(\mathbf{x}^k - \tau \mathbf{v}_x^k), \\ \mathbf{u}_i^{k+1} &= P_{\tilde{U}^i}(\mathbf{u}_i^k - \theta_i \mathbf{v}_i^k), \qquad i \in [m] \\ \mathbf{w}^{k+1} &= P_W(\mathbf{w}^k + \theta_{\mathbf{w}}(\mathbf{A}\mathbf{x}^k - \mathbf{b})) \end{aligned}$$

Note that the algorithm can be applied whenever the projections over $\{\tilde{U}^i\}_{i\in[m]}$ and X can be easily computed. We will see in Section 4, that if any of the sets are intersections of multiple simpler sets, we can utilize splitting methods that enable the use of projections only on the components of the intersection. Moreover, note that at each iteration the steps for all $\mathbf{u}_i$, $\mathbf{x}$ and $\mathbf{w}$ can be done in parallel.

As in most algorithms solving saddle point problems, the SGSP algorithm's convergence is given in terms of the ergodic sequences, *i.e.*, denoting

$$\bar{\mathbf{x}}^N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}^k, \quad \bar{\mathbf{u}}^N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{u}^k, \quad \bar{\mathbf{w}}^N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{w}^k$$

the convergence result states the rate at which the sequence $\{\left(\bar{\mathbf{x}}^N, \bar{\mathbf{u}}^N, \bar{\mathbf{w}}^N\right)\}_{N\in\mathbb{N}}$ converges to a saddle point. Here, we present the convergence in terms of the total constraint violation and the distance from the optimal value.

Theorem 1 *Let $\{\mathbf{x}^k, \mathbf{u}^k, \mathbf{w}^k\}_{k\in\mathbb{N}}$ be the sequences generated by the SGSP algorithm with step sizes*

$$\tau := \ \tilde{\tau}/\sqrt{N}, \quad \theta_i := \tilde{\theta}_i/\sqrt{N}, \quad \theta_{\mathbf{w}} := \tilde{\theta}_{\mathbf{w}}/\sqrt{N}$$

for $i \in [m]$. Assume that X is compact and define $R_{\mathbf{x}} := \max_{\mathbf{x}\in X} \|\mathbf{x}\|$. Further assume there exist constants $G_{\mathbf{x}}$, G_i for $i \in [m]$ such that the subgradients $\{\mathbf{v}_x^k\}_{k\in\mathbb{N}}$, $\{\mathbf{v}_i^k\}_{k\in\mathbb{N}}$, $i \in [m]$ generated by the algorithm satisfy

$$\left\|\mathbf{v}_x^k\right\| \leq G_{\mathbf{x}}, \ \left\|\mathbf{v}_i^k\right\| \leq G_i, \ i \in [m], \quad \forall k \in \mathbb{N}.$$

Then, we have the following feasibility and optimality convergence guarantees.

$$\sum_{i=1}^{m}[f_i(\bar{\mathbf{x}}^N)]_+ + \|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\| \leq$$
$$\frac{\max\{2, \max_i\{1+4R_i^2\}\}}{2\sqrt{N}}\left(\frac{2\max\{\left\|\mathbf{x}^0\right\|^2, R_{\mathbf{x}}^2\}}{\tau} + \sum_{i=1}^{m}\frac{\max\{\bar{\lambda}+1, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{R_{\mathbf{w}}+1, \|\mathbf{w}^0\|\}^2}{\theta_{\mathbf{w}}} + \phi\right),$$

and

$$\left|\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*)\right| \leq$$
$$\frac{\max\{2, \max_i\{1+4R_i^2\}\}}{2\sqrt{N}}\left(\frac{2\max\{\left\|\mathbf{x}^0\right\|^2, R_{\mathbf{x}}^2\}}{\tau} + \sum_{i=1}^{m}\frac{\max\{2\bar{\lambda}, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{2R_{\mathbf{w}}, \|\mathbf{w}^0\|\}^2}{\theta_{\mathbf{w}}} + \phi\right),$$

where

$$\phi := \tilde{\tau}G_{\mathbf{x}}^2 + \sum_{i=1}^{m}\tilde{\theta}_i G_i^2 + \tilde{\theta}_{\mathbf{w}}G_{\mathbf{w}}^2, \quad G_{\mathbf{w}} := \|\mathbf{A}\|R_{\mathbf{x}} + \|\mathbf{b}\|, \quad R_i := \max_{\mathbf{z}\in Z^i}\|\mathbf{z}\|.$$

Note that almost all conditions used in Theorem 1 are also needed in order to apply the online first-order (OFO) approach of [13]. In fact, the two approaches give similar convergence results with a few differences:

1. *Assumptions.* SGSP requires the existence of a Slater point $\hat{\mathbf{x}}$ while the OFO does not. Secondly, OFO requires boundedness of the subgradients of $g(\mathbf{x}, \cdot)$ while SGSP requires boundedness of the subgradients of its prespective function. In Section 3.3.2 we show that under mild assumptions on the problems, these requirements are equivalent.
2. *Implementation.* Since the OFO approach is meant to solve a feasibility, rather than an optimality, problem, it requires to perform a binary search in order to approximate the optimal value of (2). Thus, the number of needed iteration to obtain feasibility and optimality guaranties is increased by a factor of $\log(1/\epsilon)$. In the SGSP in turn, we do not need to perform bi-section, however we must find a Slater point $\hat{\mathbf{x}}$, the values of $f_i(\hat{\mathbf{x}})$ for $i \in [m]$, and $\epsilon_{\hat{\mathbf{x}}}$, as well as a lower bound on the objective function $\underline{v}$, which are needed to compute both $\bar{\lambda}$ and $R_{\mathbf{w}}$. While in some cases it may be easy to find these quantities, in general it requires solving an auxiliary optimization problem, as we discuss in Section 3.3.1.
3. *Projections.* While OFO requires projections on sets X and Z^i, SGSP requires projection onto X and the lifted set $\tilde{U}^i$. In Section 3.3.3 we show that for standard simple sets Z^i the projections onto $\tilde{U}^i$ can be simply computed.
4. *Constants.* Although the convergence rate of both methods is $O(1/\epsilon^2)$, the constants obtained by the SGSP algorithm are worse then those of OFO, if the same first order method (subgradient/mirror decent) is used.

3.2 Proximal-Alternating Predictor Corrector

In this section, we present an algorithm with a superior rate of convergence of $O(1/\epsilon)$ which does not require boundedness of X. This algorithm requires the additional assumption that the functions $g_i(\mathbf{x}, \mathbf{z}_i)$ have a biaffine form:

$$g_i(\mathbf{x}, \mathbf{z}_i) = \mathbf{x}^\top\mathbf{Q}_i\mathbf{z}_i + \mathbf{d}_i^\top\mathbf{x} + \mathbf{q}_i^\top\mathbf{z}_i + \gamma_i. \tag{8}$$

and that the primal variable is not constrained. In Remark 1, we show how more general problems of the form $g_i(\mathbf{x}, \mathbf{z}_i) \coloneqq \mathbf{h}_i(\mathbf{x})^\top \mathbf{k}_i(\mathbf{z}_i)$ with convex and concave $\mathbf{h}_i$ and $\mathbf{k}_i$, respectively, can be reformulated to fit this case.

In order to state the algorithm, we first perform two operations on our problem: (i) simplify its form due to the bi-linear structure, and (ii) eliminate the explicit constraint $x \in X$ by dualizing it. Indeed, under (8) function $\bar{L}$ reduces to

$$\bar{L}(\mathbf{x}, \mathbf{u}) \coloneqq \mathbf{c}^\top \mathbf{x} + \sum_{i=1}^{m} \left(\mathbf{x}^\top \tilde{\mathbf{Q}}_i \mathbf{u}_i + \tilde{\mathbf{q}}_i^\top \mathbf{u}_i \right) + \mathbf{w}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{c}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{w} + \mathbf{x}^\top \tilde{\mathbf{Q}} \mathbf{u} + \tilde{\mathbf{q}}^\top \mathbf{u} - \mathbf{b}^\top \mathbf{w},$$

where,

$$\tilde{\mathbf{Q}}_i = \left[\mathbf{Q}_i \ \mathbf{d}_i \right], \ \tilde{\mathbf{q}}_i = \begin{bmatrix} \mathbf{q}_i \\ \gamma_i \end{bmatrix}, \ \tilde{\mathbf{Q}} = (\tilde{\mathbf{Q}}_1^\top, \ldots, \tilde{\mathbf{Q}}_m^\top)^\top, \ \tilde{\mathbf{q}} = (\tilde{\mathbf{q}}_1^\top, \ldots, \tilde{\mathbf{q}}_m^\top)^\top. \tag{9}$$

In order to eliminate the restriction of $\mathbf{x}$ to lies in the set X, we use the so-called *dual transportation trick*, *i.e.*, dualizing the indicator function of the domain set X. Thus, for the second transformation, we shall use the indicator function δ_X, and its dual representation through the support function σ_X given by:

$$\delta_X(\mathbf{x}) = (\sigma_X)^*(\mathbf{x}) = \sup_{\boldsymbol{\pi} \in \mathbb{R}^n} \left\{ \boldsymbol{\pi}^\top \mathbf{x} - \sigma_X(\boldsymbol{\pi}) \right\},$$

to obtain:

$$\begin{aligned}
&\min_{\mathbf{x} \in X} \max_{\mathbf{u} \in U, \mathbf{w}} \bar{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w})) \\
&= \min_{\mathbf{x} \in \mathbb{R}^n} \max_{\mathbf{u} \in U, \mathbf{w}} \mathbf{c}^\top \mathbf{x} + \delta_X(\mathbf{x}) + \mathbf{x}^\top \mathbf{A}^\top \mathbf{w} + \mathbf{x}^\top \tilde{\mathbf{Q}} \mathbf{u} + \tilde{\mathbf{q}}^\top \mathbf{u} - \mathbf{b}^\top \mathbf{w} \\
&= \inf_{\mathbf{x} \in \mathbb{R}^n} \sup_{\mathbf{u} \in U, \boldsymbol{\pi} \in \mathbb{R}^n, \mathbf{w}} \tilde{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w}, \boldsymbol{\pi})) \coloneqq \mathbf{x}^\top \left(\mathbf{c} + \mathbf{A}^\top \mathbf{w} + \tilde{\mathbf{Q}} \mathbf{u} + \boldsymbol{\pi} \right) + \tilde{\mathbf{q}}^\top \mathbf{u} - \mathbf{b}^\top \mathbf{w} - \sigma_X(\boldsymbol{\pi}) \qquad (10) \\
&= \inf_{\mathbf{x} \in \mathbb{R}^n} \sup_{\mathbf{y} = (\mathbf{u}, \mathbf{w}, \boldsymbol{\pi}), \ \mathbf{u} \in U, \boldsymbol{\pi} \in \mathbb{R}^n} \mathbf{c}^\top \mathbf{x} + \mathbf{x}^\top \bar{\mathbf{Q}} \mathbf{y} + \tilde{\mathbf{q}}^\top \mathbf{u} - \mathbf{b}^\top \mathbf{w} - \sigma_X(\boldsymbol{\pi}) \equiv \tilde{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w}, \boldsymbol{\pi}))
\end{aligned}$$

where

$$\bar{\mathbf{Q}} = \left[\tilde{\mathbf{Q}}, \mathbf{A}^\top, \mathbf{I} \right]$$

The following lemma states that since the function $\bar{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w}))$ has a saddle point, the order of inf and sup in the above problem can be changed, and the function $\tilde{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w}, \boldsymbol{\pi}))$ has a saddle point.

Lemma 1 *Let Assumptions 1-3 hold. Then,* $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*))$ *is a saddle point of function* $\bar{L}$ *over* $X \times U \times \mathbb{R}^r$ *if and only if there exists a* $\boldsymbol{\pi}^*$ *such that* $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*))$ *is a saddle point of* $\tilde{L}$ *over* $\mathbb{R}^n \times (U \times \mathbb{R}^r \times \mathbb{R}^n)$.

Having stated the necessary equivalence of saddle points result, we now state the PAPC algorithm of [10] for solving problem (10).

PAPC: Proximal Alternating Predictor–Corrector
Input: $\tau > 0, \theta_i > 0$ for $i \in [m]$, and $N \in \mathbb{N}$
Initialization. Initialize $\mathbf{x}^0 \in \mathbb{R}^n$, $\mathbf{y}^0 = (\mathbf{u}^0, \mathbf{w}^0, \boldsymbol{\pi}^0) \in U \times \mathbb{R}^r \times \mathbb{R}^m$.
General step: For $k \in [N]$

$$\begin{aligned}
\mathbf{p}^k &= \mathbf{x}^{k-1} - \tau\left(\mathbf{c} + \bar{\mathbf{Q}}\mathbf{y}^{k-1}\right), \\
\mathbf{u}_i^k &= P_{U^i}\left(\mathbf{u}_i^{k-1} + \theta_i(\tilde{\mathbf{Q}}_i^\top \mathbf{p}^k + \tilde{\mathbf{q}}_i)\right), \qquad i \in [m] \\
\mathbf{w}^k &= \mathbf{w}^{k-1} + \theta_{\mathbf{w}}(\mathbf{A}\mathbf{p}^k - \mathbf{b}) \\
\boldsymbol{\pi}^k &= \mathrm{prox}_{\theta_{\boldsymbol{\pi}}\sigma_X}(\boldsymbol{\pi}^{k-1} + \theta_{\boldsymbol{\pi}}\mathbf{p}^k) \\
\mathbf{y}^k &= (\mathbf{u}^k, \mathbf{w}^k, \boldsymbol{\pi}^k) \\
\mathbf{x}^k &= \mathbf{x}^{k-1} - \tau\left(\mathbf{c} + \bar{\mathbf{Q}}\mathbf{y}^k\right).
\end{aligned}$$

With respect to the used proximal operator, note that using the Moreau identity we have that for any $\mathbf{y} \in \mathbb{R}^n$

$$\mathrm{prox}_{\theta_{\boldsymbol{\pi}}\sigma_X}(\mathbf{y}) = \mathbf{y} - \theta_{\boldsymbol{\pi}} P_X\left(\frac{\mathbf{y}}{\theta_{\boldsymbol{\pi}}}\right).$$

Thus, similarly to the SGSP algorithm, PAPC can be applied whenever the projections over U^i and X are easily computed.

Similarly to the SGSP, convergence results for the PAPC are in terms of the ergodic sequence. To state them, we define square matrices $\mathbf{S}_i = \theta_i^{-1}\mathbf{I}_{\tilde{d}_i}$ for $i \in [m]$, $\mathbf{S}_{m+1} = \theta_{\mathbf{w}}^{-1}\mathbf{I}_r$, $\mathbf{S}_{m+2} = \theta_{\boldsymbol{\pi}}^{-1}\mathbf{I}_n$ and $\mathbf{S} = \mathrm{Diag}\left([\mathbf{S}_1, \ldots, \mathbf{S}_{m+2}]\right)$.

Theorem 2 *Let Assumptions 1 and 2 hold and let $\{\mathbf{x}^k, \mathbf{w}^k, \mathbf{v}^k\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm with some $\tau > 0$, $\theta_i > 0$, $i \in [m]$, $\theta_{\mathbf{w}} > 0$ such that $\mathbf{H} = \mathbf{S} - \tau\bar{\mathbf{Q}}^\top\bar{\mathbf{Q}} \succ 0$. Define*

$$R_{\boldsymbol{\pi}} := \|\mathbf{c}\| + \bar{\lambda}\sum_{i=1}^m \left\|\tilde{\mathbf{Q}}_i\right\|(R_i + 1)$$

Then, for any $\alpha > 0$ we have the following convergence guarantees:

$$\begin{aligned}
&\sum_{i=1}^N [f_i(\bar{\mathbf{x}}^N)]_+ + \left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| + \alpha\,\mathrm{Dist}(\bar{\mathbf{x}}^N, X) \leq \\
&\frac{\max\{2, \max_i(1 + 4R_i^2)\}}{2N}\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{\bar{\lambda} + 1, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{R_{\mathbf{w}} + 1, \|\mathbf{w}^0\|\}^2}{\theta_{\mathbf{w}}} + \frac{1 + (1+\alpha)^2}{\theta_{\boldsymbol{\pi}}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\left|\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*)\right| \leq \\
&\frac{\max\{2, \max_i(1 + 4R_i^2)\}}{2N}\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{2\bar{\lambda}, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{2R_{\mathbf{w}}, \|\mathbf{w}^0\|\}^2}{\theta_{\mathbf{w}}} + \frac{1 + R_{\boldsymbol{\pi}}^2)}{\theta_{\boldsymbol{\pi}}}\right).
\end{aligned}$$

3.3 Numerical implementation of saddle point algorithms

In this section, we discuss some technical aspects related to the implementation of the above algorithms. These aspects relate to (i) finding a Slater point for the SGSP algorithm using SGSP algorithm of an auxiliary problem, (ii) computing subgradients of the lifted functions $\tilde{g}_i$ from the subgradients of the original functions g_i, (iii) computing projections onto the lifted sets U^i.

3.3.1 Search for a Slater point in the SGSP algorithm

To run the SGSP algorithm, we require parameter values dependent on the features of a Slater point of the problem. Therefore, before SGSP is run, we first need to identify a Slater point and plug the appropriate values into the algorithm. To find such a point, we need to solve the following optimization problem

$$\begin{aligned} &\min_{\mathbf{x},t} \ t && && (11)\\ &\text{s.t.} \ \max_{\mathbf{z}_i \in Z^i} g_i(\mathbf{x}, \mathbf{z}_i) \leq t && \forall i \in [m] \\ &\quad \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

This problem satisfies the Slater condition for any $\mathbf{x}^0 \in \text{int}(X)$, $\mathbf{A}\mathbf{x}^0 = \mathbf{b}$, and $t = t^0 + \delta$ such that

$$t^0 = \max_{i \in [m]} \max_{\mathbf{z}_i \in Z^i} g_i(\mathbf{x}^0, \mathbf{z}_i).$$

For that reason, we can first apply the SGSP algorithm to this problem by transforming it to a saddle point form. To keep the domain of $(\mathbf{x}, t)$ compact, we can restrict t to belong to the interval $[-1, \bar{t}]$ where $\bar{t} = t^0 + \delta$.

If we assume that the subgradients in the SGSP algorithm for the original problem (2) are bounded, then it also holds for (11). Thus, one can run the SGSP algorithm for (11), knowing that it will converge to the optimal value of t. A complication is the moment at which the algorithm stops. Assuming that the original problem satisfies the Slater condition with constant ϵ we would like to stop when both optimality and feasibility conditions are satisfied with $\frac{\epsilon}{3}$ thus ensuring that the $(\hat{\mathbf{x}}, \hat{t})$ obtained by the procedure satisfies

$$\max_{\mathbf{z}_i \in Z^i} g_i(\hat{\mathbf{x}}, \mathbf{z}_i) \leq \hat{t} + \frac{\epsilon}{3} \leq t^* + \frac{\epsilon}{3} + \frac{\epsilon}{3} = -\frac{\epsilon}{3}.$$

However, in practice, ϵ can be unknown in advance, and therefore, we construct a search procedure as follows:

Slater Point Search
Input: $\tau > 0, \theta_i > 0$ for $i \in [m]$, $\delta > 0$, $K = 2$
Initialization. Initialize $\mathbf{x}^0 \in \text{int}(X)$, $t^0 = \max_{i\in[m]} f_i(\mathbf{x}^0) + \delta$ and $\mathbf{u}_i^0 \in U^i$, for $i \in [m]$.
General step: For $k \in 1, 2, \ldots$

1. Run SGSP for the Lagrangian form of problem (11) starting from point $(\mathbf{x}^{k-1}, t^{k-1})$ and $\mathbf{u}^{k-1}$ for K iterations, and obtain the ergodic values $(\mathbf{x}^k, t^k)$ and $\mathbf{u}^k$.

2. Update $t^k = \max_{i\in[m]} f_i(\mathbf{x}^k)$.

3. If $t^k < 0$ stop and return the Slater point $\hat{\mathbf{x}} = \mathbf{x}^k$ and the Slater value $\epsilon = -t^k$.
   Otherwise, update $\bar{t} = \min\{\bar{t}, t^k + \delta\}$, $k = k + 1$ and $K = 2K$.

As stated above, this procedure is guaranteed to converge. Moreover, since $\mathbf{x}^0 \in \text{int}(X)$ then by properties of convex sets we will also obtain that for each k the iterate $\mathbf{x}^k \in \text{int}(X)$, as the average of points in X with one of them in the interior.

3.3.2 Determining the subgradients of $\tilde{g}_i$ from subgradients of g_i

In order to prove the convergence of SGSP we require that the subgradients of the perspective function $\tilde{g}_i$ are bounded. In this section, we show that the subgradients of $\tilde{g}_i$ can be easily derived from subgradients of g_i. Moreover, we will also show that under the following mild assumptions the boundedness of the subgradients of $\tilde{g}_i$ follows from the boundedness of the subgradients of g_i.

Assumption 4 *For every Z^i there exists a constant $\epsilon_i > 0$ such that $B(\mathbf{0}, \epsilon_i) \subseteq Z^i$.*

Assumption 5 *There exists a constant $\bar{\mu}_i > 0$ such that $-g_i(\mathbf{x}, \mathbf{0}) \leq \bar{\mu}_i$ for any feasible $\mathbf{x}$ of problem (2).*

Assumption 4 is a standard RO assumption that the uncertainty set is full dimensional, note that this is always true, since under a linear transformation we can always reduce the dimension of Z^i. Assumption 5 states the following: the feasible set of the robust problem does not contain rays that make any of the constraints of the 'nominal problem' (where $\mathbf{z} = 0$) be arbitrarily satisfied, *i.e.*, make $g_i(\mathbf{x}, \mathbf{0})$ arbitrarily small. We note that this assumption can be verified by checking the following sufficient condition:

$$\min\{g_i(\mathbf{x}, \mathbf{0}) : \ \mathbf{x} \in X, \ g_j(\mathbf{x}, \mathbf{0}) \leq 0, \forall j \in [m] \setminus \{i\}\} > -\infty, \ \forall i \in [m],$$

which can, in turn, be shown to hold by solving m convex (non-robust) optimization problems.

In the course of the SGSP algorithm, one needs to compute the subgradients of the perspective functions $\tilde{g}_i(\mathbf{x}, \mathbf{u})$. Ideally, this is done using the subgradients of the original functions $g_i(\mathbf{x}, \mathbf{z})$, which should typically be available. The following lemma provides a 'recipe' for doing exactly this. The 'recipe' is based on the convex analysis results for perspective functions of [9].

Lemma 2 *Let $\mathbf{x} \in X$ and $\mathbf{u}_i = (\tilde{\mathbf{z}}_i, \lambda_i) \in U^i$ for all $i \in [m]$.*

(i) If $\mathbf{v}_x \in \partial_{\mathbf{x}} \bar{L}(\mathbf{x},\mathbf{u})$, $\mathbf{v}_i \in \partial_{\mathbf{u}_i}\left(-\bar{L}(\mathbf{x},\mathbf{u})\right)$, *then they are of the following form*

$$\mathbf{v}_{\mathbf{x}} = \mathbf{c} + \mathbf{d}_x \tag{12}$$

$$\mathbf{v}_i = \begin{cases} \left(\mathbf{d}_i, -g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right) - \frac{\tilde{\mathbf{z}}_i^\top \mathbf{d}_i}{\lambda_i}\right), & \lambda_i > 0, \\ (\mathbf{d}_i, \phi_i), & \textit{otherwise} \end{cases} \tag{13}$$

where $\mathbf{d}_x \in \sum_{i:\lambda_i>0} \lambda_i \partial_{\mathbf{x}} g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right)$, $\mathbf{d}_i \in \partial_{\mathbf{z}_i}\left(-g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right)\right)$ *for all* $i \in [m]$ *such that* $\lambda_i > 0$, *and* $\mathbf{d}_i \in \cup_{\mathbf{z}_i \in \mathbb{R}^{d_i}} \left(\partial_{\mathbf{z}_i}(-g_i)(\mathbf{x},\mathbf{z}_i)\right)$, $\phi_i + (-g_i)^*(\mathbf{x},\mathbf{d}_i) \le 0$ *for all* $i \in [m]$ *such that* $\lambda_i = 0$.

(ii) Let $\mathbf{z}_i = \tilde{\mathbf{z}}_i/\lambda_i$ *if* $\lambda_i > 0$ *and* $\mathbf{z}_i = \mathbf{0}$ *otherwise, and let* $\tilde{\mathbf{d}}_{x,i} \in \partial_{\mathbf{x}} g_i(\mathbf{x},\mathbf{z}_i)$ *and* $\tilde{\mathbf{d}}_{z,i} \in \partial_{\mathbf{z}_i}\left(-g_i(\mathbf{x},\mathbf{z}_i)\right)$ *for all* $i \in [m]$. *Then,*

$$\mathbf{v}_x = \mathbf{c} + \sum_{i=1}^m \lambda_i \mathbf{d}_{x,i} \in \partial_{\mathbf{x}} \bar{L}(\mathbf{x},\mathbf{u}) \text{ and } \mathbf{v}_i = (\mathbf{d}_{z,i}, -g_i(\mathbf{x},\mathbf{z}_i) - (\mathbf{z}_i)^\top \mathbf{d}_{z,i}) \in \partial_{\mathbf{u}_i}\left(-\bar{L}(\mathbf{x},\mathbf{u})\right).$$

3.3.3 Computing the projections

In this section, we discuss the projections needed to apply SGSP and PAPC. Specifically, we will discuss how to project on set U^i and $\tilde{U}^i$, which are more complicated than Z^i. We shall assume that the projection on set Z^i is simple and formulate the projection on U^i in its terms. Note that, in this paper, we use U^i

$$U^i = \{(\tilde{\mathbf{z}}_i, \lambda_i) : \tilde{\mathbf{z}}_i \le \lambda_i Z^i, \lambda_i \ge 0\}. \tag{14}$$

for the PAPC setting, and $\tilde{U}^i$

$$\tilde{U}^i = \{(\tilde{\mathbf{z}}_i, \lambda_i) : \tilde{\mathbf{z}}_i \le \lambda_i Z^i,\ 0 \le \lambda_i \le \bar{\lambda}\}, \tag{15}$$

which uses an additional upper bound on λ_i for the more general SGSP setting. We will start by showing a general way to compute the projection over U^i.

Proposition 3 *The projection of* $\mathbf{u}_i = (\tilde{\mathbf{z}}_i, \lambda_i) \in \mathbb{R}^{d_i+1}$ *on the set* U^i *defined by* (14) *is given by*

$$P_{U^i}(\mathbf{u}_i) = \begin{cases} \mathbf{u}_i & \mathbf{u}_i \in U^i \\ \left(\mu^* P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu^*}\right), \mu^*\right) & \mathbf{u}_i \notin U^i, \sigma_{Z^i}(\tilde{\mathbf{z}}_i) > -\lambda_i \\ \mathbf{0} & \textit{otherwise} \end{cases}$$

where $\mu^* > 0$ *is the unique solution of*

$$\mu \left\| P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)\right\|^2 - \tilde{\mathbf{z}}_i^\top P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right) + \mu - \lambda_i = 0.$$

The same technique used to prove Proposition 3 can be used for the proof of projection over $\tilde{U}^i$.

Corollary 1 *The projection of* $\mathbf{u}_i = (\tilde{\mathbf{z}}_i, \lambda_i) \in \mathbb{R}^{d_i+1}$ *on the set* $\tilde{U}^i$ *defined by* (15) *is given by*

$$P_{\tilde{U}^i}(\mathbf{u}_i) = \left(\mu^* P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu^*}\right), \mu^*\right)$$

Table 2 Examples of projections onto $\tilde{U}$

Z^i	$P_{Z^i}(\mathbf{y})$	$P_{\tilde{U}^i}(\mathbf{y},\lambda) = \left(\mu^* P_{Z^i}\left(\frac{\mathbf{y}}{\mu^*}\right), \mu^*\right)$
$\{\mathbf{z} : \|\mathbf{z}\|_2 \le 1\}$	$\begin{cases} \mathbf{y} & \mathbf{y} \in Z \\ \frac{\mathbf{y}}{\|\mathbf{y}\|} & \text{otherwise} \end{cases}$	$\mu^* = \max\left\{\min\left\{\frac{\lambda+\|\mathbf{y}\|_2}{2}, \bar{\lambda}\right\}, 0\right\}$[2]
$\{\mathbf{z} : \|\mathbf{z}\|_\infty \le 1\}$	$\min\{\mathbf{e}, \lvert\mathbf{y}\rvert\} \circ \operatorname{sign}(\mathbf{y})$	$\mu^* = \min\left\{\frac{\sum_{i\le j^*} \lvert q_{(i)}\rvert + \lambda}{j^*+1}, \bar{\lambda}\right\}$[2] $j^* = \max\left\{j \in [d] : \lvert q_{(j)}\rvert \ge \frac{\sum_{i\le j} \lvert q_{(i)}\rvert + \lambda}{j+1}\right\}$[1].
$\{\mathbf{z} : \|\mathbf{z}\|_1 \le 1\}$	$\max\{\lvert\mathbf{q}\rvert - \theta^*\mathbf{e}, \mathbf{0}\} \circ \operatorname{sign}(\mathbf{q})$ where $\theta^* = \frac{\sum_{i\le j^*} \lvert q_{(i)}\rvert - 1}{j^*}$ $j^* = \max\left\{j \in [d] : \lvert q_{(j)}\rvert \ge \frac{\sum_{i\le j} \lvert q_{(i)}\rvert - 1}{j}\right\}$[1].	$\mu^* = \min\left\{\frac{\sum_{i\le j^*} \lvert q_{(i)}\rvert + \lambda}{j^*+1}, \bar{\lambda}\right\}$[2] $j^* = \max\left\{j \in [d] : \lvert q_{(j)}\rvert \ge \frac{\sum_{i\le j} \lvert q_{(i)}\rvert - \lambda}{j+1}\right\}$[1].

[1] If the problem is not feasible $j^* = -\infty$
[2] For projection over U^i set $\bar{\lambda} = \infty$

where $\mu^* = \max\{\min\{\tilde{\mu}, \bar{\lambda}\}, 0\}$ *with* $\tilde{\mu}$ *being the unique solution of*

$$\tilde{\mu}\left\|P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\tilde{\mu}}\right)\right\|^2 - \tilde{\mathbf{z}}_i^\top P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\tilde{\mu}}\right) + \tilde{\mu} - \lambda_i = 0,$$

and $0 P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{0}\right) = 0$.

Proposition 3 and Corollary 1 suggest that we can always obtain the projection onto sets U^i and $\tilde{U}^i$, which are the conic extension of Z^i, by applying a bi-section procedure to find the value of μ that satisfies the appropriate equality constraint.

Table 2 illustrates three examples, the ℓ_2, ℓ_∞, ℓ_1 balls, for which the projection onto Z^i is obtained either analytically or in $O(n)$, and similarly the value of the optimal μ^*, and therefore the projections onto their conic extensions can also be computed in $O(n)$.

4 Convergence

4.1 EB algorithms

In order to unify analysis for all forms of the Lagrangian function and the algorithms presented in this paper, we define now sufficient properties needed for an algorithm to prove the feasibility and optimality convergence. These sufficient properties will take the form of the algorithm being an *ergodically bounded* (EB) algorithm, as the following definition states.

Definition 1 Let $\mathcal{A}$ be an iterative algorithm to solve the saddle point problem (2). We call the algorithm $\mathcal{A}$ *ψ-ergodic bounded* (ψ-EB) if there exist constant scalars $\phi, \tau, \beta \ge 0$, $\theta_i, \theta_{\mathbf{w}} \in \mathbb{R}_+$, and a function $\psi(N) : \mathbb{N} \to \mathbb{R}_+$ with $\psi(N) \downarrow 0$ as $N \to +\infty$, such that for any initial point $(\mathbf{x}^0; \boldsymbol{\lambda}^0, \mathbf{w}^0)$ the algorithm $\mathcal{A}$ generates a sequence $\{(\mathbf{x}^k, \boldsymbol{\lambda}^k, \mathbf{w}^k)\}_{k\in\mathbb{N}}$ satisfying

$$L(\bar{\mathbf{x}}^N, (\boldsymbol{\lambda}, \mathbf{w})) - L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) + \alpha \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \le$$
$$\psi(N)\left(\tau^{-1}\|\mathbf{x}^* - \mathbf{x}^0\|^2 + \sum_{i=1}^m \theta_i^{-1} \max\{\lambda_i, \lambda_i^0\}^2 + \theta_{\mathbf{w}}^{-1} \max\{\|\mathbf{w}\|, \|\mathbf{w}^0\|\}^2 + \phi(1+\alpha^2)\right),$$

and

$$L(\bar{\mathbf{x}}^N;\boldsymbol{\lambda}^*,\mathbf{w}^*) - L(\mathbf{x}^*;\boldsymbol{\lambda}^*,\mathbf{w}^*) + \beta\,\mathrm{Dist}(\bar{\mathbf{x}}^N, X) \geq 0.$$

for all iterations $N \in \mathbb{N}$, dual variables $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, $\mathbf{w} \in \mathbb{R}^r$, parameters $\alpha \geq 0$ and saddle points $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ of (4).

The following theorem shows that the ergodic sequence of any ψ-EB algorithm converges to feasibility and optimality at the rate at which ψ converges to zero.

Theorem 3 *Let problem* (2) *satisfy Assumption 1 and 2. Let $\mathcal{A}$ be an ψ-EB algorithm to solve the saddle point problem* (2) *with parameters ϕ, τ, β and $\boldsymbol{\theta}$. For a given starting point $(\mathbf{x}^0, (\boldsymbol{\lambda}^0, \mathbf{w}^0)) \in \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$ let $\bar{\mathbf{x}}^N$ be the ergodic primal sequence generated by the algorithm. Then, for any optimal solution $\mathbf{x}^*$ to* (2) *we have that*

$$\sum_{i=1}^m [f_i(\bar{\mathbf{x}}^N)]_+ + \left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| + \mathrm{Dist}(\bar{\mathbf{x}}^N, X) \leq$$
$$\psi(N)\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{\bar{\lambda}+1, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{R_{\mathbf{w}}+1, \left\|\mathbf{w}^0\right\|\}^2}{\theta_{\mathbf{w}}} + \phi(1+(1+\beta)^2)\right),$$

and

$$|\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*)| \leq$$
$$\psi(N)\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{2\bar{\lambda}, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{2R_{\mathbf{w}}, \left\|\mathbf{w}^0\right\|\}^2}{\theta_{\mathbf{w}}} + \phi(1+4\beta^2)\right).$$

Proof Let $\kappa_1 > 0$ and let $\mathbf{r} \in \mathbb{R}^m_+$ be a multiplication of indicator vector of constraint violations by κ_1, *i.e.*:

$$r_i = \begin{cases} \kappa_1, & f_i(\bar{\mathbf{x}}^N) > 0, \\ 0, & \text{otherwise}, \end{cases},\ i \in [m]$$

and let $\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + \mathbf{r}$ and let $\kappa_2 > 0$ and $\mathbf{w} = \mathbf{w}^* + \kappa_2 \frac{\left(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right)}{\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\|}$ where $(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*)$ is a saddle point of (3). Then,

$$\begin{aligned} L(\bar{\mathbf{x}}^N;\boldsymbol{\lambda},\mathbf{w}) &= \mathbf{c}^\top\bar{\mathbf{x}}^N + \sum_{i=1}^m \lambda_i f_i(\bar{\mathbf{x}}^N) + \mathbf{w}^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) \\ &= \mathbf{c}^\top\bar{\mathbf{x}}^N + \sum_{i=1}^m (\lambda_i^* + r_i) f_i(\bar{\mathbf{x}}^N) + \mathbf{w}^*\left(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right) + \kappa_2\left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| \\ &= \mathbf{c}^\top\bar{\mathbf{x}}^N + \sum_{i=1}^m \lambda_i^* f_i(\bar{\mathbf{x}}^N) + \kappa_1\sum_{i=1}^m [f_i(\bar{\mathbf{x}}^N)]_+ + \mathbf{w}^*\left(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right) + \kappa_2\left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| \\ &= L(\bar{\mathbf{x}}^N;\boldsymbol{\lambda}^*,\mathbf{w}^*) + \kappa_1\sum_{i=1}^m [f_i(\bar{\mathbf{x}}^N)]_+ + \kappa_2\left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| \end{aligned} \tag{16}$$

Since $\mathcal{A}$ is an ψ-EB algorithm then

$$L(\bar{\mathbf{x}}^N;\boldsymbol{\lambda}^*,\mathbf{w}^*) + \beta\,\mathrm{Dist}(\bar{\mathbf{x}}^N, X) \geq \mathbf{c}^\top\mathbf{x}^* = L(\mathbf{x}^*;\boldsymbol{\lambda}^*,\mathbf{w}^*), \tag{17}$$

and for all $\kappa_3 > 0$

$$L(\bar{\mathbf{x}}^N; \boldsymbol{\lambda}, \mathbf{w}) - L(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*) + \kappa_3 \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \leq$$
$$\psi(N)\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{\lambda_i, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{\|\mathbf{w}\|, \left\|\mathbf{w}^0\right\|\}^2}{\theta_{\mathbf{w}}} + \phi(1+\kappa_3^2)\right). \tag{18}$$

Since Assumptions 1 and 2 hold, it follows from Proposition 2 that $\lambda_i^* \leq \bar{\lambda}$ and thus

$$\max\{\lambda_i, \lambda_i^0\} = \max\{\lambda_i^* + r_i, \lambda_i^0\} \leq \max\{\bar{\lambda} + \kappa_1, \lambda_i^0\}.$$

Similarly, since $\|\mathbf{w}^*\| \leq R_{\mathbf{w}}$, thus,

$$\begin{aligned}\max\{\|\mathbf{w}\|, \left\|\mathbf{w}^0\right\|\} &= \max\left\{\left\|\mathbf{w}^* + \kappa_2 \frac{(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b})}{\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\|}\right\|, \left\|\mathbf{w}^0\right\|\right\} \\ &\leq \max\{\|\mathbf{w}^*\| + \kappa_2, \left\|\mathbf{w}^0\right\|\} \\ &\leq \max\{R_{\mathbf{w}} + \kappa_2, \left\|\mathbf{w}^0\right\|\}. \end{aligned} \tag{19}$$

Combining inequalities (16)-(19) we obtain

$$\begin{aligned}&\kappa_1 \sum_{i=1}^m [f_i(\bar{\mathbf{x}}^N)]_+ + \kappa_2 \|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\| + (\kappa_3 - \beta) \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \\ \leq &\psi(N)\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{\bar{\lambda} + \kappa_1, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{R_{\mathbf{w}} + \kappa_2, \left\|\mathbf{w}^0\right\|\}^2}{\theta_{\mathbf{w}}} + \phi(1+\kappa_3^2)\right). \end{aligned} \tag{20}$$

Setting $\kappa_1 = \kappa_2 = 1$ and $\kappa_3 = \beta + 1$ we obtain the first result. For the second result, we need to upper and lower bound $\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*)$. First, we have

$$\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*) = L(\bar{\mathbf{x}}^N; \mathbf{0}, \mathbf{0}) - L(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*). \tag{21}$$

where the equalities follow from $L(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*) = \mathbf{c}^\top \mathbf{x}^*$ and $L(\bar{\mathbf{x}}^N; \mathbf{0}, \mathbf{0}) = \mathbf{c}^\top \bar{\mathbf{x}}^N$. Using (18) with $\kappa_3 = 0$ we obtain

$$L(\bar{\mathbf{x}}^N; \mathbf{0}, \mathbf{0}) - L(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*) \leq \psi(N)\left(\tau^{-1}\|\mathbf{x}^* - \mathbf{x}^0\|^2 + \sum_{i=1}^m \theta_i^{-1}(\lambda_i^0)^2 + \theta_{\mathbf{w}}^{-1}\left\|\mathbf{w}^0\right\|^2 + \phi\right). \tag{22}$$

Combining (21) with (22) we have that

$$\mathbf{c}^\top(\bar{\mathbf{x}}^N - \mathbf{x}^*) \leq \psi(N)\left(\tau^{-1}\|\mathbf{x}^* - \mathbf{x}^0\|^2 + \sum_{i=1}^m \theta_i^{-1}(\lambda_i^0)^2 + \theta_{\mathbf{w}}^{-1}\left\|\mathbf{w}^0\right\|^2 + \phi\right). \tag{23}$$

To obtain the other side of the bound, we use (17) to obtain

$$\mathbf{c}^\top(\mathbf{x}^* - \bar{\mathbf{x}}^N) - \sum_{i=1}^m \lambda_i^* f_i(\bar{\mathbf{x}}^N) - (\mathbf{w}^*)^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) = L(\mathbf{x}^*; \boldsymbol{\lambda}^*, \mathbf{w}^*) - L(\bar{\mathbf{x}}^N; \boldsymbol{\lambda}^*, \mathbf{w}^*) \leq \beta \operatorname{Dist}(\bar{\mathbf{x}}^N, X),$$

which in turn implies

$$\begin{aligned}\mathbf{c}^\top(\mathbf{x}^* - \bar{\mathbf{x}}^N) &\leq \sum_{i=1}^m \lambda_i^* f_i(\bar{\mathbf{x}}^N) + (\mathbf{w}^*)^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) + \beta\,\mathrm{Dist}(\bar{\mathbf{x}}^N, X)\\ &\leq \bar{\lambda}\sum_{i=1}^m [f_i(\bar{\mathbf{x}}^N)]_+ + R_{\mathbf{w}}\left\|\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right\| + \beta\,\mathrm{Dist}(\bar{\mathbf{x}}^N, X), \end{aligned} \tag{24}$$

where the last inequality follows from the the fact that $0 \leq \lambda_i^* \leq \bar{\lambda}$ and $\|\mathbf{w}^*\| \leq R_{\mathbf{w}}$. Using (20) with $\kappa_1 = \bar{\lambda}$, $\kappa_2 = R_{\mathbf{w}}$, and $\kappa_3 = 2\beta$ to bound (24) we obtain that

$$\mathbf{c}^\top(\mathbf{x}^* - \bar{\mathbf{x}}^N) \leq \psi(N)\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + \sum_{i=1}^m \frac{\max\{2\bar{\lambda}, \lambda_i^0\}^2}{\theta_i} + \frac{\max\{2R_{\mathbf{w}}, \left\|\mathbf{w}^0\right\|\}^2}{\theta_{\mathbf{w}}} + \phi(1 + (2\beta)^2)\right) \tag{25}$$

Combining (23) and (25) we obtain the desired result. □

4.2 Generalized model

So far, we kept the problem formulations simple and stated the convergence results without proofs. In this section, we shall show the convergence of both the SGSP and PAPC algorithms for a more general model using the EB-algorithm definition.

The general model is needed to tackle the fact that projections onto sets X or Z^i are the main tools used in both algorithms, and therefore these projections should be simple. Two issues that might arise is that either the primal domain X or at least one of the sets Z^i is not 'simple', but instead, it is an intersection of several simple sets.

Example 1 One of the popular uncertainty sets in RO is the so-called budgeted uncertainty set, formulated as:

$$Z = \{\mathbf{z} :\ -1 \leq \mathbf{z} \leq 1,\ \|\mathbf{z}\|_1 \leq \Gamma\}$$

which is an intersection of two simple sets: the ℓ_∞ norm and ℓ_1 norm balls.

Our strategy for dealing with these complex sets shall be to disentangle the projections onto the intersected sets. We will achieve this by including 'copies' of the respective primal or dual variables, together with the relevant equality constraints, which are to be relaxed in the Lagrangian. The saddle-point algorithm would then be applied to the new Lagrangian problem.

Consider, for example, the case where $X = \cap_{j=1}^q X_i$ is an intersection of several 'simple' sets. Formulation 2 is ready to handle this situation easily. One can 'expand' the vector $\mathbf{x}$ by having q copies of it: $\mathbf{x} \mapsto [\mathbf{x}_1, \ldots, \mathbf{x}_q]$. In the next step, each inequality constraint in the problem is made dependent only on one of the $\mathbf{x}_i$'s and at the same time, equality constraints $\mathbf{x}_1 = \mathbf{x}_i$, $i = 2, \ldots, q$ become embedded in the $\mathbf{A}\mathbf{x} = \mathbf{b}$ system (where the rank condition is easily verified).

Similarly, we can consider the case where $Z^i = \cap_{l=1,\ldots,s_i} Z^{i,l}, \quad i = 1, \ldots, m$ with $Z^{i,l}$ being compact convex sets. In this case, the constraint $\mathbf{z}_i \in \cap_{l=1,\ldots,s_i} Z^{i,l}$ can be written using concatenated uncertain parameter vector $\hat{\mathbf{z}}_i = (\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,s_i})$, defining

$$\hat{Z}^i = \left\{(\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,s_i}) \in Z^{i,1} \times \ldots \times Z^{i,s_i} :\ \mathbf{z}_{i,l} = \mathbf{z}_{i,s_i},\ l = 1, \ldots, s_i - 1\right\},$$

and formulating the constraints as

$$\max_{\hat{\mathbf{z}}_i \in \hat{Z}^i} \hat{g}_i(\mathbf{x}_1, \hat{\mathbf{z}}_i) \equiv g_i(\mathbf{x}_1, \mathbf{z}_{i,s_i}) \leq 0$$

where $Z^{i,l}$, $l = 1, \ldots, s_i$ are 'simple' sets on which it is easy to project. Thus, the general model we are going to address is the following:

$$\begin{aligned} &\min_{\mathbf{x} \in X} \mathbf{c}^\top \mathbf{x} \\ &\text{s.t. } f_i(\mathbf{x}) \coloneqq \max_{\mathbf{z}_i \in Z^i} g_i(\mathbf{x}, \mathbf{z}_i) \leq 0 \quad i \in [m] \\ &\mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned} \tag{26}$$

where $Z^i = \cap_{l=1}^{s_i} Z^{i,l}$. For both SGSP and PAPC we shall give the problem generalizations together with the expanded Lagrangian that is to be solved in Sections 4.3 and 4.4.

4.3 SGSP convergence

Consider problem (26) where we need to formulate a saddle point problem that disentangles different components of the $\mathbf{x}$ and $\mathbf{z}_i$. We start by considering a saddle point $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*))$ of $\bar{L}(\mathbf{x}, (\mathbf{u}, \mathbf{w}))$ over the sets $\mathbf{x} \in X$ and $\mathbf{u}_i \in U^i$, which we already proved exists. Thus, by the definition of the saddle point, we have that

$$\tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*) = \sup_{\mathbf{u}_i \in U^i} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i) = \sup_{\substack{\mathbf{u}_{i,l} \in U^{i,l}, \\ u_{i,s_i} = u_{i,l}, l \in [s_i - 1]}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i})$$

where the last equality follows from the definition of $U^i = \cap_{i \in [s_i]} U^{i,l}$ where $U^{i,l} = \{(\tilde{\mathbf{z}}, \lambda_i) : \tilde{\mathbf{z}} \in \lambda_i Z^{i,l}\}$, similarly to what was explained in the previous section. Note that the leftmost supremum has a solution ($\mathbf{u}_{i,1} = \mathbf{u}_{i,2} = \ldots = \mathbf{u}_{i,s_i} = \mathbf{u}_i^*$). Moreover, since $U^i \subseteq U^{i,l}$ for $l \in [s_i]$, the sets $U^{i,l}$ have a nonempty interior, therefore, dualizing the equality constraints, we have that strong duality holds. Thus, denoting $\tilde{\mathbf{u}}_i^* = (\mathbf{u}_{i,1}^*, \ldots, \mathbf{u}_{i,s_i}^*) \equiv (\mathbf{u}_i^*, \ldots, \mathbf{u}_i^*)$ there exists $\boldsymbol{\omega}_i^* = (\boldsymbol{\omega}_{i,1}^*, \ldots, \boldsymbol{\omega}_{i,s_i-1}^*)$ such that $(\boldsymbol{\omega}_i^*, \tilde{\mathbf{u}}_i^*)$ satisfies

$$\begin{aligned} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i^*) &= \sup_{\substack{\mathbf{u}_{i,l} \in U^{i,l}, \\ u_{i,s_i} = u_{i,l}, l \in [s_i - 1]}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) \\ &= \inf_{\omega_{i,l} \in \mathbb{R}^{d_i+1}, l \in [s_i-1]} \sup_{\mathbf{u}_{i,l} \in U^{i,l}, l \in [s_i-1]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \\ &= \inf_{\boldsymbol{\omega}_{i,l},\ l \in [s_i-1]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}^*) + \sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l}^* - \mathbf{u}_{i,s_i}^*) \\ &= \sup_{\mathbf{u}_{i,l} \in U^{i,l},\ l \in [s_i-1]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1} (\boldsymbol{\omega}_{i,l}^*)^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}). \end{aligned} \tag{27}$$

Due to Proposition 2 we know that we can restrict U^i to $\tilde{U}^i$, and thus, the existence of the saddle point in this case follows the same logic where $U^{i,l}$ is replaced by $\tilde{U}^{i,l} = \{\mathbf{u}_{i,l} \equiv (\tilde{\mathbf{z}}_{i,l}, \lambda_{i,l}) \in U^{i,l} :$

$\lambda_{i,l} \leq \bar{\lambda}\}$. In the following, we will show that we can restrict the domain over which we optimize $\boldsymbol{\omega}$ and still retrieve a saddle point of the original problem.

Denoting $\boldsymbol{\chi} = (\mathbf{x}, \boldsymbol{\omega})$ and $\mathbf{y} = (\tilde{\mathbf{u}}, \mathbf{w})$ where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m)$, $\tilde{\mathbf{u}} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m)$, and $\boldsymbol{\omega}_i = (\boldsymbol{\omega}_{i1}, \ldots, \boldsymbol{\omega}_{i(s_i-1)})$, $\tilde{\mathbf{u}}_i = (\mathbf{u}_{i1}, \ldots, \mathbf{u}_{is_i})$ for all $i \in [m]$, we can define a saddle point in the lifted space. Indeed, let $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ be a saddle point of $L(\mathbf{x}, (\boldsymbol{\lambda}, \mathbf{w}))$, then, by Proposition 1, and the reasoning leading to (27) above, there exists $\tilde{\mathbf{u}}^*$ and $\boldsymbol{\omega}^*$ such that

$$\begin{aligned}
&\sup_{\boldsymbol{\lambda}\geq 0, \mathbf{w}} L(\mathbf{x}^*, (\boldsymbol{\lambda}, \mathbf{w})) \\
&= \sup_{\substack{\mathbf{w}\in W, \\ \mathbf{u}_i \in U^i,\ i\in[m]}} \mathbf{c}^\top \mathbf{x}^* + \mathbf{w}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i) \\
&= \sup_{\substack{\mathbf{w}\in W, \\ \mathbf{u}_{i,l} \in U^{i,l},\ i\in[m], l\in[s_i-1]}} \mathbf{c}^\top \mathbf{x}^* + \mathbf{w}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_{i,s_i}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} (\boldsymbol{\omega}_{i,l}^*)^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i})
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{\mathbf{x}\in X} L(\mathbf{x}, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \\
&= \inf_{\mathbf{x}\in X} \mathbf{c}^\top \mathbf{x} + (\mathbf{w}^*)^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_i^*) \\
&= \inf_{\substack{\mathbf{x}\in X \\ \boldsymbol{\omega}_{i,l},\ i\in[m], l\in[s_i-1]}} \mathbf{c}^\top \mathbf{x}^* + (\mathbf{w}^*)^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_{i,s_i}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} (\boldsymbol{\omega}_{i,l})^\top (\mathbf{u}_{i,l}^* - \mathbf{u}_{i,s_i}^*).
\end{aligned}$$

Defining

$$\breve{L}(\boldsymbol{\chi}, \mathbf{y}) := \mathbf{c}^\top \mathbf{x} + \mathbf{w}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i(\mathbf{x}, \mathbf{u}_{i,s_i}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}). \tag{28}$$

we obtain that $(\boldsymbol{\chi}^*, \mathbf{y}^*)$ is a saddle point of $\breve{L}$. Following similar logic, we can also obtain that given a saddle point $(\boldsymbol{\chi}^*, \mathbf{y}^*)$ of $\breve{L}$ with $\mathbf{u}_{i,s_i}^* = (\tilde{\mathbf{z}}_{i,s_i}^*, \lambda_{i,s_i}^*)$ we can obtain a saddle point $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ of L by defining $\lambda_i^* = \lambda_{i,s_i}^*$, $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$.

Since we proved that $\breve{L}$ has a saddle point, we can now show that SGSP algorithm applied to (28) meets the EB-algorithm assumptions. However, in order to run the algorithm and prove its convergence, we need to show that the feasible sets of the variables can be restricted without losing a saddle point. Legitimacy of bounding $\mathbf{u}_{i,l}$ and $\mathbf{w}$ follows from Proposition 2 and (27). The following proposition establishes our ability to bound $\boldsymbol{\omega}_{i,l}$.

Proposition 4 *Let Assumptions 1–5 hold. Let $i \in [m]$ and consider the saddle point problem*

$$\min_{\boldsymbol{\omega}_i} \max_{\substack{\mathbf{u}_{i,l}\in\bar{U}^{i,l},\ l\in[s_i], \\ \underline{\lambda}\leq\lambda_{i,s_i}\leq\bar{\lambda}}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \tag{29}$$

where $\mathbf{x}^*$ *is an optimal solution of* (26)*, and* $0 \le \underset{\sim}{\lambda} \le \tilde{\lambda} \le \bar{\lambda}$*. Then, there exists a saddle point* $(\boldsymbol{\omega}_i^*, \tilde{\mathbf{u}}_i^*)$ *with* $\boldsymbol{\omega}_{i,l}^* = (\boldsymbol{\nu}_{i,l}^*, \mu_{i,l}^*)$ *such that* $\boldsymbol{\omega}_{i,l}^*$ *is contained in the set*

$$\Omega^{i,l} = \left\{ \boldsymbol{\omega}_{i,l} = (\boldsymbol{\nu}_{i,l}, \mu_{i,l}) : -\bar{\mu}_i \le \mu_{i,l} \le 0,\ \|\boldsymbol{\nu}_{i,l}\| \le \frac{-\mu_{i,l}}{\epsilon_i},\ i \in [m],\ l \in [s_i - 1] \right\},$$

where ϵ_i *and* $\bar{\mu}_i$ *is given by Assumptions 4 and 5, respectively.*

Now that we have shown that we can bound all variables in the saddle point of function $\breve{L}(\boldsymbol{\chi}, \mathbf{y})$ over the sets $\mathcal{X} = X \times (\times_{i=1}^m \times_{l=1}^{s_i - 1} \Omega^{i,l})$ and $\mathcal{Y} = (\times_{i=1}^m \times_{l=1}^{s_i} U^{i,l}) \times W$, we can apply SGSP to the problem with these bounded sets as follows.

SGSP for $\breve{L}$
Input: $\tau > 0, \theta_i > 0, \theta_{\mathbf{w}} > 0$ for $i \in [m]$, and $N \in \mathbb{N}$
Initialization. Initialize $\boldsymbol{\chi}^0 \in \mathcal{X}$, and $\mathbf{y}^0 \in \mathcal{Y}$ such that $\lambda_{i,l}^0 = \lambda_{i,1}^0$ for all $l \in [s_i]$,$i \in [m]$, and define the diagonal matrix $\Theta = \mathrm{Diag}(\theta_1 \mathbf{e}; \ldots; \theta_m \mathbf{e}; \theta_{\mathbf{w}} \mathbf{e})$.
General step: For $k \in [N]$
Compute subgradients $\mathbf{v}_\chi^{k-1} = (\mathbf{v}_x^{k-1}, \mathbf{v}_\omega^{k-1}) \in \partial_{\boldsymbol{\chi}} \breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1})$, and for all $i \in [m]$ $\mathbf{v}_{\tilde{\mathbf{u}}_i}^{k-1} = (\mathbf{v}_{i,1}^{k-1}, \ldots, \mathbf{v}_{i,s_i}^{k-1}) \in \partial_{\tilde{\mathbf{u}}_i} \left(-\breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1}) \right)$.

$$\begin{aligned}
\boldsymbol{\chi}^k = P_{\mathcal{X}}(\boldsymbol{\chi}^{k-1} - \tau \mathbf{v}_\chi^{k-1}) \quad &\Longleftrightarrow \quad \mathbf{x}^k = P_X(\mathbf{x}^{k-1} - \tau \mathbf{v}_x^{k-1}), \\
&\quad\ \boldsymbol{\omega}_{i,l}^k = P_{\Omega^{i,l}}(\boldsymbol{\omega}_{i,l}^{k-1} - \tau(\mathbf{u}_{i,l}^{k-1} - \mathbf{u}_{i,s_i}^{k-1})), \quad i \in [m],\ l \in [s_i - 1] \\
\mathbf{y}^k = P_{\mathcal{Y}}(\mathbf{y}^{k-1} - \Theta \mathbf{v}_y^{k-1}) \quad &\Longleftrightarrow \quad \mathbf{u}_{i,l}^k = P_{\tilde{U}^{i,l}}(\mathbf{u}_{i,l}^{k-1} + \theta_i \boldsymbol{\omega}_{i,l}^{k-1}), \quad i \in [m],\ l \in [s_i - 1] \\
&\quad\ \mathbf{u}_{i,s_i}^k = P_{\tilde{U}^{i,s_i}}(\mathbf{u}_{i,s_i}^{k-1} - \theta_i \mathbf{v}_{i,s_i}^{k-1}), \quad i \in [m] \\
&\quad\ \mathbf{w}^k = P_W(\mathbf{w}^{k-1} + \theta_{\mathbf{w}}(\mathbf{A}\mathbf{x}^{k-1} - \mathbf{b})).
\end{aligned}$$

We will now prove that SGSP is an EB-algorithm and thus, it converges to the optimal solution.

Proposition 5 *Let Assumptions 1–5 hold. And let* $\{(\chi^k, \mathbf{y}^k)\}_{k \in \mathbb{N}}$ *be the sequence generated by SGSP algorithm with parameters* $\tau = \frac{\tilde{\tau}}{\sqrt{N}}$, $\theta_i = \frac{\tilde{\theta}_i}{\sqrt{N}}$, $\theta_{\mathbf{w}} = \frac{\tilde{\theta}_w}{\sqrt{N}}$, *where* $\boldsymbol{\chi}^k = (\mathbf{x}^k, \boldsymbol{\omega}^k)$ *and* $\mathbf{y}^k = (\tilde{\mathbf{u}}^k, \mathbf{w}^k)$*. Then,*

$$\begin{aligned}
&\breve{L}(\bar{\boldsymbol{\chi}}^N, \mathbf{y}) - \breve{L}(\boldsymbol{\chi}, \bar{\mathbf{y}}^N) \\
&\quad \le \frac{1}{2\sqrt{N}} \left(\frac{\left\| \boldsymbol{\chi}^0 - \boldsymbol{\chi} \right\|^2}{\tilde{\tau}} + \tilde{\tau} G_{\boldsymbol{\chi}}^2 + \sum_{i=1}^m \left(\frac{\left\| \tilde{\mathbf{u}}_i^0 - \tilde{\mathbf{u}}_i \right\|^2}{\tilde{\theta}_i} + \tilde{\theta}_i G_{\tilde{\mathbf{u}}_i}^2 \right) + \frac{\left\| \mathbf{w}^0 - \mathbf{w} \right\|^2}{\tilde{\theta}_{\mathbf{w}}} + \tilde{\theta}_{\mathbf{w}} G_{\mathbf{w}}^2 \right)
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
G_{\boldsymbol{\chi}} &\coloneqq \|\mathbf{c}\| + \|\mathbf{A}\| R_{\mathbf{w}} + \sum_{i \in [m]} \bar{\lambda} G_{x,i} + \bar{\lambda} \sum_{i \in [m]} \left((s_i - 1)(2 + R_{i,s_i}) + \sum_{l \in [s_i - 1]} R_{i,l} \right) \\
G_{\tilde{\mathbf{u}}_i} &\coloneqq (\bar{\lambda} + R_{i,s_i}) G_{z,i} + \bar{g}_i + 2\bar{\mu}_i (s_i - 1) \left(1 + \frac{1}{\epsilon_i} \right) \\
G_{\mathbf{w}} &\coloneqq \|\mathbf{A}\|\, R_{\mathbf{x}} + \|\mathbf{b}\|
\end{aligned}$$

Moreover, SGSP is an EB-algorithm. More precisely, let $\mathbf{x}^*$ *an optimal solution of* (26)*,* $(\boldsymbol{\lambda}^*, \mathbf{w}^*)$ *be the optimal dual variables associated with its constraints, and let* $\boldsymbol{\lambda} \in \mathbb{R}^m_+$*, and* $\mathbf{w} \in \mathbb{R}^r$ *such that* $\lambda_i \equiv \lambda_{i,s_i}$ *for all* $i \in [m]$*. Then,*

$$L(\bar{\mathbf{x}}^N, (\boldsymbol{\lambda}, \mathbf{w})) - L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \leq \frac{1}{2\sqrt{N}} \left(\tilde{\tau}^{-1} \|\mathbf{x}^* - \mathbf{x}^0\|^2 + \sum_{i=1}^m \sigma_i \tilde{\theta}_i^{-1} \max\{\lambda_i, \lambda_i^0\}^2 + 2\tilde{\theta}_{\mathbf{w}}^{-1} \max\{\|\mathbf{w}\|, \|\mathbf{w}_0\|\}^2 + \phi \right), \tag{31}$$

and

$$L(\bar{\mathbf{x}}^N; (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \geq \mathbf{c}^\top \mathbf{x}^*, \tag{32}$$

where

$$\sigma_i = \sum_{l=1}^{s_i} (1 + 4R_{i,l}^2)$$

$$\phi = \tilde{\tau} G_{\boldsymbol{\chi}}^2 + \sum_{i=1}^m \tilde{\theta}_i G_{\mathbf{u}_i}^2 + 4 \frac{\left(\sum_{i=1}^m \bar{\mu}_i (s_i - 1) \left(1 + \frac{1}{\epsilon_i} \right) \right)^2}{\tilde{\tau}} + \tilde{\theta}_{\mathbf{w}} G_{\mathbf{w}}^2$$

where $R_{\mathbf{x}}$ *is the radius of* X*,* $R_{\mathbf{w}}$ *the radius of* W*,* $R_{i,l}$ *is the radius of* $Z^{i,l}$*,* $\bar{g}_i = \max_{\mathbf{x} \in X, \mathbf{z}_i \in Z^i} |g_i(\mathbf{x}, \mathbf{z}_i)|$*,* $\bar{\lambda}$*,* $\bar{\mu}_i$ *and* ϵ_i *are defined in Proposition 2, Assumption 5 and Assumption 4, respectively, and* $G_{x,i}, G_{z,i}$ *are the bounds on the subgradients* $\mathbf{d}_{x,i}^k$*,* $\mathbf{d}_{z,i}^k$ *generated as in Lemma 2.*

4.4 PAPC convergence

In the case of PAPC, we proceed similarly to SGSP with one change – here, we need to eliminate the explicit constraint that $\mathbf{x}$ lies in X, which is also not assumed to be bounded. We achieve this, as in the simple case, by performing the dual transportation trick. Compared to (28), we do an extra step to obtain:

$$\begin{aligned}
&\sup_{\boldsymbol{\lambda} \geq 0, \mathbf{w}} \inf_{\mathbf{x} \in X} L(\mathbf{x}, (\boldsymbol{\lambda}, \mathbf{w})) \\
&= \inf_{\boldsymbol{\omega}_{i,l}, \mathbf{x}} \sup_{\boldsymbol{\pi}, \mathbf{w}, \mathbf{u}_{i,l} \in U^{i,l}} \mathbf{c}^\top \mathbf{x} + \mathbf{w}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) + \sum_{i=1}^m \tilde{g}_i (\mathbf{x}, \mathbf{u}_{i,s_i}) + \sum_{i=1}^m \sum_{l=1}^{s_i - 1} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) + \boldsymbol{\pi}^\top \mathbf{x} - \sigma_X(\boldsymbol{\pi}) \\
&= \inf_{\boldsymbol{\chi}} \sup_{\mathbf{y} = (\tilde{\mathbf{u}}, \mathbf{w}, \boldsymbol{\pi}) : \mathbf{u}_{i,l} \in U^{i,l}, l \in [s_i],\ i \in [m]} \check{L}(\boldsymbol{\chi}, \mathbf{y})
\end{aligned}$$

where $\sigma_X(\boldsymbol{\pi})$ is the support function of X and $\boldsymbol{\chi} = (\mathbf{x}, \boldsymbol{\omega})$, $\mathbf{y} = (\tilde{\mathbf{u}}, \mathbf{w}, \boldsymbol{\pi})$, and recall that where $\boldsymbol{\omega}_i = (\boldsymbol{\omega}_{i,1}, \ldots, \boldsymbol{\omega}_{i,s_i - 1})$, $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m)$, $\tilde{\mathbf{u}}_i = (\mathbf{u}_{i,1}, \ldots, \mathbf{u}_{i,s_i})$, and $\tilde{\mathbf{u}} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_m)$.

As the PAPC algorithm presumes bilinearity, we assume f_i are of the form given in (8), we obtain:

$$\begin{aligned}\check{L}(\boldsymbol{\chi},\mathbf{y}) \equiv \mathbf{x}^\top &\left(\mathbf{c} + \sum_{i\in[m]} \tilde{\mathbf{Q}}_i \mathbf{u}_{i,s_i} + \mathbf{A}^\top \mathbf{w} + \boldsymbol{\pi}\right) + \\ &\sum_{i\in[m]} \left(\tilde{\mathbf{q}}_i^\top \mathbf{u}_{i,s_i} + \sum_{l\in[s_i-1]} \boldsymbol{\omega}_{i,l}^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) - \sum_{l\in[s_i]} \delta_{\tilde{U}^{i,l}}(\mathbf{u}_{i,l})\right) - \mathbf{b}^\top \mathbf{w} - \sigma_X(\boldsymbol{\pi}) \qquad (33)\\ = \mathbf{c}^\top \mathbf{x} + \boldsymbol{\chi}^\top \bar{\mathbf{Q}} \mathbf{y} &+ \sum_{i\in[m]} \left(\tilde{\mathbf{q}}_i^\top \mathbf{u}_{i,s_i} - \mathbf{b}^\top \mathbf{w} - \sum_{l\in[s_i]} \delta_{\tilde{U}^{i,l}}(\mathbf{u}_{i,l}) - \sigma_X(\boldsymbol{\pi})\right),\end{aligned}$$

where $\tilde{\mathbf{Q}}_i$ and $\tilde{\mathbf{q}}_i$ are defined in (9), and

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \dots & & \mathbf{0} & -\mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & -\mathbf{I} \\ \vdots & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & & \\ \mathbf{0} & & \dots & \mathbf{0} & \mathbf{I} & -\mathbf{I} \end{bmatrix} \in \mathbb{R}^{(d_i+1)(s_i-1)\times(d_i+1)s_i}, \; \bar{\mathbf{Q}} = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{Q}}_1 & \mathbf{0} & \tilde{\mathbf{Q}}_2 & \dots & \mathbf{0} & \tilde{\mathbf{Q}}_m & \mathbf{A} & \mathbf{I} \\ \mathbf{P}_1 & \mathbf{0} & & & \dots & & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \dots & & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \ddots & & \vdots & & \vdots \\ \mathbf{0} & & \dots & & \mathbf{0} & \mathbf{P}_m & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}$$

Now, we give a short argument for the boundedness of the optimal $\boldsymbol{\pi}^*$. For the case of PAPC the boundedness of variables $\boldsymbol{\omega}$, $\mathbf{w}$ and $\mathbf{u}$ at the saddle point follows by the same arguments as in Propositions 2 and 4. From here, we argue that at the saddle point $\boldsymbol{\pi}$ can be bounded as well.

Proposition 6 *Under Assumptions 1-3, at the saddle point* $(\boldsymbol{\chi}^*, \boldsymbol{\nu}^*)$ *of* (33) *it holds that*

$$\|\boldsymbol{\pi}^*\| \le R_{\boldsymbol{\pi}} := \|\mathbf{c}\| + \bar{\lambda} \sum_{i=1}^m \left\|\tilde{\mathbf{Q}}_i\right\| (R_{i,s_i} + 1) \qquad (34)$$

Proof Due to optimality conditions w.r.t. $\mathbf{x}$ in (33) we must have

$$\mathbf{c} + \sum_{i=1}^m \tilde{\mathbf{Q}}_i \mathbf{u}_{i,s_i}^* + \mathbf{A}\mathbf{w} + \boldsymbol{\pi}^* = 0.$$

Thus, we have

$$\|\boldsymbol{\pi}^*\| = \left\|\mathbf{c} + \sum_{i=1}^m \tilde{\mathbf{Q}}_i \mathbf{u}_{i,s_i}^*\right\| \le \|\mathbf{c}\| + \|\mathbf{A}\| R_{\mathbf{w}} + \bar{\lambda} \sum_{i=1}^m \left\|\tilde{\mathbf{Q}}_i\right\| (R_{i,s_i} + 1)$$

where the last inequality follows from the definition of induced norm, and the boundedness of $\mathbf{u}_{i,s_i}^*$ and $\mathbf{w}$, proven in Proposition 2. □

Note again, that in the case of PAPC, while the bounds on $\boldsymbol{\omega}^*$, $\tilde{\mathbf{u}}^*$, $\mathbf{w}^*$, $\boldsymbol{\pi}^*$ exist, they are needed to prove convergence, but not to apply the algorithm. Now we can apply PAPC to generalized problem as follows.

PAPC for $\check{L}$
Input: $\tau > 0, \theta_i > 0$ for $i \in [m]$, and $N \in \mathbb{N}$
Initialization. Initialize $\boldsymbol{\chi}^0 = (\mathbf{x}^0, \boldsymbol{\omega}^0) \in \mathbb{R}^n \times \mathbb{R}^{\sum_{i\in[m]}(s_i-1)}$, $\mathbf{y}^0 = (\tilde{\mathbf{u}}^0, \mathbf{w}^0, \boldsymbol{\pi}^0) \in \times_{i\in[m]} \times_{l\in[s_i]} U^{i,l} \times \mathbb{R}^r \times \mathbb{R}^m$.
General step: For $k \in [N]$:

Predictor:

$$\mathbf{p}_{\mathbf{x}}^k = \mathbf{x}^{k-1} - \tau \left(\mathbf{c} + \sum_{i\in[m]} \tilde{\mathbf{Q}}_i \mathbf{u}_{i,s_i}^{k-1} + \mathbf{A}^\top \mathbf{w}^{k-1} + \boldsymbol{\pi}^{k-1} \right),$$
$$\mathbf{p}_{\boldsymbol{\omega}_{i,l}}^k = \boldsymbol{\omega}_{i,l}^{k-1} - \tau(\mathbf{u}_{i,l}^{k-1} - \mathbf{u}_{i,s_i}^{k-1})), \qquad i \in [m],\ l \in [s_i - 1].$$

Dual update:

$$\mathbf{u}_{i,l}^k = P_{U^{i,l}} \left(\mathbf{u}_{i,l}^{k-1} + \theta_i \mathbf{p}_{\boldsymbol{\omega}_{i,l}}^k \right), \qquad i \in [m],\ l \in [s_i - 1]$$
$$\mathbf{u}_{i,s_i}^k = P_{U^{i,s_i}} \left(\mathbf{u}_{i,s_i}^{k-1} + \theta_i \left(\tilde{\mathbf{Q}}_i^\top \mathbf{p}_{\mathbf{x}}^k + \tilde{\mathbf{q}}_i - \sum_{l\in[s_i-1]} \mathbf{p}_{\boldsymbol{\omega}_{i,l}}^k) \right) \right), \quad i \in [m]$$
$$\mathbf{w}^k = \mathbf{w}^{k-1} + \theta_{\mathbf{w}}(\mathbf{A}\mathbf{p}_{\mathbf{x}}^k - \mathbf{b}),$$
$$\boldsymbol{\pi}^k = \mathrm{prox}_{\theta_{\boldsymbol{\pi}}\sigma_X}(\boldsymbol{\pi}^{k-1} + \theta_{\boldsymbol{\pi}} \mathbf{p}_{\mathbf{x}}^k).$$

Corrector:

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \tau \left(\mathbf{c} + \sum_{i\in[m]} \tilde{\mathbf{Q}}_i \mathbf{u}_{i,s_i}^k + \mathbf{A}^\top \mathbf{w}^k + \boldsymbol{\pi}^k \right),$$
$$\boldsymbol{\omega}_{i,l}{}^k = \boldsymbol{\omega}_{i,l}^{k-1} - \tau(\mathbf{u}_{i,l}^k - \mathbf{u}_{i,s_i}^k)), \qquad i \in [m],\ l \in [s_i - 1]$$

We are now ready to prove the convergence of PAPC for the general formulation via showing that it is an EB-algorithm.

Proposition 7 *Let Assumptions 1-5 hold true. Then, applying the PAPC algorithm to $\check{L}(\boldsymbol{\chi}, \mathbf{y})$ with parameters $\tau, \theta_{\lambda_1}, \ldots, \theta_{\lambda_m}, \theta_{\mathbf{w}}, \theta_{\boldsymbol{\pi}} > 0$ such that $\mathbf{H} = \mathbf{S} - \tau \bar{\mathbf{Q}}^\top \bar{\mathbf{Q}} \succeq \mathbf{0}$ where*

$$\mathbf{S} = Diag(\theta_1^{-1}\mathbf{I}_{s_1(d_1+1)}, \ldots, \theta_m^{-1}\mathbf{I}_{s_m(d_m+1)}, \theta_{\mathbf{w}}^{-1}\mathbf{I}_r, \theta_{\boldsymbol{\pi}}^{-1}\mathbf{I}_n)$$

result in

$$\check{L}(\bar{\boldsymbol{\chi}}^N, \mathbf{y}) - \check{L}(\boldsymbol{\chi}, \bar{\mathbf{y}}^N) \leq \frac{\tau^{-1}\|\boldsymbol{\chi} - \boldsymbol{\chi}^0\|^2 + \|\mathbf{y} - \mathbf{y}^0\|_{\mathbf{H}}^2}{2N}. \tag{35}$$

Furthermore, SGSP is an EB algorithm. More precisely, let $\mathbf{x}^*$ be an optimal solution of (26) and let $(\boldsymbol{\lambda}^*, \mathbf{w}^*)$ be the optimal dual variables associated with its constraints. Then, for any $\alpha > 0$ and

$\boldsymbol{\lambda} \in \mathbb{R}^m_+$ *and* $\mathbf{w}$ *such that* $\lambda_i \equiv \lambda_{i,s_i}$ *for all* $i \in [m]$.

$$L(\bar{\mathbf{x}}^N, (\mathbf{w}, \boldsymbol{\lambda})) - L(\mathbf{x}^*, (\mathbf{w}^*, \boldsymbol{\lambda}^*) + \alpha \operatorname{Dist}(\bar{\mathbf{x}}^N, X)$$
$$\leq \frac{1}{2N}\left(\frac{\|\mathbf{x}^* - \mathbf{x}^0\|^2}{\tau} + 2\sum_{i=1}^m \frac{\sigma_i}{\theta_i}\max\{\lambda_i, \lambda_i^0\}^2 + \frac{2\max\{\|\mathbf{w}\|, \|\mathbf{w}^0\|\}^2}{\theta_{\mathbf{w}}} + \phi(1+\alpha^2)\right). \tag{36}$$

and

$$L(\bar{\mathbf{x}}^N, \boldsymbol{\lambda}^*) + \beta \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \geq \mathbf{c}^\top \mathbf{x}^*,$$

where $\sigma_i = \sum_{l\in[s_i]}(1+4R_{i,l}^2)$, $R_{i,l}$ *is the radius of* $Z^{i,l}$, $\phi = \max\left\{\frac{2}{\theta_{\boldsymbol{\pi}}}, 4\sum_{i=1}^m s_i\bar{\mu}_i^2\left(1+\frac{1}{\epsilon_i}\right)^2 + \frac{2\|\pi^0\|^2}{\theta_{\boldsymbol{\pi}}}\right\}$ *with* $\bar{\mu}_i$ *and* ϵ_i *are defined in Assumptions 5 and 4, respectively, and* $\beta = R_\pi$ *with* R_π *defined in* (34).

*Remark 1 (***PAPC: the non-biaffine case)** The PAPC algorithm can be extends to the case where $g_i(\mathbf{x}, \mathbf{z}_i)$ is not bilinear, but rather has the following general form:

$$g_i(\mathbf{x}, \mathbf{z}_i) \coloneqq \mathbf{h}_i(\mathbf{x})^\top \boldsymbol{\ell}_i(\mathbf{z}_i),$$

where $\mathbf{h}_i(\cdot) : X \to \mathbb{R}^{k_i}$ and $\boldsymbol{\ell}_i(\cdot) : Z^i \to \mathbb{R}^{k_i}$, where each element of the mapping $\mathbf{h}_i(\cdot)$ is a convex function, *i.e.*, $h_{ij}(\cdot)$ is convex for all $j \in [l_i]$, and each element of $\boldsymbol{\ell}_i(\cdot)$ is a concave function, *i.e.*, $k_{ij}(\cdot)$ is concave for all $j \in [l_i]$. To maintain the convex-concave structure, we assume that

$$\min_{j\in[k_i]}\min_{\mathbf{z}_i\in Z^i} \ell_{ij}(\mathbf{z}_i) \geq 0, \quad \min_{j\in[k_i]}\min_{\mathbf{x}\in X} h_{ij}(\mathbf{x}) \geq 0 \quad \forall i \in [m].$$

We will show that we can transform such problems to the biaffine form of Section 3.2. For this, we introduce vectors $\boldsymbol{\varpi}_i$ ($\boldsymbol{\zeta}_i$ respectively) whose entries upper (lower) bound the entries of $\boldsymbol{\ell}_i$ ($\mathbf{h}_i$). With these extra variables, the constraint $g_i(\mathbf{x}, \mathbf{z}_i) \leq 0,\ \forall \mathbf{z}_i \in Z^i$ from problem (2) can be reformulated as

$$\begin{aligned} &\boldsymbol{\varpi}_i^\top \boldsymbol{\zeta}_i \leq 0,\ \forall (\mathbf{z}_i, \boldsymbol{\zeta}_i) \in \Xi^i \\ &\mathbf{h}_i(\mathbf{x}) \leq \boldsymbol{\varpi}_i \end{aligned}$$

where $\Xi^i = \left\{\boldsymbol{\xi}_i = (\mathbf{z}_i, \boldsymbol{\zeta}_i) : \mathbf{z}_i \in Z^i,\ \boldsymbol{\ell}_i(\mathbf{z}_i) \geq \boldsymbol{\zeta}_i\right\}$. As we see, the first constraint becomes biaffine in the respective variables. In order to decouple the constraint in $\mathbf{x}$, one can also duplicate $\mathbf{x}$ to $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_m$, and add equality constraints $\mathbf{x}_i = \mathbf{x}_0$ for all $i \in [m]$.

Accordingly, we can define an extended primal variable vector $\boldsymbol{\chi} = (\mathbf{x}_0, \ldots, \mathbf{x}_m, \boldsymbol{\varpi}_1, \ldots, \boldsymbol{\varpi}_m) \in \mathbb{R}^{n+\sum_{i=1}^m l_i}$ with feasible set

$$\mathcal{X} = \{\boldsymbol{\chi} = (\mathbf{x}_0, \ldots, \mathbf{x}_m, \boldsymbol{\varpi}_1, \ldots, \boldsymbol{\varpi}_m) : \mathbf{x}_0 \in X,\ \mathbf{h}_i(\mathbf{x}_i) \leq \boldsymbol{\varpi}_i\},$$

and an extended new uncertain parameter $\boldsymbol{\xi}_i = (\mathbf{z}_i, \boldsymbol{\zeta}_i)$ for constraint i such that $\boldsymbol{\xi}_i \in \Xi^i$. The i-th constraint of problem (2) becomes then:

$$\tilde{g}_i(\boldsymbol{\chi}, \boldsymbol{\xi}) = \boldsymbol{\chi}^\top \mathbf{Q}_i \boldsymbol{\xi}_i$$

where

$$\mathbf{Q}_i = \begin{bmatrix} \mathbf{0}_{(n(m+1)+\sum_{j=1}^{i-1} k_j)\times d_i} & \mathbf{0}_{(n(m+1)+\sum_{j=1}^{i-1} k_j)\times k_i} \\ \mathbf{0}_{k_i\times d_i} & \mathbf{I}_{k_i} \\ \mathbf{0}_{\sum_{j=i+1}^m k_j\times d_i} & \mathbf{0}_{\sum_{j=i+1}^m l_j\times k_i}. \end{bmatrix}$$

In the end, problem (2) can be shortly written as

$$\begin{aligned}\min_{\boldsymbol{\chi}\in\mathcal{X}}\ & \tilde{\mathbf{c}}^\top\boldsymbol{\chi}\\ \text{s.t.}\ & \sup_{\boldsymbol{\xi}_i\in\Xi^i}\tilde{g}_i(\boldsymbol{\chi},\boldsymbol{\xi}_i)\le 0,\quad i\in[m],\\ & \mathbf{A}\boldsymbol{\chi}=0\end{aligned}$$

where $\tilde{\mathbf{c}}=(\mathbf{c},\mathbf{0})$ and

$$\mathbf{A}=\begin{bmatrix}\mathbf{I}_{(n)} & -\mathbf{I}_{(n)} & \mathbf{0}_{(n\times n)} & & \dots & \mathbf{0}_{(n\times n)} & \mathbf{0}_{(n\times\sum_{i\in[m]}k_i)}\\ \mathbf{I}_{(n)} & \mathbf{0}_{(n\times n)} & -\mathbf{I}_{(n)} & \mathbf{0}_{(n\times n)} & \dots & \mathbf{0}_{(n\times n)} & \mathbf{0}_{(n\times\sum_{i\in[m]}k_i)}\\ \vdots & \vdots & \ddots & \ddots & & & \vdots\\ \vdots & \vdots & & \ddots & \ddots & & \vdots\\ \mathbf{I}_{(n)} & \mathbf{0}_{(n\times n)} & & \dots & \mathbf{0}_{(n\times n)} & -\mathbf{I}_{(n)} & \mathbf{0}_{(n\times\sum_{i\in[m]}k_i)}\end{bmatrix}$$

Thus, we are back at the biaffine case for the lifted variables $(\boldsymbol{\chi},\boldsymbol{\xi})$, and PAPC can be applied as long as the projections over Ξ^i and $\mathcal{X}$ are easily attainable.

5 Numerical experiment: robust quadratic programming

5.1 Introduction

In this section, we compare the numerical performance of our SGSP algorithm to the various approaches of [13] and the standard cutting-plane algorithm. To do this, we consider an extension of the experiment in [13], solving problems

$$\begin{aligned}\min_{\mathbf{x}\in X}\ & \sup_{\mathbf{z}\in Z} g_0(\mathbf{x},\mathbf{z})\\ \text{s.t.}\ & g_i(\mathbf{x},\mathbf{z})\le 0 \qquad\qquad \forall\mathbf{z}\in Z,\ i\in[m]\end{aligned}\tag{37}$$

where the objective and constraint functions are:

$$g_i(\mathbf{x},\mathbf{z})=\left\|\left(\mathbf{P}_{i0}+\sum_{k=1}^{K}\mathbf{P}_{ik}z_k\right)\mathbf{x}\right\|^2+\mathbf{b}_i^\top\mathbf{x}+c_i.$$

with $\mathbf{P}_{ik}\in\mathbb{R}^{L\times n}$, $\mathbf{b}_i\in\mathbb{R}$ and $c_i\in\mathbb{R}$. We assume that $Z=\{\mathbf{z}\in\mathbb{R}^K:\|\mathbf{z}\|_2\le 1\}$ and $X=\{\mathbf{x}\in\mathbb{R}^n:\|\mathbf{x}\|_2\le 1\}$. The experiment is an extension of the one performed in [13], since the original considers problem with uncertainty only in the objective, while ours captures the more general setting with uncertain constraints.

Note that in (37), the functions $g_i(\mathbf{x},\cdot)$ are convex, and therefore the problem does not readily fit our framework. However, in the pessimization problem

$$\sup_{\|\mathbf{z}\|\le 1} g_i(\mathbf{x},\mathbf{z}),\tag{38}$$

known as the *trust region problem*, the function $g_i(x,\cdot)$ can be replaced by equivalent concave function, where the equivalence is in the sense of the same maximal value. For this, g_i is first rewritten as

$$g_i(\mathbf{x},\mathbf{z}) = \mathbf{z}^\top Q_i(\mathbf{x})\mathbf{z} + 2\mathbf{r}_i(\mathbf{x})^\top \mathbf{z} + s_i(\mathbf{x})$$

where $Q_i(\mathbf{x}) = P_i(\mathbf{x})^\top P_i(\mathbf{x})$ with $P_i(\mathbf{x}) \in \mathbb{R}^{n\times K}$ being a matrix whose columns are the vectors $\mathbf{P}_{ik}\mathbf{x}$ for $k \in [K]$, $\mathbf{r}_i(\mathbf{x}) = P_i(\mathbf{x})^\top P_{i0}\mathbf{x}$ and $s_i(\mathbf{x}) = \|P_{i0}\mathbf{x}\|_2^2 - \mathbf{b}_i^\top \mathbf{x} - c_i$. For this function, by result of [14] problem (38) can be reformulated as

$$\sup_{\mathbf{z}:\|\mathbf{z}\|\le 1} g_i(\mathbf{x},\cdot) = \sup_{\mathbf{z}:\|\mathbf{z}\|\le 1} \mathbf{z}^\top (Q_i(\mathbf{x}) - \lambda_{\max}(Q_i(\mathbf{x}))\mathbf{I})\mathbf{z} + 2\mathbf{r}_i(\mathbf{x})^\top \mathbf{z} + s_i(\mathbf{x}) + \lambda_{\max}(Q_i(\mathbf{x})) \tag{39}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of the argument matrix. Since

$$\bar{g}_i(\mathbf{x},\mathbf{z}) := \mathbf{z}^\top (Q_i(\mathbf{x}) - \lambda_{\max}(Q_i(\mathbf{x}))\mathbf{I})\mathbf{z} + 2\mathbf{r}_i(\mathbf{x})^\top \mathbf{z} + s_i(\mathbf{x}) + \lambda_{\max}(Q_i(\mathbf{x})),$$

is convex-concave, using $\bar{g}_i(\mathbf{x},\mathbf{z})$ instead of $g_i(x,z)$ in the robust formulation is equivalent and our setting applies. Therefore, in this set of experiments, whenever using first-order methods, we solve the modified problem:

$$\begin{aligned} &\min_{\mathbf{x}\in X} \sup_{\mathbf{z}\in Z} \bar{g}_0(\mathbf{x},\mathbf{z}) \\ &\text{s.t. } \bar{g}_i(\mathbf{x},\mathbf{z}) \le 0 \qquad\qquad \forall \mathbf{z}\in Z,\ i\in[m] \end{aligned} \tag{40}$$

We note that with respect $\mathbf{x}$, problem (40) is semidefinite optimization problem. We are ready to state the four algorithms we compare:

- **Cutting planes algorithm** applied directly to (37), where pessimization subproblem (38) is solved using a generic solver to find $\mathbf{z}$ violating the constraints.
- **The SGSP algorithm** applied to (40). As part of this method, we first apply SGSP to find a Slater point for (40), and given the obtained Slater point we run SGSP to solve problem (40).
- **The OCO algorithm** of [13], applied to (40), where both the variables $\mathbf{x}$ and $\mathbf{z}$ are solved using online gradient descent (OGD).
- **The FO-pessimization approach** of [13] applied to (40), where the primal problem is solved using OGD and for each constraint the worst-case $\mathbf{z}$ is found by solving (39) using a generic solver.

We will compare the algorithms on their speed of reducing the feasibility and optimality gaps in the sense of Theorem 1. In the following Section, we describe the exact numerical setup and the results.

Remark 2 In the presented experiments we do not compare to the 'nominal' approach suggested in [6]. Indeed, in the implementation of [6] presented in [13], the authors suggests that in the k-th iteration, the following problem will be solved:

$$\begin{aligned} &\min_{\mathbf{x}} g_0(\mathbf{x},\mathbf{z}_0^k) \\ &\text{s.t. } g_i(\mathbf{x},\mathbf{z}_i^k) \le 0 \qquad\qquad i\in[m]. \end{aligned}$$

However, since the convergence of such a method requires for there to be a saddle point

$$\inf_{\mathbf{x}\in X}\sup_{\mathbf{z}_i\in Z} g_i(\mathbf{x},\mathbf{z}_i) = \sup_{\mathbf{z}_i\in Z}\inf_{\mathbf{x}\in X} g_i(\mathbf{x},\mathbf{z}_i),$$

which does not necessarily exist due to the convexity of $g_i(\mathbf{x},\cdot)$. Thus, in order to solve the (37) problem using some kind of a 'nominal' approach with updated $\mathbf{z}_i^k$, one would need to either (i) use the semidefinite program (40) as the nominal oracle which contradicts the idea of solving 'simple' problems per iteration, or (ii) use the dual-subgradient meta-algorithm of [6] where $\mathbf{z}_i$ is lifted to a semidefinite matrix and the original nominal oracle (37) is used w.r.t. $\mathbf{x}$ – however, running the dual step would require projections on the spectahedron which, again, contradicts the idea of solving 'simple' problems, or (iii) use the dual-perturbation meta-algorithm of [6] – however, as we focus on deterministic algorithms, we do not include an implementation of [6].

5.2 Expriment setting

We explored different problem sizes, with respect to n the dimension of $\mathbf{x}$, m the number of constraints, K the dimension of uncertainty $\mathbf{z}_i$, and ℓ the dimension of the vector in the norm. For each problem size, we sampled the problem data for 50 problem instances as follows. First, each entry of $\mathbf{P}_{ik}$ and $\mathbf{b}_i$ is sampled uniformly from interval $[-1,1]$. Fixed value $c_i=-0.05$ is chosen deterministically to ensure Slater feasibility of the problem. Next, $\mathbf{P}_{ik}$ and $\mathbf{b}_i$ are normalized as

$$\begin{aligned}\mathbf{P}_{ik} &= \frac{\mathbf{P}_{ik}}{S_{i1}} && \text{where } S_{i1} = \left\| \left[\mathbf{P}_{i0}^\top \ \cdots \ \mathbf{P}_{iK}^\top\right]^\top \right\|_{2,2}\\ \mathbf{b}_i &= \frac{\mathbf{b}_i}{S_{i2}} && \text{where } S_{i2} = \|\mathbf{b}_i\|_2\end{aligned}$$

To compare the algorithms in a fair way, we will use the same starting point for all of them. This point will be the optimal solution to the nominal problem

$$\begin{aligned}&\min_{\mathbf{x}}\ g_0(\mathbf{x},\mathbf{0})\\ &\text{s.t. } g_i(\mathbf{x},\mathbf{0})\le 0 && i\in[m].\end{aligned}$$

The values $\mathbf{z}_i$ are initialized as the zero vectors. We consider an ϵ tolerance of 0.001 for both feasibility and optimality.

Note that the first-order algorithms require a choice of step-size. The step-size for OGD is chosen to be $2/(\|\nabla_{\mathbf{x}} g_i(\mathbf{x},\mathbf{z}_i)\|\sqrt{k})$ and $2/(\|\nabla_{\mathbf{z}_i} g_i(\mathbf{x},\mathbf{z}_i)\|\sqrt{k})$ for the primal and dual steps at iteration k, respectively. These stepsizes correspond with the analysis of OGD given in [12, Theorem 3.4]. Similarly, the step sizes for SGSP is given by $\tau^k = 2/(\|\nabla_{\mathbf{x}} L(\mathbf{x},\mathbf{u})\|\sqrt{k})$ and $\theta_i^k = 2/(\|\nabla_{\mathbf{u}_i} L(\mathbf{x},\mathbf{u})\|\sqrt{k})$ for the primal and dual steps at iteration k, respectively [1]. In both cases, the average solution $\bar{\mathbf{x}}^N$ is computed as a weighted sum of the iterates $\mathbf{x}^k$ with weights corresponding to the step-sizes. We note that these step-sizes were chosen instead of constant step-sizes, since they produced better results for all methods while retaining theoretical convergence guarantees.

Table 5.2 describes the different settings in which the algorithms were run. We consider small, medium and large problem sizes, with either no constraints or three constraints for each. For each problem size we specify the time limit we gave the methods as well as the sampling frequency for the output (see explanation below).

In order to measure the feasibility and the optimality gap for each method, for each parameter realization, we first define a constant $\tilde{N}$, such that every $\tilde{N}$ iterations statistics on the solution

[1] We note that although the analysis of [16] was done for a constant step-size, a similar analysis to the one shown in [12, Theorem 3.4] can be done for SGSP, with similar theoretical results.

Table 3 Sizes of tested problems

Name	n	K	L	m	maximum allowed time (seconds)	Output frequency ($\tilde{N}$ iterations)
Small	10	10	10	0 3	600	100 [2]
Medium	600	25	15	0 3	1200	100
Large	3600	30	16	0 3	3600	100

are gathered. Specifically, for all $k \in \mathbb{N}$, let $\mathbf{x}^{k\tilde{N}}$ be the solution obtained after iteration $k\tilde{N}$ of the algorithm, and let T_k be the time it took to run these $k\tilde{N}$ iterations. The feasibility gap at iteration $k\tilde{N}$ is given by

$$\mathrm{FG}_k := \max_{i\in[m]} \max_{\mathbf{z}_i \in Z} g_i(\mathbf{x}^{k\tilde{N}}, \mathbf{z}_i).$$

Defining the optimality gap to be infinity if the feasibility gap is larger than the defined ϵ, the optimality gap ratio at iteration $k\tilde{N}$ is given by

$$\mathrm{OGR}_k := \frac{\delta_{\{p:p\leq\epsilon\}}(\mathrm{FG}_k)\max_{\mathbf{z}_i\in Z} g_0(\mathbf{x}^{k\tilde{N}}, \mathbf{z}_i) - \mathrm{LB}}{\mathrm{LB}}.$$

In the above formula, $\delta_{\{p:p\leq\epsilon\}}$ is the indicator function, and LB is the lower bound on the optimal solution obtained at the end of the cutting planes algorithm, by only considering the cuts added during the algorithm. Thus, for each time $t \geq 0$ we can record the minimal feasibility gap up to time t as

$$\min_{k:T_k\leq t} \mathrm{FG}_k,$$

and the minimal optimality gap ratio up to time t as

$$\min_{k:T_k\leq t} \mathrm{OGR}_k.$$

All the code was run using Python 3.7, with the CasADi package [1] as the optimization interface. The optimization problems were solved using Gurobi 9.0.0 [17], with the exception of the trust-region subproblem (38) that due to numerical difficulties was solved using the IPOPT solver [18]. The code was run on a PowerEdge R740xd server with two Intel Xeon Cold 6254 3.1GHz processors, each with 18 cores, and a total RAM of 384GB.

5.3 Results

In Figure 1, we show the optimality convergence for all methods for the small, medium, and large instances without constraints. In the small instances, the computational and memory requirements of the cutting planes algorithm are negligible, and its performance dominates the other methods. Among the first order methods, however, the SGSP algorithm attains the fastest convergence.

When the problem instances become larger, the memory and computational requirements of the cutting planes algorithm become more significant, making its performance similar (medium instances) and then worse than the performance of the first-order methods. Among the first-order methods, our SGSP algorithm consistently dominates the other methods, although the differences

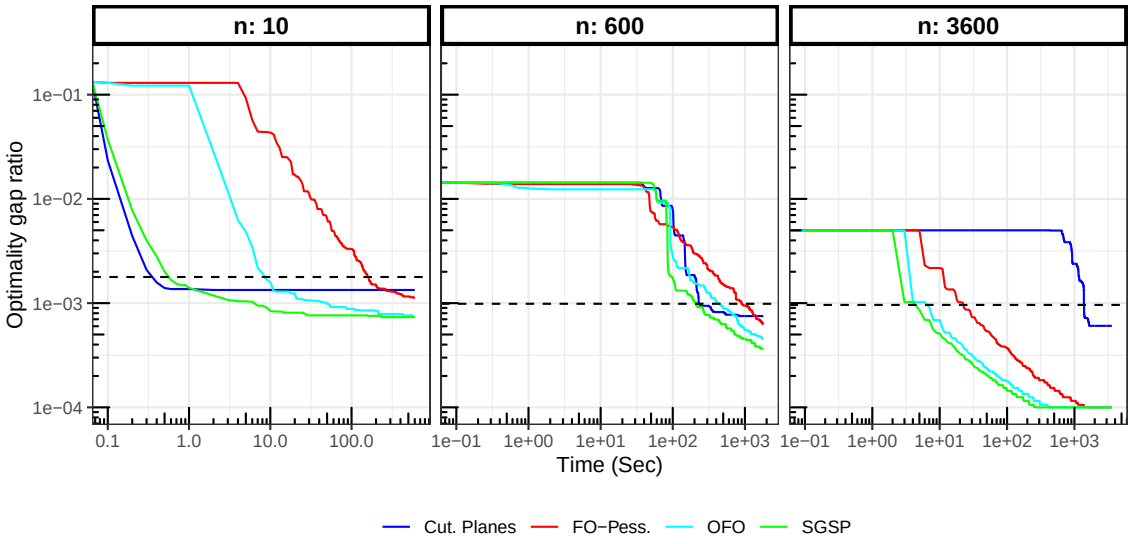

Fig. 1 Optimality gap for all problem sizes with $m = 0$.

with the OCO are rather small. Interestingly, we observe that for all methods it is the medium instances that keep their optimality gaps rather large for the longest. This might be related to the way we sample the problems where up to a certain point the problem size growth effect outweighs the 'averaging out' effect of the large matrices that make it easier to find a high-quality solution.

We now proceed to discuss the constrained problems. In Figures 2-4 we show the results for the small, medium, and large instances, respectively. We start with discussing the small examples. Similarly to the unconstrained case, the cutting planes algorithm is the fastest of all. Among the first-order methods, OFO is better than SGSP at finding feasible solutions fast, but having found them, it gets 'stuck' on improving optimality due to the need to run binary search on the objective value. Specifically, the problem is with the bi-section iterations which do not have a feasible solution, but this infeasibility can only be identified after performing a very large number of iterations. The FO-Pess method is the slowest across all three measures.

For the medium instances in Figure 3, we would expect the first order methods to already perform better than the cutting planes algorithm. Indeed, the cutting planes algorithm becomes worse in decreasing the feasibility gap and the number of infeasible instances compared to all first-order methods. Among the feasible first-order methods, again, it is the SGSP algorithm that manages to reach optimality guarantees within the prescribed amount of time which are equivalent to those of the cutting planes, while OFO and FO-Pess do get stuck due to the need for running the binary search on the objective function value.

Finally, for the largest instances in Figure 4, predictably, the cutting planes does not only have inferior performance to all first order based methods. We therefore focus on the comparison of computational performance of the three first-order methods. We observe that all first order methods are able to find feasible solutions faster than in the medium instances. We again believe that this is due to the problem generation process. Among the first-order methods, SGSP is relatively the slowest one to find a feasible solution for all the instances. With respect to the optimality gap,

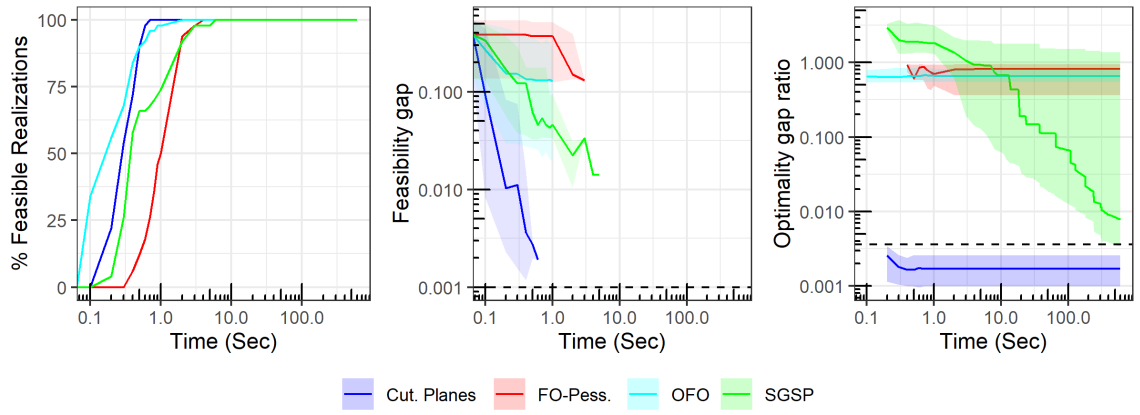

Fig. 2 Small instances, $m = 3$. Percentage of instances with a feasible solution found, feasibility gap among infeasible instances, and optimality gap among the feasible instances.

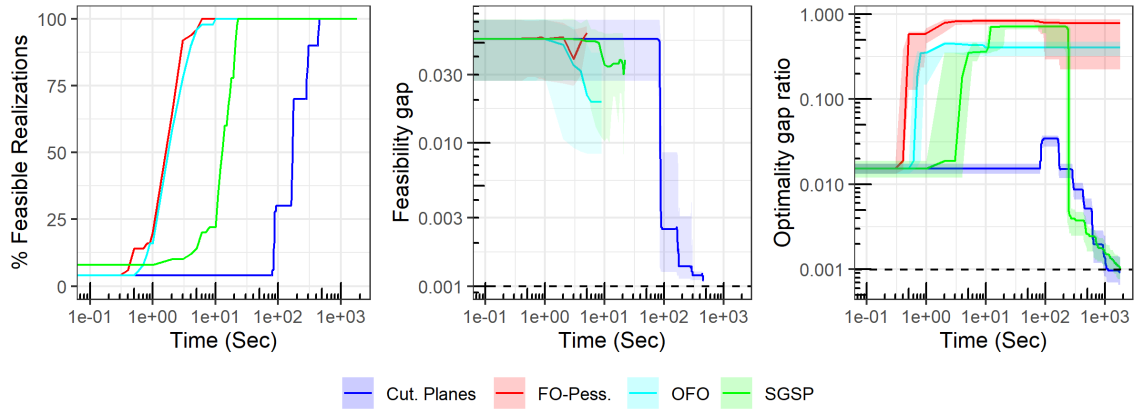

Fig. 3 Medium instances, $m = 3$. Percentage of instances with a feasible solution found, feasibility gap among infeasible instances, and optimality gap among the feasible instances.

we observe that OFO slightly dominates the SGSP. The SGSPs performance on these instances is affected by the fact that at each iteration, it requires gradient computations with respect to all the constraints, whereas for the OFO this is done only for the constraint with largest current value of the left-hand side. Since for large dimensional problems these computations are substantial, the SGSP performs slightly slower than OFO, while dominating over the FO-Pess method.

6 Conclusions

In this paper, we have proposed a first-order optimization approach to robust optimization problems based on a convex-concave saddle-point reformulation of the problem's Lagrangian. Our approach recovers the $\mathcal{O}(1/\epsilon^2)$ convergence rate for general problems considered also by [6,13], and offers an improved $\mathcal{O}(1/\epsilon)$ convergence guarantee for problems with a biaffine function structure. Similar to

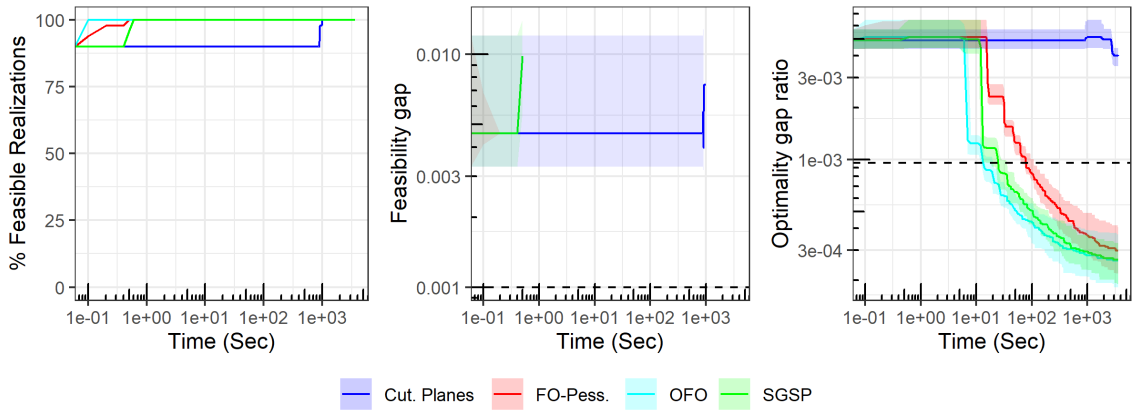

Fig. 4 Large instances, $m = 3$. Percentage of instances with a feasible solution found, feasibility gap among infeasible instances, and optimality gap among the feasible instances.

those algorithms, our method allows for a convenient parallelization of the computations related to different constraint functions and avoids problem size increase typical for the cutting planes and robust counterparts approaches. At the same time, our approach has the numerical benefit of avoiding a binary-search procedure for the optimal value of the objective as in [13], while providing a deterministic algorithm which does not have to solve the nominal problem, contrary to [6].

Acknowledgements

We thank Fatma Kılınç-Karzan and Nam Ho-Nguyen for their discussions and sharing with us the implementation of their experiments. We also thank Shoham Sabach for his work at the early stage of this work and valuable suggestions.

A Proofs

Proof of Proposition 1. **Proof of $\Rightarrow$.** We first note that from the definition of f_i and the construction of the lifted uncertainty set U^i we have that for any $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ and $\mathbf{x} \in X$ and $i \in [m]$.

$$\lambda_i f_i(\mathbf{x}) = \lambda_i \sup_{\mathbf{z}_i \in Z^i} g_i(\mathbf{x}, \mathbf{z}_i) = \sup_{\tilde{\mathbf{z}}_i : (\tilde{\mathbf{z}}_i, \lambda_i) \in U^i} \lambda_i g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right).$$

Indeed, if $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*))$ is a saddle point of $\bar{L}$ then, defining $\mathbf{z}_i^* = \tilde{\mathbf{z}}_i^*/\lambda_i^*$ if $\lambda^* > 0$ and $\mathbf{z}_i^* = 0$ otherwise, gives:

$$L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) = \bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)), \tag{41}$$

and

$$\begin{aligned}
&\bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) \\
&= \sup_{\mathbf{u}\in U, \mathbf{w}\in\mathbb{R}^r} \bar{L}(\mathbf{x}^*, (\mathbf{u}, \mathbf{w})) \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \tilde{\mathbf{z}}_i\in\lambda_i Z^i,\, i\in[m], \mathbf{w}\in\mathbb{R}^r} \mathbf{c}^\top\mathbf{x}^* + \sum_{i=1}^m \lambda_i g_i\left(\mathbf{x}^*, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right) + \mathbf{w}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+, \mathbf{w}\in\mathbb{R}^r} \mathbf{c}^\top\mathbf{x}^* + \sum_{i=1}^m \lambda_i \sup_{\mathbf{z}_i\in Z^i} g_i(\mathbf{x}^*, \mathbf{z}_i) + \mathbf{w}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+, \mathbf{w}\in\mathbb{R}^r} \mathbf{c}^\top\mathbf{x}^* + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \mathbf{w}^\top(\mathbf{A}\mathbf{x}^* - \mathbf{b}) \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+, \mathbf{w}\in\mathbb{R}^r} L(\mathbf{x}^*, (\boldsymbol{\lambda}, \mathbf{w})),
\end{aligned} \tag{42}$$

where the subsequent equalities follow from (i) the definition of $\bar{L}$, (ii) definition of U^i, (iii) the definition of f_i, and (iv) the definition of L. Moreover,

$$\begin{aligned}
\bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) &= \inf_{\mathbf{x}\in X} \bar{L}(\mathbf{x}, (\mathbf{u}^*, \mathbf{w}^*)) \\
&= \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \lambda_i^* g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i^*}{\lambda_i^*}\right) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&\le \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \sum_{i=1}^m \sup_{\tilde{\mathbf{z}}_i:(\tilde{\mathbf{z}}_i, \lambda_i^*)\in U^i} \lambda_i^* g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i^*}\right) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \sum_{i=1}^m \sup_{\mathbf{z}_i\in Z^i} \lambda_i^* g_i(\mathbf{x}, \mathbf{z}_i) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \inf_{\mathbf{x}\in X} L(\mathbf{x}, (\boldsymbol{\lambda}^*, \mathbf{w}^*)),
\end{aligned} \tag{43}$$

where the subsequent steps follow from (i) the definition of $\bar{L}$, (ii) the definition of U^i, and (iii) the definition of L. Combining (41), (42) and (43) we obtain that

$$\begin{aligned}
\inf_{\mathbf{x}\in X} \bar{L}(\mathbf{x}, (\mathbf{u}^*, \mathbf{w}^*)) &= \inf_{\mathbf{x}\in X} L(\mathbf{x}, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \\
&= L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) \\
&= \bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) \\
&= \sup_{\mathbf{u}\in U, \mathbf{w}\in\mathbb{R}^r} \bar{L}(\mathbf{x}^*, (\mathbf{u}, \mathbf{w})) \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^m_+, \mathbf{w}\in\mathbb{R}^r} L(\mathbf{x}^*, (\boldsymbol{\lambda}, \mathbf{w})),
\end{aligned}$$

i.e., $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ is a saddle point of L.

Proof of $\Leftarrow$. We shall show that $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ can be extended to a saddle point of the lifted Lagrangian $\bar{L}$. Note that defining $\tilde{\mathbf{z}}_i = \lambda_i \mathbf{z}_i$ we have that

$$\begin{aligned}
\mathbf{c}^\top\mathbf{x}^* + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x}^* - \mathbf{b}) &= \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \sum_{i=1}^m \lambda_i^* \max_{(\tilde{\mathbf{z}}_i, \lambda_i^*)\in U^i} g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i^*}\right) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \max_{(\tilde{\mathbf{z}}_i, \lambda_i^*)\in U^i,\, i\in[m]} \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \sum_{i=1}^m \lambda_i^* g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i^*}\right) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \inf_{\mathbf{x}\in X} \mathbf{c}^\top\mathbf{x} + \sum_{i=1}^m \lambda_i^* g_i\left(\mathbf{x}, \frac{\hat{\mathbf{z}}_i}{\lambda_i^*}\right) + \mathbf{w}^{*\top}(\mathbf{A}\mathbf{x} - \mathbf{b}).
\end{aligned}$$

The first equality is due to $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ being a saddle point of L and the definition of f_i. The third equality follows from Sion's theorem, applicable due to boundedness of $\{(\tilde{\mathbf{z}},\lambda_i)\in U^i:\ \lambda_i=\lambda_i^*\}$ and where we define $\hat{\mathbf{z}}_i$ as a (necessarily-existing) maximizer:

$$(\hat{\mathbf{z}}_i)_{i\in[m]}\in\arg\max_{\tilde{\mathbf{z}}_i:(\tilde{\mathbf{z}}_i,\lambda_i^*)\in U^i}\left\{\inf_{\mathbf{x}\in X}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i^* g_i\left(\mathbf{x},\frac{\tilde{\mathbf{z}}_i}{\lambda_i^*}\right)+\mathbf{w}^{*\top}(\mathbf{A}\mathbf{x}-\mathbf{b})\right\}.$$

Thus, defining $\mathbf{u}_i^*=(\hat{\mathbf{z}}_i^*,\lambda_i^*)$ we have that

$$L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))=\bar{L}(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))=\inf_{\mathbf{x}\in X}\mathbf{c}^\top\mathbf{x}+\sum_{i=1}^m\lambda_i^* g_i\left(\mathbf{x},\frac{\hat{\mathbf{z}}_i^*}{\lambda_i^*}\right)+\mathbf{w}^{*\top}(\mathbf{A}\mathbf{x}-\mathbf{b})=\inf_{\mathbf{x}\in X}\bar{L}(\mathbf{x},(\mathbf{u}^*,\mathbf{w}^*))$$

Moreover,

$$\begin{aligned}\bar{L}(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))=L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))&=\mathbf{c}^\top\mathbf{x}^*+\sum_{i=1}^m\lambda_i^* f_i(\mathbf{x}^*)+\mathbf{w}^{*\top}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\\&=\mathbf{c}^\top\mathbf{x}^*+\sup_{\boldsymbol{\lambda}\geq\mathbb{R}_+^m,\mathbf{w}\in\mathbb{R}^r}\sum_{i=1}^m\lambda_i\sup_{\mathbf{z}_i\in Z^i}g_i(\mathbf{x}^*,\mathbf{z})+\mathbf{w}^\top(\mathbf{A}\mathbf{x}^*-\mathbf{b})\\&=\mathbf{c}^\top\mathbf{x}^*+\sup_{\mathbf{u}_i=(\lambda_i,\tilde{\mathbf{z}}_i)\in U^i,\ i\in[m],\mathbf{w}}\sum_{i=1}^m\lambda_i g_i\left(\mathbf{x}^*,\frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right)+\mathbf{w}^\top(\mathbf{A}\mathbf{x}^*-\mathbf{b})\\&=\max_{\mathbf{u}_i\in U^i,\ i\in[m],\mathbf{w}\in\mathbb{R}^r}\bar{L}(\mathbf{x}^*,(\mathbf{u},\mathbf{w})),\end{aligned}$$

where we used the fact that $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ is a saddle point of the original Lagrangian L, and the definition of f_i, L, and $\bar{L}$. Thus, we showed that $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))$ is a saddle point of $\bar{L}$. □

Proof of Lemma 1. From Assumptions 1 and 2 we have that problem (4) has a saddle point, and due to Proposition 1 so does problem (6). Note that for any $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{u}\in U$, and $\boldsymbol{\pi}^*\in\mathbb{R}^n$ we have that

$$\tilde{L}(\mathbf{x},(\mathbf{u},\mathbf{w},\boldsymbol{\pi}^*))\leq\sup_{\boldsymbol{\pi}\in\mathbb{R}^n}\tilde{L}(\mathbf{x},(\mathbf{u},\mathbf{w},\boldsymbol{\pi}))=\bar{L}(\mathbf{x},(\mathbf{u},\mathbf{w}))+\delta_X(\mathbf{x}).\tag{44}$$

Let $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))$ be a saddle point of $\bar{L}$, and let $\boldsymbol{\pi}^*=-\mathbf{c}-\tilde{\mathbf{Q}}\mathbf{u}^*-\mathbf{A}^\top\mathbf{w}^*$. Then,

$$\begin{aligned}\inf_{\mathbf{x}\in X}\tilde{L}(\mathbf{x},(\mathbf{u}^*,\mathbf{w}^*,\boldsymbol{\pi}^*))&=\inf_{\mathbf{x}\in X}\mathbf{x}^\top\left(\mathbf{c}+\mathbf{A}^\top\mathbf{w}^*+\tilde{\mathbf{Q}}\mathbf{u}^*+\boldsymbol{\pi}^*\right)+\tilde{\mathbf{q}}^\top\mathbf{u}^*-\mathbf{b}^\top\mathbf{w}^*-\sigma_X(\boldsymbol{\pi}^*)\\&=\tilde{\mathbf{q}}^\top\mathbf{u}^*-\mathbf{b}^\top\mathbf{w}^*-\sigma_X(\boldsymbol{\pi}^*)\\&=\tilde{\mathbf{q}}^\top\mathbf{u}^*-\mathbf{b}^\top\mathbf{w}^*+\inf_{\mathbf{x}\in X}-\mathbf{x}^\top\boldsymbol{\pi}^*\\&=\inf_{\mathbf{x}\in X}\mathbf{x}^\top\left(\mathbf{c}+\mathbf{A}^\top\mathbf{w}^*+\tilde{\mathbf{Q}}\mathbf{u}^*\right)-\mathbf{b}^\top\mathbf{w}^*+\tilde{\mathbf{q}}^\top\mathbf{u}^*=\inf_{\mathbf{x}\in X}\bar{L}(\mathbf{x},(\mathbf{u}^*,\mathbf{w}^*)),\end{aligned}\tag{45}$$

where the first equality follows from the definition of $\tilde{L}$, the second equality follows from the definition of $\boldsymbol{\pi}^*$, the third equality follows from the definition of the support function, the fourth equality follows from the definition of $\boldsymbol{\pi}^*$, and the last equality follows from the definition of $\bar{L}$. Combining (44) and (45) we have that

$$\tilde{L}(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*,\boldsymbol{\pi}^*))=\bar{L}(\mathbf{x}^*,(\mathbf{w}^*,\mathbf{u}^*))=\inf_{\mathbf{x}\in X}\bar{L}(\mathbf{x},(\mathbf{u}^*,\mathbf{w}^*))=\inf_{\mathbf{x}\in X}\tilde{L}(\mathbf{x},(\mathbf{u}^*,\mathbf{w}^*,\boldsymbol{\pi}^*))$$

Moreover,

$$\tilde{L}(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*,\boldsymbol{\pi}^*))=\bar{L}(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))=\sup_{\mathbf{u}\in U,\mathbf{w}}\bar{L}(\mathbf{x}^*,(\mathbf{u},\mathbf{w}))=\sup_{\boldsymbol{\pi}\in\mathbb{R}^n,\mathbf{u}\in U,\mathbf{w}}\tilde{L}(\mathbf{x}^*,(\mathbf{u},\mathbf{w},\boldsymbol{\pi}))$$

where the second equality follows from $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*))$ being a saddle point of $\bar{L}$, and the last equality from (44) and $\mathbf{x}^*\in X$. Thus, $(\mathbf{x}^*,(\mathbf{u}^*,\mathbf{w}^*,\boldsymbol{\pi}^*))$ is a saddle point of $\tilde{L}$.

Now, let $(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*))$ be a saddle point of $\tilde{L}$ on $\mathbb{R}^n \times (U \times \mathbb{R}^r \times \mathbb{R}^n)$. Then, by definition

$$\tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*)) = \max_{\boldsymbol{\pi} \in \mathbb{R}^n} \tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi})) = \bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) + \delta_X(\mathbf{x}^*), \tag{46}$$

and

$$\begin{aligned}\tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*)) &= \inf_{\mathbf{x} \in \mathbb{R}^n} \tilde{L}(\mathbf{x}, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*)) \\ &= \inf_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \left(\mathbf{c} + \mathbf{A}^\top \mathbf{w}^* + \tilde{\mathbf{Q}}\mathbf{u}^* + \boldsymbol{\pi}^*\right) - \mathbf{b}^\top \mathbf{u}^* + \tilde{\mathbf{q}}^\top \mathbf{u}^* - \sigma_X(\boldsymbol{\pi}^*) < \infty,\end{aligned}$$

implying that $\boldsymbol{\pi}^* = -\mathbf{c} - \tilde{\mathbf{Q}}\mathbf{u}^* - \mathbf{A}^\top \mathbf{w}^*$, and thus, following the same arguments as in (45)

$$\bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) = \tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*)) = \inf_{\mathbf{x} \in X} \bar{L}(\mathbf{x}, (\mathbf{u}^*, \mathbf{w}^*)),$$

where the first equality is due to (46) and the finiteness of $\tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*))$, implying that $\delta_X(\mathbf{x}^*) = 0$ and thus $\mathbf{x}^* \in X$. Furthermore, (46) implies that

$$\begin{aligned}&\bar{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*)) = \tilde{L}(\mathbf{x}^*, (\mathbf{u}^*, \mathbf{w}^*, \boldsymbol{\pi}^*)) \\ &= \sup_{\mathbf{u} \in U, \mathbf{w}, \boldsymbol{\pi} \in \mathbb{R}^n} \tilde{L}(\mathbf{x}^*, (\mathbf{u}, \mathbf{w}, \boldsymbol{\pi})) = \sup_{\mathbf{u} \in U, \mathbf{w}} \bar{L}(\mathbf{x}^*, (\mathbf{u}, \mathbf{w})) + \delta_X(\mathbf{x}^*) \\ &= \sup_{\mathbf{u} \in U, \mathbf{w}} \bar{L}(\mathbf{x}^*, (\mathbf{u}, \mathbf{w})).\end{aligned}$$

Thus, $(\mathbf{x}^*, \mathbf{u}^*)$ is a saddle point of $\bar{L}$ on $X \times U \times \mathbb{R}^r$. □

Proof of Lemma 2. (i) Applying [9, Lemma 2.3] and using the fact that for any $\mathbf{u}_i = (\tilde{\mathbf{z}}_i, \lambda_i) \in U^i$ we have that either both $\lambda_i = 0$ and $\tilde{\mathbf{z}}_i = 0$, or $\mathbf{z}_i = \tilde{\mathbf{z}}_i / \lambda_i \in Z^i$.

(ii) We use the fact that $\mathbf{z}_i = \tilde{\mathbf{z}}_i / \lambda_i$ for all i such that $\lambda_i > 0$, and define

$$\mathbf{d}_x \coloneqq \sum_{i=1}^m \lambda_i \tilde{\mathbf{d}}_{x,i} = \sum_{i:\lambda_i > 0} \lambda_i \tilde{\mathbf{d}}_{x,i} \in \sum_{i:\lambda_i > 0} \lambda_i \partial_{\mathbf{x}} g_i(\mathbf{x}, \mathbf{z}_i) = \sum_{i:\lambda_i > 0} \lambda_i \partial_{\mathbf{x}} g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right).$$

Thus, we see that (12) holds, i.e., $\mathbf{v}_x \in \partial_{\mathbf{x}} \bar{L}(\mathbf{x}, \mathbf{u})$. For all i such that $\lambda_i > 0$ define

$$\mathbf{d}_i \coloneqq \tilde{\mathbf{d}}_{z,i} \in \partial_{\mathbf{z}_i} \left(-g_i(\mathbf{x}, \mathbf{z}_i)\right) = \partial_{\mathbf{z}_i} \left(-g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right)\right).$$

For such $\mathbf{d}_i$ we have that

$$\mathbf{v}_i = (\mathbf{d}_{z,i}, -g_i(\mathbf{x}, \mathbf{z}_i) - \mathbf{z}_i^\top \mathbf{d}_{z,i}) = \left(\mathbf{d}_i, -g_i\left(\mathbf{x}, \frac{\tilde{\mathbf{z}}_i}{\lambda_i}\right) - \frac{\tilde{\mathbf{z}}_i^\top \mathbf{d}_i}{\lambda_i}\right) \in \partial_{\mathbf{u}_i} \left(-\bar{L}(\mathbf{x}, \mathbf{u})\right).$$

i.e, the first case of (13) holds.
Finally, for all i such that $\lambda_i = 0$, using the fact that $\mathbf{z}_i = \mathbf{0}$ define

$$\mathbf{d}_i \coloneqq \mathbf{d}_{z,i} \in \partial_{\mathbf{z}_i} \left(-g_i(\mathbf{x}, \mathbf{0})\right) \subseteq \cup_{\mathbf{z}_i \in Z^i} \partial_{\mathbf{z}_i} \left(-g_i\left(\mathbf{x}, \mathbf{z}_i\right)\right) + T_{Z^i}(\mathbf{z}_i)^*,$$

where the tangent cone $T_{Z^i}(\mathbf{z}_i)^*$ always includes the zero vector. Moreover, defining $\phi_i \coloneqq -g_i(\mathbf{x}, \mathbf{0})$ we have that

$$\begin{aligned}(-g_i)^*(\mathbf{x}, \mathbf{d}_i) &= \sup_{\boldsymbol{\zeta}_i} \{\boldsymbol{\zeta}_i^\top \mathbf{d}_i + g(\mathbf{x}, \boldsymbol{\zeta}_i)\} \\ &= \sup_{\boldsymbol{\zeta}_i} \{\mathbf{d}_i^\top (\boldsymbol{\zeta}_i - \mathbf{0}) - g_i(\mathbf{x}, \mathbf{0}) + g_i(\mathbf{x}, \boldsymbol{\zeta}_i)\} + g_i(\mathbf{x}, \mathbf{0}) \\ &\leq g_i(\mathbf{x}, \mathbf{0}) = -\phi_i\end{aligned}$$

where the first equality follows from the definition of convex conjugate, and the inequality follows from the convexity of $-g_i(\mathbf{x}, \cdot)$ and the fact that $\mathbf{d}_i \in \partial_{\mathbf{z}_i} \left(-g_i(\mathbf{x}, \mathbf{0})\right)$. Thus, meeting the definition (13) we have that for the case where $\lambda_i = 0$ we obtain that $\mathbf{v}_i = (\mathbf{d}_i, \phi_i) \in \partial_{\mathbf{u}_i} \left(-\bar{L}(\mathbf{x}, \mathbf{u})\right)$. □

Proof of Proposition 3. The projection over set U^i is given by computing the minimizer of the following optimization problem

$$\min\left\{\frac{1}{2}\left\|\mathbf{u}_i-\mathbf{v}\right\|^2 : \mathbf{v}\in U^i\right\} = \min\left\{\frac{1}{2}\left\|\tilde{\mathbf{z}}_i-\boldsymbol{\zeta}\right\|^2+\frac{1}{2}(\lambda_i-\mu)^2 : \boldsymbol{\zeta}\in\mu Z^i,\ \mu\geq 0\right\} \tag{47}$$

It is clear that if $\mathbf{u}_i\in U^i$ its projection onto U^i is the vector $\mathbf{u}_i$ itself. Otherwise, we can rewrite the projection problem as follows

$$\min\left\{\left\|\tilde{\mathbf{z}}_i-\boldsymbol{\zeta}\right\|^2+(\lambda_i-\mu)^2 : \boldsymbol{\zeta}\in\mu Z^i,\ \mu\geq 0\right\} = \min\left\{\left\|\tilde{\mathbf{z}}_i-\mu\mathbf{z}_i\right\|^2+(\lambda_i-\mu)^2 : \mathbf{z}_i\in Z^i,\ \mu\geq 0\right\}.$$

Computing this minimum first over $\mathbf{z}_i\in Z^i$ we obtain that if $\mu>0$ then

$$\arg\min\left\{\left\|\tilde{\mathbf{z}}_i-\mu\mathbf{z}_i\right\|^2 : \mathbf{z}_i\in Z^i\right\} = \arg\min\left\{\left\|\frac{\tilde{\mathbf{z}}_i}{\mu}-\mathbf{z}_i\right\|^2 : \mathbf{z}_i\in Z^i\right\} = P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)$$

in which case the optimal $\boldsymbol{\zeta}$ is given in by $\mu P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)$. Otherwise, if $\mu=0$, then all points $\mathbf{z}_i\in Z^i$ are optimal and the optimal $\boldsymbol{\zeta}$ is $\mathbf{0}$. Moreover,

$$\min\{\left\|\mathbf{u}_i-\mathbf{v}\right\|^2 : \mathbf{v}\in U^i\} = \min\left\{\inf_{\mu>0}\psi_i(\mu), \left\|\mathbf{u}_i\right\|^2\right\}$$

since

$$\begin{aligned}\psi_i(\mu) &= \inf_{\boldsymbol{\zeta}\in\mu Z^i}\frac{1}{2}\left\|\tilde{\mathbf{z}}_i-\boldsymbol{\zeta}\right\|^2+\frac{1}{2}(\lambda_i-\mu)^2\\ &= \frac{\mu^2}{2}\left\|\frac{\tilde{\mathbf{z}}_i}{\mu}-P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)\right\|^2+\frac{1}{2}(\lambda_i-\mu)^2\\ &= \frac{1}{2}\mu^2\operatorname{Dist}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}, Z^i\right)^2+\frac{1}{2}(\lambda_i-\mu)^2.\end{aligned}$$

We first show that ψ_i is convex on the domain $\mu>0$. Indeed, since (47) is jointly convex in $\boldsymbol{\zeta}$ and $\mu>0$, ψ_i is a convex function as a partial minimization of a convex problem. In particular, ψ_i is strongly convex since it is a sum of a convex and strongly convex functions. Using [2, Proposition 18.22] and the chain rule, we obtain the following derivative of ψ_i

$$\psi_i'(\mu) = \mu\operatorname{Dist}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}, Z^i\right)^2-\tilde{\mathbf{z}}_i^\top\left(\frac{\tilde{\mathbf{z}}_i}{\mu}-P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)\right)+\mu-\lambda_i = \mu\left\|P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)\right\|^2-\tilde{\mathbf{z}}_i^\top P_{Z^i}\left(\frac{\tilde{\mathbf{z}}_i}{\mu}\right)+\mu-\lambda_i.$$

Due to the strong convexity of $\psi(\mu)$, its infimum over $\mu>0$ is attained if and only if there exists $\mu^*>0$ such that $\psi_i'(\mu^*)=0$. Since $\lim_{\mu\to\infty}\psi_i'(\mu)=\infty$, and due the monotonicity of the gradient of convex functions, such a $\mu^*>0$ exists if and only if $\lim_{\mu\to 0^+}\psi_i'(\mu)<0$ or equivalently $\lim_{\alpha\to\infty}\tilde{\mathbf{z}}_i^\top P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)>-\lambda_i$. In the rest of the proof we show that the latter is always true.

We first note that since $\mathbf{0}\in Z^i$ it follows from [3, Theorem 9.9] that for any $\alpha>0$

$$\alpha\tilde{\mathbf{z}}_i^\top P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)\geq\left\|P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)\right\|^2\geq 0,$$

and thus, $\lim_{\alpha\to\infty}\tilde{\mathbf{z}}_i^\top P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)\geq 0$ which implies that the condition is satisfied for all $\lambda_i>0$. Moreover, it follows from the definition of the support function that $\tilde{\mathbf{z}}_i^\top P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)\leq\sigma_{Z^i}(\tilde{\mathbf{z}}_i)$ for any $\alpha>0$. We will now show that $\lim_{\alpha\to\infty}P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)\in\partial\sigma_{Z^i}(\tilde{\mathbf{z}}_i)=\arg\max_{\mathbf{p}\in Z^i}\tilde{\mathbf{z}}_i^\top\mathbf{p}$. From optimality of the projection we have that $\mathbf{q}=P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)$ if and only if $\mathbf{y}\in\alpha\tilde{\mathbf{z}}_i-\partial\delta_{Z^i}(\mathbf{y})$, so $\alpha\tilde{\mathbf{z}}_i-\mathbf{y}\in\partial\delta_{Z^i}(\mathbf{y})$. Moreover, note that by definition of the indicator function and the subdifferential, if $\mathbf{y}\in\partial\delta_{Z^i}(\mathbf{y})$ the $\alpha\mathbf{y}\in\partial\delta_{Z^i}(\mathbf{y})$ for all $\alpha>0$, and thus $\tilde{\mathbf{z}}_i-\frac{\mathbf{y}}{\alpha}\in\partial\delta_{Z^i}(\mathbf{y})$. Since $\mathbf{p}\in Z^i$ is bounded, it follows that as $\alpha\to\infty$ we have that $\tilde{\mathbf{z}}_i\in\partial\delta_{Z^i}(\mathbf{y})$, which is equivalent to $\mathbf{y}\in\arg\max_{\mathbf{p}\in Z^i}\tilde{\mathbf{z}}_i^\top\mathbf{p}$. Thus, we established that the condition $\lim_{\alpha\to\infty}\tilde{\mathbf{z}}_i^\top P_{Z^i}(\alpha\tilde{\mathbf{z}}_i)>-\lambda_i$ is equivalent to $\sigma_{Z^i}(\tilde{\mathbf{z}}_i)>-\lambda$, concluding our proof. □

Proof of Proposition 4. We look at the saddle point problem (29) where the optimization problem is done over the nonrestricted sets $U^{i,l}$, that is

$$\min_{\boldsymbol{\omega}_i} \max_{\substack{\mathbf{u}_{i,l}\in U^{i,l},\ l\in[s_i]\\ \underline{\lambda}\le\lambda_{i,s_i}\le\tilde{\lambda}}} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top(\mathbf{u}_{i,l}-\mathbf{u}_{i,s_i}). \tag{48}$$

We will run the proof in three steps:

- Proving that (29) has a saddle point.
- Proving that any saddle point of (48) is also a saddle point of (29).
- Proving the boundedness of $\boldsymbol{\omega}^*$ for the problem without the restriction.

It will therefore follow that after restricting $\boldsymbol{\omega}$ problem (29) still has saddle points. We begin with the first step. Indeed, since

$$\sup_{\mathbf{u}_i\in U^i,\ \underline{\lambda}\le\lambda_i\le\tilde{\lambda}} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_i) \le \sup_{\mathbf{u}_i\in U^i} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_i) < \infty$$

by the same logic as (27), (48) must have a saddle point.

Moving to the second step, we use the necessary and sufficient optimality conditions of the saddle point formulation to obtain that $(\boldsymbol{\omega}^\dagger,\tilde{\mathbf{u}}^\dagger)$ is a saddle point of (48) where $\mathbf{u}_{i,l}\in U^{i,l}$ if and only if

$$\begin{aligned}
&\mathbf{u}_{i,l}^\dagger = \mathbf{u}_{i,s_i}^\dagger && l\in[s_i-1]\\
&\quad 0\in\boldsymbol{\omega}_{i,l}^\dagger - \partial\delta_{U^{i,l}}(\mathbf{u}_{i,l}^\dagger), && l\in[s_i-1]\\
&\quad 0\in -\partial_{\mathbf{u}}(-\tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}^\dagger)) - \sum_{l\in[s_i-1]} \boldsymbol{\omega}_{i,l}^\dagger - \partial\delta_{\bar{U}^{i,s_i}}(\mathbf{u}_{i,s_i}^\dagger),
\end{aligned}$$

where $\bar{U}^{i,s_i} = \{\mathbf{u}_{i,s_i}\in U_{i,s} : \underline{\lambda}\le\lambda_{i,s_i}\le\tilde{\lambda}\}\subseteq\tilde{U}^{i,s_i}$. Since $0\le\underline{\lambda}\le\tilde{\lambda}\le\bar{\lambda}$ it follows that $\mathbf{u}_{i,l}^\dagger\in\tilde{U}^{i,l}\subseteq U^{i,l}$ and thus, $\boldsymbol{\omega}_{i,l}^\dagger\in\partial\delta_{U^{i,l}}(\mathbf{u}_{i,l}^\dagger)\subseteq\partial\delta_{\tilde{U}^{i,l}}(\mathbf{u}_{i,l}^\dagger)$ and $-\sum_{l\in[s_i-1]}\boldsymbol{\omega}_{i,l}^\dagger\in\partial\delta_{\bar{U}^{i,s_i}}(\mathbf{u}_{i,s_i}^\dagger)+\partial_{\mathbf{u}}(-\tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}^\dagger))$. Therefore,

$$\begin{aligned}
&\mathbf{u}_{i,l}^\dagger = \mathbf{u}_{i,s_i}^\dagger && l\in[s_i-1]\\
&\quad 0\in\boldsymbol{\omega}_{i,l}^\dagger - \partial\delta_{\tilde{U}^{i,l}}(\mathbf{u}_{i,l}^\dagger), && l\in[s_i-1]\\
&\quad 0\in -\partial_{\mathbf{u}}(-\tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}^\dagger)) - \sum_{l\in[s_i-1]} \boldsymbol{\omega}_{i,l}^\dagger - \partial\delta_{\bar{U}^{i,s_i}}(\mathbf{u}_{i,s_i}^\dagger),
\end{aligned}$$

which are exactly the optimality conditions when using $\tilde{U}^i, l$ instead of $U^{i,l}$, and so $(\boldsymbol{\omega}_i^\dagger,\tilde{\mathbf{u}}_i^\dagger)$ is also a saddle point of the restricted problem.

We now move to the last step, showing that $\boldsymbol{\omega}^\dagger$ must be bounded. Note that if for some $i\in[m]$ we have that $\lambda_{i,s_i}^\dagger = 0$, then the pair $(\boldsymbol{\omega}_i^\dagger,\tilde{\mathbf{u}}_i^\dagger) = (\mathbf{0},\mathbf{0})$ is a saddle point for the ith element of the sum, and so restricting the norm of $\boldsymbol{\omega}_i$ is possible. We therefore continue with bounding $\boldsymbol{\omega}_i$ for the case where $\tilde{\lambda}>0$. By definition, we have that $\mathbf{u}_{i,s_i}^\dagger = (\mathbf{z}_{i,s_i}^\dagger\lambda_{i,s_i}^\dagger,\lambda_{i,s_i}^\dagger)$ where $0\le\underline{\lambda}\le\lambda_{i,s_i}^\dagger\le\tilde{\lambda}\le\bar{\lambda}$. Also note that it must be that $\mathbf{u}_{i,s_i}^\dagger = \mathbf{u}_{i,l}^\dagger$, otherwise the minimization over $\boldsymbol{\omega}_i$ would yield minus infinity. Thus, we have that $\mathbf{z}_{i,s_i}^\dagger\in\cap_{l=1}^{s_i}Z^{i,l}$. Finally, we have that

$$\begin{aligned}
0 &\ge \lambda_{i,s_i}^\dagger g_i(\mathbf{x}^*,\mathbf{z}_{i,s_i}^\dagger)\\
&= \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}^\dagger) + \sum_{l=1}^{s_i-1}(\boldsymbol{\omega}_{i,l}^\dagger)^\top(\mathbf{u}_{i,l}^\dagger-\mathbf{u}_{i,s_i}^\dagger)\\
&= \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\\ \lambda_{i,s_i}=\lambda_{i,s_i}^\dagger}} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1}(\boldsymbol{\omega}_{i,l}^\dagger)^\top(\mathbf{u}_{i,l}-\mathbf{u}_{i,s_i}) && (49)\\
&\ge \max_{\mathbf{z}_{i,l}\in Z^{i,l}} \lambda_{i,s_i}^\dagger\left(g_i(\mathbf{x}^*,\mathbf{z}_{i,s_i}) + \sum_{l=1}^{s_i-1}(\boldsymbol{\nu}_{i,l}^\dagger)^\top(\mathbf{z}_{i,l}-\mathbf{z}_{i,s_i})\right)\\
&\ge \lambda_{i,s_i}^\dagger\left(g_i(\mathbf{x}^*,\mathbf{0}) + \epsilon_i\left\|\boldsymbol{\nu}_{i,l'}^\dagger\right\|\right), \forall l'\in[s_i-1] && (50)
\end{aligned}$$

where the first inequality comes from the feasibility of $\mathbf{x}^*$ for all $\mathbf{z}_i \in Z^i = \cap_{l=1}^{s_i} Z^{i,l}$ and the non-negativity of $\lambda^\dagger_{i,s_i}$, the first equality follows from $\mathbf{u}^\dagger_{i,s_i} = \mathbf{u}^\dagger_{i,l}$, the second equality follows from the optimality of $\mathbf{u}^\dagger$ and $\boldsymbol{\omega}^\dagger$ for the saddle point problem, the second inequality follows from choosing $\mathbf{u}_{i,l} = (\mathbf{z}_{i,l}\lambda^\dagger_{i,s_i}, \lambda^\dagger_{i,s_i})$, and the third inequality follows from choosing $\mathbf{z}_{i,l} = 0$ for all $l \in [s_i]/\{l'\}$ and $\mathbf{z}_{i,l'} = \epsilon_i \boldsymbol{\nu}^\dagger_{i,l'}$. Therefore, if $\lambda^\dagger_{i,s_i} > 0$ it follows from (50) that $\left\|\boldsymbol{\nu}^\dagger_{i,l}\right\| \leq -g_i(\mathbf{x}^*, \mathbf{0})/\epsilon_i \leq \bar{\mu}_i/\epsilon_i$.

Similarly, taking an arbitrary $l' \in [s_i - 1]$, we can bound (49) from below by choosing $\mathbf{u}_{i,l} = (\mathbf{0}, \lambda^\dagger_{i,s_i})$ for all $l \in [s_i] \setminus \{l'\}$ and $\mathbf{u}_{i,l'} = (\mathbf{0}, \lambda^\star_{i,l'})$.

$$\begin{aligned}
0 &\geq \max_{\mathbf{u}_{i,l} \in U_{i,l},\ \underline{\lambda} \leq \lambda_{i,s_i} \leq \tilde{\lambda}} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) + \sum_{l=1}^{s_i-1} (\boldsymbol{\omega}^\dagger_{i,l})^\top (\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \\
&\geq \lambda^\dagger_{i,s_i} g_i(\mathbf{x}^*, \mathbf{0}) + \sum_{l=1}^{s_i-1} \mu^\dagger_{i,l}(\lambda_{i,l} - \lambda^\dagger_{i,s_i}) \\
&= \lambda^\dagger_{i,s_i}(g_i(\mathbf{x}^*, \mathbf{0}) - \mu^\dagger_{i,l'}) + \mu^\dagger_{i,l'}\lambda^\star_{i,l'}.
\end{aligned}$$

Since $\lambda^\star_{i,l'}$ can be taken to infinity, it follows that the equality holds only if $\mu^\dagger_{i,l'} \leq 0$. Moreover, choosing $\lambda^\star_{i,l'} = 0$ implies that $-\mu^\dagger_{i,l'} \leq -g_i(\mathbf{x}^*, \mathbf{0}) \leq \bar{\mu}_i$. Since l' was arbitrarily chosen the proof is complete. □

Proof of Proposition 5. In the proof the the proposition, claim (30) plays the key role, from which (31) and (32) follow.

Proof of (30) Denoting $G_{\boldsymbol{\chi}}, G_{\mathbf{u},i}, G_{\mathbf{w}}$ as bounds on the subgradients $\mathbf{v}^k_{\boldsymbol{\chi}} \in \partial_\chi \breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1})$, $\mathbf{v}^k_{\tilde{\mathbf{u}}} \in \partial_{\tilde{\mathbf{u}}}(-\breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1}))$, $\mathbf{v}^k_{\mathbf{w}} \in \partial_{\mathbf{w}} \breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1})$ used throughout the algorithm, the first of the corollary follows directly from [16, Lemmas 3.1 and 3.2]. It is left to prove that under the chosen assumptions these bounds exist and are equal to the stated values.

We begin with the primal variables χ. Denoting $\mathbf{u}^{k-1}_{i,l} = (\tilde{\mathbf{z}}^{k-1}_{i,l}, \lambda^{k-1}_{i,l})$ and defining $\mathbf{z}^{k-1}_{i,s_i}$ as

$$\mathbf{z}^k_i = \begin{cases} \frac{\tilde{\mathbf{z}}^k_i}{\lambda^k_i}, & \lambda_i > 0 \\ \mathbf{0}, & \lambda_i = 0. \end{cases}$$

we have that

$$\mathbf{v}^k_\chi = \begin{bmatrix} \mathbf{c} + \mathbf{A}^\top \mathbf{w}^{k-1} + \sum_{i=1}^m \lambda^{k-1}_{i,s_i} \mathbf{d}^k_{x,i} \\ (\mathbf{u}^{k-1}_{1,s_1} - \mathbf{u}^{k-1}_{1,s_1-1}) \\ \vdots \\ (\mathbf{u}^{k-1}_{m,s_m} - \mathbf{u}^{k-1}_{m,s_m-1}) \end{bmatrix} \in \partial_\chi \breve{L}(\boldsymbol{\chi}^{k-1}, \mathbf{u}^{k-1}),$$

where $\mathbf{d}^k_{x,i} \in \partial_{\mathbf{x}} g_i(\mathbf{x}^{k-1}, \mathbf{z}^{k-1}_{i,s_i})$. By boundedness of the subgradients, λ_i and $\mathbf{w}$ we can bound the first component

$$\left\|\mathbf{c} + \mathbf{A}^\top \mathbf{w}^{k-1} + \sum_{i=1}^m \lambda^{k-1}_i \mathbf{d}^k_{x,i}\right\| \leq \|\mathbf{c}\| + \|\mathbf{A}\| R_{\mathbf{w}} + \sum_{i=1}^m \bar{\lambda} G_{x,i}.$$

By the definition of $\tilde{U}^{i,l}$ we have that for all $i \in [m]$ and $l \in [s_i - 1]$

$$\begin{aligned}
\left\|\mathbf{u}^{k-1}_{i,s_i} - \mathbf{u}^{k-1}_{i,l}\right\| &\leq (\lambda^{k-1}_{i,s_i} - \lambda^{k-1}_{i,l}) + \left\|\tilde{\mathbf{z}}_{i,s_i} - \tilde{\mathbf{z}}_{i,l}\right\| \\
&\leq 2\bar{\lambda} + \bar{\lambda} \max_{\mathbf{z}_{i,s_i} \in Z^{i,s_i}, \mathbf{z}_{i,l} \in Z^{i,l}} \left\|\mathbf{z}_{i,s_i} - \mathbf{z}_{i,l}\right\| \leq \bar{\lambda}(2 + R_{i,s_i} + R_{i,l}).
\end{aligned}$$

Adding these two bounds, we obtain the following bound:

$$\left\|\mathbf{v}^k_\chi\right\| \leq \|\mathbf{c}\| + \|\mathbf{A}\| R_{\mathbf{w}} + \sum_{i \in [m]} \bar{\lambda} G_{x,i} + \bar{\lambda} \sum_{i \in [m]} \left((s_i - 1)(2 + R_{i,s_i}) + \sum_{l \in [s_i - 1]} R_{i,l} \right) = G_{\boldsymbol{\chi}}$$

Now, we will bound the norms of the subgradients corresponding to the dual variables $\mathbf{u}_{i,l}$, $\mathbf{w}$:

$$\mathbf{v}^k_{\tilde{\mathbf{u}}_i} = \begin{bmatrix} \mathbf{v}^k_{i,1} \\ \vdots \\ \mathbf{v}^k_{i,s_i} \end{bmatrix}$$

First, consider the subgradients $\mathbf{v}_{i,s_i}$ with respect to $\mathbf{u}_{i,s_i}$. According to Lemma 2

$$\mathbf{v}^k_{i,s_i} = \begin{bmatrix} \lambda^{k-1}_{i,s_i}\mathbf{d}^k_{z,i} \\ (-g_i)(\mathbf{x}^{k-1}, \mathbf{z}^{k-1}_{i,s_i}) + (\mathbf{d}^k_{z,i})^\top \mathbf{z}^{k-1}_{i,s_i} \end{bmatrix} - \sum_{l\in[s_i-1]} \boldsymbol{\omega}^{k-1}_{i,l} \in \partial_{\mathbf{u}_{i,s_i}}(-\breve{L})(\boldsymbol{\chi}^{k-1}, \mathbf{u}^{k-1}),$$

where $\mathbf{d}^k_{z,i} = \partial_{\mathbf{z}_i}(-g_i)(\mathbf{x}^{k-1}, \mathbf{z}^{k-1}_{i,s_i})$. From the definitions of the sets $\tilde{U}^{i,l}$,$\Omega^{i,l}$, and W we get:

$$\begin{aligned} \left\|\mathbf{v}^k_{i,s_i}\right\| &\le \left\|\lambda^{k-1}_{i,s_i}\mathbf{d}^k_{z,i}\right\| + \left|(-g_i)(\mathbf{x}^{k-1}, \mathbf{z}^{k-1}_{i,s_i})\right| + \left\|(\mathbf{d}^k_{z,i})\right\| \left\|\mathbf{z}^{k-1}_{i,s_i}\right\| + \sum_{l\in[s_i-1]} \left\|\boldsymbol{\omega}^{k-1}_{i,l}\right\| \\ &\le (\bar{\lambda} + R_{i,s_i})G_{z,i} + \bar{g}_i + \bar{\mu}_i(s_i-1)\left(1+\frac{1}{\epsilon_i}\right). \end{aligned}$$

Moreover, for any $l \in [s_i - 1]$ it follows from the definition $\Omega^{i,l}$ that

$$\mathbf{v}^k_{i,l} = -\boldsymbol{\omega}^{k-1}_{i,l} \in \partial_{\mathbf{u}_{i,l}}(-\breve{L})(\boldsymbol{\chi}^{k-1}, \mathbf{y}^{k-1}) \quad \Rightarrow \quad \left\|\mathbf{v}^k_{i,l}\right\| \le \bar{\mu}_i\left(1+\frac{1}{\epsilon_i}\right).$$

In the end, for the subgradients w.r.t. $\mathbf{w}$, we have $\mathbf{v}^k_{\mathbf{w}} = -\mathbf{A}\mathbf{x}^{k-1} + \mathbf{b}$ so by boundedness of X we obtain the bound $\left\|\mathbf{v}^k_{\mathbf{w}}\right\| \le \|\mathbf{A}\| R_{\mathbf{x}} + \|\mathbf{b}\|$. Using the defined $\mathbf{v}^k_{\chi}$, $\mathbf{v}^k_{i,l}$ and $\mathbf{v}^k_{\mathbf{w}}$ in the algorithm we obtain the desired result. □

Proof of (31) To prove this claim, we will use (30). First, define $\mathbf{u}^\dagger_i = \lambda_i \mathbf{z}^\dagger_i$ where $\mathbf{z}^\dagger_i = \arg\max_{\mathbf{z}_i \in Z^i} g(\bar{\mathbf{x}}^N, \mathbf{z}_i)$. Following (27) we have that

$$\begin{aligned} L(\bar{\mathbf{x}}^N; (\boldsymbol{\lambda}, \mathbf{w})) &= \mathbf{c}^\top \bar{\mathbf{x}}^N + \sum_{i\in[m]} \tilde{g}_i(\bar{\mathbf{x}}^N, \mathbf{u}^\dagger_i) + \mathbf{w}^\top\left(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}\right) \\ &= \min_{\boldsymbol{\omega}} \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ \lambda_{i,l}=\lambda_i \\ l\in[s_i],\ i\in[m]}} \mathbf{c}^\top \bar{\mathbf{x}}^N + \sum_{i\in[m]} \tilde{g}_i(\bar{\mathbf{x}}^N, \mathbf{u}_{i,s_i}) + \mathbf{w}^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} \boldsymbol{\omega}^\top_{i,l}(\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \\ &\le \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ \lambda_{i,l}=\lambda_i \\ l\in[s_i],\ i\in[m]}} \mathbf{c}^\top \bar{\mathbf{x}}^N + \sum_{i\in[m]} \tilde{g}_i(\bar{\mathbf{x}}^N, \mathbf{u}_{i,s_i}) + \mathbf{w}^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} (\bar{\boldsymbol{\omega}}^N_{i,l})^\top(\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \\ &= \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ \lambda_{i,l}=\lambda_i \\ l\in[s_i],\ i\in[m]}} \breve{L}((\bar{\mathbf{x}}^N, \bar{\boldsymbol{\omega}}^N), (\tilde{\mathbf{u}}, \mathbf{w})). \end{aligned} \tag{51}$$

We move on to lower bounding $L(\mathbf{x}^*, (\bar{\boldsymbol{\lambda}}^N, \bar{\mathbf{w}}^N))$ with a term of the form $\breve{L}((\mathbf{x}^*, \boldsymbol{\omega}), \bar{\mathbf{y}}^N)$:

$$\begin{aligned} &L(\mathbf{x}^*, (\bar{\boldsymbol{\lambda}}^N, \bar{\mathbf{w}}^N)) \\ =& \max_{\mathbf{u}_i\in\bar{U}_i,\ \bar{\lambda}^N_i = \bar{\lambda}^N_{i,s_i}} \mathbf{c}^\top \mathbf{x}^* + \sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_i) \\ =& \min_{\boldsymbol{\omega}} \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ l\in[s_i] \\ \lambda_{i,s_i} = \bar{\lambda}^N_{i,s_i}}} \mathbf{c}^\top \mathbf{x}^* + \bar{\mathbf{w}}^{N\top}(\mathbf{A}\mathbf{x}^* - \mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*, \mathbf{u}_{i,s_i}) + \sum_{i=1}^m \sum_{l=1}^{s_i-1} \boldsymbol{\omega}^\top_{i,l}(\mathbf{u}_{i,l} - \mathbf{u}_{i,s_i}) \end{aligned} \tag{52}$$

$$
\begin{aligned}
&= \min_{\boldsymbol{\omega}\in\Omega} \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ l\in[s_i] \\ \lambda_{i,s_i}=\bar{\lambda}^N_{i,s_i}}} \mathbf{c}^\top\mathbf{x}^* + \bar{\mathbf{w}}^{N\top}(\mathbf{A}\mathbf{x}^*-\mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}) + \sum_{i=1}^m\sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top(\mathbf{u}_{i,l}-\mathbf{u}_{i,s_i}) \\
&= \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ l\in[s_i] \\ \lambda_{i,s_i}=\bar{\lambda}^N_{i,s_i}}} \min_{\boldsymbol{\omega}\in\Omega} \mathbf{c}^\top\mathbf{x}^* + \bar{\mathbf{w}}^{N\top}(\mathbf{A}\mathbf{x}^*-\mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*,\mathbf{u}_{i,s_i}) + \sum_{i=1}^m\sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top(\mathbf{u}_{i,l}-\mathbf{u}_{i,s_i}) \\
&\geq \min_{\boldsymbol{\omega}\in\Omega} \mathbf{c}^\top\mathbf{x}^* + \bar{\mathbf{w}}^{N\top}(\mathbf{A}\mathbf{x}^*-\mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\mathbf{x}^*,\bar{\mathbf{u}}^N_{i,s_i}) + \sum_{i=1}^m\sum_{l=1}^{s_i-1} \boldsymbol{\omega}_{i,l}^\top(\bar{\mathbf{u}}^N_{i,l}-\bar{\mathbf{u}}^N_{i,s_i}) \\
&= \min_{\boldsymbol{\omega}\in\Omega} \breve{L}((\mathbf{x}^*,\boldsymbol{\omega}),\bar{\mathbf{y}}^N)
\end{aligned} \tag{53}
$$

where the first equality follows (27), the second equality follows from Proposition 4, the third equality follows from the fact that existence of a saddle point, established in (27), and the final inequality follows from the definition of $\breve{L}$. Moreover, for any $\bar{\boldsymbol{\lambda}}^N_i \equiv \bar{\boldsymbol{\lambda}}^N_{i,s_i}$, and $\bar{\mathbf{w}}^N$), it follows from $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ being a saddle point of L that

$$
L(\mathbf{x}^*,(\bar{\boldsymbol{\lambda}}^N,\bar{\mathbf{w}}^N)) \leq L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*)). \tag{54}
$$

Combining (51), (53), and (54) we have that for any $\lambda_i \in [0,\bar{\lambda}]$ and any $\mathbf{w}\in W$ the following holds

$$
\begin{aligned}
&L(\bar{\mathbf{x}}^N,(\boldsymbol{\lambda},\mathbf{w})) - L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*)) \\
\leq& L(\bar{\mathbf{x}}^N,(\boldsymbol{\lambda},\mathbf{w})) - L(\mathbf{x}^*,(\bar{\boldsymbol{\lambda}}^N,\bar{\mathbf{w}}^N)) \\
\leq& \max_{\boldsymbol{\omega}\in\Omega} \max_{\substack{\mathbf{y}=(\tilde{\mathbf{u}},\mathbf{w}): \\ \mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ \lambda_{i,l}=\lambda_i \\ l\in[s_i],\ i\in[m]}} \left(\breve{L}((\bar{\mathbf{x}}^N,\bar{\boldsymbol{\omega}}^N),\mathbf{y}) - \breve{L}((\mathbf{x}^*,\boldsymbol{\omega}),\bar{\mathbf{y}}^N)\right) \\
\leq& \max_{\boldsymbol{\omega}\in\Omega} \max_{\substack{\mathbf{u}_{i,l}\in\tilde{U}^{i,l},\ \lambda_{i,l}=\lambda_i \\ l\in[s_i],\ i\in[m]}} \frac{1}{2\sqrt{N}} \left(\frac{\left\|\mathbf{x}^0-\mathbf{x}^*\right\|^2 + \left\|\boldsymbol{\omega}^0-\boldsymbol{\omega}\right\|^2}{\tilde{\tau}} + \tilde{\tau}G^2_{\boldsymbol{\mathcal{X}}} \right. \\
&\qquad \left. + \sum_{i=1}^m \left(\sum_{l=1}^{s_i} \frac{\left\|\mathbf{u}^0_{i,l}-\mathbf{u}_{i,l}\right\|^2}{\tilde{\theta}_i} + \tilde{\theta}_i G^2_{\tilde{\mathbf{u}}_i} \right) + \frac{\left\|\mathbf{w}^0-\mathbf{w}\right\|^2}{\tilde{\theta}_{\mathbf{w}}} + \tilde{\theta}_{\mathbf{w}} G^2_{\mathbf{w}} \right),
\end{aligned} \tag{55}
$$

where the last inequality follows from (30). Now we can use the fact that for any $\boldsymbol{\omega},\boldsymbol{\omega}^0\in\Omega$

$$
\left\|\boldsymbol{\omega}^0-\boldsymbol{\omega}\right\| \leq 2\sum_{i=1}^m \bar{\mu}_i(s_i-1)(1+1/\epsilon_i)
$$

and for any $\mathbf{u}^0_{i,l},\mathbf{u}_{i,l}\in\tilde{U}^{i,l}$ such that $\mathbf{u}_{i,l}=(\tilde{\mathbf{z}}_{i,l},\lambda_{i,l}),\ \lambda_{i,l}=\lambda_i$

$$
\left\|\mathbf{u}^0_{i,l}-\mathbf{u}_{i,l}\right\|^2 \leq \max\{\lambda^0_i,\lambda_i\}^2(1+4R^2_{i,l}).
$$

In the end we have

$$
\left\|\mathbf{w}^0-\mathbf{w}\right\|^2 \leq \frac{2\max\{\|\mathbf{w}\|,\|\mathbf{w}_0\|\}^2}{\tilde{\theta}_{\mathbf{w}}}
$$

We can use these inequalities and the bounds on variables to bound (55) and the definition of ϕ and σ_i we obtain the desired result.

Proof of (32) For the last claim we use the fact that if $((\mathbf{x}^*,\boldsymbol{\omega}^*),(\tilde{\mathbf{u}}^*,\mathbf{w}^*))$ is a saddle point of $\breve{L}$ then $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ where $\lambda^*_i=\lambda^*_{i,s_i}$ is a saddle point of L. Thus,

$$
L(\bar{\mathbf{x}}^N,(\boldsymbol{\lambda}^*,\mathbf{w}^*)) \geq L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*)) = \mathbf{c}^\top\mathbf{x}^*,
$$

where the inequality follows from $(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))$ being a saddle point, and the equality follows from the saddle point value for the Lagrangian being equal to the optimal primal objective value. □

Proof of Proposition 7. The first part of the proposition is a direct result of [10, Theorem 3]. To prove the second part, the strategy is the same as in the SGSP case: we upper (lower) bound the first (second) terms in the left-hand side of (36) with terms that are like the ones in the LHS of (35) albeit over a bounded domain, and then suprimize the RHS of (35) over that domain.

We first note that

$$\begin{aligned}\alpha \operatorname{Dist}(\bar{\mathbf{x}}^N, X) &= \inf_{\mathbf{x}\in X} \alpha \left\|\bar{\mathbf{x}}^N - \mathbf{x}\right\| \\ &= \inf_{\mathbf{x}\in X} \max_{\|\boldsymbol{\pi}\|\le\alpha} \boldsymbol{\pi}^\top(\bar{\mathbf{x}}^N - \mathbf{x}) \\ &= \max_{\|\boldsymbol{\pi}\|\le\alpha} \boldsymbol{\pi}^\top \bar{\mathbf{x}}^N - \sup_{\mathbf{x}\in X} \boldsymbol{\pi}^\top \mathbf{x} \\ &= \max_{\|\boldsymbol{\pi}\|\le\alpha} \boldsymbol{\pi}^\top \bar{\mathbf{x}}^N - \sigma_X(\boldsymbol{\pi}), \end{aligned} \tag{56}$$

where the third inequality follows from Sion's theorem, and the fourth equality follows from definition of the support function. Defining $\mathbf{u}^\dagger_{i,l} = \lambda_i(\mathbf{z}^\dagger_i, 1)$ for all $l = 1, \ldots, s_i$, where $\mathbf{z}^\dagger_i := \arg\max_{\mathbf{z}_i\in Z^i} g_i(\bar{\mathbf{x}}^N, \mathbf{z}_i)$, we then have:

$$\begin{aligned}&L(\bar{\mathbf{x}}^N, (\boldsymbol{\lambda}, \mathbf{w})) + \alpha \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \\ =&\mathbf{c}^\top \bar{\mathbf{x}}^N + \mathbf{w}^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\bar{\mathbf{x}}^N, \mathbf{u}^\dagger_i) + \max_{\|\boldsymbol{\pi}\|\le\alpha} \left\{\boldsymbol{\pi}^\top \bar{\mathbf{x}}^N - \sigma_X(\boldsymbol{\pi})\right\} \\ =&\mathbf{c}^\top \bar{\mathbf{x}}^N + \mathbf{w}^\top(\mathbf{A}\bar{\mathbf{x}}^N - \mathbf{b}) + \sum_{i\in[m]} \tilde{g}_i(\bar{\mathbf{x}}^N, \mathbf{u}^\dagger_{i,s_i}) + \sum_{l\in[s_i-1]} (\bar{\boldsymbol{\omega}}^N_{i,l})^\top(\mathbf{u}^\dagger_{i,l} - \mathbf{u}^\dagger_{i,s_i}) + \max_{\|\boldsymbol{\pi}\|\le\alpha} \left\{\boldsymbol{\pi}^\top \bar{\mathbf{x}}^N - \sigma_X(\boldsymbol{\pi})\right\} \\ =&\sup_{\boldsymbol{\pi}:\|\boldsymbol{\pi}\|\le\alpha} \check{L}(\bar{\boldsymbol{\chi}}^N, (\tilde{\mathbf{u}}^\dagger, \mathbf{w}, \boldsymbol{\pi})). \end{aligned} \tag{57}$$

where the first equality follows from the definition of L and (56), the second equality follows from $\mathbf{u}^\dagger_{i,l} = \mathbf{u}^\dagger_{i,s_i}$ for all $l \in [s_i - 1]$ and $i \in [m]$, and the third equality follows from definition of $\check{L}$. Since $(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*))$ is a saddle point of L, it can be extended to $(\boldsymbol{\chi}^*, \mathbf{y}^*)$ which is a saddle point of $\check{L}$, and by definition

$$L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) = \check{L}(\boldsymbol{\chi}^*, \mathbf{y}^*) \ge \check{L}(\boldsymbol{\chi}^*, \bar{\mathbf{y}}^N),$$

which combined with inequalities (35) and (57) gives

$$L(\bar{\mathbf{x}}^N, (\boldsymbol{\lambda}, \mathbf{w})) - L(\mathbf{x}^*, (\boldsymbol{\lambda}^*, \mathbf{w}^*)) + \alpha \operatorname{Dist}(\bar{\mathbf{x}}^N, X) \le \sup_{\substack{\mathbf{y}=(\tilde{\mathbf{u}}^\dagger, \mathbf{w}, \boldsymbol{\pi}):\mathbf{u}^\dagger_{i,l}=\mathbf{u}^\dagger_{i,s_i} \\ \mathbf{u}^\dagger_{i,s_i}=\lambda_i(\mathbf{z}^\dagger_i,1) \\ \mathbf{z}^\dagger_i\in\arg\max_{\mathbf{z}\in Z^i} g_i(\bar{\mathbf{x}}^N,\mathbf{z}) \\ \|\boldsymbol{\pi}\|\le\alpha}} \frac{\tau^{-1}\|\boldsymbol{\chi}^* - \boldsymbol{\chi}^0\|^2 + \|\mathbf{y} - \mathbf{y}^0\|^2_{\mathbf{H}}}{2N}. \tag{58}$$

From Proposition 4 we have that

$$\left\|\boldsymbol{\chi}^* - \boldsymbol{\chi}^0\right\|^2 = \left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2 + \left\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^0\right\|^2 \le \left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2 + 4\sum_{i=1}^m s_i \bar{\mu}_i^2 \left(1 + \frac{1}{\epsilon_i}\right)^2. \tag{59}$$

The boundedness of Z^i implies that

$$\sup_{\substack{\mathbf{y}=(\tilde{\mathbf{u}}^\dagger, \mathbf{w}, \boldsymbol{\pi}):\mathbf{u}^\dagger_{i,l}=\mathbf{u}^\dagger_{i,s_i} \\ \mathbf{u}^\dagger_{i,s_i}=\lambda_i(\mathbf{z}^\dagger_i,1) \\ \mathbf{z}^\dagger_i\in\arg\max_{\mathbf{z}\in Z^i} g_i(\bar{\mathbf{x}}^N,\mathbf{z}) \\ \|\boldsymbol{\pi}\|\le\alpha}} \left\|\mathbf{y} - \mathbf{y}^0\right\|^2_{\mathbf{H}} \le \sup_{\substack{\mathbf{y}=(\tilde{\mathbf{u}}^\dagger, \mathbf{w}, \boldsymbol{\pi}):\mathbf{u}^\dagger_{i,l}=\mathbf{u}^\dagger_{i,s_i} \\ \mathbf{u}^\dagger_{i,s_i}=\lambda_i(\mathbf{z}^\dagger_i,1) \\ \mathbf{z}^\dagger_i\in\arg\max_{\mathbf{z}\in Z^i} g_i(\bar{\mathbf{x}}^N,\mathbf{z}) \\ \|\boldsymbol{\pi}\|\le\alpha}} \left\|\mathbf{y} - \mathbf{y}^0\right\|^2_{\mathbf{S}} \le$$

$$\sum_{i\in[m]}\frac{1}{\theta_i}\sum_{l\in[s_i]}\max_{\mathbf{u}_{i,l}=\lambda_i\mathbf{z}_{i,l}:\mathbf{z}_{i,l}\in Z^{i,l}}\left\|\mathbf{u}_{i,l}-\mathbf{u}_{i,l}^0\right\|^2+\max_{\boldsymbol{\pi}:\|\boldsymbol{\pi}\|\leq\alpha}\frac{1}{\theta_{\boldsymbol{\pi}}}\left\|\boldsymbol{\pi}-\boldsymbol{\pi}^0\right\|^2+\frac{1}{\theta_{\mathbf{w}}}\left\|\mathbf{w}-\mathbf{w}^0\right\|^2\leq$$

$$\sum_{i\in[m]}\frac{1}{\theta_i}\sum_{l\in[s_i]}(1+4R_{i,l}^2)\max\{\lambda_i,\lambda_i^0\}^2+\frac{2(\alpha+\left\|\boldsymbol{\pi}^0\right\|^2)}{\theta_{\boldsymbol{\pi}}}+\frac{2}{\theta_{\mathbf{w}}}\max\{\|\mathbf{w}\|,\left\|\mathbf{w}^0\right\|\}^2=$$

$$\sum_{i\in[m]}\frac{\sigma_i}{\theta_i}\max\{\lambda_i,\lambda_i^0\}^2+\frac{2(\alpha^2+\left\|\boldsymbol{\pi}^0\right\|^2)}{\theta_{\boldsymbol{\pi}}}+\frac{2}{\theta_{\mathbf{w}}}\max\{\|\mathbf{w}\|,\left\|\mathbf{w}^0\right\|\}^2 \tag{60}$$

Combining (58), (59), and (60) we obtain (36). To prove the last part, we use the fact that $\|\boldsymbol{\pi}^*\|\leq R_\pi$ and so by (57) choosing $\alpha=R_{\boldsymbol{\pi}}$ we have

$$L(\bar{\mathbf{x}}^N,(\boldsymbol{\lambda}^*,\mathbf{w}^*))+R_{\boldsymbol{\pi}}\operatorname{Dist}(\bar{\mathbf{x}}^N,X)=\sup_{\substack{(\tilde{\mathbf{u}}^\dagger,\boldsymbol{\pi}):\|\boldsymbol{\pi}\|\leq R_{\boldsymbol{\pi}}\\ \mathbf{u}_{i,l}^\dagger=\mathbf{u}_{i,s_i}^\dagger,l\in[s_i-1]\\ \mathbf{z}_i^\dagger\in\arg\max_{\mathbf{z}\in Z^i}g_i(\bar{\mathbf{x}}^N,\mathbf{z})\\ \mathbf{u}_{i,s_i}^\dagger=\lambda_i^*(1,\mathbf{z}_i^\dagger)\ ,i\in[m]}}\check{L}(\bar{\boldsymbol{\chi}}^N,(\boldsymbol{\pi},\tilde{\mathbf{u}}^\dagger,\mathbf{w}^*))\geq\check{L}(\bar{\boldsymbol{\chi}}^N,\mathbf{y}^*). \tag{61}$$

Moreover, from the definition of the saddle point we have that

$$\check{L}(\bar{\boldsymbol{\chi}}^N,\mathbf{y}^*)\geq\check{L}(\boldsymbol{\chi}^*,\mathbf{y}^*)=L(\mathbf{x}^*,(\boldsymbol{\lambda}^*,\mathbf{w}^*))=\mathbf{c}^\top\mathbf{x}^*. \tag{62}$$

Combining (61) and (62) concludes the proof. □

References

1. J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi – A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36, 2019.
2. H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
3. A. Beck. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. Siam, 2014.
4. A. Ben-Tal, D. den Hertog, and J.-Ph. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1):265–299, Feb 2015.
5. A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.
6. A. Ben-Tal, E. Hazan, T. Koren, and S. Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015.
7. D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.
8. D. Bienstock. Histogram models for robust portfolio optimization. *Journal of Computational Finance*, 11(1):1, 2007.
9. P. L. Combettes and Ch. L Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, 457(2):1283–1306, 2018.
10. Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209 – 214, 2015.
11. V. Gabrel, C. Murat, and A. Thiele. Recent advances in robust optimization: An overview. *European Journal of Operational Research*, 235(3):471–483, 2014.
12. E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.
13. N. Ho-Nguyen and F. Kılınç-Karzan. Online first-order framework for robust convex optimization. *Operations Research*, 66(6):1670–1692, 2018.
14. V. Jeyakumar and G. Li. Trust-region problems with linear inequality constraints: exact sdp relaxation, global optimality and robust optimization. *Mathematical Programming*, 147(1-2):171–206, 2014.
15. A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.
16. A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.
17. Gurobi Optimization. Gurobi optimizer reference manual, 2020.
18. A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.