Alternating Direction Method of Multipliers for Decomposable Saddle-Point Problems

Mustafa O. Karabag
Electrical & Computer Engineering
The University of Texas at Austin
Austin, USA
karabag@utexas.edu

David Fridovich-Keil
Aerospace Engineering
The University of Texas at Austin
Austin, USA
dfk@utexas.edu

Ufuk Topcu
Aerospace Engineering
The University of Texas at Austin
Austin, USA
utopcu@utexas.edu

2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3503-9998-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ALLERTON49937.2022.9929349

***Abstract*—Saddle-point problems appear in various settings including machine learning, zero-sum stochastic games, and regression problems. We consider decomposable saddle-point problems and study an extension of the alternating direction method of multipliers to such saddle-point problems. Instead of solving the original saddle-point problem directly, this algorithm solves smaller saddle-point problems by exploiting the decomposable structure. We show the convergence of this algorithm for convex-concave saddle-point problems under a mild assumption. We also provide a sufficient condition for which the assumption holds. We demonstrate the convergence properties of the saddle-point alternating direction method of multipliers with numerical examples on a power allocation problem in communication channels and a network routing problem with adversarial costs.**

***Index Terms*—Saddle-point problems, decomposable optimization, alternating direction method of multipliers**

I. Introduction

Saddle-point problems consider optimization of an objective function simultaneously by a minimizer and maximizer. These problems appear, for example, in zero-sum stochastic games [1], adversarial training of machine learning models [2], [3], regression problems [4], and maximum-margin estimation of structured output models [5].

We focus on decomposable saddle-point problems of the following form that have a decomposable objective function with *complicating* global constraints:

$$\min_{x_a} \max_{x_b} \quad \sum_{i=1}^{N} f_i(x_{a,i}, x_{b,i}) \tag{1a}$$

$$\text{subject to} \quad x_a \in \mathcal{X}_a \tag{1b}$$

$$x_b \in \mathcal{X}_b \tag{1c}$$

$$x_{a,i} \in \mathcal{X}_{a,i}, \quad \text{for all } i \in 1, \dots, N \tag{1d}$$

$$x_{b,i} \in \mathcal{X}_{b,i}, \quad \text{for all } i \in 1, \dots, N \tag{1e}$$

where $x_a = [x_{a,1}, \dots, x_{a,N}]$ and $x_b = [x_{b,1}, \dots, x_{b,N}]$ are each concatenations of N vectors, and $\mathcal{X}_a \subseteq \mathbb{R}^{n_a}$, $\mathcal{X}_b \subseteq \mathbb{R}^{n_b}$, $\mathcal{X}_{a,i} \subseteq \mathbb{R}^{n_{a,i}}$, and $\mathcal{X}_{b,i} \subseteq \mathbb{R}^{n_{b,i}}$ are closed, convex sets such that $\sum_{i=1}^{N} n_{a,i} = n_a$ and $\sum_{i=1}^{N} n_{b,i} = n_b$. In particular, we are interested in the convex-concave case, i.e., $f_i : \mathcal{X}_{a,i} \times \mathcal{X}_{b,i} \to \mathbb{R}$ is convex in $x_{a,i}$ and concave in $x_{b,i}$. This problem structure arises, for example, in power allocation problems for communication channels with adversarial noise [6] and optimal network routing problems with adversarial costs.

This work was supported in part by NSF 1652113 and ARO W911NF-20-1-0140.

The paper [7] proposed the alternating direction method of multipliers (ADMM) to solve an optimization problem with decomposable nonconvex-concave objective functions. In this paper, we analyze the convergence properties of this method, saddle-point ADMM (SP-ADMM), for the decomposable convex-concave objective functions. The iterative SP-ADMM preserves the separable structure of (1) and consists of three steps. In the first step, SP-ADMM solves a saddle-point problem separately for every block. It performs projections onto the global constraints (1b)–(1c) in the next step, and performs the dual variable updates in the last step.

SP-ADMM has several advantages. Each individual saddle-point problem has a lower number of dimensions compared to the original problem and hence can be solved more efficiently. For some objective functions, for example bilinear functions of two one-dimensional variables, these individual saddle-point problems can be solved analytically. Since the individual saddle-point problems have no coupling, they can be solved in parallel. SP-ADMM performs the projection onto the global constraints without considering the individual constraints. For some global constraints such as unit ball or probability simplex, this projection step can be performed more efficiently compared to the case that takes the individual constraints into account.

The contributions of this paper are threefold. The paper [7] demonstrated the performance of SP-ADMM for a specific robust optimization problem without any theoretical guarantees. We analyze the performance of SP-ADMM. We first show that for the convex-concave case SP-ADMM converges to the saddle point of the problem under a mild assumption. Secondly, we provide a sufficient condition for convergence by considering standard conditions of the minimax theorem [8] and Slater's constraint qualification [9]. Finally, we demonstrate and evaluate the performance of SP-ADMM for a power allocation problem for communication channels with adversarial noise [6] and an optimal network routing problem with adversarial costs.

II. Related Work

Saddle-point problems: Convergent variants of gradient descent-ascent methods such as the extra gradient method [10], optimistic gradient descent-ascent method [11], and subgradient descent-ascent method [12] have been proposed for convex-concave saddle-point problems. The paper [13] extended the Frank-Wolfe (conditional gradient) method to solve strongly convex-strongly concave saddle-point problems.

We remark that first-order methods can also exploit the decomposable structure during gradient computation. However, the coexistence of local and global constraints for the projection step may result in harder optimization problems compared to SP-ADMM that decouples the projection step and local constraints.

SP-ADMM solves saddle-point problems as a subroutine and one can employ these methods to solve the individual saddle-problems. The quadratic penalties introduced in the SP-ADMM results in strongly convex-strongly concave objective functions that often increase the rate of convergence.

Decomposable optimization: Decomposable optimization studies optimization problems that can be decomposed into smaller sub-problems once the complicating constraints (or variables) are removed. Seminal Dantzig–Wolfe [14] and Benders [15] decomposition methods solve block decomposable linear programs. ADMM [16], [17] solves general decomposable convex optimization problems. ADMM has convergence guarantees for convex problems [18], [19] and also for some nonconvex problems [20], [21]. In practice, ADMM often generates acceptable solutions in a few iterations, however it behaves like a first-order method and suffers from slow convergence in the long run [18].

Decentralized saddle-point problems: Decentralized saddle-point problems [22]–[26] consider the optimization of a separable objective function subject to the communication constraints (usually defined with a graph). Unlike the decomposable setting that we consider, these works consider that each component of the objective function is a function of a global variable. The paper [24] also consider local variables as a part of the objective functions, however these local variables do not have complicating constraints that we have in (1).

III. Notation and Preliminaries for Decomposable Optimization

A. Notation

We use subscripts a and b with colors blue and red to denote the variables/constants of the minimizer and maximizer, respectively. The subscript i denotes the i^{th} block (element) of the object with the subscript. With an abuse of notation we also use the subscript i, for $\mathcal{X}_{a,i}$ and $\mathcal{X}_{b,i}$ that the sets for $x_{a,i}$ and $x_{b,i}$, and are not blocks of $\mathcal{X}_a$ and $\mathcal{X}_b$, respectively. The superscript k denotes the value of the variable with the superscript at the k^{th} iteration of the algorithms. The superscript 2 is used as an exponent. $I_{\mathcal{X}}(x)$ is the indicator function of set $\mathcal{X}$ such that $I_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and ∞ otherwise.

B. Preliminaries for Decomposable Optimization

While the objective function is block-decomposable for (1), the constraints $x_a \in \mathcal{X}_a$ and $x_b \in \mathcal{X}_b$ are not separable. The potential existence of these constraints result in different problem structures:

a) Fully separable case: In the absence of both $x_a \in \mathcal{X}_a$ and $x_b \in \mathcal{X}_b$, we can solve the saddle-point problem separately for every block i. The saddle points of these individual problems are jointly a saddle point for the global problem.

b) Maximizer separable case: In this case, the inner maximization problem is a function of x_a, i.e., $\hat{f}_i(x_{a,i}) = \max_{x_{b,i} \in \mathcal{X}_{b,i}} f_i(x_{a,i}, x_{b,i})$. If $\hat{f}_i$ can be derived, we get a minimization problem with a block-separable objective. However, the global minimization problem still contains constraint $x_b \in \mathcal{X}_b$. This optimization problem can be solved with decomposable optimization methods such as ADMM.

c) Inseparable case: If both $x_a \in \mathcal{X}_a$ and $x_b \in \mathcal{X}_b$ are present, we cannot use $\hat{f}_i(x_{a,i}) = \max_{x_{b,i} \in \mathcal{X}_{b,i}} f_i(x_{a,i}, x_{b,i})$ due to the globally bounding constraint $x_b \in \mathcal{X}_b$. We may attempt to derive $\hat{f}(x_a) = \max_{x_b \in \mathcal{X}_b \cap (\cap_{i=1}^N \mathcal{X}_{b,i})} f_i(x_{a,i}, x_{b,i})$. However, this process (potentially) removes the separability of the objective function. Hence, decomposable optimization methods are not directly applicable to this case. We are interested in separable solutions for this case by preserving the minimax formulation.

IV. Alternating Direction Method of Multipliers for Decomposable Optimization

The alternating direction method of multipliers (ADMM) [16], [17] is an optimization method to solve optimization problems with separable objectives and complicating constraints. Consider the problem

$$\begin{aligned} \min_{x_a} \quad & \sum_{i=1}^{N} g_i(x_{a,i}) && \text{(2a)} \\ \text{subject to} \quad & x_a \in \mathcal{X}_a && \text{(2b)} \\ & x_{a,i} \in \mathcal{X}_{a,i}, \quad \text{for all } i \in 1, \dots, N. && \text{(2c)} \end{aligned}$$

To solve this problem using ADMM, we use an auxiliary variable z_a and rewrite (2) as

$$\begin{aligned} \min_{x_a} \quad & \sum_{i=1}^{N} g_i(x_{a,i}) && \text{(3a)} \\ \text{subject to} \quad & z_a \in \mathcal{X}_a && \text{(3b)} \\ & z_{a,i} = x_{a,i}, \quad \text{for all } i \in 1, \dots, N && \text{(3c)} \\ & x_{a,i} \in \mathcal{X}_{a,i}, \quad \text{for all } i \in 1, \dots, N. && \text{(3d)} \end{aligned}$$

For (3), we define the Lagrangian $\mathcal{K}(x_a, z_a, \lambda_a)$ as

$$\sum_{i=1}^{N} \left(g_i(x_{a,i}) + I_{\mathcal{X}_{a,i}}(x_{a,i}) \right) + {\lambda_a}^\top (x_a - z_a) + I_{\mathcal{X}_a}(z_a)$$

and the augmented Lagrangian $\hat{\mathcal{K}}(x_a, z_a, \lambda_a)$ as

$$\mathcal{K}(x_a, z_a, \lambda_a) + \frac{\rho_a}{2} \|x_a - z_a\|_2^2$$

Algorithm 1: Alternating Direction Method of Multipliers (ADMM) for decomposable optimization

1 Initialize $x_a^0, z_a^0, \lambda_a^0$ such that $x_a^0 \in \mathcal{X}_{a,1} \times \ldots \times \mathcal{X}_{a,N}$, $z_a^0 \in \mathcal{X}_a$.
2 **for** $k = 0, 1, \ldots$ **do**
3 $\quad x_a^{k+1} = \arg\min_{x_a} \hat{\mathcal{K}}(x_a, z_a^k, \lambda_a^k)$.
4 $\quad z_a^{k+1} = \arg\min_{z_a} \hat{\mathcal{K}}(x_a^{k+1}, z_a, \lambda_a^k)$.
5 $\quad \lambda_a^{k+1} = \lambda_a^k + \rho_a(x_a^{k+1} - z_a^{k+1})$

where $\rho_a > 0$ is the penalty parameter.

ADMM for decomposable optimization, Algorithm 1, consists of three steps: primal variable x_a, auxiliary primal variable z_a, and dual variable λ_a updates. We note that Line 3 of Algorithm 1 is separable and is the same with assigning

$$\arg \min_{x_{a,i} \in \mathcal{X}_{a,i}} g_i(x_{a,i}) + (\lambda_{a,i}^k)^\top (x_{a,i} - z_{a,i}^k) + \frac{\rho_a}{2} \left\| x_{a,i} - z_{a,i}^k \right\|_2^2$$

to $x_{a,i}^{k+1}$ for every $i \in 1, \ldots N$. Line 4 is the convex projection step and is equal to letting

$$z_a^{k+1} = \arg \min_{z_a \in \mathcal{X}_a} (\lambda_a^k)^\top (x_a^{k+1} - z_a) + \frac{\rho_a}{2} \left\| x_a^{k+1} - z_a \right\|_2^2$$

which is equal to

$$z_a^{k+1} = \arg \min_{z_a \in \mathcal{X}_a} \left\| x_a^{k+1} + \lambda_a^k / \rho_a - z_a \right\|_2^2 .$$

Let x_a^* be an optimal solution of (2). The iterates of ADMM converge to an optimal solution [18], i.e., $z_a^k \to x_a^*$, if there exists $(x_a^*, z_a^*, \lambda_a^*)$ for all x_a, z_a, and λ_a such that

$$\mathcal{K}(x_a^*, z_a^*, \lambda_a) \leq \mathcal{K}(x_a^*, z_a^*, \lambda_a^*) \leq \mathcal{K}(x_a, z_a, \lambda_a^*). \tag{4}$$

V. Saddle-Point Alternating Direction Method of Multipliers

In this section, we describe the alternating direction method of multipliers (ADMM) for saddle-point problems that was first introduced in [7]. The method shares the same steps with standard ADMM and enjoys the same convergence guarantees.

To apply ADMM to saddle-point problem (1) we first rewrite the problem using the auxiliary variables $z_{a,i}$ and $z_{b,i}$:

$$\min_{x_a, z_a} \max_{x_b, z_b} \quad \sum_{i=1}^N f_i(x_{a,i}, x_{b,i}) \tag{5a}$$

$$\text{subject to} \quad z_a \in \mathcal{X}_a \tag{5b}$$

$$z_b \in \mathcal{X}_b \tag{5c}$$

$$z_{a,i} = x_{a,i}, \quad \text{for all } i \in 1, \ldots, N, \tag{5d}$$

$$z_{b,i} = x_{b,i}, \quad \text{for all } i \in 1, \ldots, N, \tag{5e}$$

$$x_{a,i} \in \mathcal{X}_{a,i}, \quad \text{for all } i \in 1, \ldots, N, \tag{5f}$$

$$x_{b,i} \in \mathcal{X}_{b,i}, \quad \text{for all } i \in 1, \ldots, N. \tag{5g}$$

Algorithm 2: Saddle-Point Alternating Direction Method of Multipliers (SP-ADMM) for decomposable optimization

1 Initialize $x_a^0, x_b^0, z_a^0, z_b^0, \lambda_a^0, \lambda_b^0$ such that $x_a^0 \in \mathcal{X}_{a,1} \times \ldots \times \mathcal{X}_{a,N}$, $x_b^0 \in \mathcal{X}_{b,1} \times \ldots \times \mathcal{X}_{b,N}$, $z_a^0 \in \mathcal{X}_a$, and $z_b^0 \in \mathcal{X}_b$.
2 **for** $k = 0, 1, \ldots$ **do**
3 $\quad x_a^{k+1}, x_b^{k+1} = \arg\min_{x_a} \max_{x_b} \hat{\mathcal{L}}(x_a, x_b, z_a^k, z_b^k, \lambda_a^k, \lambda_b^k)$
4 $\quad z_a^{k+1} = \arg\min_{z_a} \hat{\mathcal{L}}(x_a^{k+1}, x_b^{k+1}, z_a, z_b^k, \lambda_a^k, \lambda_b^k)$
5 $\quad z_b^{k+1} = \arg\max_{z_b} \hat{\mathcal{L}}(x_a^{k+1}, x_b^{k+1}, z_a^{k+1}, z_b, \lambda_a^k, \lambda_b^k)$
6 $\quad \lambda_a^{k+1} = \lambda_a^k + \rho_a(x_a^{k+1} - z_a^{k+1})$
$\quad\quad \lambda_b^{k+1} = \lambda_b^k + \rho_b(x_b^{k+1} - z_b^{k+1})$

For (5), we define the Lagrangian

$$\mathcal{L}(x_a, x_b, z_a, z_b, \lambda_a, \lambda_b) = \sum_{i=1}^N \left(f_i(x_{a,i}, x_{b,i}) + I_{\mathcal{X}_{a,i}}(x_{a,i}) - I_{\mathcal{X}_{b,i}}(x_{b,i}) \right) + {\lambda_a}^\top (x_a - z_a) + I_{\mathcal{X}_a}(z_a) - {\lambda_b}^\top (x_b - z_b) - I_{\mathcal{X}_b}(z_b)$$

and the augmented Lagrangian

$$\hat{\mathcal{L}}(x_a, x_b, z_a, z_b, \lambda_a, \lambda_b) = \mathcal{L}(x_a, x_b, z_a, z_b, \lambda_a, \lambda_b) + \frac{\rho_a}{2} \left\| x_a - z_a \right\|_2^2 - \frac{\rho_b}{2} \left\| x_b - z_b \right\|_2^2 .$$

where $\rho_a > 0$ is the penalty parameter for the minimizer, and $\rho_b > 0$ is the penalty parameter for the maximizer.

Saddle-point ADMM for decomposable optimization, Algorithm 2, also consists of three steps: primal variable x_a, x_b updates, auxiliary primal variable z_a, z_b updates, and dual variable λ_a, λ_b updates. Line 3 of Algorithm 2 is separable: This step assigns

$$\arg \min_{x_{a,i} \in \mathcal{X}_{a,i}} \max_{x_{b,i} \in \mathcal{X}_{b,i}} f_i(x_{a,i}, x_{b,i}) + (\lambda_{a,i}^k)^\top (x_{a,i} - z_{a,i}^k) + \frac{\rho_a}{2} \left\| x_{a,i} - z_{a,i}^k \right\|_2^2 - (\lambda_{b,i}^k)^\top (x_{b,i} - z_{b,i}^k) - \frac{\rho_b}{2} \left\| x_{b,i} - z_{b,i}^k \right\|_2^2$$

to $x_{a,i}^{k+1}$ and $x_{b,i}^{k+1}$ for every $i \in 1, \ldots N$. These sub-problems have a significantly lower number of dimensions compared to the original saddle-point problem (1) and can be solved in parallel. The sub-problems can be solved using existing saddle-point optimization methods and for some objective functions such as bilinear functions of two one-dimensional variables, they have analytical solutions. Lines 4–5 are the convex projection steps and are equal to letting

$$z_a^{k+1} = \arg \min_{z_a \in \mathcal{X}_a} \left\| x_a^{k+1} + \lambda_a^k / \rho_a - z_a \right\|_2^2$$

and

$$z_b^{k+1} = \arg \min_{z_b \in \mathcal{X}_b} \left\| x_b^{k+1} + \lambda_b^k / \rho_b - z_b \right\|_2^2 ,$$

which can be solved using convex optimization methods.

We show the convergence of SP-ADMM, under a similar assumption of standard ADMM. We assume that there exists a saddle-point where strong duality holds for the minimizer's problem when the maximizer is fixed, and vice versa.

Assumption 1. *There exists* $(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b^*)$ *such that*

$$\begin{aligned} &\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b^*) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b^*) \\ &\leq \mathcal{L}(x_a, x_b^*, z_a, z_b^*, \lambda_a^*, \lambda_b^*) \end{aligned} \tag{6}$$

and

$$\begin{aligned} &\mathcal{L}(x_a^*, x_b, z_a^*, z_b, \lambda_a^*, \lambda_b^*) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b^*) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b) \end{aligned} \tag{7}$$

for all $x_a, x_b, z_a, z_b, \lambda_a$, *and* λ_b.

Note that $x_a^* = z_a^*$ and $x_b^* = z_b^*$ for the saddle-point since $\sup_{\lambda_a} \lambda_a{}^\top (x_a^* - z_a^*) = \infty$ and $\inf_{\lambda_b} \lambda_b{}^\top (x_b^* - z_b^*) = -\infty$ otherwise. Also note that $x_a^* \in \mathcal{X}_a \cap (\mathcal{X}_{a,1} \times \ldots \times \mathcal{X}_{a,N})$ and $x_b^* \in \mathcal{X}_b \cap (\mathcal{X}_{b,1} \times \ldots \times \mathcal{X}_{b,N})$ due to the indicator functions.

Despite its complicated nature, the assumption is satisfied for the convex-concave saddle-point point problems where Slater's condition [9] is satisfied.

Proposition 1 (Sufficient condition for a saddle-point)**.** *There exists a saddle point* $(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b^*)$ *for* $\mathcal{L}$ *that satisfies Assumption 1 if*

1) *Every* f_i *is a convex function of* $x_{a,i}$ *and concave function of* $x_{b,i}$ *in* $\mathcal{X}_{a,i} \times \mathcal{X}_{b,i}$.
2) *Every* f_i *is continuous.*
3) $\mathcal{X}_a$, $\mathcal{X}_b$, *and every* $\mathcal{X}_{a,i}$, $\mathcal{X}_{b,i}$ *are compact, convex polytopes.*

We give the proof of Proposition 1 in Appendix A.

Note that the conditions given in Proposition 1 imply that the saddle-point problem satisfies Slater's condition for the minimizer and maximizer. In the proposition, we use polytope constraints for simplicity; the proposition can be improved to general convex sets $\mathcal{X}_{a,i}$, $\mathcal{X}_{b,i}$, $\mathcal{X}_a$, and $\mathcal{X}_b$ as long as there is a saddle point for (1) that satisfies Slater's condition.

Under Assumption 1, the iterates of SP-ADMM converges to a saddle point of (1). If every f_i is Lipschitz continuous, the proposition also implies the convergence of value.

Proposition 2. *Under Assumption 1, the iterates of SP-ADMM converge to a saddle point for* (1)*, i.e.,* $z_a^k \to x_a^*$ *and* $z_b^k \to x_b^*$ *where* (x_a^*, x_b^*) *is a saddle-point of* (1).

We give the proof of Proposition 2 in Appendix B.

Proposition 2 shows convergence in the limit. As in the standard ADMM [18], one can use the magnitude of primal and dual residuals as the stopping criterion in practice: Terminate when $\left\|x_a^k - z_a^k\right\|_2 + \left\|x_b^k - z_b^k\right\|_2 \leq \epsilon^{\text{primal}}$ and $\rho_a \left\|z_a^k - z_a^{k-1}\right\|_2 + \rho_b \left\|z_b^k - z_b^{k-1}\right\|_2 \leq \epsilon^{\text{dual}}$ where $x_a^k - z_a^k$ and $x_b^k - z_b^k$ are the primal residuals, and $\rho_a(z_a^k - z_a^{k-1})$ and $\rho_b(z_b^k - z_b^{k-1})$ are the dual residuals after iteration k.

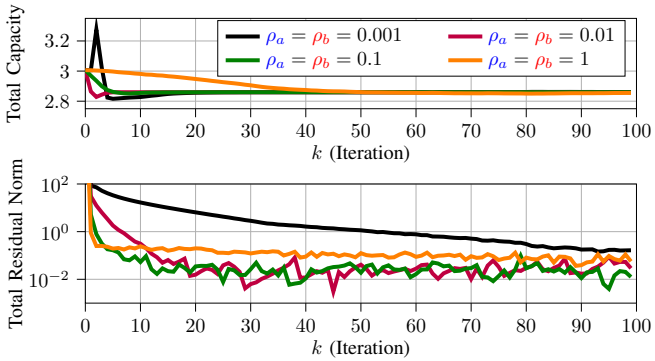

Fig. 1. (Top) Total capacity of the communication channels with (z_a^k, z_b^k). (Bottom) Total residual norm is $\left\|x_a^k - z_a^k\right\|_2 + \left\|x_b^k - z_b^k\right\|_2 + \rho_a \left\|z_a^k - z_a^{k-1}\right\|_2 + \rho_b \left\|z_b^k - z_b^{k-1}\right\|_2$.

VI. Numerical Examples

In this section, we give numerical examples for SP-ADMM and compare it with saddle-point Frank-Wolfe (SP-FW) method [13]. The implementations are given at https://github.com/mustafakarabag/SP-ADMM.

A. Power Allocation Game for Communication Channels

In this example from [6], we consider a power allocation problem in Gaussian communication channels. The total communication capacity is $\sum_{i=1}^N \log\left(1 + \frac{x_{b,i}}{\sigma_i + x_{a,i}}\right)$ where $x_{b,i}$ is the signal power allocated to the i^{th} channel, σ_i is the receiver noise for the i^{th} channel, and $x_{a,i}$ is the noise of the i^{th} channel.

We consider a game between a maximizer that allocates signal powers and a minimizer that adversarially chooses the noise levels for $N = 10$ channels. The global constraints are $\sum_{i=1}^N x_{b,i} = 20$ for the the maximizer and $\sum_{i=1}^N x_{a,i} = 10$ for the minimizer. Players have individual constraints $x_{a,i} \geq 0$ and $x_{b,i} \geq 0$. The receiver noise level is $\sigma = [2, 6, 5, 8, 3, 9, 5, 6, 7, 3]$. The equilibrium value of the problem instance is 2.860 [6].

For the implementation of SP-ADMM, we use SP-FW to solve the sub-saddle-point problems that are in the form of

$$\begin{aligned} &\min_{x_{a,i}} \max_{x_{b,i}} \sum_{i=1}^N \log\left(1 + \frac{x_{b,i}}{\sigma_i + x_{a,i}}\right) + \lambda_{a,i}^k (x_{a,i} - z_{a,i}^k) \\ &+ \rho_a (x_{a,i} - z_{a,i}^k)^2 - \lambda_{b,i}^k (x_{b,i} - z_{b,i}^k) - \rho_b (x_{b,i} - z_{b,i}^k)^2. \end{aligned}$$

We initialize x_a and x_b with a vector of zeros. The variables z_a and z_b are initialized with the projections of x_a and x_b onto their global constraints, respectively.

In Figure 1, we show the output of SP-ADMM for different penalty parameters. Similar to the standard ADMM, SP-ADMM generates acceptable solutions within a few iterations: The total capacity converges to the equilibrium value 2.860. The total residual norm decay as the number of iterations increase. However, similar to the standard ADMM, the rate of convergence is slow. We suspect that the fluctuations of the total residual norm is due to the dynamic competition between

the players and the fact that sub-problems are solved with a finite accuracy. When we compare the effects of the penalty parameters ρ_a and ρ_b, we observe that mild penalties such as 0.1 lead to both faster objective and residual convergences.

B. Network Routing Game with Adversarial Agents

In this example, we consider a network routing problem represented with a Markov decision process (MDP). The MDP is deterministic, i.e., it is a directed graph with N edges. Players choose a policy for this MDP that induces a Markov chain. The players' policies control the density of atomic agents that are transitioning in the Markov chain. The variables, x_a and x_b, of the players represent the stationary distributions induced by the players over the edges of the Markov chain. We generate the underlying directed graph of the MDP using a random Erdos-Renyi graph such that every node has 5 edges in expectation.

The network has a price function for every edge i that is equal to $x_{a,i} + x_{b,i}$, i.e., the total demand for edge i. The cost of an edge i, for the minimizer is $x_{a,i}(x_{a,i} + x_{b,i})$ that is the density of minimizer times the price of the edge. The minimizer's goal is to minimize the total cost $\sum_{i=1}^{N} x_{a,i}(x_{a,i}+ x_{b,i})$. The maximizer is an adversary trying to maximize the same cost. The minimizer and maximizer control a unit density each. The individual constraints are $0 \leq x_{a,i} \leq 1$ and $0 \leq x_{b,i} \leq 1$ for every edge i. The global contraints are enforced by the dynamics of the MDP: The players' stationary distributions have to be valid. In addition, the maximizer's density at state 1 has to be at least 0.1, i.e., $\sum_{i \in E} x_{a,i} \geq 0.1$ where E is the incoming edges of state 1.

We compare the performance of SP-ADMM with SP-FW for different sizes of MDPs. For the initialization of both SP-ADMM with SP-FW, we use the valid stationary distribution that is closest to the uniform distribution in L_2 distance. We solve the sub-saddle-point problems of SP-ADMM using an analytical solution exploiting the bilinear structure of sub-problems. This step has $\mathcal{O}(N)$ time complexity. The gradients for SP-FW are also computed using analytical solutions, which has $\mathcal{O}(N)$ time complexity. The projection step of SP-ADMM and the maximization step of SP-FW are both computed using ECOS solver [27] with CVXPY [28] interface. For SP-ADMM, we use $\rho_a = \rho_b = 1$, and for SP-FW, we use the step size $2/(2+k)$ at iteration k as suggested in [13].

For both algorithms, we compute a bound on the optimality gap in the following way. Let $z_a^{*,k}$ be the optimal response of the minimizer against the maximizer's z_b^k action, and $z_b^{*,k}$ be the optimal response of the maximizer against the minimizer's z_a^k action. We compute the best action of a player by solving a convex optimization problem where the other player's action is fixed. By the definition of a saddle-point, we have

$$\sum_{i=1}^{N} f_i(z_{a,i}^{*,k}, z_{b,i}^k) \leq \sum_{i=1}^{N} f_i(z_{a,i}^*, z_{b,i}^*) \leq \sum_{i=1}^{N} f_i(z_{a,i}^k, z_{b,i}^{*,k}).$$

The best lower bound is $l^k = \max_{1 \leq j \leq k} \sum_{i=1}^{N} f_i(z_{a,i}^{*,j}, z_{b,i}^j)$ and best upper bound is $u^k = \min_{1 \leq j \leq k} \sum_{i=1}^{N} f_i(z_{a,i}^j, z_{b,i}^{*,j})$ at iteration k. The optimality gap of an iterative algorithm at iteration k is bounded by $u^k - l^k$.

TABLE I
COMPARISON OF SP-ADMM AND SP-FW FOR DIFFERENT MDP SIZES

Network size		SP-ADMM		SP-FW	
# nodes	# edges N	Opt. gap $u^k - l^k$	Time (s)	Opt. gap $u^k - l^k$	Time (s)
10	49	1.36e-9[a]	5.48	1.36e-9[a]	5.17
20	93	2.49e-7	9.39	5.13e-3	9.09
50	282	1.87e-6	28.06	2.35e-3	25.67
100	494	1.35e-6	51.18	1.62e-3	48.17

[a] Both algorithms fail to improve on the initialization point due to numerical precision issues.

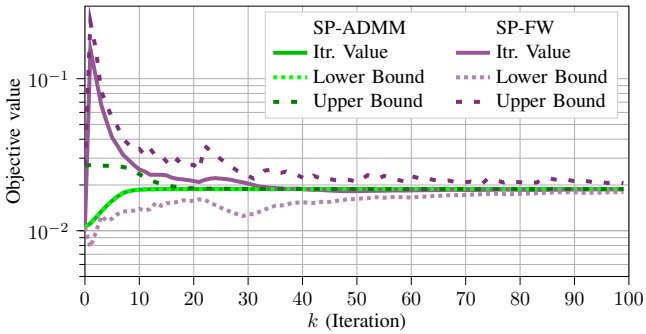

Fig. 2. The objective values for SP-ADMM and SP-FW. For each algorithm, 'Itr. Value' refers to the value with the variables from the current iterate, i.e., $(z_a^k, z_b^{k,*})$. 'Lower Bound' refers to the value with the maximizer's variable from the current iterate and the minimizer's best response to it, i.e., $(z_a^{k,*}, z_b^k)$. 'Upper Bound' refers to the value with the minimizer's variable from the current iterate and the maximizer's best response to it, i.e., $(z_a^k, z_b^{k,*})$.

We compare SP-ADMM and SP-FW in Table I and Figure 2. In Figure 2, we observe that SP-ADMM performs better than SP-FW for objective convergence. In addition, the upper and lower bounds are closer for SP-ADMM, which shows a better convergence to the saddle-point solution. In Table I, we observe that the solution time for SP-ADMM is slightly worse since we solve a quadratic program for SP-ADMM whereas we solve a linear program of the same size for SP-FW. On the other hand, the optimality gap $u^k - l^k$ is orders of magnitude better for SP-ADMM with similar solution times.

VII. Conclusion

We demonstrated saddle-point alternating direction method of multipliers (SP-ADMM) to solve decomposable saddle-point problems. We show that SP-ADMM has convergence guarantees under a saddle-point assumption. This assumption is satisfied for convex-concave problems that satisfy Slater's conditions. While we show that SP-ADMM converges asymptotically, we suspect that it also enjoys the non-asymptotic guarantees of standard ADMM [19], for example, in the strongly convex-strongly concave setting.

References

[1] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
[3] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.
[4] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *Journal of machine learning research*, vol. 10, no. 7, 2009.
[5] B. Taskar, S. Lacoste-Julien, M. I. Jordan, K. P. Bennett, and E. Parrado-Hernández, "Structured prediction, dual extragradient and bregman projections," *Journal of Machine Learning Research*, vol. 7, no. 7, 2006.
[6] A. Ghosh and S. Boyd, "Minimax and convex-concave games," *lecture notes for course EE392o:"Optimization Projects" Stanford Univ., Stanford, CA*, 2003.
[7] M. O. Karabag, M. Ornik, and U. Topcu, "Deception in supervisory control," *IEEE Transactions on Automatic Control*, vol. 67, no. 2, pp. 738–753, 2022.
[8] K. C. Border, *Fixed point theorems with applications to economics and game theory*. Cambridge university press, 1985.
[9] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[10] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matecon*, vol. 12, pp. 747–756, 1976.
[11] C. Daskalakis and I. Panageas, "The limit points of (optimistic) gradient descent in min-max optimization," *Advances in neural information processing systems*, vol. 31, 2018.
[12] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of optimization theory and applications*, vol. 142, no. 1, pp. 205–228, 2009.
[13] G. Gidel, T. Jebara, and S. Lacoste-Julien, "Frank-wolfe algorithms for saddle point problems," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 362–371.
[14] G. B. Dantzig and P. Wolfe, "Decomposition principle for linear programs," *Operations research*, vol. 8, no. 1, pp. 101–111, 1960.
[15] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numerische mathematik*, vol. 4, no. 1, pp. 238–252, 1962.
[16] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & mathematics with applications*, vol. 2, no. 1, pp. 17–40, 1976.
[17] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires," *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, vol. 9, no. R2, pp. 41–76, 1975.
[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.
[19] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan, "A general analysis of the convergence of admm," in *International Conference on Machine Learning*. PMLR, 2015, pp. 343–352.
[20] K. Guo, D. Han, D. Z. Wang, and T. Wu, "Convergence of admm for multi-block nonconvex separable optimization models," *Frontiers of Mathematics in China*, vol. 12, no. 5, pp. 1139–1162, 2017.
[21] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.
[22] W. Liu, A. Mokhtari, A. Ozdaglar, S. Pattathil, Z. Shen, and N. Zheng, "A decentralized proximal point-type method for saddle point problems," *arXiv preprint arXiv:1910.14380*, 2019.
[23] C. Hou, K. K. Thekumparampil, G. Fanti, and S. Oh, "Efficient algorithms for federated saddle point optimization," *arXiv preprint arXiv:2102.06333*, 2021.
[24] A. Rogozin, A. Beznosikov, D. Dvinskikh, D. Kovalev, P. Dvurechensky, and A. Gasnikov, "Decentralized distributed optimization for saddle point problems," *arXiv preprint arXiv:2102.07758*, 2021.
[25] P. Sharma, R. Panda, G. Joshi, and P. K. Varshney, "Federated minimax optimization: Improved convergence analyses and algorithms," *arXiv preprint arXiv:2203.04850*, 2022.
[26] D. Mateos-Núnez and J. Cortés, "Distributed subgradient methods for saddle-point problems," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5462–5467.
[27] A. Domahidi, E. Chu, and S. Boyd, "Ecos: An socp solver for embedded systems," in *2013 European Control Conference (ECC)*. IEEE, 2013, pp. 3071–3076.
[28] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.
[29] J. Burke, "Nonlinear optimization," *Lecture Notes*, 2020.

APPENDIX A
PROOF OF PROPOSITION 1

We show the existence of a saddle point for the augmented Lagrangian by considering the minimax theorem [8] and Slater's constraint qualification for convex duality [9]. Since $\sum_{i=1}^N f_i(x_{a,i}, x_{b,i})$ is a continous, convex-concave function and the feasible spaces are compact, convex for both minimizer and maximizer, there exists a saddle point (x_a^*, x_b^*) for (1) by the minimax theorem [8]. Consequently, $(x_a^*, x_b^*, z_a^*, z_b^*)$ is a saddle point of (5) where $z_a^* = x_a^*$ and $x_b^* = z_b^*$.

Since the feasible spaces are compact, convex polytopes,

- there exist $G_{a,i}$ and $h_{a,i}$ such that $G_{a,i}x_{a,i} + h_{a,i} \leq 0$ is equal to $x_{a,i} \in \mathcal{X}_{a,i}$,
- there exist $G_{b,i}$ and $h_{b,i}$ such that $G_{b,i}x_{b,i} + h_{b,i} \leq 0$ is equal to $x_{b,i} \in \mathcal{X}_{b,i}$,
- there exist G_a and h_a such that $G_a z_a + h_a \leq 0$ is equal to $z_a \in \mathcal{X}_a$, and
- there exist G_b and h_b such that $G_b z_b + h_b \leq 0$ is equal to $z_b \in \mathcal{X}_b$.

Define the Lagrangian for (5)

$$\begin{aligned}
&\bar{\mathcal{L}}(x_a, x_b, z_a, z_b, \lambda_a, \lambda_b, \mu_a, \mu_b, [\mu_{a,i}]_{i=1}^N, [\mu_{b,i}]_{i=1}^N) \\
&= \sum_{i=1}^N f_i(x_{a,i}, x_{b,i}) \\
&+ {\lambda_a}^\top (x_a - z_a) + \mu_a^\top (G_a z_a + h_a) + \sum_{i=1}^N \mu_{a,i}^\top (G_{a,i}x_{a,i} + h_{a,i}) \\
&- {\lambda_b}^\top (x_b - z_b) - \mu_b^\top (G_b z_b + h_b) - \sum_{i=1}^N \mu_{b,i}^\top (G_{b,i}x_{b,i} + h_{b,i})
\end{aligned}$$

where $\mu_a, \mu_b, \mu_{a,1}, \ldots, \mu_{a,N}, \mu_{b,1}, \ldots, \mu_{b,N} \geq 0$.

For fixed x_b^* and z_b^*, $\sum_{i=1}^N f_i(x_{a,i}, x_{b,i}^*)$ is a continuous, jointly convex function of x_a and λ_a and the constraints of (5) satisfies Slater's condition. Note that x_a^* and z_a^* is optimal for fixed x_b^* and z_b^*. By the saddle point theorem [29], there exists $(x_a^*, z_a^*, \lambda_a^*, \mu_a^*, \mu_{a,1}^*, \ldots, \mu_{a,N}^*)$ such that

$$\begin{aligned}
&\bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b, \mu_a, \mu_b, [\mu_{a,i}]_{i=1}^N, [\mu_{b,i}]_{i=1}^N) && \text{(8a)} \\
&\leq \bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b, \mu_a^*, \mu_b, [\mu_{a,i}^*]_{i=1}^N, [\mu_{b,i}]_{i=1}^N) && \text{(8b)} \\
&\leq \bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b, \mu_a^*, \mu_b, [\mu_{a,i}^*]_{i=1}^N, [\mu_{b,i}]_{i=1}^N) && \text{(8c)}
\end{aligned}$$

for any $\lambda_b, \mu_b, [\mu_{b,i}]_{i=1}^N$. Let $\lambda_b, \mu_b, [\mu_{b,i}]_{i=1}^N = 0$. Note that

$$\begin{aligned}
&\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b) \\
&\leq \bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b, \mu_a, 0, [\mu_{a,i}]_{i=1}^N, [0]_{i=1}^N)
\end{aligned}$$

since $G_{a,i}x_{a,i}^* + h_{a,i} \leq 0$, $G_{a,i}z_a^* + h_{a,i} \leq 0$, and $\mu_a, \mu_{a,1}, \ldots, \mu_{a,N} \geq 0$. We also have

$$\begin{aligned}
&\bar{\mathcal{L}}(x_a, x_b^*, z_a, z_b^*, \lambda_a^*, \lambda_b, \mu_a^*, 0, [\mu_{a,i}^*]_{i=1}^N, [0]_{i=1}^N) \\
&\leq \mathcal{L}(x_a, x_b^*, z_a, z_b^*, \lambda_a^*, \lambda_b)
\end{aligned}$$

since $I_{\mathcal{X}_{a,i}}(x_{a,i}) \geq \mu_{a,i}^\top(G_{a,i}x_{a,i}+h_{a,i})$, $I_{\mathcal{X}_a}(z_a) \geq \mu_a^\top(G_a x_a + h_a)$, $x_{b,i}^* \in \mathcal{X}_{b,i}$, and $z_b^* \in \mathcal{X}_b$.

We established

$$\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b) \leq \mathcal{L}(x_a, x_b^*, z_a, z_b^*, \lambda_a^*, \lambda_b).$$

We now show

$$\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b) \leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b).$$

Note that the optimization problem

$$\min_{\lambda_a, \mu_a, [\mu_{a,i}]_{i=1}^N} \bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b, \mu_a, \mu_b, [\mu_{a,i}]_{i=1}^N, [\mu_{b,i}]_{i=1}^N)$$

is separable: the optimal values of λ_a and $\mu_a, \mu_{a,1}, \ldots, \mu_{a,N}$ can be computed independently. Consequently, since ${\lambda_a}^*$ is a maximizer for

$$\bar{\mathcal{L}}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b, \mu_a, \mu_b, [\mu_{a,i}]_{i=1}^N, [\mu_{b,i}]_{i=1}^N),$$

it is also a maximizer for $\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b)$, and we have $\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b) \leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b)$. Combining these results, we get

$$\begin{aligned} &\mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b) \\ &\leq \mathcal{L}(x_a, x_b^*, z_a, z_b^*, \lambda_a^*, \lambda_b) \end{aligned} \tag{9}$$

for arbitrary λ_b. By symmetry, we can repeat the same arguments and get

$$\begin{aligned} &\mathcal{L}(x_a^*, x_b, z_a^*, z_b, \lambda_a, \lambda_b^*) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b^*) \\ &\leq \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a, \lambda_b) \end{aligned} \tag{10}$$

for arbitrary λ_a. Finally, by letting $\lambda_b = \lambda_b^*$ in (9) and $\lambda_a = \lambda_a^*$ in (10), we get the desired result.

Appendix B
Proof of Proposition 2

The proof follows the same steps of the proof for convergence for the standard ADMM algorithm [18]. The work [18] proves convergence of standard ADMM by considering only the properties of minimizer updates. To prove the convergence of SP-ADMM, we consider the properties of both minimizer and maximizer updates.

We define the value function of the algorithm

$$V^k = \frac{\left\|\lambda_a^k - \lambda_a^*\right\|_2^2}{\rho_a} + \frac{\left\|\lambda_b^k - \lambda_b^*\right\|_2^2}{\rho_b} + \frac{\left\|z_a^k - z_a^*\right\|_2^2}{1/\rho_a} + \frac{\left\|z_b^k - z_b^*\right\|_2^2}{1/\rho_b}.$$

We will show that the value decreases at every step, i.e.,

$$\begin{aligned} V^{k+1} \leq & V^k - \rho_a \left\|r_a^{k+1}\right\|_2^2 - \rho_b \left\|r_b^{k+1}\right\|_2^2 \\ & - \rho_a \left\|z_a^{k+1} - z_a^k\right\|_2^2 - \rho_b \left\|z_b^{k+1} - z_b^k\right\|_2^2. \end{aligned} \tag{11}$$

where $r_a^k = x_a^k - z_a^k$ is the primal residual for the minimizer and $r_b^k = x_b^k - z_b^k$ is the primal residual for the maximizer. By telescoping sum over k, we get

$$V^0 \geq \sum_{i=1}^{\infty} \rho_a \left\|r_a^k\right\|_2^2 - \rho_b \left\|r_b^k\right\|_2^2 - \rho_a \left\|z_a^k - z_a^*\right\|_2^2 - \rho_b \left\|z_b^k - z_b^*\right\|_2^2.$$

Since V^0 is finite, and ρ_a and ρ_b are strictly positive, we must have $\lim_{k\to\infty} \left\|r_a^k\right\|_2^2 = 0$, $\lim_{k\to\infty} \left\|r_b^k\right\|_2^2 = 0$, $\lim_{k\to\infty} \left\|z_a^k - z_a^*\right\|_2^2 = 0$, and $\lim_{k\to\infty} \left\|z_b^k - z_b^*\right\|_2^2 = 0$. Consequently, $x_a^k \to x_a^*$, $z_a^k \to x_a^*$, $x_b^k \to x_b^*$, and $z_b^k \to x_b^*$,

For ease of notation, we also define the following quantities:

- Equilibrium value $p^* = \sum_{i=1}^N f_i(x_{a,i}^*, x_{b,i}^*)$. Note that $p^* = \mathcal{L}(x_a^*, x_b^*, z_a^*, z_b^*, \lambda_a^*, \lambda_b^*)$ since $x_a^* = z_a^*$, $x_b^* = z_b^*$
- $p^k = \sum_{i=1}^N f_i(x_{a,i}^k, x_{b,i}^k)$, $(p_b^*)^k = \sum_{i=1}^N f_i(x_{a,i}^k, x_{b,i}^*)$, $(p_a^*)^k = \sum_{i=1}^N f_i(x_{a,i}^*, x_{b,i}^k)$.

To prove (11), we will show

$$p^* - (p_b^*)^{k+1} \leq (\lambda_a^*)^\top r_a^{k+1}, \tag{12}$$

$$(p_a^*)^{k+1} - p^* \leq (\lambda_b^*)^\top r_b^{k+1}, \tag{13}$$

$$\begin{aligned} p^{k+1} - (p_a^*)^{k+1} \leq & \rho_a (z_a^{k+1} - z_a^k)^\top (r_a^{k+1} + z_a^{k+1} - z_a^*) \\ & - (\lambda_a^{k+1})^\top r_a^{k+1}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} (p_b^*)^{k+1} - p^{k+1} \leq & \rho_b (z_b^{k+1} - z_b^k)^\top (r_b^{k+1} + z_b^{k+1} - z_b^*) \\ & - (\lambda_b^{k+1})^\top r_b^{k+1}. \end{aligned} \tag{15}$$

We, for now, assume that these inequalities hold.

A. Proof of (11)

Adding (12), (13), (14), and (15), and multiplying by 2, we get

$$\begin{aligned} 0 \leq & 2(\lambda_a^* - \lambda_a^{k+1})^\top r_a^{k+1} + 2(\lambda_b^* - \lambda_b^{k+1})^\top r_b^{k+1} \\ & + 2\rho_a (z_a^{k+1} - z_a^k)^\top (r_a^{k+1} + z_a^{k+1} - z_a^*) \\ & + 2\rho_b (z_b^{k+1} - z_b^k)^\top (r_b^{k+1} + z_b^{k+1} - z_b^*). \end{aligned} \tag{16}$$

We use the definitions to rewrite (16).

Using $\lambda_a^{k+1} = \lambda_a^k + \rho_a r_a^{k+1}$, $r_a^{k+1} = (\lambda_a^{k+1} - \lambda_a^k)/\rho_a$, $\lambda_a^{k+1} - \lambda_a^k = \lambda_a^{k+1} - \lambda_a^* + \lambda_a^* - \lambda_a^k$, we get

$$\begin{aligned} & 2(\lambda_a^* - \lambda_a^{k+1})^\top r_a^{k+1} = 2(\lambda_a^* - \lambda_a^k)^\top r_a^{k+1} - 2\rho_a \left\|r_a^{k+1}\right\|_2^2 \\ = & \frac{2}{\rho_a}(\lambda_a^* - \lambda_a^k)^\top (\lambda_a^{k+1} - \lambda_a^*) - \frac{1}{\rho_a}\left\|\lambda_a^{k+1} - \lambda_a^k\right\|_2^2 - \rho_a \left\|r_a^{k+1}\right\|_2^2 \\ = & \frac{1}{\rho_a}\left\|\lambda_a^k - \lambda_a^*\right\|_2^2 - \frac{1}{\rho_a}\left\|\lambda_a^{k+1} - \lambda_a^*\right\|_2^2 - \rho_a \left\|r_a^{k+1}\right\|_2^2. \end{aligned} \tag{17}$$

By the symmetry of the definitions, we also get

$$2(\lambda_a^* - \lambda_a^{k+1})^\top r_a^{k+1} \tag{18}$$

$$= \frac{1}{\rho_a}\left\|\lambda_a^k - \lambda_a^*\right\|_2^2 - \frac{1}{\rho_a}\left\|\lambda_a^{k+1} - \lambda_a^*\right\|_2^2 - \rho_a \left\|r_a^{k+1}\right\|_2^2 \tag{19}$$

Using $z_a^{k+1} - z_a^* = z_a^{k+1} - z_a^k + z_a^k - z_a^*$ and $z_a^{k+1} - z_a^k = z_a^{k+1} - z_a^* - z_a^k + z_a^*$, we get

$$\begin{aligned} & 2\rho_a (z_a^{k+1} - z_a^k)^\top (r_a^{k+1} + z_a^{k+1} - z_a^*) - \left\|r_a^{k+1}\right\|_2^2 \\ = & -\rho_a \left\|r_a^{k+1} + z_a^{k+1} - z_a^k\right\|_2^2 \\ & - \rho_a (\left\|z_a^{k+1} - z_a^*\right\|_2^2 - \left\|z_a^k - z_a^*\right\|_2^2) \end{aligned} \tag{20}$$

By the symmetry of the definitions, we also get

$$\begin{aligned}&2\rho_b(z_b^{k+1}-z_b^k)^\top(r_b^{k+1}+z_b^{k+1}-z_b^*)-\left\|r_b^{k+1}\right\|_2^2\\&=-\rho_b\left\|r_b^{k+1}+z_b^{k+1}-z_b^k\right\|_2^2\\&\quad-\rho_b(\left\|z_b^{k+1}-z_b^*\right\|_2^2-\left\|z_b^k-z_b^*\right\|_2^2).\end{aligned}\tag{21}$$

By substituting (17), (19), (20), and (21) in (16), we get

$$\begin{aligned}V^{k+1}&\le V^k-\rho_a\left\|r_a^{k+1}+z_a^{k+1}-z_a^k\right\|_2^2\\&\quad-\rho_b\left\|r_b^{k+1}+z_b^{k+1}-z_b^k\right\|_2^2\\&\le V^k-\rho_a\left\|r_a^{k+1}\right\|_2^2-\rho_b\left\|r_b^{k+1}\right\|_2^2\\&\quad-\rho_a\left\|z_a^{k+1}-z_a^k\right\|_2^2-\rho_b\left\|z_b^{k+1}-z_b^k\right\|_2^2\\&\quad-2\rho_a(\lambda_a^{k+1})^\top(z_a^{k+1}-z_a^k)-2\rho_b(\lambda_b^{k+1})^\top(z_b^{k+1}-z_b^k)\end{aligned}\tag{22}\tag{23}$$

As shown in the proofs of (14) and (15), z_a^{k+1} minimizes $-(\lambda_a^{k+1})^\top z_a$ in $\mathcal{X}_a$, and z_b^{k+1} maximizes $(\lambda_b^{k+1})^\top z_b$ in $\mathcal{X}_b$. Consequently, we have $-2\rho_a(\lambda_a^{k+1})^\top z_a^{k+1}\le -2\rho_a(\lambda_a^{k+1})^\top z_a^k$, and $2\rho_b(\lambda_b^{k+1})^\top z_b^{k+1}\le -2\rho_b(\lambda_b^{k+1})^\top z_b^k$. Combining these with (23), we get (11).

B. Proofs of (12) and (13)

Due to the saddle point assumption, we have

$$\mathcal{L}(x_a^*,x_b^*,z_a^*,z_b^*,\lambda_a^*,\lambda_b^*)\le\mathcal{L}(x_a^{k+1},x_b^*,z_a^{k+1},z_b^*,\lambda_a^*,\lambda_b^*).$$

Since $p^*=\mathcal{L}(x_a^*,x_b^*,z_a^*,z_b^*,\lambda_a^*,\lambda_b^*)$ and $x_b^*=z_b^*$, we have

$$p^*\le(p_b^*)^{k+1}+(\lambda_a^*)^\top(x_a^{k+1}-z_a^{k+1}).$$

Using $r_a^{k+1}=x_a^{k+1}-z_a^{k+1}$ and rearranging the terms, we get

$$p^*-(p_b^*)^{k+1}\le(\lambda_a^*)^\top r_a^{k+1}.\tag{24}$$

The proof of (13) has the same steps with the proof of (12).

C. Proofs of (14) and (15)

We note that $\hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k)$ is a convex function of x_a and a concave function of x_b, and (x_a^{k+1},x_b^{k+1}) is a solution to

$$\min_{x_a\in\mathcal{X}_{a,1}\times\ldots\times\mathcal{X}_{a,N}}\ \max_{x_b\in\mathcal{X}_{b,1}\times\ldots\times\mathcal{X}_{b,N}}\ \hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k).$$

Define

$$\begin{aligned}g(x_a,x_b)=&\sum_{i=1}^N f_i(x_{a,i},x_{b,i})+(\lambda_a^k-\rho_a(z_a^{k+1}-z_a^k))^\top x_a\\&-(\lambda_b^k-\rho_b(z_b^{k+1}-z_b^k))^\top x_b\end{aligned}$$

Using $\lambda_a^{k+1}\in\mathcal{X}_{a,1}\times\ldots\times\mathcal{X}_{a,N}$ and $\lambda_a^{k+1}=\lambda_a^k+\rho_a(x_a^{k+1}-z_a^{k+1})$, we get

$$\left.\frac{\partial\hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k)}{\partial x_a}\right|_{x_a=x_a^{k+1}}=\left.\frac{\partial g(x_a,x_b)}{\partial x_a}\right|_{x_a=x_a^{k+1}}$$

Similarly, we get

$$\left.\frac{\partial\hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k)}{\partial x_b}\right|_{x_b=x_b^{k+1}}=\left.\frac{\partial g(x_a,x_b)}{\partial x_b}\right|_{x_b=x_b^{k+1}}.$$

Since $\hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k)$ and $g(x_a,x_b)$ share the same gradient field for x_a and x_b, and (x_a^{k+1},x_b^{k+1}) is a saddle point of $\hat{\mathcal{L}}(x_a,x_b,z_a^k,z_b^k,\lambda_a^k,\lambda_b^k)$, (x_a^{k+1},x_b^{k+1}) is also a saddle point of $g(x_a,x_b)$. Using the saddle point property we have,

$$\begin{aligned}&\sum_{i=1}^N f_i(x_{a,i}^{k+1},x_{b,i}^{k+1})+(\lambda_a^{k+1}-\rho_a(z_a^{k+1}-z_a^k))^\top x_a^{k+1}\\&\quad-(\lambda_b^{k+1}-\rho_b(z_b^{k+1}-z_b^k))^\top x_b^{k+1}\\&\le\sum_{i=1}^N f_i(x_{a,i}^*,x_{b,i}^{k+1})+(\lambda_a^{k+1}-\rho_a(z_a^{k+1}-z_a^k))^\top x_a^*\\&\quad-(\lambda_b^{k+1}-\rho_b(z_b^{k+1}-z_b^k))^\top x_b^{k+1}\end{aligned}$$

By definitions of $(p_a^*)^{k+1}$ and p^{k+1}, we get

$$\begin{aligned}p^{k+1}+&(\lambda_a^{k+1}-\rho_a(z_a^{k+1}-z_a^k))^\top x_a^{k+1}\\&\le(p_a^*)^{k+1}+(\lambda_a^{k+1}-\rho_a(z_a^{k+1}-z_a^k))^\top x_a^*.\end{aligned}\tag{25}$$

By the saddle point property, we also get

$$\begin{aligned}p^{k+1}-&(\lambda_b^{k+1}-\rho_b(z_b^{k+1}-z_b^k))^\top x_b^{k+1}\\&\ge(p_b^*)^{k+1}-(\lambda_b^{k+1}-\rho_b(z_b^{k+1}-z_b^k))^\top x_b^*.\end{aligned}\tag{26}$$

Define $h_a(z_a)=-(\lambda_a^{k+1})^\top z_a$. and $h_b(z_b)=(\lambda_b^{k+1})^\top z_b$. We have

$$\left.\frac{\partial\hat{\mathcal{L}}(x_a{}^{k+1},x_b{}^{k+1},z_a,z_b^k,\lambda_a^k,\lambda_b^k)}{\partial z_a}\right|_{z_a=z_a^{k+1}}=\left.\frac{\partial h_a(z_a)}{\partial x_a}\right|_{x_a=z_a^{k+1}}$$

and similarly

$$\left.\frac{\partial\hat{\mathcal{L}}(x_a{}^{k+1},x_b{}^{k+1},z_a^{k+1},z_b,\lambda_a^k,\lambda_b^k)}{\partial z_b}\right|_{z_b=z_b^{k+1}}=\left.\frac{\partial h_b(z_b)}{\partial x_b}\right|_{x_b=z_b^{k+1}}.$$

Since $\hat{\mathcal{L}}(x_a{}^{k+1},x_b{}^{k+1},z_a,z_b^k,\lambda_a^k,\lambda_b^k)$ and $h_a(z_a)$ has the same gradient field in $\mathcal{X}_a$, z_a^{k+1} is also a minimizer of $h_a(z_a)$ in $\mathcal{X}_a$. Similarly, z_b^{k+1} is also a maximizer of $h_b(z_b)$ in $\mathcal{X}_b$. Due to these we have

$$-(\lambda_a^{k+1})^\top z_a^{k+1}\le-(\lambda_a^{k+1})^\top z_a^*\tag{27}$$

and

$$(\lambda_b^{k+1})^\top z_b^{k+1}\ge(\lambda_b^{k+1})^\top z_b^*.\tag{28}$$

By combining (25) and (27), and noting that $x_a^*=z_a^*$ and $r_a^{k+1}=x_a^{k+1}-z_a^{k+1}$, we get

$$\begin{aligned}p^{k+1}-(p_a^*)^{k+1}\le&\rho_a(z_a^{k+1}-z_a^k)^\top(r_a^{k+1}+z_a^{k+1}-z_a^*)\\&-(\lambda_a^{k+1})^\top r_a^{k+1}\end{aligned}\tag{29}$$

Similarly, by combining (26) and (28), and noting that $x_b^*=z_b^*$ and $r_b^{k+1}=x_b^{k+1}-z_b^{k+1}$, we get

$$\begin{aligned}(p_b^*)^{k+1}-p^{k+1}\le&\rho_b(z_b^{k+1}-z_b^k)^\top(r_b^{k+1}+z_b^{k+1}-z_b^*)\\&-(\lambda_b^{k+1})^\top r_b^{k+1}.\end{aligned}\tag{30}$$