Accelerated Methods for Solving Nonlinear Generalized Nash Equilibrium Problems with Convergence Guarantees

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

August 30, 2024

Abstract

The nonlinear *generalized Nash equilibrium problem* (GNEP) extends the classical Nash equilibrium by incorporating strategy-dependent constraints, making it applicable in fields like machine learning, economics, and engineering. However, solving nonlinear GNEPs efficiently remains challenging due to the complex interdependencies among players' strategies. In this paper, we introduce two first-order algorithms—*Accelerated Mirror-Prox Quadratic Penalty* (AMP-QP) and *Accelerated Mirror-Prox Augmented Lagrangian* (AMP-AL)—designed to solve nonlinear GNEPs characterized by monotonicity. These methods achieve optimal global convergence rates by integrating the accelerated mirror-prox scheme with quadratic penalty and augmented Lagrangian techniques. We provide theoretical guarantees for global convergence and demonstrate the practical efficiency of our algorithms through numerical experiments, showing their effectiveness in large-scale nonlinear GNEPs relevant to machine learning and game theory.

Keywords: Generalized Nash Equilibrium Problem (GNEP); First-Order Algorithms; Global Convergence Rates; Monotone and Strongly Monotone Problems; Accelerated Mirror-Prox (AMP) Algorithm; Quadratic Penalty Method

1 Introduction

The *Nash equilibrium problem* (NEP), first introduced by John Nash in the 1950s, has become a cornerstone of game theory, influencing fields ranging from economics to computer science. A Nash equilibrium represents a stable state in which no player can unilaterally improve their payoff by changing their strategy, given the strategies of the other players. Traditionally, the NEP assumes that each player's strategy set is independent of the others, leading to a well-defined solution concept that has been extensively studied and applied.

However, real-world scenarios often involve more complex interactions where a player's feasible strategies depend not only on their own choices but also on the strategies selected by others. This interdependence is captured by the *generalized Nash equilibrium problem* (GNEP), where the strategy set of each player is influenced by the decisions of all players. The GNEP thus extends the classical NEP by incorporating shared constraints, leading to a more realistic and flexible modeling framework. Applications of GNEPs have emerged in various domains, including power allocation in telecommunication networks, environmental regulation, and adversarial machine learning, underscoring their practical significance.

Despite their theoretical appeal and practical relevance, GNEPs pose significant computational challenges. The coupling of strategy sets introduces nontrivial dependencies that complicate the design and analysis of algorithms. While various methods have been proposed to solve GNEPs, including relaxation, Newton-type methods, and penalty approaches, there remains a critical need for algorithms that not only guarantee convergence but also offer scalable performance suitable for large-scale problems.

In this paper, we address these challenges by developing first-order algorithms tailored to nonlinear GNEPs. Specifically, we focus on GNEPs characterized by monotonicity—a property that ensures the feasibility of deriving global convergence guarantees. Building on the *accelerated mirror-prox* (AMP) method, we propose two novel algorithmic frameworks: the *Accelerated Mirror-Prox Quadratic Penalty* (AMP-QP) method and the *Accelerated Mirror-Prox Augmented Lagrangian* (AMP-AL) method. These methods combine classical optimization techniques with modern advancements in variational inequality theory to achieve optimal convergence rates.

Backgrounds. The NEP [Nash(1950), Nash(1951)] is a central topic in mathematics, economics and computer science. NEP problems have begun to play an important role in machine learning as researchers begin to focus on decisions, incentives and the dynamics of multi-agent learning. In a classical NEP, the payoff to each player depends upon the strategies chosen by all, but the domains from which the strategies are to be chosen for each player are independent of the strategies chosen by other players. The goal is to arrive at a joint optimal outcome where no player can do better by deviating unilaterally [Osborne and Rubinstein(1994), Myerson(2013)].

The GNEP is a natural generalization of an NEP where the choice of an action by one agent affects both the payoff and the *domain* of actions of all other agents [Arrow and Debreu(1954)]. Its introduction in the 1950's provided the foundation for a rigorous theory of economic equilibrium [Debreu(1952), Arrow and Debreu(1954), Debreu(1959)]. More recently, the GNEP problem has emerged as a powerful paradigm in a range of engineering applications involving noncooperative games. In particular, in the survey of [Facchinei and Kanzow(2010a)], three general classes of problems were developed in detail: the abstract model of general equilibrium, power allocation in a telecommunication system, and environmental pollution control. Further applications of the GNEP problem in recent years have included adversarial classification [Brückner and Scheffer(2009), Brückner et al.(2012) wireless communication and networks [Pang et al.(2008), Pang et al.(2010), Han et al.(2012), Scutari et al.(2014)], power grids [Jing-Yuan and Smeers(1999), Hobbs and Pang(2007)], cloud computing [Ardagna et al.(2011), Ardagna et al.(2015)], modern traffic systems with e-hailing services [Ban et al.(2019)], supply and demand constraints for transportation systems [Stein and Sudermann-Merx(2018)] and pollution quotas for environmental applications [Krawczyk(2005), Breton et al.(2006)]. For an overview of GNEP theory and applications, we refer to more detailed surveys [Cominetti et al.(2012), Facchinei and Pang(2007) Facchinei and Kanzow(2010a), Fischer et al.(2014)] and the references therein.

It is of significant interest to bring the GNEP framework and its decision-making applications into contact with machine learning. By analogy with the bridge between machine learning and smooth nonlinear optimization that has been so productive in the recent years, the major challenges that arise for GNEP formulations of machine-learning problems are computational, including the development of scalable gradient-based algorithms. The first attempts to design such algorithms are due to [Robinson(1993a), Robinson(1993b)]. Further progress has been made over the ensuing three decades by considering generic algorithms based on either the Nikaido-Isoda function [Uryas' ev and Rubinstein(1994), Krawczyk and Uryasev(2000), Von Heusinger and Kanzow(2009), Facchinei et al.(2009), Dreves et al.(2011), von Heusinger et al.(2012), Dreves et al.(2013), Izmailov and Solodov(2 Fischer et al.(2016)] or penalty functions [Pang and Fukushima(2005), Facchinei and Lampariello(2011), Fukushima(2011), Facchinei and Kanzow(2010b), Kanzow(2016), Kanzow and Steck(2016), Kanzow and Steck(201 Ba and Pang(2022)]. In terms of theoretical guarantees, global convergence and local convergence rates have been established for some of these algorithms under suitable assumptions. For an overview of recent progress on penalty-type algorithms, we refer to [Ba and Pang(2022)].

Despite this progress on the algorithmic aspects of GNEPs, significant computational challenges remain and they hinder the the practical impact of the GNEP concept in machine learning. Most notably, global convergence rate characterizations are not yet available for any of the algorithms that target the solution of nonlinear GNEPs. Thus, we summarize the focus of our work: ***Can we design efficient algorithms for solving GNEPs that have global convergence rate guarantees? what rates can we obtain and in what class of GNEPs?***

Brief contributions. Our contributions are manifold. First, we rigorously analyze the global convergence properties of the proposed algorithms under monotone and strongly monotone conditions, filling a gap in the existing literature. Second, we implement these algorithms and conduct extensive numerical experiments to validate their effectiveness in solving GNEPs of varying complexity. Our results demonstrate that the AMP-QP and AMP-AL methods not only provide strong theoretical guarantees but also exhibit robust performance in practice, making them well-suited for real-world applications where GNEPs play a critical role.

1.1 Related work

To appreciate the broad scope of our agenda, we start by reviewing the algorithms for computing the solution of GNEPs. [Rosen(1965)] studied the jointly convex GNEP and proposed an algorithm based on gradient projection. Subsequently, other algorithms were developed, including the relaxation method [Uryas' ev and Rubinstein(1994), Krawczyk and Uryasev(2000), Von Heusinger and Kanzow(2009)], Newton-type methods [Facchinei et al.(2009), von Heusinger et al.(2012), Dreves et al.(2013), Izmailov and Solodo Fischer et al.(2016)] and the interior-point potential reduction method [Dreves et al.(2011)]. Most of these approaches are based on the Nikaido-Isoda function and thus are very expensive computationally. By exploiting the special structure of certain GNEPs, some specific algorithms were developed with improved theoretical guarantees; these include the VI approaches of [Facchinei et al.(2007)] and [Nabetani et al.(2011)], which are restricted to jointly convex GNEPs, and Lemke's method from [Schiro et al.(2013)], which is specifically designed to solve a class of linear GNEPs.

More recently, progress has been made on developing penalty-type algorithms for GNEPs. Representatives of this class of algorithms are exact and inexact penalty methods [Facchinei and Lampariello(2011), Fukushima(2011), Facchinei and Kanzow(2010b), Kanzow and Steck(2018), Ba and Pang(2022)] and exact and inexact augmented Lagrangian method [Pang and Fukushima(2005), Kanzow(2016), Kanzow and Steck(2016), Kanzow and Steck(2018)]. The exact penalty method was introduced in [Fukushima(2011)], and its variants have been proposed where either all of the constraints are penalized [Facchinei and Kanzow(2010b)] or some of the constraints are penalized [Facchinei and Lampariello(2011)]. These algorithms achieve exactness results under suitable assumptions but suffer from the nonsmoothness of penalized subproblems, thereby leading to numerical difficulties from a practical point of view. [Pang and Fukushima(2005)] proposed the augmented Lagrangian method to solve quasi-variational inequalities (QVIs) and [Kanzow(2016)] improved this scheme with a global convergence guarantee. They discussed the GNEP within their framework by treating it as a special QVI. Later on, [Kanzow and Steck(2016), Kanzow and Steck(2018)] applied a similar idea directly to GNEPs and proved theoretical results which are stronger than those that arise from the QVI framework. Even though an asymptotic global convergence has been established under suitable assumptions, the nonasymptotic global convergence rate (aka, the iteration complexity bound) of these algorithms is open. For a brief overview of algorithmic results, we refer to the survey of [Fischer et al.(2014)].

Another line of relevant work is concerned with optimality conditions and constraint qualifications (CQs). In nonlinear optimization, the Karush-Kuhn-Tucker (KKT) condition provides a general characterization of local optimality under various CQs [Kuhn and Tucker(1951), Karush(1939)]. Both the KKT condition and CQs have been extended to GNEPs. In particular, [Pang and Fukushima(2005)] considered the generalization of Mangasarian-Fromovitz condition [Mangasarian and Fromovitz(1967)] for QVIs; see also [Kanzow(2016)]. Subsequently, [Facchinei and Kanzow(2010b)] specialized this notion to GNEPs and designed exact penalty methods with asymptotic global convergence guarantees. [Kanzow and Steck(2016)] and [Bueno et al.(2019)] studied the augmented Lagrangian methods for GNEPs and established their asymptotic global convergence guarantee under CQs, including the constant positive linear dependence property [Qi and Wei(2000), Andreani et al.(2005)] and the cone continuity property [Andreani et al.(2016)]. Despite the significance of these results for theory, their practical impact has been limited. This is because all of these CQs are defined at or around a solution of the GNEP and it is hard to verify them for a nonlinear GNEP where the candidate solution set is generally unavailable.

1.2 Contribution

We tackle the problem of designing efficient first-order algorithms for a class of nonlinear GNEPs and proving optimal global convergence rates in various settings. More specifically, we start by defining a class of *monotone and strongly monotone* GNEPs which cover a range of machine-learning applications. Leveraging the recent progress on the iteration complexity analysis of first-order algorithms for nonlinear optimization and *variational inequality* (VI) problems [Lan and Monteiro(2013), Lan and Monteiro(20 Chen et al.(2017), Xu(2021)], we develop first-order algorithms for computing the solutions of GNEPs with global convergence rate estimates. The algorithms that we study incorporate the *accelerated mirror-prox* (AMP) scheme into a *quadratic penalty method* (QPM) or an *augmented Lagrangian method* (ALM). At a high level, the following informal theorems summarize the main results of our paper:

Informal Theorem 1.1 (Theorem 4)**.** *The required number of gradient evaluations for accelerated mirror-prox quadratic penalty (AMP-QP) method to reach an ϵ-solution is bounded by $\widetilde{O}(\epsilon^{-1})$ and $\widetilde{O}(\epsilon^{-1/2})$ in monotone and strongly monotone nonlinear GNEPs.*

Informal Theorem 1.2 (Theorem 5)**.** *The required number of gradient evaluations for the accelerated mirror-prox augmented Lagrangian (AMP-AL) method to reach an ϵ-solution is bounded by $\widetilde{O}(\epsilon^{-1})$ and $\widetilde{O}(\epsilon^{-1/2})$ in monotone and strongly monotone nonlinear GNEPs.*

Although the algorithmic frameworks based on QPM and ALM are classical for GNEPs, we remark that combining these frameworks with AMP is new in the literature. For inexact ALM, the convergence guarantees are asymptotic for GNEPs while the rate results are only derived for an optimization setting when combined with Nesterov's accelerated gradient (NAG) method. These analyses cannot be directly extended to nonlinear GNEPs since NAG does not apply to the VI subproblem.

Building on background in linearly constrained optimization [Lan and Monteiro(2013), Lan and Monteiro(2016) our results are new in the GNEP setting. In particular, we highlight our analysis of the AMP algorithm for strongly monotone VIs and the demonstration of its applicability for nonlinear GNEPs, which is established via the derivation of an optimal convergence rate in monotone settings. We also prove the equivalence between KKT points and solutions for nonlinear GNEPs. Noting that [Bueno et al.(2019)] have presented a counterexample for general GNEPs, our result is obtained by identifying and exploiting a special structure of nonlinear GNEPs. This equivalence forms the basis of our development of first-order algorithms for solving nonlinear GNEPs.

We make some further comments. Note that there are many instances of GNEPs in specialized problems where global convergence rate guarantees have been established [Nesterov and Scrimali(2011), Nabetani et al.(2011), Bianchi et al.(2022), Franci and Grammatico(2022)]. These GNEPs are either monotone or satisfy an error bound condition that is analogous to strong monotonicity. This suggests that the notion of monotonicity or strongly monotonicity is key to global convergence rate estimates for the algorithms. Further, the local convergence rates were derived under local strong monotonicity or more general local error bounds [Facchinei et al.(2009), Facchinei et al.(2015)]. There are also several natural applications for monotone GNEPs, motivating the investigation of *simple*, *optimal* and *implementable* first-order algorithms for monotone GNEPs. In particular, the real-world GNEPs application problems arising from machine learning, e.g., team games [Celli and Gatti(2018), Celli et al.(2019), Farina et al.(2021), Carminati et al.(2022), Kalogiannis et al.(2022)], are extremely large and make significant demands with respect to computational feasibility. This necessitates the investigation of first-order algorithms for GNEPs. In summary, although the notion of global monotonicity or strong monotonicity rules out some interesting application problems, it encompasses a rather rich class of GNEPs and leads to global convergence rate guarantees.

Organization. The remainder of this paper is organized as follows. In Section 2, we define the problem setting, review relevant background on *generalized Nash equilibrium problems* (GNEPs) and monotonicity, and introduce the basic setup for nonlinear GNEPs, including motivating examples and an ϵ-solution concept. In Section 3, we provide new results for the *accelerated mirror-prox* (AMP) algorithm in the strongly monotone case, complementing the current iteration complexity analysis. In Section 4, we propose the AMP-QP and AMP-AL algorithms, detailing their design and theoretical analysis, and establish bounds on their iteration complexity for solving monotone and strongly monotone nonlinear GNEPs. Finally, in Section 5, we conclude the paper and discuss potential avenues for future research.

Notation. We let $[N] = \{1, 2, \ldots, N\}$ and let $\mathbb{R}^n_+$ denote the set of all vectors in $\mathbb{R}^n$ with nonnegative entries. For a vector $x \in \mathbb{R}^n$ and $p \in [1, +\infty]$, we denote $\|x\|_p$ as its ℓ_p-norm and $\|x\|$ as its ℓ_2-norm. For a matrix $A \in \mathbb{R}^{n \times n}$, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as largest and smallest eigenvalues and $\|A\|$ as the spectral norm. For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, we let the subdifferential of f at x be $\partial f(x)$. If f is differentiable, we have $\partial f(x) = \{\nabla f(x)\}$ where $\nabla f(x)$ denotes the gradient of f at x and $\nabla_\nu f$ denotes the partial gradient of f with respect to ν-th block of x. Given $n \geq 1$ and desired tolerance $\epsilon \in (0, 1)$, the notation $a = O(b(n, \epsilon))$ stands for $a \leq C \cdot b(n, \epsilon)$ where C is independent of n and $1/\epsilon$, and $a = \widetilde{O}(b(n, \epsilon))$ indicates the same inequality where C depends on log factors of n and $1/\epsilon$.

2 Preliminaries

We start by providing the definitions for monotone generalized Nash equilibrium problems. Moreover, we define an ϵ-solution concept based on a variational characterization of GNEPs and review the accelerated mirror-prox algorithm and its iteration complexity.

2.1 Problem setup

We consider nonlinear generalized Nash equilibrium problems (GNEPs) with a finite set of players $\nu \in \mathcal{N} = [N]$. The ν-th player selects a strategy x^ν from a subset $X_\nu(\cdot)$ of a finite-dimensional vector space $\mathbb{R}^{n_\nu}$ and the incurred cost is determined by a function $\theta_\nu(\cdot)$. The profile $x = (x^1, x^2, \ldots, x^N)$ denotes all players' actions and $x^{-\nu}$ denotes the joint action of all players but player ν. We rewrite the overall joint action as $x = (x^\nu; x^{-\nu}) \in \mathbb{R}^n$ where $n = n_1 + n_2 + \ldots + n_N$.

Definition 1. *We define a class of nonlinear GNEPs by a tuple $G = (\mathcal{N}, \{X_\nu(\cdot)\}_{\nu\in\mathcal{N}}, \{\theta_\nu(\cdot)\}_{\nu\in\mathcal{N}})$, where $X_\nu(\cdot) : \mathbb{R}^{n-n_\nu} \rightrightarrows \mathbb{R}^{n_\nu}$ is a point-to-set mapping representing the strategy set of player ν and $\theta_\nu : \mathbb{R}^n \mapsto \mathbb{R}$ is the ν-th player's cost function. The following conditions are satisfied:*

(i) $\mathcal{N} = \cup_{s=1}^S \mathcal{N}_s$, where $\mathcal{N}_i$ and $\mathcal{N}_j$ are not necessarily disjoint for $i \neq j$. Each of $\mathcal{N}_s$ is associated with $(A_s, b_s) \in \mathbb{R}^{m_s \times (\sum_{i\in\mathcal{N}_s} n_i)} \times \mathbb{R}^{m_s}$ and $(E_s, d_s) \in \mathbb{R}^{e_s \times (\sum_{i\in\mathcal{N}_s} n_i)} \times \mathbb{R}^{e_s}$.

(ii) $X_\nu(\cdot)$ consists of a simple,[1] convex and compact set $\widehat{X}_\nu$ intersected with the constraints that correspond to $\mathcal{I}_\nu = \{s : \nu \in \mathcal{N}_s\}$. Indeed, we have $X_\nu(x^{-\nu}) = \{x^\nu \in \widehat{X}_\nu : A_s x^{\mathcal{N}_s} \leq b_s,\ E_s x^{\mathcal{N}_s} = d_s,\ \forall s \in \mathcal{I}_\nu\}$ where $x^{\mathcal{N}}$ is the concatenation of all x^i for $i \in \mathcal{N}$.

(iii) $\theta_\nu(\cdot)$ is continuously differentiable and $\nabla_\nu\theta_\nu(\cdot)$ is ℓ_θ-Lipschitz, i.e., we have $\|\nabla_\nu\theta_\nu(x) - \nabla_\nu\theta_\nu(x')\| \leq \ell_\theta\|x - x'\|$ for all $x, x' \in \widehat{X}$.[2]

Definition 2. *The solution set of an nonlinear GNEP contains strategy profiles that discourage unilateral deviations. Formally, $\overline{x} \in \mathbb{R}^n$ is a solution if the following statement holds:*

$$\theta_\nu(\overline{x}^\nu, \overline{x}^{-\nu}) \leq \theta_\nu(x^\nu, \overline{x}^{-\nu}), \quad \text{for all } x^\nu \in X_\nu(\overline{x}^{-\nu}) \text{ and all } \nu \in \mathcal{N}$$

The nonlinear GNEP is a generalization of Nash equilibrium problems (NEPs) and Definition 2 extends the notion of Nash equilibrium [Rosen(1965)] to the setting where the strategy set of each player depends on the strategies of the rival players through equality and inequality constraints. If the nonlinear GNEP is convex, we have a variational characterization for nonlinear GNEPs. Indeed, if $\theta_\nu(\cdot, x^{-\nu})$ is convex for any fixed $x^{-\nu}$, an equilibrium can be computed by solving a quasi-variational inequality (QVI) [Chan and Pang(1982), Harker(1991)]. More specifically, letting $v(x) = (v_1(x), \ldots, v_N(x))$ be the profile of all players' individual cost gradients, $v_\nu(x) = \nabla_\nu\theta_\nu(x)$, for all $\nu \in \mathcal{N}$, we have the following proposition.

Proposition 1. *If the nonlinear GNEP is convex, $\overline{x} \in \mathbb{R}^n$ is a solution if and only if $\overline{x}^\nu \in X_\nu(\overline{x}^{-\nu})$ and the following QVI holds: $(x - \overline{x})^\top v(\overline{x}) \geq 0$ for all $x \in X_\nu(\overline{x})$.*

Proof. If the QVI holds: $(x - \overline{x})^\top v(\overline{x}) \geq 0$ for all $x \in X_\nu(\overline{x})$, we obtain that $\overline{x}^\nu \in X_\nu(\overline{x}^{-\nu})$ and $(x^\nu - \overline{x}^\nu)^\top v_\nu(\overline{x}) \geq 0$ for all $x^\nu \in X_\nu(\overline{x}^{-\nu})$ by setting $x^{-\nu} = \overline{x}^{-\nu}$. Since the nonlinear GNEP is convex, we have $\theta_\nu(\overline{x}^\nu, \overline{x}^{-\nu}) \leq \theta_\nu(x^\nu, \overline{x}^{-\nu})$ for all $x^\nu \in X_\nu(\overline{x}^{-\nu})$. Thus, $\overline{x}$ is a solution.

Conversely, if $\overline{x}$ is a solution of the nonlinear GNEP, the convexity of the nonlinear GNEP guarantees that $\overline{x}^\nu \in X_\nu(\overline{x}^{-\nu})$ and $(x^\nu - \overline{x}^\nu)^\top v_\nu(\overline{x}) \geq 0$ for all $x^\nu \in X_\nu(\overline{x}^{-\nu})$. Summing these inequalities over $\nu \in \mathcal{N}$, we obtain the desired QVI. □

Proposition 1 is a generalization of standard results for NEPs; see, e.g., Proposition 2.1 [Mertikopoulos and Zhou and such a characterization ensures existence results; see [Chan and Pang(1982)], Theorem 5.2

[1] "Simple" means that the projection operator admits a closed-form solution or can be computed efficiently.

[2] Throughout this paper, we define the base strategy set $\widehat{X}$ by $\widehat{X} = \Pi_{\nu=1}^N \widehat{X}_\nu$.

or [Harker(1991)], Theorem 2. Uniqueness results, however, are another matter. While the *diagonally strict concavity* (DSC) condition is sufficient for uniqueness of solutions in NEPs [Rosen(1965)], in the GNEP setting it does not even guarantee a connected solution set; see Figure 1 in [Harker(1991)] and the comments after his Corollary 3.1. The condition can be augmented to ensure uniqueness but only under additional conditions that are mainly of theoretical interest; see [Harker(1991), Eq. (33)]. In this paper, we do not view uniqueness as a reasonable goal but instead argue that the DSC condition itself—which is also referred to as *strict monotonicity* in convex analysis [Bauschke and Combettes(2017)]—is strong enough for global convergence rate estimation for the algorithms. Strict monotonicity and the related notions of *monotonicity* and *strong monotonicity* are useful assumptions for analyzing models and algorithms in the variational inequality literature [Facchinei and Pang(2007)].

Definition 3. *An nonlinear GNEP (i.e., $G = (\mathcal{N}, \{X_\nu(\cdot)\}_{\nu=1}^N, \{\theta_\nu(\cdot)\}_{\nu=1}^N)$) is said to be*

(i) monotone *if $(x - x')^\top(v(x) - v(x')) \geq 0$ for all $x, x' \in \widehat{X}$.*

(ii) α-strongly monotone *if $(x - x')^\top(v(x) - v(x')) \geq \alpha\|x - x'\|^2$ for all $x, x' \in \widehat{X}$.*

Remark. Proposition 1 can be directly applied to these monotone nonlinear GNEPs. Indeed, $(x - x')^\top(v(x) - v(x')) \geq 0$ for all $x, x' \in \widehat{X}$ guarantees that $(x^\nu - (x')^\nu)^\top(v_\nu(x^\nu, x^{-\nu}) - v_\nu((x')^\nu, x^{-\nu})) \geq 0$ for all $x^\nu, (x')^\nu \in \widehat{X}_\nu$ and all $\nu \in \mathcal{N}$. Since $v_\nu(x) = \nabla_\nu \theta_\nu(x)$, we have $\theta_\nu(\cdot, x^{-\nu})$ is convex over $\widehat{X}_\nu$ for all $\nu \in \mathcal{N}$. Moreover, the equilibrium existence is guaranteed here if $\widehat{X}_\nu$ is convex and compact for all $\nu \in \mathcal{N}$ (Theorem 2 in [Harker(1991)]). See [Chan and Pang(1982), Harker(1991), Pang and Fukushima(2005)] for further results that flow from the monotonicity assumption.

Strongly monotone nonlinear GNEPs are a subclass of monotone nonlinear GNEPs that admit a unique solution if $X_\nu(x^{-\nu})$ does not depend on the rival's choices $x^{-\nu}$. This feature is important from an algorithmic viewpoint since a natural quantity for measuring the iteration complexity is the distance between the iterates generated by an algorithm and a unique solution. However, such solution uniqueness is not guaranteed in simplest two-player strongly monotone GNEPs [Harker(1991)], demonstrating that, even with the stringent condition of strong monotonicity, it has been unclear which quantity is suitable for measuring the iteration complexity of the algorithms.

Remark. There have been many algorithms for solving the GNEPs, including relaxation methods [Uryas' ev and Rubinstein(1994), Krawczyk and Uryasev(2000), Von Heusinger and Kanzow(2009)], penalty methods and augmented Lagrangian methods [Pang and Fukushima(2005), Facchinei and Lampariello(201 Fukushima(2011), Facchinei and Kanzow(2010b), Kanzow(2016), Kanzow and Steck(2016), Kanzow and Steck(201 Newton-type methods [Facchinei et al.(2009), von Heusinger et al.(2012), Dreves et al.(2013), Izmailov and Solodo Fischer et al.(2016)], the interior-point potential reduction method [Dreves et al.(2011)], Lemke's method [Schiro et al.(2013)], and the trust-region method [Galli et al.(2018)]. Since nonlinear GNEPs are a special class of GNEPs, some of these algorithms can be directly applied to solve monotone nonlinear GNEPs and provide a global convergence guarantee. On the other hand, nonasymptotic convergence rates (or iteration complexity bounds) are unknown for these algorithms to the best of our knowledge.

2.2 Examples of monotone GNEPs

We provide a few examples of monotone GNEPs to give a sense of their expressivity. Two examples are linear [Stein and Sudermann-Merx(2016), Dreves and Sudermann-Merx(2016), Dreves(2017), Stein and Sudermann-Merx(2018)] and the other two are nonlinear.

Example 1 (Basic economic market model)**.** Consider a set of firms indexed by $\nu \in \mathcal{N}$ that offer the same product in a common market. Each firm acts as a price taker and sells the quantity $x_k^\nu \geq 0$ in the price category p_k^ν, where $k \in \mathcal{K}$ for an index set $\mathcal{K}$ and where the prices are given. We impose the allocation constraint $\sum_{k\in\mathcal{K}} x_k^\nu \leq C^\nu$ for $C^\nu > 0$ and also a public constraint, $\sum_{\nu\in\mathcal{N}} x_k^\nu \leq D_k$, which assures that the total offering does not exceed the total demand within each price category. Then, the cost function of firm ν is given by

$$\theta_\nu(x^\nu, x^{-\nu}) = c_\nu \sum_{k\in\mathcal{K}} x_k^\nu - \sum_{k\in\mathcal{K}} p_k^\nu x_k^\nu$$

where $c_\nu \geq 0$ is the marginal cost of firm ν for producing one unit of product.

Example 2 (Extended optimal transportation)**.** Consider a set of competing shippers indexed by $\nu \in \mathcal{N}$ who hope to transport the quantity x_{rt}^ν from manufacturers $r \in \mathcal{R}$ to customers indexed by $t \in \mathcal{T}$. Manufacturer r has a production capacity, $C_r \geq 0$, and customer t needs at least $D_t \geq 0$ units of this good with $\sum_{r\in\mathcal{R}} C_r = \sum_{t\in\mathcal{T}} D_t$. This implies the supply constraint $\sum_{\nu\in\mathcal{N}} \sum_{t\in\mathcal{T}} x_{rt}^\nu = C_r$ and the demand constraint $\sum_{\nu\in\mathcal{N}} \sum_{r\in\mathcal{R}} x_{rt}^\nu = D_t$. The cost function of firm ν is then given by

$$\theta_\nu(x^\nu, x^{-\nu}) = \sum_{r\in\mathcal{R}} \sum_{t\in\mathcal{T}} c_{rt}^\nu x_{rt}^\nu$$

where $c_{rt}^\nu \geq 0$ is the transportation cost from manufacturer r to consumer t by shipper ν.

Example 3 (Cournot competition with capacity constraint)**.** Consider a set of firms indexed by $\nu \in \mathcal{N}$, each supplying the market with a quantity $x^\nu \in [0, C_\nu]$ up to the firm's capacity $C_\nu \geq 0$. Suppose that we have $\mathcal{N} = \cup_{s=1}^S \mathcal{N}_s$ and assume *capacity constraints* $C_s \geq 0$ such that $\sum_{\nu\in\mathcal{N}_s} x^\nu \leq C_s$. By the law of supply and demand, the good is priced as a decreasing function of the total amount $\overline{x} = \sum_{\nu\in\mathcal{N}} x^\nu$ where the common choice is the linear form of $a - b\overline{x}$ for some $a, b > 0$. Then, the cost function of firm ν is given by

$$\theta_\nu(x^\nu, x^{-\nu}) = c_\nu(x^\nu) - x^\nu(a - b\overline{x})$$

which captures the cost of producing x^ν units of the good (the function $c_\nu(\cdot)$ represents a marginal cost function of firm ν and is assumed to be convex) minus the total revenue of such production.

Example 4 (Resource allocation auctions with bid capacity)**.** Suppose that there is a service provider with resources $s \in \mathcal{S}$. These resources are leased to a set of bidders indexed by $\nu \in \mathcal{N}$ who place monetary bids x_s^ν for the utilization of each resource $s \in \mathcal{S}$ up to each player's budget: $\sum_{s\in\mathcal{S}} x_s^\nu \leq b^\nu$. There is a *bid capacity* $C_s \geq 0$ for resource $s \in \mathcal{S}$ such that $\sum_{\nu\in\mathcal{N}} x_s^\nu \leq C_s$. Once all bids are tendered, the unit of resource s allocated to bidder ν is

$$\rho_s^\nu = \frac{q_s x_s^\nu}{d_s + \sum_{\nu\in\mathcal{N}} x_s^\nu}$$

where q_s denotes the total amount of resource s and $d_s > 0$ is the "entry barrier" for bidding on it). Then, the cost function of bidder ν is given by

$$\theta_\nu(x^\nu, x^{-\nu}) = \sum_{s\in\mathcal{S}} (x_s^\nu - c_\nu \rho_{\mathbf{s}}^\nu)$$

where $c_\nu \geq 0$ is the marginal gain to bidder ν from acquiring a unit of resources.

In addition to the above examples, the class of monotone nonlinear GNEPs contains all general convex-concave zero-sum games with linear constraints, monotone NEPs, and all convex potential games with private convex constraints (there exists a convex function $f : \widehat{X} \mapsto \mathbb{R}$ such that $v_\nu(x) = \nabla_\nu f(x)$ for all $\nu \in \mathcal{N}$). They are also a natural generalization of convex programming problems with linear constraints which constitute the backbone of nonlinear optimization [Ben-Tal and Nemirovski(2001)]. As such, the notion of (strong) monotonicity, which will play a crucial role in the analysis of this paper, is not limiting in practice but encompasses a wide range of application problems arising from economics and online decision making [Facchinei and Pang(2007)].

2.3 Solution concept

Following up Definition 2 and Proposition 1, we define the notion of ϵ-solution of monotone nonlinear GNEPs. In particular, if $\overline{x}$ is a solution of nonlinear GNEPs, then the feasibility condition at $\overline{x}$ holds true:

$$\overline{x} \in \widehat{X} \cap \left\{x \in \mathbb{R}^n : A_s x^{\mathcal{N}_s} \leq b_s, \ E_s x^{\mathcal{N}_s} = d_s, \text{ for all } s \in [S]\right\} \tag{1}$$

and $(\overline{x} - x)^\top v(\overline{x}) \leq 0$ for all $x \in \widehat{X}$ satisfying the following condition:

$$\text{for all } \nu \in \mathcal{N} \ \begin{cases} A_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} A_s^j \overline{x}^j - b_s \leq 0, & \forall s \in \mathcal{I}_\nu \\ E_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} E_s^j \overline{x}^j - d_s = 0, & \forall s \in \mathcal{I}_\nu \end{cases} \tag{2}$$

Inspired by the gap function and the weak solution concepts that are commonly used for iteration complexity analysis of algorithms in the VI literature [Facchinei and Pang(2007)], we define an ϵ-solution of monotone nonlinear GNEPs as follows.

Definition 4 (ϵ-solution concept)**.** *We say a point $\overline{x} \in \widehat{X}$ is an ϵ-solution of monotone nonlinear GNEPs if the following ϵ-feasibility condition at $\overline{x}$ holds true:*

$$\|\max\{0, A_s \overline{x}^{\mathcal{N}_s} - b_s\}\| \leq \epsilon, \quad \|E_s \overline{x}^{\mathcal{N}_s} - d_s\| \leq \epsilon, \quad \text{for all } s \in [S]$$

and an ϵ-quasi-variational inequality at $\overline{x}$ holds true: for all $\nu \in \mathcal{N}$, we have

$$(\overline{x}^\nu - x^\nu)^\top v_\nu(x^\nu, \overline{x}^{-\nu}) = \widetilde{O}(\epsilon)$$

for all $x^\nu \in \widehat{X}_\nu$ satisfying the following condition:

$$\|\max\{0, A_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} A_s^j \overline{x}^j - b_s\}\| \leq \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} E_s^j \overline{x}^j - d_s\| \leq \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu$$

This definition reduces to standard ϵ-optimality notions in special settings. For example, monotone nonlinear GNEPs reduce to monotone VIs when $A_s = 0$, $E_s = 0$, $b_s = 0$ and $d_s = 0$ for all $s \in [S]$. If $\overline{x}$ is a ϵ-weak solution of the VI such that $(\widetilde{x} - x)^\top v(x) \leq \epsilon$ for all $x \in \widehat{X}$, it is an ϵ-solution. Indeed, letting $x^{-\nu} = \overline{x}^{-\nu}$ for each $\nu \in \mathcal{N}$ yields the desired result. Thus, our notion in Definition 4 is a generalization of the ϵ-weak solution concept, which has been adopted for measuring the iteration complexity in the VI setting [Nemirovski(2004), Nesterov(2007), Malitsky(2015), Kotsalis et al.(2022)]. Moreover, monotone nonlinear GNEPs reduce to linearly constrained convex problems for the case of $N = 1$, where the similar notions have been adopted for linearly constrained nonsmooth nonconvex problems [Jiang et al.(2019)] as well as nonlinearly constrained nonsmooth convex problems [Rockafellar(1976), Yu and Neely(2017), Xu(2021)].

Remark. We show that an ϵ-solution is a solution of monotone nonlinear GNEPs when $\epsilon = 0$. Indeed, we have $\|\max\{0, A_s\overline{x}^{\mathcal{N}_s} - b_s\}\| \leq 0$ and $\|E_s\overline{x}^{\mathcal{N}_s} - d_s\| \leq 0$ for all $s \in [S]$ and $\overline{x} \in \widehat{X}$. Putting these pieces together yields Eq. (1). Furthermore, we have $(\overline{x} - x)^\top v(x) \leq 0$ for all $x \in \widehat{X}$ satisfying the following condition:

$$\text{for all } \nu \in \mathcal{N} \left\{ \begin{array}{l} A_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} A_s^j \overline{x}^j - b_s \leq 0, \quad \forall s \in \mathcal{I}_\nu \\ E_s^\nu x^\nu + \sum_{j \in \mathcal{N}_s, j \neq \nu} E_s^j \overline{x}^j - d_s = 0, \quad \forall s \in \mathcal{I}_\nu \end{array} \right. \tag{3}$$

Now it suffices to show that $(\overline{x} - x)^\top v(\overline{x}) \leq 0$ for all $x \in \widehat{X}$ satisfying the conditions in Eq. (3). Indeed, we fix $x \in \widehat{X}_\nu$ such that Eq. (3) holds at x and let $x(t) = tx + (1-t)\overline{x}$ be a function of $t \in [0, 1]$. Since Eq. (1) holds and $\widehat{X}$ is convex, we have $x(t) \in \widehat{X}$ and $x(t)$ satisfies Eq. (3) for any $t \in [0, 1]$. Therefore, we have

$$(\overline{x} - x)^\top v(tx + (1-t)\overline{x}) = \frac{1}{t}(\overline{x} - x(t))^\top v(x(t)) \leq 0$$

which implies that (by letting $t \to 0$) $(\overline{x} - x)^\top v(\overline{x}) \leq 0$. Since $x \in \widehat{X}$ is chosen as any point satisfying Eq. (3), we obtain the claimed result.

An alternative way to characterize the solution concept in the context of GNEPs is based on the introduction of Lagrangian multipliers and Karush-Kuhn-Tucker (KKT) conditions tailored to GNEPs [Pang and Fukushima(2005), Facchinei and Kanzow(2010b), Kanzow(2016), Bueno et al.(2019)]. In particular, we say a point $\overline{x} \in \widehat{X}$ is a KKT point of nonlinear GNEPs when each block $\overline{x}^\nu$ is a KKT point for minimizing the function $\theta_\nu(\cdot, \overline{x}^{-\nu})$ over the set $X_\nu(\overline{x}^{-\nu})$ for all $\nu \in \mathcal{N}$; that is, the following feasibility condition at $\overline{x}$ holds true:

$$\overline{x} \in \widehat{X} \cap \left\{x \in \mathbb{R}^n : A_s x^{\mathcal{N}_s} \leq b_s, \ E_s x^{\mathcal{N}_s} = d_s, \text{ for all } s \in [S]\right\}$$

and there exist Lagrangian multipliers $\overline{\lambda}^{\nu,s} \geq 0$ and $\overline{\mu}^{\nu,s}$ such that

$$\text{for all } \nu \in \mathcal{N} \left\{ \begin{array}{r} (\overline{x}^\nu - x^\nu)^\top (v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu}((A_s^\nu)^\top \overline{\lambda}^{\nu,s} + (E_s^\nu)^\top \overline{\mu}^{\nu,s})) \leq 0, \quad \forall x^\nu \in \widehat{X}_\nu \\ (\overline{\lambda}^{\nu,s})^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0, \quad \forall s \in \mathcal{I}_\nu \end{array} \right.$$

It is natural to investigate the relationship between solutions and KKT points in the context of GNEPs. In general, constraint qualifications (CQs) are necessary for ensuring that a solution is a KKT point [Bueno et al.(2019)]. However, monotone nonlinear GNEPs have special structure in the sense that $\theta_\nu(\cdot, x^{-\nu})$ is convex for any fixed $x^{-\nu}$, $\widehat{X}$ is simple and all the coupled constraints are linear. The features guarantee the equivalence between the solution concept in Definition 2 and the above KKT conditions tailored to nonlinear GNEPs. A proof based on Farkas's lemma and Proposition 1 is presented in Section 2.4.

Theorem 1. *If the nonlinear GNEP is monotone, $\overline{x} \in \widehat{X}$ is a solution (cf. Definition 2) if and only if it is a KKT point of this nonlinear GNEP.*

For completeness, we also present the notion of an ϵ-KKT point.

Definition 5 (ϵ-KKT)**.** *We say a point $\overline{x} \in \widehat{X}$ is an ϵ-*KKT point *for a monotone nonlinear GNEP if the following ϵ-feasibility condition at $\overline{x}$ holds true:*

$$\|\max\{0, A_s\overline{x}^{\mathcal{N}_s} - b_s\}\| \leq \epsilon, \quad \|E_s\overline{x}^{\mathcal{N}_s} - d_s\| \leq \epsilon, \quad \text{for all } s \in [S]$$

and there exist some Lagrangian multipliers $\overline{\lambda}^{\nu,s} \geq 0$ and $\overline{\mu}^{\nu,s}$ such that

$$\text{for all } \nu \in \mathcal{N} \left\{ \begin{array}{r} (\overline{x}^\nu - x^\nu)^\top (v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu}((A_s^\nu)^\top \overline{\lambda}^{\nu,s} + (E_s^\nu)^\top \overline{\mu}^{\nu,s})) \leq \epsilon, \quad \forall x^\nu \in \widehat{X}_\nu \\ \|\min\{\overline{\lambda}^{\nu,s}, -(A_s \overline{x}^{\mathcal{N}_s} - b_s)\}\| \leq \epsilon, \quad \forall s \in \mathcal{I}_\nu \end{array} \right.$$

Remark. The notion of ϵ-KKT point has been introduced in Definition 5.2 in [Bueno et al.(2019)] for more general GNEPs and Definition 5 is that notion applied to nonlinear GNEPs. For the special case of monotone VIs with $\overline{\lambda}^{\nu,s} = 0$ and $\overline{\mu}^{\nu,s} = 0$, Definition 5 corresponds to the ϵ-strong solution concept which is different from the ϵ-weak solution concept from an algorithmic point of view (see [Diakonikolas(2020)] for recent progress). Can we design some variants of the algorithms in this paper to pursue an ϵ-KKT point as in Definition 5 and prove their global convergence rate? This is an interesting open question.

2.4 Proof of Theorem 1

We provide the complete proof of Theorem 1. We first present a simple lemma and prove the theorem by appealing to Farkas' lemma [Farkas(1902)].

Lemma 1. *Let $A \in \mathbb{R}^{m \times n}$, $E \in \mathbb{R}^{e \times n}$ and $b \in \mathbb{R}^n$. Exactly one of the following systems has a solution:*

- *$Ax \leq 0$, $Ex = 0$ and $b^\top x > 0$;*
- *$A^\top y + E^\top z - b = 0$ and $y \geq 0$.*

Proof. Recall that Farkas' lemma states that exactly one of the following systems has a solution:

- $\overline{A}x \geq 0$ and $\overline{b}^\top x < 0$;
- $\overline{A}^\top y = b$ and $y \geq 0$.

We thus rewrite two linear systems from Lemma 1 as follows,

- $\begin{bmatrix} -A & E & -E \end{bmatrix} x \geq 0$ and $(-b)^\top x < 0$.
- $(-A)^\top y + E^\top \max\{0, -z\} + (-E)^\top \max\{0, z\} = -b$ and $y \geq 0$.

Applying Farkas' lemma with $\overline{A} = \begin{bmatrix} -A & E & -E \end{bmatrix}$ and $\overline{b} = -b$ yields the desired result. □

Necessity. By Proposition 1, a point $\overline{x} \in \widehat{X}$ being a solution of a monotone nonlinear GNEP implies that $\overline{x} \in \widehat{X}$ satisfies that $\overline{x}^\nu \in X_\nu(\overline{x}^{-\nu})$ and $(x - \overline{x})^\top v(\overline{x}) \geq 0$ for all $x \in X_\nu(\overline{x})$. By the definition of $v(\cdot)$ and $X_\nu(\cdot)$, we have

$$(x^\nu - \overline{x}^\nu)^\top \nabla_\nu \theta_\nu(\overline{x}) \geq 0$$

for all $x^\nu \in \widehat{X}_\nu$ satisfying that

$$\left\{ \begin{array}{ll} A_s^\nu x^\nu + \sum_{j \neq \nu, j \in \mathcal{N}_s} A_s^j \overline{x}^j - b_s \leq 0, & \forall s \in \mathcal{I}_\nu \\ E_s^\nu x^\nu + \sum_{j \neq \nu, j \in \mathcal{N}_s} E_s^j \overline{x}^j - d_s = 0, & \forall s \in \mathcal{I}_\nu \end{array} \right. \tag{4}$$

Equivalently, by letting $\overline{\xi}^\nu$ be an element of the normal cone of $\widehat{X}_\nu$ at an $\overline{x}^\nu$, we have

$$(x^\nu - \overline{x}^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \geq 0$$

for all $x^\nu \in \mathbb{R}^{n_\nu}$ satisfying the feasibility conditions in Eq. (4). By the change of variables $y^\nu = x^\nu - \overline{x}^\nu$, we have

$$(y^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \geq 0$$

for all $y^\nu \in \mathbb{R}^{n_\nu}$ satisfying

$$A_s^\nu y^\nu \le b_s - A_s \overline{x}^{\mathcal{N}_s}, \quad E_s^\nu y^\nu = 0, \quad \text{for all } s \in \mathcal{I}_\nu$$

Denote the index set of active inequality constraints by $I_s^\nu(\overline{x}) = \{i : (A_s \overline{x}^{\mathcal{N}_s} - b_s)_i = 0\}$ for all $s \in \mathcal{I}_\nu$, where $(\cdot)_i$ stands for the i-th element of a vector. Then, we have

$$A_s^\nu y^\nu \le b_s - A_s \overline{x}^{\mathcal{N}_s} \iff \begin{cases} (A_s^\nu y^\nu)_i \le 0, & \forall i \in I_s^\nu(\overline{x}) \\ (A_s^\nu y^\nu)_i \le (b_s - A_s \overline{x}^{\mathcal{N}_s})_i, & \forall i \notin I_s^\nu(\overline{x}) \end{cases}$$

Putting these pieces together yields that, if a point $\overline{x} \in \widehat{X}$ is a solution of nonlinear GNEP, then the following statement is true:

$$(y^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \ge 0$$

for all $y^\nu \in \mathbb{R}^{n_\nu}$ satisfying the following conditions:

$$\begin{cases} (A_s^\nu y^\nu)_i \le 0, & \forall i \in I_s^\nu(\overline{x}), \quad \forall s \in \mathcal{I}_\nu \\ (A_s^\nu y^\nu)_i \le (b_s - A_s \overline{x}^{\mathcal{N}_s})_i, & \forall i \notin I_s^\nu(\overline{x}), \quad \forall s \in \mathcal{I}_\nu \\ E_s^\nu y^\nu = 0, & \forall s \in \mathcal{I}_\nu \end{cases} \tag{5}$$

Suppose that $z^\nu \in \mathbb{R}^{n_\nu}$ satisfies that

$$\begin{cases} (A_s^\nu z^\nu)_i \le 0, & \forall i \in I_s^\nu(\overline{x}), \quad \forall s \in \mathcal{I}_\nu \\ E_s^\nu z^\nu = 0, & \forall s \in \mathcal{I}_\nu \end{cases}$$

Since $(b_s - A_s \overline{x}^{\mathcal{N}_s})_i > 0$ for all $i \notin I_s^\nu(\overline{x})$ and all $s \in \mathcal{I}_\nu$, it follows there exists a sufficiently small scalar $\tau > 0$ such that $\tau(A_s^\nu z^\nu)_i \le (b_s - A_s \overline{x}^{\mathcal{N}_s})_i$. This together with the definition of z^ν yields that $\overline{z}^\nu = \tau z^\nu$ satisfies Eq. (5). As such, we have $(\tau z^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \ge 0$. Since $\tau > 0$, we have $(z^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \ge 0$. In summary, we have

$$\begin{cases} (A_s^\nu z^\nu)_i \le 0, & \forall i \in I_s^\nu(\overline{x}), \quad \forall s \in \mathcal{I}_\nu \\ E_s^\nu z^\nu = 0, & \forall s \in \mathcal{I}_\nu \end{cases} \implies (z^\nu)^\top (\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu) \ge 0$$

This implies that the following system does not have a solution:

$$\begin{cases} (z^\nu)^\top (-\nabla_\nu \theta_\nu(\overline{x}) - \overline{\xi}^\nu) > 0, & \\ (A_s^\nu z^\nu)_i \le 0, & \forall i \in I_s^\nu(\overline{x}), \quad \forall s \in \mathcal{I}_\nu \\ E_s^\nu z^\nu = 0, & \forall s \in \mathcal{I}_\nu \end{cases}$$

Thus, by Lemma 1, there exist $\overline{\mu}^{\nu,s}$ and $\overline{\lambda}_i^{\nu,s} \ge 0$ such that for all $i \in I_s^\nu(\mathbf{x})$ and all $s \in \mathcal{I}_\nu$ the following equality holds true:

$$\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu + \sum_{s \in \mathcal{I}_\nu} \left(\sum_{i \in I_s^\nu(\overline{x})} \overline{\lambda}_i^{\nu,s} a_i^{\nu,s} + (E_s^\nu)^\top \overline{\mu}^{\nu,s} \right) = 0$$

where $a_i^{\nu,s}$ is the i^{th} column of $(A_s^\nu)^\top$. Equivalently, by letting $\overline{\lambda}_i^{\nu,s} = 0$ for all $i \notin I_s^\nu(\overline{x})$ and all $s \in \mathcal{I}_\nu$, we have

$$\nabla_\nu \theta_\nu(\overline{x}) + \overline{\xi}^\nu + \sum_{s \in \mathcal{I}_\nu} ((A_s^\nu)^\top \overline{\lambda}^{\nu,s} + (E_s^\nu)^\top \overline{\mu}^{\nu,s}) = 0, \quad (\overline{\lambda}^{\nu,s})^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0, \quad \text{for all } s \in \mathcal{I}_\nu$$

By the definition of $\overline{\xi}^{\nu}$, we have

$$(x^{\nu} - \overline{x}^{\nu})^{\top} \left(\nabla_{\nu}\theta_{\nu}(\overline{x}) + \sum_{s \in \mathcal{I}_{\nu}} ((A_s^{\nu})^{\top}\overline{\lambda}^{\nu,s} + (E_s^{\nu})^{\top}\overline{\mu}^{\nu,s}) \right) \geq 0, \text{ for all } x^{\nu} \in \widehat{X}_{\nu}$$

Since $\overline{x}^{\nu} \in X_{\nu}(\overline{x}^{-\nu})$, we have

$$A_s\overline{x}^{\mathcal{N}_s} - b_s \leq 0, \quad E_s\overline{x}^{\mathcal{N}_s} - d_s = 0, \quad \text{for all } s \in \mathcal{I}_{\nu}$$

Putting these pieces together yields that $\overline{x} \in \widehat{X}$ is a KKT point.

Sufficiency. Suppose that a point $\overline{x} = (x^1, x^2, \ldots, x^N) \in \widehat{X}$ is a KKT point so that there exist Lagrangian multipliers $\overline{\lambda}^{\nu,s} \geq 0$ and $\overline{\mu}^{\nu,s}$ satisfying that

$$\begin{aligned} (x^{\nu} - \overline{x}^{\nu})^{\top}(\nabla_{\nu}\theta_{\nu}(\overline{x}) + \sum_{s \in \mathcal{I}_{\nu}} ((A_s^{\nu})^{\top}\overline{\lambda}^{\nu,s} + (E_s^{\nu})^{\top}\overline{\mu}^{\nu,s})) \geq 0 & \quad \forall x^{\nu} \in \widehat{X}_{\nu} \\ A_s\overline{x}^{\mathcal{N}_s} - b_s \leq 0 & \quad \forall s \in \mathcal{I}_{\nu} \\ E_s\overline{x}^{\mathcal{N}_s} - d_s = 0 & \quad \forall s \in \mathcal{I}_{\nu} \\ (\overline{\lambda}^{\nu,s})^{\top}(A_s\overline{x}^{\mathcal{N}_s} - b_s) = 0 & \quad \forall s \in \mathcal{I}_{\nu} \end{aligned} \tag{6}$$

Define the function $L_{\nu}(x^{\nu})$ (which is convex due to the monotonicity) by

$$\begin{aligned} L_{\nu}(x^{\nu}) &= \theta_{\nu}(x^{\nu}; \overline{x}^{-\nu}) \\ &+ \sum_{s \in \mathcal{I}_{\nu}} \left(\left(A_s^{\nu}x^{\nu} + \sum_{j \neq \nu, j \in \mathcal{N}_s} A_s^j\overline{x}^j - b_s \right)^{\top} \overline{\lambda}^{\nu,s} + \left(E_s^{\nu}x^{\nu} + \sum_{j \neq \nu, j \in \mathcal{N}_s} E_s^j\overline{x}^j - d_s \right)^{\top} \overline{\mu}^{\nu,s} \right) \end{aligned}$$

By the definition of L_{ν} and using the third and fourth equations in Eq. (6), we have

$$L_{\nu}(\overline{x}^{\nu}) = \theta_{\nu}(\overline{x}) + \sum_{s \in \mathcal{I}_{\nu}} ((\overline{\lambda}^{\nu,s})^{\top}(A_s\overline{x}^{\mathcal{N}_s} - b_s) + (\overline{\mu}^{\nu,s})^{\top}(E_s\overline{x}^{\mathcal{N}_s} - d_s)) = \theta_{\nu}(\overline{x})$$

The first inequality in Eq. (6) implies that $(x^{\nu} - \overline{x}^{\nu})^{\top}\nabla L_{\nu}(\overline{x}^{\nu}) \geq 0$ for all $x^{\nu} \in \widehat{X}_{\nu}$ and thus we have $L_{\nu}(\overline{x}^{\nu}) = \min_{x^{\nu} \in \widehat{X}_{\nu}} L_{\nu}(x^{\nu})$. Putting these pieces together yields that

$$\theta_{\nu}(\overline{x}) \leq L_{\nu}(x^{\nu}), \quad \text{for all } x^{\nu} \in \widehat{X}_{\nu}$$

Suppose that $x^{\nu} \in \widehat{X}_{\nu}$ is a feasible point in the sense that $x^{\nu} \in X_{\nu}(\overline{x}^{-\nu}) \subseteq \widehat{X}_{\nu}$, we have

$$\begin{aligned} rcll A_s^{\nu}x^{\nu} + \sum_{j \neq \nu, j \in \mathcal{N}_s} A_s^j\overline{x}^j - b_s \leq 0 & \quad \forall s \in \mathcal{I}_{\nu} \\ E_s^{\nu}x^{\nu} + \sum_{j \neq \nu, j \in \mathcal{N}_s} E_s^j\overline{x}^j - d_s = 0 & \quad \forall s \in \mathcal{I}_{\nu} \end{aligned}$$

By the definition of L_{ν} and some simple calculations, we have $L_{\nu}(x^{\nu}) \leq \theta_{\nu}(x^{\nu}, \overline{x}^{-\nu})$. Thus, we have

$$\theta_{\nu}(\overline{x}^{\nu}, \overline{x}^{-\nu}) \leq \theta_{\nu}(x^{\nu}, \overline{x}^{-\nu}), \quad \text{for all } x^{\nu} \in X_{\nu}(\overline{x}^{-\nu})$$

Repeating the above argument for each player $\nu \in \mathcal{N}$, we conclude from Definition 2 that $\overline{x} \in \widehat{X}$ is a solution.

3 Accelerated Mirror-Prox for Strongly Monotone Operators

We review the deterministic accelerated mirror-prox (AMP) algorithm [Chen et al.(2017)], which provides a multi-step acceleration scheme for solving *composite* smooth and monotone VIs with an optimal complexity bound guarantee. More specifically, the AMP algorithm aims at solving the following class of VIs:

$$\text{Find } z^\star \in Z: \quad (z - z^\star)^\top (F(z^\star) + \nabla G(z^\star)) \geq 0, \quad \text{for all } z \in Z \tag{7}$$

where $F : Z \mapsto \mathbb{R}^n$ is ℓ_F-Lipschitz and monotone, $G : X \mapsto \mathbb{R}$ is ℓ_G-smooth and convex, and and Z is a closed and convex nonempty set with diameter $D_Z > 0$. With the initialization $z_1 = z_1^{\text{ag}} = w_1 \in Z$, a typical iteration of the AMP algorithm with Euclidean projection is

$$\begin{aligned}
z_k^{\text{md}} &= (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k w_k \\
z_{k+1} &= \operatorname*{argmin}_{z \in Z} \gamma_k (z - w_k)^\top (F(w_k) + \nabla G(z_k^{\text{md}})) + \frac{1}{2}\|z - w_k\|^2 \\
w_{k+1} &= \operatorname*{argmin}_{z \in X} \gamma_k (z - w_k)^\top (F(z_{k+1}) + \nabla G(z_k^{\text{md}})) + \frac{1}{2}\|z - w_k\|^2 \\
z_{k+1}^{\text{ag}} &= (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k z_{k+1}
\end{aligned}$$

An iteration complexity bound of the AMP algorithm in terms of the number of gradient evaluations is presented in Corollary 1 in [Chen et al.(2017)]. We summarize the result in the following theorem.[3]

Theorem 2 (The AMP Algorithm, Corollary 1 in [Chen et al.(2017)])**.** *Suppose that $F : Z \mapsto \mathbb{R}^n$ is ℓ_F-Lipschitz and monotone and $G : Z \mapsto \mathbb{R}$ is ℓ_G-smooth and convex, and we set the parameters $\alpha_k = \frac{2}{k+1}$ and $\gamma_k = \frac{k}{4\ell_G + 3k\ell_F}$. Then, the following inequality holds for all $k \geq 2$,*

$$\max_{z \in Z} \left\{ G(z_k^{\text{ag}}) - G(z) + (z_k^{\text{ag}} - z)^\top F(z) \right\} \leq \frac{16\ell_G D_Z^2}{k(k-1)} + \frac{12\ell_F D_Z^2}{k-1}$$

and the required number of gradient evaluations to return the point $\widehat{z} \in Z$ satisfying that

$$\max_{z \in Z} \{G(\widehat{z}) - G(z) + (\widehat{z} - z)^\top F(z)\} \leq \epsilon$$

is bounded by $O\left(\sqrt{\frac{\ell_G D_Z^2}{\epsilon}} + \frac{\ell_F D_Z^2}{\epsilon}\right)$.

Remark. There has been considerable interest in the development of algorithms for solving monotone VI problems; among them we highlight the extragradient method [Korpelevich(1976), Tseng(2008)], mirror-prox [Nemirovski(2004)], dual extrapolation [Nesterov(2007)], reflected gradient [Malitsky(2015)], accelerated mirror-prox [Chen et al.(2017)], Halpern iteration [Diakonikolas(2020)] and operator extrapolation [Kotsalis et al.(2022)]. For a general introduction, see [Facchinei and Pang(2007)].

Remark. The subproblems in our frameworks are in the form of Eq. (7) with $Z = \widehat{X}$, $\ell_G = O(\frac{1}{\epsilon})$ for some $\epsilon > 0$ and $\ell_F = O(1)$. As such, the AMP algorithm is more suitable than other algorithms for playing a role in the subroutine for solving monotone nonlinear GNEPs in the sense that it achieves an complexity bound with better dependence on $1/\epsilon$.

[3] We only focus on the deterministic algorithm with Euclidean projection. As such, we provide a simplified version of Corollary 1 in [Chen et al.(2017)] in the Euclidean setting and $\sigma = 0$ in Theorem 2.

We provide the iteration complexity bound of the AMP algorithm for solving the VI in Eq. (7) in terms of the number of gradient evaluations in which $F: Z \mapsto \mathbb{R}^n$ is further assumed to be α-strongly monotone. We summarize the result in Theorem 3 and refer interested readers to Section 3.1 for the proof details.

Theorem 3 (The AMP Algorithm: strongly monotone)**.** *Suppose that $F : X \mapsto \mathbb{R}^n$ is ℓ_F-Lipschitz and α-strongly monotone and $G : Z \mapsto \mathbb{R}$ is ℓ_G-smooth and convex, and we set the parameters $\alpha_k = \frac{1}{4}\min\{\frac{\alpha}{\ell_F}, \sqrt{\frac{\alpha}{\ell_G}}\}$ and $\gamma_k = \frac{\alpha_k}{\alpha}$. Then, the following inequality holds for all $k \geq 2$:*

$$\max_{z \in Z} \left\{G(z_k^{\text{ag}}) - G(z) + (z_k^{\text{ag}} - z)^\top F(z)\right\} \leq \left(1 - \frac{1}{4}\min\left\{\frac{\alpha}{\ell_F}, \sqrt{\frac{\alpha}{\ell_G}}\right\}\right)^{k-1} \left(\ell_F + \frac{\ell_G + \alpha}{2}\right) D_Z^2$$

and the required number of gradient evaluations to return the point $\widehat{z} \in Z$ satisfying that

$$\max_{z \in Z}\{G(\widehat{z}) - G(z) + (\widehat{z} - z)^\top F(z)\} \leq \epsilon$$

is bounded by $O\left(\left(\sqrt{\frac{\ell_G}{\alpha}} + \frac{\ell_F}{\alpha}\right)\log\left(\frac{(\ell_F + \ell_G)D_Z^2}{\epsilon}\right)\right)$.

Remark. Theorem 3 complements the results in [Chen et al.(2017)] and provides a more complete picture for the optimal iteration complexity bounds of deterministic first-order algorithms for composite smooth and (strongly) monotone VIs in the form of Eq. (7). An open problem is to consider extensions to the stochastic setting.

3.1 Proof of Theorem 3

We provide the proof of Theorem 3 that delineates an iteration complexity bound of the AMP algorithm for solving *composite* smooth and strongly monotone VIs. More specifically, the scheme aims at solving the following class of VIs:

$$\text{Find } z^\star \in Z: \quad (z - z^\star)^\top (F(z^\star) + \nabla G(z^\star)) \geq 0, \quad \text{for all } z \in Z \tag{8}$$

where $F : Z \mapsto \mathbb{R}^n$ is ℓ_F-Lipschitz and α-strongly monotone, $G : Z \mapsto \mathbb{R}$ is ℓ_G-smooth and convex, and Z is an nonempty, closed and convex set with diameter $D_Z > 0$. With the initialization $z_1 = z_1^{\text{ag}} = w_1 \in Z$, a typical iteration of the algorithm is given by

$$\begin{aligned}
z_k^{\text{md}} &= (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k w_k \\
z_{k+1} &= \operatorname*{argmin}_{z \in Z} \gamma_k (z - w_k)^\top (F(w_k) + \nabla G(z_k^{\text{md}})) + \frac{1}{2}\|z - w_k\|^2 \\
w_{k+1} &= \operatorname*{argmin}_{z \in Z} \gamma_k (z - w_k)^\top (F(z_{k+1}) + \nabla G(z_k^{\text{md}})) + \frac{1}{2}\|z - w_k\|^2 \\
z_{k+1}^{\text{ag}} &= (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k z_{k+1}
\end{aligned} \tag{9}$$

where $\alpha_k \in (0, 1)$ and $\gamma_k > 0$ are parameters to be determined.

We present some technical lemmas which are important for the subsequent analysis. To state these lemmas it is useful to define the following *gap function*:

$$Q(\widetilde{z}, z) = G(\widetilde{z}) - G(z) + (\widetilde{z} - z)^\top F(z)$$

Our first lemma gives a descent inequality for the iterates generated by the AMP algorithm (cf. Eq. (9)). This is a consequence of Lemma 6.3 in [Juditsky et al.(2011)]; we provide a proof for completeness.

Lemma 2. *Suppose that the iterates $\{(z_k^{\mathrm{md}}, z_k, w_k, z_k^{\mathrm{ag}})\}_{k \geq 1}$ are generated by the AMP algorithm (cf. Eq. (9)). Then, the following statement holds true for all $z \in Z$:*

$$\gamma_k(z_{k+1} - z)^\top (F(z_{k+1}) + \nabla G(z_k^{\mathrm{md}})) \leq \frac{1}{2}(\|z - w_k\|^2 - \|z - w_{k+1}\|^2) - (\frac{1}{2} - \frac{\gamma_k^2 \ell_F^2}{2})\|z_{k+1} - w_k\|^2$$

The second measures the progress on $Q(z_k^{\mathrm{ag}}, z)$:

Lemma 3. *Suppose that the iterates $\{(z_k^{\mathrm{md}}, z_k, w_k, z_k^{\mathrm{ag}})\}_{k \geq 1}$ are generated by the AMP algorithm (cf. Eq. (9)). Then, the following statement holds true for all $z \in Z$:*

$$\begin{aligned} & Q(z_{k+1}^{\mathrm{ag}}, z) - (1 - \alpha_k) Q(z_k^{\mathrm{ag}}, z) \\ \leq \;& \alpha_k (z_{k+1} - z)^\top (F(z_{k+1}) + \nabla G(z_k^{\mathrm{md}})) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 - \alpha_k \alpha \|z_{k+1} - z\|^2 \end{aligned}$$

We are ready to present the

Proof of Theorem 3. Combining Lemma 2 and Lemma 3, we have

$$\begin{aligned} & Q(z_{k+1}^{\mathrm{ag}}, z) - (1 - \alpha_k) Q(z_k^{\mathrm{ag}}, z) \\ \leq \;& \frac{\alpha_k}{\gamma_k}\left(\frac{1}{2}(\|z - w_k\|^2 - \|z - w_{k+1}\|^2) - (\frac{1}{2} - \frac{\gamma_k^2 \ell_F^2}{2})\|z_{k+1} - w_k\|^2\right) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 - \alpha_k \alpha \|z_{k+1} - z\|^2 \\ = \;& \frac{\alpha_k}{2\gamma_k}\|z - w_k\|^2 - \frac{\alpha_k}{2\gamma_k}\|z - w_{k+1}\|^2 - (\frac{\alpha_k}{2\gamma_k} - \frac{\alpha_k \gamma_k \ell_F^2}{2} - \frac{\alpha_k^2 \ell_G}{2})\|z_{k+1} - w_k\|^2 - \alpha_k \alpha \|z_{k+1} - z\|^2 \end{aligned}$$

Since $\alpha_k \equiv \alpha_0 = \frac{1}{4}\min\{\frac{\alpha}{\ell_F}, \sqrt{\frac{\alpha}{\ell_G}}\}$ and $\gamma_k = \frac{\alpha_k}{\alpha}$ for all $k \geq 1$, we remove the subscript in the above inequality and define the energy term $E_k = Q(z_k^{\mathrm{ag}}, z) + \frac{\alpha}{2}\|z - w_k\|^2$. We have

$$E_{k+1} - (1 - \alpha_0)E_k \leq \frac{\alpha_0 \alpha}{2}\|z - w_k\|^2 - (\frac{\alpha}{2} - \frac{\alpha_0^2 \ell_F^2}{2\alpha} - \frac{\alpha_0^2 \ell_G}{2})\|z_{k+1} - w_k\|^2 - \alpha_0 \alpha \|z_{k+1} - z\|^2$$

By Young's inequality, we have

$$E_{k+1} - (1 - \alpha_0)E_k \leq -(\frac{\alpha}{2} - \alpha_0 \alpha - \frac{\alpha_0^2 \ell_F^2}{2\alpha} - \frac{\alpha_0^2 \ell_G}{2})\|z_{k+1} - w_k\|^2$$

By the definition of $\alpha_0 = \frac{1}{4}\min\{\frac{\alpha}{\ell_F}, \sqrt{\frac{\alpha}{\ell_G}}\}$, we have

$$\alpha \alpha_0 \leq \frac{\alpha}{4}, \quad \frac{\alpha_0^2 \ell_F^2}{2\alpha} \leq \frac{\alpha}{8}, \quad \frac{\alpha_0^2 \ell_G}{2} \leq \frac{\alpha}{8}$$

Putting these pieces together yields that $E_{k+1} \leq (1 - \alpha_0)E_k$ for all $k \geq 1$. Then, by using the definition of E_k, we have

$$\max_{z \in Z} Q(z_k^{\mathrm{ag}}, z) \leq \left(1 - \frac{1}{4}\min\left\{\frac{\alpha}{\ell_F}, \sqrt{\frac{\alpha}{\ell_G}}\right\}\right)^{k-1} \max_{z \in Z} \{Q(z_1^{\mathrm{ag}}, z) + \frac{\alpha}{2}\|z - w_1\|^2\}$$

This together with the fact that the diameter of Z is D_Z implies the desired inequality. $\square$

4 Proposed Methods for Solving (Strongly) Monotone Nonlinear GNEPs

We propose two first-order algorithms—one based on *quadratic penalty method* (QPM) and the other based on *augmented Lagrangian method* (ALM)—where each iteration consists in a subroutine for approximately solving a smooth and monotone penalized VI using the AMP algorithm. We carry out an estimation of a global convergence rate for these algorithms in monotone and strongly monotone settings and present iteration complexity bounds in terms of the number of gradient evaluations. Our algorithms update penalty parameters adaptively and thus are favorable from a practical viewpoint.

4.1 Quadratic penalty method

Our first framework is a natural combination of the quadratic penalty method and the AMP algorithm. We refer to it as the *accelerated mirror-prox quadratic penalty method* (AMP-QP). The idea is to transform an nonlinear GNEP into a sequence of structured quadratic penalized NEPs and approximately solve each one in the inner loop using a subroutine based on the AMP algorithm. Indeed, we consider a quadratic penalization of the nonlinear GNEP where each player's minimization problem is given by

$$\min_{x^\nu \in \widehat{X}^\nu} \theta_\nu(x^\nu, x^{-\nu}) + \left(\sum_{s \in \mathcal{I}_\nu} \frac{\beta^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s\}\|^2 + \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s\|^2\right) \tag{10}$$

and where $(\beta^s, \rho^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ stand for penalty parameters associated with inequality and equality constraints. Since the function $\theta_\nu(\cdot, x^{-\nu})$ is convex with Lipschitz gradient and the constraint set $\widehat{X}^\nu$ does not depend on the rival's choices, the quadratic penalization of the nonlinear GNEP in Eq. (10) is an NEP.

For simplicity, we define the functions $g_1 : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_1 : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$g_1(x) = \sum_{s=1}^{S} \frac{\beta^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s\}\|^2, \quad h_1(x) = \sum_{s=1}^{S} \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s\|^2$$

By concatenating the first-order optimality conditions of Eq. (10) for each player, we aim at solving the following smooth and monotone penalized VI:

$$\text{Find } x \in \widehat{X} : \quad (x' - x)^\top(v(x) + \nabla g_1(x) + \nabla h_1(x)) \geq 0, \text{ for all } x' \in \widehat{X} \tag{11}$$

It is worth mentioning that this VI is in the form of Eq. (7) and can be approximately solved by the AMP algorithm with the theoretical guarantees that we have explicated.

The introduction of quadratic penalization based on squared Euclidean norm was a milestone in optimization [Bertsekas(1976), Aybat and Iyengar(2011), Lan and Monteiro(2013), Necoara et al.(2019)]. It has been, however, less explored for GNEPs possibly because the quadratic penalty method cannot recover an exact solution for a finite penalty parameter; see [Ba and Pang(2022)], Example 4(b). Nevertheless, if our goal is finding an ϵ-approximate solution and estimating the iteration complexity bounds in terms of the number of gradient evaluations, the quadratic penalty method has a clear advantage over other exact penalty methods; indeed, each subproblem in exact penalty methods is nondifferentiable and problematic from the numerical viewpoint [Facchinei and Kanzow(2010b)]. Although some algorithms for NEPs work without differentiability, e.g., the relaxation method [Uryas' ev and Rubin

Algorithm 1 Accelerated Mirror-Prox Quadratic Penalty Method (AMP-QP)

Input: $\gamma > 1$, $\delta_0 \in (0,1)$ and $x_0 \in \widehat{X}$
Initialization: $(\beta_0^s, \rho_0^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ for all $s \in [S]$
for $k = 0, 1, 2, \ldots, T-1$ **do**
  Update $(\beta_{k+1}^s, \rho_{k+1}^s) = (\gamma\beta_k^s, \gamma\rho_k^s)$ for all $s \in [S]$ and $\delta_{k+1} = \frac{\delta_k}{\gamma}$
  Compute $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, \delta_{k+1}, x_k)$
end for
Output: x_T

Von Heusinger and Kanzow(2009)] or the proximal-like method [Flåm and Antipin(1996)], they require solving certain constrained optimization problems which are only tractable for sufficiently smooth data. Another approach is to use nonlinear system solvers after smoothing the nondifferentiable terms; see Section 3 of [Facchinei and Kanzow(2010b)]. To the best of our knowledge, the global convergence rates of these approaches are unknown and the nondifferentiability makes the iteration complexity analysis difficult. In contrast, the quadratic penalization in Eq. (11) is smooth and monotone such that a standard iteration complexity analysis can be applied.

To further enhance the practical viability of the algorithm, there are two issues that need to be addressed: (i) when to stop the AMP algorithm for each inner loop; and (ii) how to update the penalty parameters (β^s, ρ^s). In principle, there are adaptive strategies adopted in the exact penalty methods [Facchinei and Lampariello(2011), Facchinei and Kanzow(2010b), Fukushima(2011)] such that global convergence guarantees can be achieved given a convergent algorithm for each nondifferentiable NEP. However, all of these theoretical results are asymptotic and the global convergence rates of these adaptive approaches are unknown.

We summarize the pseudocode of the AMP-QP algorithm in Algorithm 1. In particular, we follow an adaptive strategy akin to those of [Facchinei and Kanzow(2010b), Facchinei and Lampariello(2011), Fukushima(2011)], who suggest updating $(\beta_{k+1}^s, \rho_{k+1}^s) = (\gamma\beta_k^s, \gamma\rho_k^s)$ with $\gamma > 1$ at the k^{th} iteration. Also, δ_{k+1} is updated accordingly and we let the AMP subroutine return a δ_{k+1}-approximate solution $\widehat{x} \in \widehat{X}$ to the VI in Eq. (11) with $(\beta^s, \rho^s) = (\beta_{k+1}^s, \rho_{k+1}^s)$ as follows:

$$\max_{x \in \widehat{X}} \left\{ g_1(\widehat{x}) + h_1(\widehat{x}) - g_1(x) - h_1(x) + (\widehat{x} - x)^\top v(x) \right\} \leq \delta_{k+1} \tag{12}$$

Our subroutine implements the AMP algorithm with an input x_k from previous iteration and returns x_{k+1} which is a δ_{k+1}-approximate solution. For simplicity, we write the subroutine in the compact form as $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, \delta_{k+1}, x_k)$.

Intuitively, the idea of Algorithm 1 can be described as follows. Suppose that we know the "correct" values of the penalty parameters $\{\beta^s, \rho^s\}_{s\in[S]}$ and threshold δ such that a δ-approximate solution of the penalized VI with these parameters is an ϵ-solution of nonlinear GNEPs. Then, a single application of the AMP algorithm would give us an ϵ-solution. However, the correct values are unknown beforehand and thus we need to design the adaptive strategies in Algorithm 1 to handle this issue. Even though we stated for simplicity that the subroutine $\textsc{amp}(\cdot)$ should produce x_{k+1} based only on x_k, it is plausible to use any information gathered in previous iterations as well. Nevertheless, after the penalty parameters are updated, we have changed the subproblem and all old information should be discarded (since it is related to a different problem). As such, we prefer to write $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, \delta_{k+1}, x_k)$, where x_k is regarded as a starting point for the application of the AMP algorithm.

Notably, it is intractable to verify if x_T is an ϵ-solution using Definition 4. In practice, an alternative approach is to set the appropriate stopping criteria. For example, we can run the algorithm until either of the following conditions hold true: (i) a theoretically derived maximum number of iterations is reached, (ii) the iterative gap $\|x_T - x_{T-1}\|$ is sufficiently small, or (iii) the minimum of penalty parameters, $\min_{1\leq s\leq S}\{\beta_T^s, \rho_T^s\}$, is sufficiently large. The same practical issues hold for verifying if x_{k+1} is an δ_{k+1}-solution of the VI subproblem using Eq. (12). In our implementation, we set the stopping criteria for the AMP subroutine if either a theoretically derived maximum number of iterations (cf. Theorem 2 and 3) is reached or the iterative gap is sufficiently small.

4.2 Augmented Lagrangian method

Our second framework combines the inexact augmented Lagrangian method with the AMP algorithm. We refer to it as the *accelerated mirror-prox augmented Lagrangian method* (AMP-AL). The idea is again to transform an nonlinear GNEP into a sequence of quadratic penalized NEPs but using the augmented Lagrangian function and to approximately solve the VI-type subproblem in the inner loop using the AMP algorithm. Indeed, we define the following augmented Lagrangian function for each player's minimization problem:

$$\mathcal{L}_\nu(x,\lambda,\mu) = \theta_\nu(x^\nu, x^{-\nu}) + \left(\sum_{s\in\mathcal{I}_\nu} \frac{\beta^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda^s}{\beta^s}\}\|^2 + \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu^s}{\rho^s}\|^2\right)$$

where $(\beta^s, \rho^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ and $(\lambda^s, \mu^s) \in \mathbb{R}^{m_s} \times \mathbb{R}^{e_s}$ stand for penalty parameters and Lagrangian multipliers associated with inequality and equality constraints. The augmented Lagrangian function has some important advantages: (i) it is convex in the variable x^ν and concave in the variable $\{(\lambda^s, \mu^s)\}_{s\in\mathcal{I}_\nu}$; (ii) its gradient with respect to x^ν is Lipschitz and the constraint set $\widehat{X}^\nu$ does not depend on the rival's choices. Putting these pieces together implies that a joint minimization of the augmented Lagrangian function for all players forms an NEP. For simplicity, we define the functions $g_2 : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_2 : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$g_2(x) = \sum_{s=1}^S \frac{\beta^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda^s}{\beta^s}\}\|^2, \quad h_2(x) = \sum_{s=1}^S \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu^s}{\rho^s}\|^2$$

By concatenating the first-order optimality conditions of minimizing the augmented Lagrangian function for each player, we aim at solving the following smooth and monotone penalized VI:

$$\text{Find } x \in \widehat{X} : \quad (x'-x)^\top(v(x) + \nabla g_2(x) + \nabla h_2(x)) \geq 0, \text{ for all } x' \in \widehat{X} \tag{13}$$

This VI is in the form of Eq. (7) and can be approximately solved by the AMP algorithm with the theoretical guarantees that we have explicated.

The augmented Lagrangian method was proposed by [Hestenes(1969)] and [Powell(1969)] and its convergence has been studied extensively in the optimization literature. Milestones include early results on asymptotic convergence and local convergence [Rockafellar(1973a), Rockafellar(1973b), Rockafellar(1976)] and recent results on the nonasymptotic global convergence rate [Lan and Monteiro(2016), Xu(2021)]. Each iteration of the algorithm consists in first minimizing the augmented Lagrangian function with respect to primal variables while fixing the dual variables and then performing a dual gradient ascent update. It is, however, generally intractable to exactly minimize the augmented Lagrangian function with respect to primal variables, leading to the inexact variant [Xu(2021)] where

Algorithm 2 Accelerated Mirror-Prox Augmented Lagrangian Method (AMP-AL)

Input: $\gamma > 1$, $\delta_0 \in (0, 1)$ and $x_0 \in \widehat{X}$
Initialization: $(\beta_0^s, \rho_0^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ and $(\lambda_0^s, \mu_0^s) \in \mathbb{R}^{m_s} \times \mathbb{R}^{e_s}$ for all $s \in [S]$
for $k = 0, 1, 2, \ldots, T-1$ **do**
  Update $(\beta_{k+1}^s, \rho_{k+1}^s) = (\gamma\beta_k^s, \gamma\rho_k^s)$ for all $s \in [S]$ and $\delta_{k+1} = \frac{\delta_k}{\gamma}$
  Compute $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, (\lambda_k^s, \mu_k^s)_{s\in[S]}, \delta_{k+1}, x_k)$
  Update λ_{k+1}^s and μ_{k+1}^s for all $s \in [S]$ by

$$\lambda_{k+1}^s = \max\{0, \lambda_k^s + \beta_{k+1}^s(A_s x_{k+1}^{\mathcal{N}_s} - b_s)\}, \quad \mu_{k+1}^s = \mu_k^s + \rho_{k+1}^s(E_s x_{k+1}^{\mathcal{N}_s} - d_s)$$

end for
Output: $\widehat{x}_T = \frac{1}{T}\sum_{k=1}^T x_k$

each subproblem is solved within a desired tolerance. This idea was later extended to GNEPs and QVIs [Kanzow and Steck(2016), Kanzow and Steck(2018), Bueno et al.(2019)] and the resulting algorithms were shown to outperform the penalty-type methods.

As pointed out by [Kanzow and Steck(2016)], the practical advantages come from the smoothness of each subproblem and the safeguarding effect of multipliers. In this context, one often solves each subproblem by appealing to well-known second-order methods, e.g., semismooth Newton methods [Kummer(1988), Qi and Sun(1993)] and Levenberg-Marquardt methods [Levenberg(1944), Marquardt(1963), Yamashita and Fukushima(2001)]. However, a singularity arises in GNEPs when some players share the same constraints [Facchinei et al.(2009)] and this makes the iteration complexity analysis of these approaches difficult.

We summarize the pseudocode of the AMP-AL algorithm in Algorithm 2. In particular, there are several heuristics for adapting penalty parameters in the augmented Lagrangian methods [Kanzow and Steck(2016), Kanzow and Steck(2018), Bueno et al.(2019)] and some of them have been proven very efficient in practice; see [Kanzow and Steck(2016)] Section 6. We use these strategies in the AMP-AL algorithm together with updating $(\beta_{k+1}^s, \rho_{k+1}^s) = (\gamma\beta_k^s, \gamma\rho_k^s)$. Also, δ_{k+1} is updated accordingly and we let the AMP subroutine return a δ_{k+1}-approximate solution $\widehat{x} \in \widehat{X}$ to the VI in Eq. (13) with $(\beta^s, \rho^s) = (\beta_{k+1}^s, \rho_{k+1}^s)$:

$$\max_{x\in\widehat{X}} \{g_2(\widehat{x}) + h_2(\widehat{x}) - g_2(x) - h_2(x) + (\widehat{x} - x)^\top v(x)\} \leq \delta_{k+1}$$

Similar to the AMP-QP algorithm, our subroutine implements the AMP algorithm with an input x_k from the previous iteration and returns x_{k+1} which is a δ_{k+1}-approximate solution. We write the subroutine as $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, (\lambda_k^s, \mu_k^s)_{s\in[S]}, \delta_{k+1}, x_k)$ and update the Lagrangian multipliers $(\lambda_{k+1}^s, \mu_{k+1}^s)_{s\in[S]}$ using x_{k+1} and $(\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}$. In practice, we set the appropriate stopping criteria rather than directly verifying if x_T is an ϵ-solution using Definition 4 and if x_{k+1} is a δ_{k+1}-approximate solution using Eq. (12).

4.3 Iteration complexity bound

Our first result is the iteration complexity bound for Algorithm 1 when applied to monotone and strongly monotone nonlinear GNEPs.

Theorem 4 (Iteration Complexity Analysis of the AMP-QP Algorithm)**.** *For a given tolerance $\epsilon > 0$, the required number of gradient evaluations for Algorithm 1 to return an ϵ-solution (cf. Definition 4) is upper bounded by*

$$N_{\text{grad}} = \begin{cases} O(\epsilon^{-1}), & \text{if } \alpha = 0 \\ O(\epsilon^{-1/2}\log(1/\epsilon)), & \text{if } \alpha > 0 \end{cases}$$

where the two lines refer to the case of monotone and strongly monotone nonlinear GNEPs.

For both monotone and strongly monotone nonlinear GNEPs, our iteration complexity bounds are new and generalize classical results. Indeed, a monotone nonlinear GNEP reduces to a linearly constrained convex problem when $N = 1$ and our algorithm achieves the lower bound for any deterministic first-order algorithms; see Theorem 4.1 in [Xu(2021)]. A monotone nonlinear GNEP also reduces to a monotone VI when $A_s = 0$, $E_s = 0$, $b_s = 0$ and $d_s = 0$ for all $s \in [S]$ and our algorithm also achieves the lower bound for any deterministic first-order algorithms; see Lemma 16 in [Diakonikolas(2020)]. Moreover, a strongly monotone nonlinear GNEP reduces to a linearly constrained strongly convex problem when $N = 1$ and the lower bound for any deterministic first-order algorithm is $\Omega(\epsilon^{-1/2})$; see Theorem 4.2 [Ouyang and Xu(2021)] which is attained by Algorithm 1 up to log factors. Although existing lower bounds for linearly constrained convex problems and monotone VIs are derived using *slightly different* notions, the matching orders demonstrate the theoretical efficiency of Algorithm 1.

Our second result is the iteration complexity bound for Algorithm 2 when applied to solve monotone and strongly monotone nonlinear GNEPs.

Theorem 5 (Iteration Complexity Analysis of the AMP-AL Algorithm)**.** *For a given tolerance $\epsilon > 0$, the required number of gradient evaluations for Algorithm 2 to return an ϵ-solution (cf. Definition 4) is upper bounded by*

$$N_{\text{grad}} = \begin{cases} O(\epsilon^{-1}\log(1/\epsilon)), & \text{if } \alpha = 0 \\ O(\epsilon^{-1/2}\log(1/\epsilon)), & \text{if } \alpha > 0 \end{cases}$$

where the two lines refer to the case of monotone and strongly monotone nonlinear GNEPs.

Theorem 5 shows that Algorithm 2 recovers the optimal iteration complexity bound of the inexact augmented Lagrangian methods for solving linearly constrained (strongly) convex problems; c.f. [Lan and Monteiro(2016), Xu(2021)]. Our results show that Algorithm 2 matches the lower bound for any deterministic first-order algorithm up to log factors.

4.4 Proof of Theorem 4

We provide the proof of Theorem 4 on the iteration complexity bound of the AMP-QP algorithm (cf. Algorithm 1) for solving monotone and strongly monotone nonlinear GNEPs.

We present some technical lemmas which are important to the subsequent analysis. Consider a quadratic penalization of an nonlinear GNEP where each player's minimization problem is given by

$$\min_{x^\nu \in \widehat{X}^\nu} \theta_\nu(x^\nu, x^{-\nu}) + \left(\sum_{s \in \mathcal{I}_\nu} \frac{\beta^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s\}\|^2 + \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s\|^2\right)$$

where $(\beta^s, \rho^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ stand for penalty parameters associated with inequality and equality constraints. We define the functions $g_1 : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_1 : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$g_1(x) = \sum_{s=1}^S \frac{\beta^s}{2} \|\max\{0, A_s x^{\mathcal{N}_s} - b_s\}\|^2, \quad h_1(x) = \sum_{s=1}^S \frac{\rho^s}{2} \|E_s x^{\mathcal{N}_s} - d_s\|^2$$

Concatenating the first-order optimality conditions of Eq. (10) for each player, we aim to solve the following VI ($\widehat{X}$ is convex and compact with a diameter $D > 0$):

$$\text{Find } x \in \widehat{X} : \quad (x' - x)^\top (v(x) + \nabla g_1(x) + \nabla h_1(x)) \geq 0, \quad \text{for all } x' \in \widehat{X} \tag{14}$$

We first state the following well-known result which guarantees that the distance function has a Lipschitz-continuous gradient; see [Lan et al.(2011)] Proposition 5.

Proposition 2. *Given a closed convex set $\mathcal{C} \subseteq \mathbb{R}^n$, we let $d_{\mathcal{C}} : \mathbb{R}^n \mapsto \mathbb{R}$ be the distance function to $\mathcal{C}$ with respect to $\|\cdot\|$ on $\mathbb{R}^n$. Then, the function $\psi(\cdot) = (d_{\mathcal{C}}(\cdot))^2$ is convex and its gradient is given by $\nabla\psi(x) = 2(x - \mathcal{P}_{\mathcal{C}}(x))$ for all $x \in \mathbb{R}^n$. In addition, the gradient is Lipschitz continuous; that is, $\|\nabla\psi(\widetilde{x}) - \nabla\psi(x)\| \leq 2\|\widetilde{x} - x\|$ for all $\widetilde{x}, x \in \mathbb{R}^n$.*

As an immediate consequence of Proposition 2, we obtain the following lemma.

Lemma 4. *The gradients of the functions g_1 and h_1 are ℓ_β-Lipschitz continuous and ℓ_ρ-Lipschitz continuous respectively, where $\ell_\beta := \sum_{s=1}^S \beta^s \|A_s\|^2$ and $\ell_\rho := \sum_{s=1}^S \rho^s \|E_s\|^2$.*

Combining the above lemma with Theorem 2 and 3, we obtain the following lemma.

Lemma 5. *Suppose that $x_{k+1} = \textsc{amp}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, \delta_{k+1}, x_k)$ in the sense that x_{k+1} is a δ_{k+1}-approximate solution to the structured VI in Eq. (14) with $(\beta^s, \rho^s) = (\beta_{k+1}^s, \rho_{k+1}^s)$ such that*

$$\max_{x \in \widehat{X}} \left\{ g_1(x_{k+1}) + h_1(x_{k+1}) - g_1(x) - h_1(x) + (x_{k+1} - x)^\top v(x) \right\} \leq \delta_{k+1}$$

The required number of gradient evaluations at the k^{th} iteration (for $k \geq 1$) is bounded by

$$N_k = \begin{cases} O\left(\sqrt{\frac{\sum_{s=1}^S (\beta_k^s \|A_s\|^2 + \rho_k^s \|E_s\|^2) D^2}{\delta_k}} + \frac{\sqrt{N}\ell_\theta D^2}{\delta_k} \right), & \text{if } \alpha = 0 \\ O\left(\left(\sqrt{\frac{\sum_{s=1}^S (\beta_k^s \|A_s\|^2 + \rho_k^s \|E_s\|^2)}{\alpha}} + \frac{\sqrt{N}\ell_\theta}{\alpha} \right) \log\left(\frac{(\sqrt{N}\ell_\theta + \ell_G) D^2}{\delta_k} \right) \right), & \text{if } \alpha > 0 \end{cases}$$

We continue with proving that x_T is an $\widehat{\epsilon}$-solution of nonlinear GNEPs for some $\widehat{\epsilon} > 0$.

Lemma 6. *Suppose that x_T is a δ_T-approximate solution to the structured VI in Eq. (14) with $(\beta^s, \rho^s) = (\beta_T^s, \rho_T^s)$ in the sense that*

$$\max_{x \in \widehat{X}} \left\{ g_1(x_T) + h_1(x_T) - g_1(x) - h_1(x) + (x_T - x)^\top v(x) \right\} \leq \delta_T$$

Then, $x_T \in \widehat{X}$ satisfies

$$\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \leq \widehat{\epsilon}_1, \quad \|E_s x_T^{\mathcal{N}_s} - d_s\| \leq \widehat{\epsilon}_1, \quad \text{for all } s \in [S] \tag{15}$$

where $\widehat{\epsilon}_1 > 0$ *is defined by*

$$\widehat{\epsilon}_1 = 2\sqrt{\frac{\delta_T + 10S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_T^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_T^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_T^s, \rho_T^s\}}}$$

and for all $\nu \in \mathcal{N}$*, we have*

$$(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, x_T^{-\nu}) \le \widehat{\epsilon}_2$$

for all $x^\nu \in \widehat{X}_\nu$ *satisfying that*

$$\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \le \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \le \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu \tag{16}$$

where $\widehat{\epsilon}_2 > 0$ *is defined by*

$$\widehat{\epsilon}_2 = \delta_T + \frac{\epsilon^2}{2}\left(\sum_{s=1}^S \beta_T^s + \rho_T^s\right)$$

We are ready for the

Proof of Theorem 4. Fixing a sufficiently small $\epsilon \in (0, 1)$, we have $x_T \in \widehat{X}$ is an ϵ-solution of an nonlinear GNEP if

$$\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \le \epsilon, \quad \|E_s x_T^{\mathcal{N}_s} - d_s\| \le \epsilon, \quad \text{for all } s \in [S]$$

and, for all $\nu \in \mathcal{N}$, we have

$$(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, \overline{x}^{-\nu}) \le C\epsilon$$

for all $x^\nu \in \widehat{X}_\nu$ satisfying that

$$\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \le \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \le \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu$$

By Lemma 6, it suffices to guarantee that $\delta_T > 0$ and $(\beta_T^s, \rho_T^s)_{s\in[S]}$ satisfy the following conditions:

$$\frac{4\delta_T + 40S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_T^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_T^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_T^s, \rho_T^s\}} \le \epsilon^2$$

$$\delta_T + \frac{\epsilon^2}{2}\left(\sum_{s=1}^S \beta_T^s + \rho_T^s\right) \le C\epsilon$$

From the update rule in Algorithm 1, we have $\beta_T^s = \gamma^T \beta_0^s$ and $\rho_T^s = \gamma^T \rho_0^s$ for all $s \in [S]$ and $\delta_T = \frac{\delta_0}{\gamma^T}$ in which $\gamma > 1$. Putting these pieces together, it suffices to guarantee that

$$\begin{aligned} \frac{\epsilon}{2} &\ge \frac{1}{\gamma^T}\sqrt{\frac{\delta_0 + 10S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_0^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_0^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_0^s, \rho_0^s\}}} \\ C\epsilon &\ge \frac{\delta_0}{\gamma^T} + \frac{\gamma^T\epsilon^2}{2}\left(\sum_{s=1}^S \beta_0^s + \rho_0^s\right) \end{aligned} \tag{17}$$

Suppose that we set $T > 0$ as

$$T = 1 + \left\lfloor \frac{1}{\log(\gamma)} \left\{ \log\left(\frac{2}{\epsilon}\right) + \frac{1}{2}\log\left(\frac{\delta_0 + 10S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_0^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_0^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_0^s, \rho_0^s\}} \right) \right\} \right\rfloor$$

Then, Eq. (17) holds true with a positive constant $C > 0$ given by

$$\begin{aligned} C &= 2 + \left[\left(\sum_{s=1}^S \beta_0^s + \rho_0^s\right) \left(\frac{\delta_0 + 10S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_0^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_0^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_0^s, \rho_0^s\}} \right)^{1/2} \right. \\ & \qquad \left. + \frac{\delta_0}{2} \left(\frac{\delta_0 + 10S\left(\max_{1\le s\le S}\left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_0^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_0^s}\right\}\right)}{\min_{1\le s\le S}\{\beta_0^s, \rho_0^s\}} \right)^{-1/2} \right] \end{aligned}$$

Since $\beta_k^s = \gamma^k \beta_0^s$ and $\rho_k^s = \gamma^k \rho_0^s$ for all $s \in [S]$ and $\delta_k = \frac{\delta_0}{\gamma^k}$, Lemma 5 guarantees that the required number of gradient evaluations at the k^{th} iteration is bounded by

$$N_k = \left\{ \begin{array}{ll} O\left(\gamma^k \left(\sqrt{\frac{\sum_{s=1}^S (\beta_0^s\|A_s\|^2 + \rho_0^s\|E_s\|^2)D^2}{\delta_0}} + \frac{\sqrt{N}\ell_\theta D^2}{\delta_0} \right)\right), & \text{if } \alpha = 0 \\ O\left(\left(\gamma^{\frac{k}{2}} \sqrt{\frac{\sum_{s=1}^S (\beta_0^s\|A_s\|^2 + \rho_0^s\|E_s\|^2)}{\alpha}} + \frac{\sqrt{N}\ell_\theta}{\alpha}\right) \log\left(\frac{\gamma^k(\sqrt{N}\ell_\theta + \ell_G)D^2}{\delta_0}\right)\right), & \text{if } \alpha > 0 \end{array} \right.$$

Therefore, we conclude that the total number of gradient evaluations required to return an ϵ-solution of the nonlinear GNEP is

$$N_{\text{grad}} = \sum_{k=1}^{T-1} N_k = \left\{ \begin{array}{ll} O\left(\epsilon^{-1}\right), & \text{if } \alpha = 0 \\ O\left(\epsilon^{-1/2}\log(1/\epsilon)\right), & \text{if } \alpha > 0 \end{array} \right.$$

This completes the proof. □

4.5 Proof of Theorem 5

We provide the proof of Theorem 5 on the iteration complexity bound of the AMP-AL algorithm (cf. Algorithm 2) for solving monotone and strongly monotone nonlinear GNEPs.

We present some technical lemmas which are important to the subsequent analysis. For the ease of presentation, we first review the quadratic penalization of nonlinear GNEPs based on the augmented Lagrangian function. In particular, we consider the augmented Lagrangian function for each player's minimization problem given by

$$\mathcal{L}_\nu(x, \lambda, \mu) = \theta_\nu(x) + \left(\sum_{s \in \mathcal{I}_\nu} \frac{\beta^s}{2} \|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda^s}{\beta^s}\}\|^2 + \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu^s}{\rho^s}\|^2 \right)$$

where $(\beta^s, \rho^s) \in \mathbb{R}_+ \times \mathbb{R}_+$ and $(\lambda^s, \mu^s) \in \mathbb{R}^{m_s} \times \mathbb{R}^{e_s}$ stand for penalty parameters and Lagrangian multipliers associated with inequality and equality constraints. We define the functions $g_2 : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_2 : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$g_2(x) = \sum_{s=1}^S \frac{\beta^s}{2} \|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda^s}{\beta^s}\}\|^2, \quad h_2(x) = \sum_{s=1}^S \frac{\rho^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu^s}{\rho^s}\|^2$$

Concatenating the first-order optimality conditions of minimizing the augmented Lagrangian function for each player, we aim at solving the following VI ($\widehat{X}$ is convex and compact with a diameter $D > 0$):

$$\text{Find } x \in \widehat{X}: \quad (x' - x)^\top (v(x) + \nabla g_2(x) + \nabla h_2(x)) \geq 0, \quad \text{for all } x' \in \widehat{X} \tag{18}$$

As an immediate consequence of Proposition 2, we obtain the following lemma. We omit the proof since it is the same as that of Lemma 4.

Lemma 7 (Lemma 4 restated)**.** *The gradients of the functions g_2 and h_2 are also ℓ_β-Lipschitz continuous and ℓ_ρ-Lipschitz continuous respectively, where $\ell_\beta := \sum_{s=1}^S \beta^s \|A_s\|^2$ and $\ell_\rho := \sum_{s=1}^S \rho^s \|E_s\|^2$.*

Combining this lemma with Theorem 2 and 3, we get the following lemma for the subroutine in Algorithm 2. The proof is omitted since it is the same as that of Lemma 5.

Lemma 8 (Lemma 5 restated)**.** *Suppose that $x_{k+1} = \text{AMP}((\beta^s_{k+1}, \rho^s_{k+1})_{s\in[S]}, (\lambda^s_k, \mu^s_k)_{s\in[S]}, \delta_{k+1}, x_k)$ in the sense that x_{k+1} is a δ_{k+1}-approximate solution to the structured VI in Eq.* (18) *with $(\beta^s, \rho^s) = (\beta^s_{k+1}, \rho^s_{k+1})$ and $(\lambda^s, \mu^s) = (\lambda^s_k, \mu^s_k)$ such that*

$$\max_{x\in\widehat{X}} \left\{ g_2(x_{k+1}) + h_2(x_{k+1}) - g_2(x) - h_2(x) + (x_{k+1} - x)^\top v(x) \right\} \leq \delta_{k+1}$$

The required number of gradient evaluations at the k^{th} iteration (for $k \geq 1$) is bounded by

$$N_k = \begin{cases} O\left(\sqrt{\frac{\sum_{s=1}^S (\beta^s_k \|A_s\|^2 + \rho^s_k \|E_s\|^2) D^2}{\delta_k}} + \frac{\sqrt{N}\ell_\theta D^2}{\delta_k} \right), & \text{if } \alpha = 0 \\ O\left(\left(\sqrt{\frac{\sum_{s=1}^S (\beta^s_k \|A_s\|^2 + \rho^s_k \|E_s\|^2)}{\alpha}} + \frac{\sqrt{N}\ell_\theta}{\alpha} \right) \log\left(\frac{(\sqrt{N}\ell_\theta + \ell_G) D^2}{\delta_k} \right) \right), & \text{if } \alpha > 0 \end{cases}$$

To establish the convergence of the AMP-AL algorithm, we provide the following lemma.

Lemma 9. *Suppose that $\{(\lambda^s_k, \mu^s_k)_{s\in[S]}\}_{k\geq 1}$ is generated by the AMP-AL algorithm. Then, for any $(\lambda^s, \mu^s) \in \mathbb{R}^{m_s}_+ \times \mathbb{R}^{e_s}$, we have*

$$\frac{1}{2\beta^s_{k+1}} (\|\lambda^s_{k+1} - \lambda^s\|^2 - \|\lambda^s_k - \lambda^s\|^2 + \|\lambda^s_{k+1} - \lambda^s_k\|^2) \leq (\lambda^s_{k+1} - \lambda^s)^\top (A_s x^{\mathcal{N}_s}_{k+1} - b_s)$$
$$\frac{1}{2\rho^s_{k+1}} (\|\mu^s_{k+1} - \mu^s\|^2 - \|\mu^s_k - \mu^s\|^2 + \|\mu^s_{k+1} - \mu^s_k\|^2) = (\mu^s_{k+1} - \mu^s)^\top (E_s x^{\mathcal{N}_s}_{k+1} - d_s)$$

We continue with studying the per-iteration progress of the AMP-AL algorithm.

Lemma 10. *Suppose that $x_{k+1} = \text{AMP}((\beta^s_{k+1}, \rho^s_{k+1})_{s\in[S]}, (\lambda^s_k, \mu^s_k)_{s\in[S]}, \delta_{k+1}, x_k)$ in the sense that x_{k+1} is a δ_{k+1}-approximate solution to the structured VI in Eq.* (18) *with $(\beta^s, \rho^s) = (\beta^s_{k+1}, \rho^s_{k+1})$ and $(\lambda^s, \mu^s) = (\lambda^s_k, \mu^s_k)$ such that*

$$\max_{x\in\widehat{X}} \left\{ g_2(x_{k+1}) + h_2(x_{k+1}) - g_2(x) - h_2(x) + (x_{k+1} - x)^\top v(x) \right\} \leq \delta_{k+1}$$

Then, $x_T \in \widehat{X}$ satisfies that

$$\|\max\{0, A_s x^{\mathcal{N}_s}_T - b_s\}\| \leq \widehat{\epsilon}_1, \quad \|E_s x^{\mathcal{N}_s}_T - d_s\| \leq \widehat{\epsilon}_1, \quad \text{for all } s \in [S] \tag{19}$$

where $\widehat{\epsilon}_1 > 0$ *is defined by*

$$\widehat{\epsilon}_1 = \frac{1}{\gamma^T}\left(\delta_0 T + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s}\right)\right)$$

and for all $\nu \in \mathcal{N}$*, we have*

$$(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, x_T^{-\nu}) \leq \widehat{\epsilon}_2$$

for all $x^\nu \in \widehat{X}_\nu$ *satisfying that*

$$\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \leq \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \leq \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu \quad (20)$$

where $\widehat{\epsilon}_2 > 0$ *is defined by*

$$\widehat{\epsilon}_2 = \widehat{\epsilon}_1 + \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s=1}^S \beta_0^s + \rho_0^s\right) + \epsilon^2\left(2\delta_0(T-1) + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2 + \|\overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2 + \|\overline{\lambda}^s\|^2}{\beta_0^s}\right)\right)$$

We are ready for the

Proof of Theorem 5. Fixing a sufficiently small $\epsilon \in (0,1)$, we have $x_T \in \widehat{X}$ is an ϵ-solution of an nonlinear GNEP if

$$\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \leq \epsilon, \quad \|E_s x_T^{\mathcal{N}_s} - d_s\| \leq \epsilon, \quad \text{for all } s \in [S]$$

and for all $\nu \in \mathcal{N}$, we have

$$(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, \overline{x}^{-\nu}) \leq C\epsilon$$

for all $x^\nu \in \widehat{X}_\nu$ satisfying that

$$\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \leq \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \leq \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu$$

By Lemma 10, it suffices to guarantee that $T > 0$ satisfies the following conditions:

$$\begin{aligned} \epsilon &\geq \frac{1}{\gamma^T}\left(\delta_0 T + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s}\right)\right) \\ C\epsilon &\geq \epsilon^2\left(2\delta_0(T-1) + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2 + \|\overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2 + \|\overline{\lambda}^s\|^2}{\beta_0^s}\right)\right) \\ &\quad + \widehat{\epsilon}_1 + \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s=1}^S \beta_0^s + \rho_0^s\right) \end{aligned} \quad (21)$$

Suppose that we set $T > 0$ as

$$T = 1 + \left\lfloor \frac{1}{\log(\gamma)}\left\{\log\log\left(\frac{2+2\delta_0}{\epsilon}\right) + \log\left(\frac{2+2\delta_0}{\epsilon}\right) + \log\left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{\beta_0^s}\right)\right\}\right\rfloor$$

Then, Eq. (21) holds true with a positive constant $C > 0$ given by

$$C = 1 + \left\lceil \left(\frac{1}{2} + (1+\delta_0)\left(\sum_{s=1}^{S} \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{\beta_0^s}\right)\log\left(\frac{2+2\delta_0}{\epsilon}\right)\right)\left(\sum_{s=1}^{S} \beta_0^s + \rho_0^s\right)\right\rceil$$
$$+ \left\lceil \frac{1}{\log(\gamma)}\left\{2\delta_0\left(\log\log\left(\frac{2+2\delta_0}{\epsilon}\right) + \log\left(\frac{2+2\delta_0}{\epsilon}\right) + \log\left(\sum_{s=1}^{S} \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{\beta_0^s}\right)\right)\right\}\right.$$
$$\left. + \left(\sum_{s=1}^{S} \frac{\|\mu_0^s - \overline{\mu}^s\|^2 + \|\overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2 + \|\overline{\lambda}^s\|^2}{\beta_0^s}\right)\right\rceil$$

Since $\beta_k^s = \gamma^k \beta_0^s$ and $\rho_k^s = \gamma^k \rho_0^s$ for all $s \in [S]$ and $\delta_k = \frac{\delta_0}{\gamma^k}$, Lemma 8 guarantees that the required number of gradient evaluations at the k^{th} iteration is bounded by

$$N_k = \left\{ \begin{array}{ll} O\left(\gamma^k \left(\sqrt{\frac{\sum_{s=1}^S (\beta_0^s \|A_s\|^2 + \rho_0^s \|E_s\|^2) D^2}{\delta_0}} + \frac{\sqrt{N}\ell_\theta D^2}{\delta_0}\right)\right), & \text{if } \alpha = 0 \\ O\left(\left(\gamma^{\frac{k}{2}} \sqrt{\frac{\sum_{s=1}^S (\beta_0^s \|A_s\|^2 + \rho_0^s \|E_s\|^2)}{\alpha}} + \frac{\sqrt{N}\ell_\theta}{\alpha}\right)\log\left(\frac{\gamma^k(\sqrt{N}\ell_\theta + \ell_G) D^2}{\delta_0}\right)\right), & \text{if } \alpha > 0 \end{array} \right.$$

Therefore, we conclude that the total number of gradient evaluations required to return an ϵ-solution of an nonlinear GNEP is

$$N_{\text{grad}} = \sum_{k=1}^{T-1} N_k = \left\{ \begin{array}{ll} O\left(\epsilon^{-1}\log(1/\epsilon)\right), & \text{if } \mu = 0 \\ O\left(\epsilon^{-1/2}\log(1/\epsilon)\right), & \text{if } \mu > 0 \end{array} \right.$$

This completes the proof. □

5 Conclusions

We have presented an inquiry into equilibrium computation in a special class of nonlinear generalized Nash equilibrium problems (GNEPs). Focusing on simple gradient-based schemes, we investigated algorithms based on a quadratic penalty method and an augmented Lagrangian method, respectively. Both of these algorithms make use of the accelerated mirror-prox algorithm as an inner loop. We established global convergence rate estimates for both algorithms for solving monotone and strongly monotone nonlinear GNEPs: the complexity bounds are $\widetilde{O}(\epsilon^{-1})$ and $\widetilde{O}(\epsilon^{-1/2})$ in monotone and strongly monotone cases, respectively, in terms of the number of gradient evaluations. We highlighted the practical relevance of these theoretical rates in experiments with several real-world datasets.

It is worth mentioning that there are many numerical variations of our work that might lead to quantitative improvements. Aside from fine-tuning of parameters, a more detailed analysis of penalized NEPs which occur in our method may shed light on the possibility of a more effective subproblem solver than accelerated mirror-prox. Future directions include the identification of other nonlinear GNEPs which can be solved by an algorithm with a global convergence rate estimate and the investigation of the sequential quadratic programming (SQP) method [Tolle(1995)] for solving GNEPs.

References

[Andreani et al.(2005)] R. Andreani, J. M. Martínez, and M. L. Schuverdt. On the relation between constant positive linear dependence condition and quasinormality constraint qualification. *Journal of Optimization Theory and Applications*, 125(2):473–483, 2005.

[Andreani et al.(2016)] R. Andreani, J. M. Martínez, A. Ramos, and P. J. S. Silva. A cone-continuity constraint qualification and algorithmic consequences. *SIAM Journal on Optimization*, 26(1):96–110, 2016.

[Ardagna et al.(2011)] D. Ardagna, B. Panicucci, and M. Passacantando. A game theoretic formulation of the service provisioning problem in cloud systems. In *WWW*, pages 177–186, 2011.

[Ardagna et al.(2015)] D. Ardagna, M. Ciavotta, and M. Passacantando. Generalized Nash equilibria for the service provisioning problem in multi-cloud systems. *IEEE Transactions on Services Computing*, 10(3):381–395, 2015.

[Arrow and Debreu(1954)] K. J. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22:265–290, 1954.

[Aybat and Iyengar(2011)] N. S. Aybat and G. Iyengar. A first-order smoothed penalty method for compressed sensing. *SIAM Journal on Optimization*, 21(1):287–313, 2011.

[Ba and Pang(2022)] Q. Ba and J-S. Pang. Exact penalization of generalized Nash equilibrium problems. *Operations Research*, 70(3):1448–1464, 2022.

[Ban et al.(2019)] X. J. Ban, M. Dessouky, J-S. Pang, and R. Fan. A general equilibrium model for transportation systems with e-hailing services and flow congestion. *Transportation Research Part B: Methodological*, 129:273–304, 2019.

[Bauschke and Combettes(2017)] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

[Ben-Tal and Nemirovski(2001)] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

[Bertsekas(1976)] D. P. Bertsekas. On penalty and multiplier methods for constrained minimization. *SIAM Journal on Control and Optimization*, 14(2):216–235, 1976.

[Bianchi et al.(2022)] M. Bianchi, G. Belgioioso, and S. Grammatico. Fast generalized Nash equilibrium seeking under partial-decision information. *Automatica*, 136:110080, 2022.

[Breton et al.(2006)] M. Breton, G. Zaccour, and M. Zahaf. A game-theoretic formulation of joint implementation of environmental projects. *European Journal of Operational Research*, 168(1):221–239, 2006.

[Brückner and Scheffer(2009)] M. Brückner and T. Scheffer. Nash equilibria of static prediction games. In *NIPS*, pages 171–179, 2009.

[Brückner et al.(2012)] M. Brückner, C. Kanzow, and T. Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

[Bueno et al.(2019)] L. F. Bueno, G. Haeser, and F. N. Rojas. Optimality conditions and constraint qualifications for generalized Nash equilibrium problems and their practical implications. *SIAM Journal on Optimization*, 29(1):31–54, 2019.

[Carminati et al.(2022)] L. Carminati, F. Cacciamani, M. Ciccone, and N. Gatti. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In *ICML*, pages 2638–2657. PMLR, 2022.

[Celli and Gatti(2018)] A. Celli and N. Gatti. Computational results for extensive-form adversarial team games. In *AAAI*, pages 965–972, 2018.

[Celli et al.(2019)] A. Celli, M. Ciccone, R. Bongo, and N. Gatti. Coordination in adversarial sequential team games via multi-agent deep reinforcement learning. *ArXiv Preprint: 1912.07712*, 2019.

[Chan and Pang(1982)] D. Chan and J. S. Pang. The generalized quasi-variational inequality problem. *Mathematics of Operations Research*, 7(2):211–222, 1982.

[Chen et al.(2017)] Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[Cominetti et al.(2012)] R. Cominetti, F. Facchinei, and J. B. Lasserre. *Modern Optimization Modelling Techniques*. Springer Science & Business Media, 2012.

[Debreu(1952)] G. Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

[Debreu(1959)] G. Debreu. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium.* Yale University Press, 1959.

[Diakonikolas(2020)] J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *COLT*, pages 1428–1451. PMLR, 2020.

[Dreves(2017)] A. Dreves. Computing all solutions of linear generalized Nash equilibrium problems. *Mathematical Methods of Operations Research*, 85(2):207–221, 2017.

[Dreves and Sudermann-Merx(2016)] A. Dreves and N. Sudermann-Merx. Solving linear generalized Nash equilibrium problems numerically. *Optimization Methods and Software*, 31(5):1036–1063, 2016.

[Dreves et al.(2011)] A. Dreves, F. Facchinei, C. Kanzow, and S. Sagratella. On the solution of the KKT conditions of generalized Nash equilibrium problems. *SIAM Journal on Optimization*, 21(3):1082–1108, 2011.

[Dreves et al.(2013)] A. Dreves, A. von Heusinger, C. Kanzow, and M. Fukushima. A globalized Newton method for the computation of normalized Nash equilibria. *Journal of Global Optimization*, 56(2): 327–340, 2013.

[Facchinei and Kanzow(2009)] F. Facchinei and C. Kanzow. Penalty methods for the solution of generalized Nash equilibrium problems (with complete test problems). *Sapienza University of Rome*, 2009.

[Facchinei and Kanzow(2010a)] F. Facchinei and C. Kanzow. Generalized Nash equilibrium problems. *Annals of Operations Research*, 175(1):177–211, 2010a.

[Facchinei and Kanzow(2010b)] F. Facchinei and C. Kanzow. Penalty methods for the solution of generalized Nash equilibrium problems. *SIAM Journal on Optimization*, 20(5):2228–2253, 2010b.

[Facchinei and Lampariello(2011)] F. Facchinei and L. Lampariello. Partial penalization for the solution of generalized Nash equilibrium problems. *Journal of Global Optimization*, 50(1):39–57, 2011.

[Facchinei and Pang(2007)] F. Facchinei and J-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Science & Business Media, 2007.

[Facchinei et al.(2007)] F. Facchinei, A. Fischer, and V. Piccialli. On generalized Nash games and variational inequalities. *Operations Research Letters*, 35(2):159–164, 2007.

[Facchinei et al.(2009)] F. Facchinei, A. Fischer, and V. Piccialli. Generalized Nash equilibrium problems and Newton methods. *Mathematical Programming*, 117(1):163–194, 2009.

[Facchinei et al.(2015)] F. Facchinei, C. Kanzow, S. Karl, and S. Sagratella. The semismooth Newton method for the solution of quasi-variational inequalities. *Computational Optimization and Applications*, 62(1): 85–109, 2015.

[Fan and Yuan(2005)] J. Fan and Y. Yuan. On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption. *Computing*, 74(1):23–39, 2005.

[Farina et al.(2021)] G. Farina, A. Celli, N. Gatti, and T. Sandholm. Connecting optimal ex-ante collusion in teams to extensive-form correlation: Faster algorithms and positive complexity results. In *ICML*, pages 3164–3173. PMLR, 2021.

[Farkas(1902)] J. Farkas. Theorie der einfachen ungleichungen. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1902(124):1–27, 1902.

[Fischer et al.(2014)] A. Fischer, M. Herrich, and K. Schönefeld. Generalized Nash equilibrium problems: Recent advances and challenges. *Pesquisa Operacional*, 34(3):521–558, 2014.

[Fischer et al.(2016)] A. Fischer, M. Herrich, A. F. Izmailov, and M. V. Solodov. A globally convergent LP-Newton method. *SIAM Journal on Optimization*, 26(4):2012–2033, 2016.

[Flåm and Antipin(1996)] S. D. Flåm and A. S. Antipin. Equilibrium programming using proximal-like algorithms. *Mathematical Programming*, 78(1):29–41, 1996.

[Franci and Grammatico(2022)] B. Franci and S. Grammatico. Stochastic generalized nash equilibrium seeking under partial-decision information. *Automatica*, 137:110101, 2022.

[Fukushima(2011)] M. Fukushima. Restricted generalized Nash equilibria and controlled penalty algorithm. *Computational Management Science*, 8(3):201–218, 2011.

[Galli et al.(2018)] L. Galli, C. Kanzow, and M. Sciandrone. A nonmonotone trust-region method for generalized Nash equilibrium and related problems with strong convergence properties. *Computational Optimization and Applications*, 69(3):629–652, 2018.

[Han et al.(2012)] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes. *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2012.

[Harker(1991)] P. T. Harker. Generalized Nash games and quasi-variational inequalities. *European Journal of Operational Research*, 54(1):81–94, 1991.

[Hestenes(1969)] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[Hobbs and Pang(2007)] B. F. Hobbs and J-S. Pang. Nash-Cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints. *Operations Research*, 55(1):113–127, 2007.

[Izmailov and Solodov(2014)] A. F. Izmailov and M. V. Solodov. On error bounds and Newton-type methods for generalized Nash equilibrium problems. *Computational Optimization and Applications*, 59(1-2): 201–218, 2014.

[Jiang et al.(2019)] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72 (1):115–157, 2019.

[Jing-Yuan and Smeers(1999)] W. Jing-Yuan and Y. Smeers. Spatial oligopolistic electricity models with Cournot generators and regulated transmission prices. *Operations Research*, 47(1):102–112, 1999.

[Juditsky et al.(2011)] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Kalogiannis et al.(2022)] F. Kalogiannis, I. Anagnostides, I. Panageas, E-V. Vlatakis-Gkaragkounis, V. Chatziafratis, and S. Stavroulakis. Efficiently computing Nash equilibria in adversarial team Markov games. *ArXiv Preprint: 2208.02204*, 2022.

[Kanzow(2016)] C. Kanzow. On the multiplier-penalty-approach for quasi-variational inequalities. *Mathematical Programming*, 160(1):33–63, 2016.

[Kanzow and Steck(2016)] C. Kanzow and D. Steck. Augmented Lagrangian methods for the solution of generalized Nash equilibrium problems. *SIAM Journal on Optimization*, 26(4):2034–2058, 2016.

[Kanzow and Steck(2018)] C. Kanzow and D. Steck. Augmented Lagrangian and exact penalty methods for quasi-variational inequalities. *Computational Optimization and Applications*, 69(3):801–824, 2018.

[Karush(1939)] W. Karush. Minima of functions of several variables with inequalities as side conditions. *Master's Thesis, Department of Mathematics, University of Chicago*, 1939.

[Korpelevich(1976)] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Kotsalis et al.(2022)] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[Krawczyk(2005)] J. B. Krawczyk. Coupled constraint Nash equilibria in environmental games. *Resource and Energy Economics*, 27(2):157–181, 2005.

[Krawczyk and Uryasev(2000)] J. B. Krawczyk and S. Uryasev. Relaxation algorithms to find Nash equilibria with economic applications. *Environmental Modeling & Assessment*, 5(1):63–73, 2000.

[Kuhn and Tucker(1951)] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[Kummer(1988)] B. Kummer. Newton's method for non-differentiable functions. *Advances in Mathematical Optimization*, 45(1988):114–125, 1988.

[Lan and Monteiro(2013)] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1):115–139, 2013.

[Lan and Monteiro(2016)] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

[Lan et al.(2011)] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $o(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[Levenberg(1944)] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

[Malitsky(2015)] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[Mangasarian and Fromovitz(1967)] O. L. Mangasarian and S. Fromovitz. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17(1):37–47, 1967.

[Marquardt(1963)] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[Mertikopoulos and Zhou(2019)] P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019.

[Myerson(2013)] R. B. Myerson. *Game Theory*. Harvard University Press, 2013.

[Nabetani et al.(2011)] K. Nabetani, P. Tseng, and M. Fukushima. Parametrized variational inequality approaches to generalized Nash equilibrium problems with shared constraints. *Computational Optimization and Applications*, 48(3):423–452, 2011.

[Nash(1950)] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[Nash(1951)] J. F. Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951.

[Necoara et al.(2019)] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, 34(2):305–335, 2019.

[Nemirovski(2004)] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[Nesterov(2007)] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[Nesterov and Scrimali(2011)] Y. Nesterov and L. Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems*, 31(4):1383, 2011.

[Osborne and Rubinstein(1994)] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[Ouyang and Xu(2021)] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

[Pang and Fukushima(2005)] J-S. Pang and M. Fukushima. Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games. *Computational Management Science*, 2(1):21–56, 2005.

[Pang et al.(2008)] J-S. Pang, G. Scutari, F. Facchinei, and C. Wang. Distributed power allocation with rate constraints in Gaussian parallel interference channels. *IEEE Transactions on Information Theory*, 54(8):3471–3489, 2008.

[Pang et al.(2010)] J-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei. Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach. *IEEE Transactions on Signal Processing*, 58(6):3251–3271, 2010.

[Powell(1969)] M. J. D. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

[Qi and Sun(1993)] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1):353–367, 1993.

[Qi and Wei(2000)] L. Qi and Z. Wei. On the constant positive linear dependence condition and its application to SQP methods. *SIAM Journal on Optimization*, 10(4):963–981, 2000.

[Robinson(1993a)] S. M. Robinson. Shadow prices for measures of effectiveness, I: Linear model. *Operations Research*, 41(3):518–535, 1993a.

[Robinson(1993b)] S. M. Robinson. Shadow prices for measures of effectiveness, II: General model. *Operations Research*, 41(3):536–548, 1993b.

[Rockafellar(1973a)] R. T. Rockafellar. A dual approach to solving nonlinear programming problems by unconstrained optimization. *Mathematical Programming*, 5(1):354–373, 1973a.

[Rockafellar(1973b)] R. T. Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and applications*, 12(6):555–562, 1973b.

[Rockafellar(1976)] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

[Rosen(1965)] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

[Schiro et al.(2013)] D. A. Schiro, J-S. Pang, and U. V. Shanbhag. On the solution of affine generalized Nash equilibrium problems with shared constraints by Lemke's method. *Mathematical Programming*, 142(1):1–46, 2013.

[Scutari et al.(2014)] G. Scutari, F. Facchinei, J-S. Pang, and D. P. Palomar. Real and complex monotone communication games. *IEEE Transactions on Information Theory*, 60(7):4197–4231, 2014.

[Stein and Sudermann-Merx(2016)] O. Stein and N. Sudermann-Merx. The cone condition and nonsmoothness in linear generalized Nash games. *Journal of Optimization Theory and Applications*, 170(2): 687–709, 2016.

[Stein and Sudermann-Merx(2018)] O. Stein and N. Sudermann-Merx. The noncooperative transportation problem and linear generalized Nash games. *European Journal of Operational Research*, 266(2):543–553, 2018.

[Tolle(1995)] J. W. Tolle. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1995.

[Tseng(2008)] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2:3, 2008.

[Uryas' ev and Rubinstein(1994)] S. Uryas' ev and R. Y. Rubinstein. On relaxation algorithms in computation of noncooperative equilibria. *IEEE Transactions on Automatic Control*, 39(6):1263–1267, 1994.

[Von Heusinger and Kanzow(2009)] A. Von Heusinger and C. Kanzow. Relaxation methods for generalized Nash equilibrium problems with inexact line search. *Journal of Optimization Theory and Applications*, 143(1):159–183, 2009.

[von Heusinger et al.(2012)] A. von Heusinger, C. Kanzow, and M. Fukushima. Newton's method for computing a normalized equilibrium in the generalized Nash game through fixed point formulation. *Mathematical Programming*, 132(1):99–123, 2012.

[Xu(2021)] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming*, 185(1):199–244, 2021.

[Yamashita and Fukushima(2001)] N. Yamashita and M. Fukushima. On the rate of convergence of the Levenberg-Marquardt method. In *Topics in Numerical Analysis*, pages 239–249. Springer, 2001.

[Yu and Neely(2017)] H. Yu and M. J. Neely. A simple parallel algorithm with an o(1/t) convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.

A Variational Equilibrium

There is an important subclass of solutions, referred to as *variational solutions* (or normalized equilibrium), which is characterized by the condition that the same Lagrange multipliers are associated with the constraints in each player's problem. This condition is motivated by economic models where the Lagrange multipliers in each player's problem have an interpretation as shadow prices of the resources and are thus equal. Computational algorithms have been developed that specifically target the variational solutions of GNEPs, and these algorithms have been analyzed in terms of asymptotic global convergence. But iteration complexity analyses have not yet been provided for these algorithms.

Definition 6. *We say a point* $\overline{\mathbf{x}}$ *is a variational solution of an nonlinear GNEP if*

$$\overline{x} \in \widehat{X} \cap \left\{x \in \mathbb{R}^n : A_s x^{\mathcal{N}_s} \leq b_s, \ E_s x^{\mathcal{N}_s} = d_s, \ \text{for all } s \in [S]\right\}$$

and there exist some Lagrangian multipliers $\overline{\lambda}^s \geq 0$ *and* $\overline{\mu}^s$ *such that*

$$\text{for all } \nu \in \mathcal{N} \text{ we have } \left\{ \begin{array}{r} (\overline{x}^\nu - x^\nu)^\top (v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu}((A_s^\nu)^\top \overline{\lambda}^s + (E_s^\nu)^\top \overline{\mu}^s)) \leq 0, \quad \forall x^\nu \in \widehat{X}_\nu \\ (\overline{\lambda}^s)^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0, \quad \forall s \in \mathcal{I}_\nu \end{array} \right.$$

Based on the KKT conditions, the set of variational equilibria of monotone nonlinear GNEPs coincides with the solution set of the following VI:

$$\text{Find } \overline{x} \in X : \quad (x - \overline{x})^\top v(\overline{x}) \geq 0, \ \text{for all } x \in X \tag{22}$$

where X is defined by

$$X = \widehat{X} \cap \left\{x \in \mathbb{R}^n : A_s x^{\mathcal{N}_s} \leq b_s, \ E_s x^{\mathcal{N}_s} = d_s, \ \text{for all } s \in [S]\right\}$$

By reading off classical results from the VI literature, we can derive an existence result for variational solutions of nonlinear GNEPs (cf. Definition 1) due to the compactness of $\widehat{X}$. Formally, we have

Proposition 3. *There exists at least one variational solution of any nonlinear GNEP.*

Proof. Since at least one solution of the nonlinear GNEP exists, X will be an nonempty, convex and compact set. Therefore, the VI in Eq. (22) must have a solution and hence there exists at least one variational equilibrium of the nonlinear GNEP in the sense of Definition 6. □

B Deferred Proofs of Auxiliary Lemmas

B.1 Proof of Lemma 2

Proof of Lemma 2. From the update rule for z_{k+1} and w_{k+1}, we have

$$\gamma_k(z_{k+1} - z)^\top (F(w_k) + \nabla G(z_k^{\text{md}})) \leq \frac{1}{2}(\|z - w_k\|^2 - \|z_{k+1} - w_k\|^2 - \|z - z_{k+1}\|^2)$$
$$\gamma_k(w_{k+1} - z)^\top (F(z_{k+1}) + \nabla G(z_k^{\text{md}})) \leq \frac{1}{2}(\|z - w_k\|^2 - \|w_{k+1} - w_k\|^2 - \|z - w_{k+1}\|^2)$$

Setting $z = w_{k+1}$ in the first inequality and adding the resulting inequality to the second inequality yields that

$$\begin{aligned} & \gamma_k(z_{k+1} - w_{k+1})^\top (F(w_k) + \nabla G(z_k^{\text{md}})) + \gamma_k(w_{k+1} - z)^\top (F(z_{k+1}) + \nabla G(z_k^{\text{md}})) \\ \leq \; & \frac{1}{2}(\|z - w_k\|^2 - \|z - w_{k+1}\|^2 - \|z_{k+1} - w_k\|^2 - \|w_{k+1} - z_{k+1}\|^2) \end{aligned}$$

Equivalently, we have

$$\begin{aligned} \gamma_k(z_{k+1} - z)^\top (F(z_{k+1}) + \nabla G(z_k^{\text{md}})) \leq \gamma_k(z_{k+1} - w_{k+1})^\top (F(z_{k+1}) - F(w_k)) \\ + \frac{1}{2}(\|z - w_k\|^2 - \|z - w_{k+1}\|^2 - \|z_{k+1} - w_k\|^2 - \|w_{k+1} - z_{k+1}\|^2) \end{aligned}$$

By Young's inequality and the fact that F is ℓ_F-Lipschitz, we have

$$(z_{k+1} - w_{k+1})^\top (F(z_{k+1}) - F(w_k)) \leq \frac{1}{2\gamma_k}\|z_{k+1} - w_{k+1}\|^2 + \frac{\gamma_k \ell_F^2}{2}\|z_{k+1} - w_k\|^2$$

Putting these pieces together yields the desired inequality. $\square$

B.2 Proof of Lemma 3

Proof of Lemma 3. From the update rule for z_k^{md} and z_{k+1}^{ag}, we have $z_{k+1}^{\text{ag}} - z_k^{\text{md}} = \alpha_k(z_{k+1} - w_k)$. Since G is ℓ_G-smooth, we have

$$\begin{aligned} G(z_{k+1}^{\text{ag}}) &\leq G(z_k^{\text{md}}) + (z_{k+1}^{\text{ag}} - z_k^{\text{md}})^\top \nabla G(z_k^{\text{md}}) + \frac{\ell_G}{2}\|z_{k+1}^{\text{ag}} - z_k^{\text{md}}\|^2 \\ &\leq \; G(z_k^{\text{md}}) + (z_{k+1}^{\text{ag}} - z_k^{\text{md}})^\top \nabla G(z_k^{\text{md}}) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 \end{aligned}$$

Since $z_{k+1}^{\text{ag}} = (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k z_{k+1}$, we have

$$\begin{aligned} & G(z_{k+1}^{\text{ag}}) \leq (1 - \alpha_k)\left(G(z_k^{\text{md}}) + (z_k^{\text{ag}} - z_k^{\text{md}})^\top \nabla G(z_k^{\text{md}})\right) \\ & \quad + \alpha_k\left(G(z_k^{\text{md}}) + (z - z_k^{\text{md}})^\top \nabla G(z_k^{\text{md}})\right) + \alpha_k(z_{k+1} - z)^\top \nabla G(z_k^{\text{md}}) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 \end{aligned}$$

Since G is convex, we have

$$G(z_{k+1}^{\text{ag}}) \leq (1 - \alpha_k) G(z_k^{\text{ag}}) + \alpha_k G(z) + \alpha_k(z_{k+1} - z)^\top \nabla G(z_k^{\text{md}}) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2$$

Combining the definition of Q with the above inequality and $z_{k+1}^{\text{ag}} = (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k z_{k+1}$, we have

$$\begin{aligned} & Q(z_{k+1}^{\text{ag}}, z) - (1 - \alpha_k) Q(z_k^{\text{ag}}, z) \\ \leq \; & G(z_{k+1}^{\text{ag}}) - (1 - \alpha_k) G(z_k^{\text{ag}}) - \alpha_k G(z) + (z_{k+1}^{\text{ag}} - z)^\top F(z) - (1 - \alpha_k)(z_k^{\text{ag}} - z)^\top F(z) \\ \leq \; & \alpha_k(z_{k+1} - z)^\top \nabla G(z_k^{\text{md}}) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 + (z_{k+1}^{\text{ag}} - z - (1 - \alpha_k)(z_k^{\text{ag}} - z))^\top F(z) \\ = \; & \alpha_k(z_{k+1} - z)^\top \nabla G(z_k^{\text{md}}) + \frac{\alpha_k^2 \ell_G}{2}\|z_{k+1} - w_k\|^2 + \alpha_k(z_{k+1} - z)^\top F(z) \end{aligned}$$

Since F is α-strongly monotone, we have

$$(z_{k+1} - z)^\top F(z) \leq (z_{k+1} - z)^\top F(z_{k+1}) - \alpha\|z_{k+1} - z\|^2$$

Putting these pieces together yields the desired inequality. $\square$

B.3 Proof of Lemma 4

Proof of Lemma 4. By the definition, the differentiability of g_1 and h_1 comes from Proposition 2. By the chain rule and letting $d_{\mathcal{C}}$ be the distance function to $\mathcal{C}$ with respect to $\|\cdot\|$, we have

$$\nabla_\nu g_1(x) = \sum_{s=1}^S \beta^s (A_s^\nu)^\top (\max\{0, A_s x^{\mathcal{N}_s} - b_s\}), \quad \nabla_\nu h_1(x) = \sum_{s=1}^S \rho^s (E_s^\nu)^\top (E_s x^{\mathcal{N}_s} - d_s)$$

Together with Proposition 2, we have

$$\|\nabla g_1(\widetilde{x}) - \nabla g_1(x)\| \leq \sum_{s=1}^S \beta^s \|A_s\| \|A_s \widetilde{x}^{\mathcal{N}_s} - A_s x^{\mathcal{N}_s}\| \leq \ell_\beta \|\widetilde{x} - x\|$$

$$\|\nabla h_1(\widetilde{x}) - \nabla h_1(x)\| \leq \sum_{s=1}^S \rho^s \|E_s\| \|E_s \widetilde{x}^{\mathcal{N}_s} - E_s x^{\mathcal{N}_s}\| \leq \ell_\rho \|\widetilde{x} - x\|$$

This completes the proof. $\square$

B.4 Proof of Lemma 5

Proof of Lemma 5. By Lemma 4, the VI in Eq. (14) is in the form of Eq. (8) with $\ell_F = \sqrt{N}\ell_\theta$ and $\ell_G = \sum_{s=1}^S (\beta_k^s \|A_s\|^2 + \rho_k^s \|E_s\|^2)$. Applying Theorem 2 and 3, we obtain the desired upper bound for the required number of gradient evaluations at the k^{th} iteration. $\square$

B.5 Proof of Lemma 6

Proof of Lemma 6. By Proposition 3, there exists a triple $(\overline{x} \in \widehat{X}, \overline{\lambda} \geq 0, \overline{\mu})$ such that

$$A_s \overline{x}^{\mathcal{N}_s} \leq b_s, \ E_s \overline{x}^{\mathcal{N}_s} = d_s, \ (\overline{\lambda}^s)^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0, \text{ for all } s \in [S] \tag{23}$$

and the following VI holds true for all $\nu \in \mathcal{N}$,

$$(\overline{x}^\nu - x^\nu)^\top \left(v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu} ((A_s^\nu)^\top \overline{\lambda}^s + (E_s^\nu)^\top \overline{\mu}^s) \right) \leq 0, \quad \text{for all } x^\nu \in \widehat{X}_\nu \tag{24}$$

Note that Eq. (23) implies that $g_1(\overline{x}) = 0$ and $h_1(\overline{x}) = 0$ and we have

$$g_1(x_T) + h_1(x_T) - g_1(x) - h_1(x) + (x_T - x)^\top v(x) \leq \delta_T, \quad \text{for all } x \in \widehat{X} \tag{25}$$

Plugging $x = \overline{x}$ into Eq. (25), we have

$$g_1(x_T) + h_1(x_T) + (x_T - \overline{x})^\top v(\overline{x}) \leq \delta_T \tag{26}$$

Plugging $x^\nu = x_T^\nu$ into Eq. (24) yields that

$$(\overline{x}^\nu - x_T^\nu)^\top \left(v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu} (A_s^\nu)^\top \overline{\lambda}^s + (E_s^\nu)^\top \overline{\mu}^s \right) \leq 0, \quad \text{for all } \nu \in \mathcal{N}$$

Summing over $\nu \in \mathcal{N}$ and rearranging yields that

$$(\overline{x} - x_T)^\top v(\overline{x}) + \sum_{s=1}^S ((A_s \overline{x}^{\mathcal{N}_s} - A_s x_T^{\mathcal{N}_s})^\top \overline{\lambda}^s + (E_s \overline{x}^{\mathcal{N}_s} - E_s x_T^{\mathcal{N}_s})^\top \overline{\mu}^s) \leq 0$$

Combining this inequality with Eq. (23) and the fact that $\overline{\lambda}^s \geq 0$, we have

$$\begin{aligned}(\overline{x} - x_T)^\top v(\overline{x}) &\leq \sum_{s=1}^{S} (\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \|\overline{\lambda}^s\| + \|E_s x_T^{\mathcal{N}_s} - d_s\| \|\overline{\mu}^s\|) \\ &\leq \left(\max_{1\leq s\leq S} \|\overline{\lambda}^s\|\right) \sum_{s=1}^{S} \|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| + \left(\max_{1\leq s\leq S} \|\overline{\mu}^s\|\right) \sum_{s=1}^{S} \|E_s x_T^{\mathcal{N}_s} - d_s\| \\ &\leq \left(\max_{1\leq s\leq S} \sqrt{\frac{2\|\overline{\lambda}^s\|^2 S}{\beta_T^s}}\right) \sqrt{g_1(x_T)} + \left(\max_{1\leq s\leq S} \sqrt{\frac{2\|\overline{\mu}^s\|^2 S}{\rho_T^s}}\right) \sqrt{h_1(x_T)}\end{aligned} \tag{27}$$

Proof of Eq. (15)**.** Combining Eq. (27) with Eq. (26), we have

$$g_1(x_T) + h_1(x_T) \leq \delta_T + \left(\sqrt{2S} \max_{1\leq s\leq S} \left\{\sqrt{\frac{\|\overline{\lambda}^s\|^2}{\beta_T^s}} + \sqrt{\frac{\|\overline{\mu}^s\|^2}{\rho_T^s}}\right\}\right) \sqrt{g_1(x_T) + h_1(x_T)}$$

After some simple calculations, we have

$$g_1(x_T) + h_1(x_T) \leq 2\delta_T + 20S \left(\max_{1\leq s\leq S} \left\{\frac{\|\overline{\lambda}^s\|^2}{\beta_T^s} + \frac{\|\overline{\mu}^s\|^2}{\rho_T^s}\right\}\right)$$

Then, Eq. (15) follows from the definition of g_1, h_1 and $\widehat{\epsilon}_1$.

Proof of Eq. (16)**.** It follows from $g_1(x_T) \geq 0$, $h_1(x_T) \geq 0$ and Eq. (25) that $-g_1(x) - h_1(x) + (x_T - x)^\top v(x) \leq \delta_T$ for all $x \in \widehat{X}$. Fixing $\nu \in \mathcal{N}$, we plug $x^{-\nu} = x_T^{-\nu}$ into the above inequality and obtain that

$$\begin{aligned}&(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, x_T^{-\nu}) - \sum_{s\in\mathcal{I}_\nu} \frac{\beta_T^s}{2} \|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\|^2 \\ &- \sum_{s\in\mathcal{I}_\nu} \frac{\rho_T^s}{2} \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\|^2 \leq \delta_T, \quad \text{for all } x^\nu \in \widehat{X}_\nu\end{aligned}$$

Suppose that $x^\nu \in \widehat{X}_\nu$ satisfies Eq. (16), we obtain from $\mathcal{I}_\nu \subseteq [S]$ that

$$(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, x_T^{-\nu}) \leq \delta_T + \frac{\epsilon^2}{2} \left(\sum_{s=1}^{S} \beta_T^s + \rho_T^s\right)$$

Then, Eq. (16) follows from the definition of $\widehat{\epsilon}_2$. □

B.6 Proof of Lemma 9

Proof of Lemma 9. By the update of λ_k^s and μ_k^s in Algorithm 2, we have

$$\begin{aligned}&\frac{1}{2\beta_{k+1}^s} (\|\lambda_{k+1}^s - \lambda^s\|^2 - \|\lambda_k^s - \lambda^s\|^2 + \|\lambda_{k+1}^s - \lambda_k^s\|^2) \\ =& \frac{1}{\beta_{k+1}^s} (\lambda_{k+1}^s - \lambda^s)^\top \max\{-\lambda_k^s, \beta_{k+1}^s (A_s x_{k+1}^{\mathcal{N}_s} - b_s)\}\end{aligned}$$

and

$$\frac{1}{2\rho_{k+1}^s}(\|\mu_{k+1}^s - \mu^s\|^2 - \|\mu_k^s - \mu^s\|^2 + \|\mu_{k+1}^s - \mu_k^s\|^2) = (\mu_{k+1}^s - \mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s)$$

where the second equality is one of desired results. Let us denote I_k^s as the set of coordinates of $\lambda_k^s + \beta_{k+1}^s(A_s x_{k+1}^{\mathcal{N}_s} - b_s) \in \mathbb{R}^{m^s}$ whose values are nonnegative. This together with the update of λ_{k+1}^s yields that $(\lambda_{k+1}^s)_j = 0$ for all $j \in [m^s] \setminus I_k^s$. Then, we have

$$\begin{aligned}
&\frac{1}{\beta_{k+1}^s}(\lambda_{k+1}^s - \lambda^s)^\top \max\{-\lambda_k^s, \beta_{k+1}^s(A_s x_{k+1}^{\mathcal{N}_s} - b_s)\} \\
&= \sum_{j \in I_k^s} (\lambda_{k+1}^s - \lambda^s)_j (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j + \sum_{j \in [m^s] \setminus I_k^s} (-\lambda^s)_j \left(-\frac{\lambda_k^s}{\beta_{k+1}^s}\right)_j
\end{aligned}$$

Since $\lambda^s \geq 0$ and $\left(-\frac{\lambda_k^s}{\beta_{k+1}^s}\right)_j > (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j$ for all $j \in [m^s] \setminus I_k^s$, we have

$$\sum_{j \in [m^s] \setminus I_k^s} (-\lambda^s)_j \left(-\frac{\lambda_k^s}{\beta_{k+1}^s}\right)_j \leq \sum_{j \in [m^s] \setminus I_k^s} (-\lambda^s)_j (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j = \sum_{j \in [m^s] \setminus I_k^s} (\lambda_{k+1}^s - \lambda^s)_j (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j$$

Putting these pieces together yields that

$$\frac{1}{\beta_{k+1}^s}(\lambda_{k+1}^s - \lambda^s)^\top \max\{-\lambda_k^s, \beta_{k+1}^s(A_s x_{k+1}^{\mathcal{N}_s} - b_s)\} \leq (\lambda_{k+1}^s - \lambda^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s)$$

This completes the proof. □

B.7 Proof of Lemma 10

Proof of Lemma 10. Since $x_{k+1} = \text{AMP}((\beta_{k+1}^s, \rho_{k+1}^s)_{s\in[S]}, (\lambda_k^s, \mu_k^s)_{s\in[S]}, \delta_{k+1}, x_k)$, we have

$$\begin{aligned}
\delta_{k+1} &\geq g_2(x_{k+1}) + h_2(x_{k+1}) - g_2(x) - h_2(x) + (x_{k+1} - x)^\top v(x) \qquad (28)\\
&= \left(\sum_{s=1}^S \left(\frac{\rho_{k+1}^s}{2}\|E_s x_{k+1}^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_{k+1}^s}\|^2 - \frac{\rho_{k+1}^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_{k+1}^s}\|^2\right)\right) \\
&\quad + \left(\sum_{s=1}^S \frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x_{k+1}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2\right) \\
&\quad - \left(\sum_{s=1}^S \frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2\right) \\
&\quad + (x_{k+1} - x)^\top v(x), \text{ for all } x \in \widehat{X}
\end{aligned}$$

Using the update of μ_k^s in Algorithm 2 together with the second equality in Lemma 9, we have

$$\begin{aligned}
&\frac{\rho_{k+1}^s}{2}\|E_s x_{k+1}^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_{k+1}^s}\|^2 \qquad (29)\\
&= (\mu_k^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + \frac{\rho_{k+1}^s}{2}\|E_s x_{k+1}^{\mathcal{N}_s} - d_s\|^2 + \frac{\|\mu_k^s\|^2}{2\rho_{k+1}^s} \\
&= (\mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + (\mu_{k+1}^s - \mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) - \frac{1}{2\rho_{k+1}^s}\|\mu_{k+1}^s - \mu_k^s\|^2 + \frac{\|\mu_k^s\|^2}{2\rho_{k+1}^s} \\
&= (\mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + \frac{1}{2\rho_{k+1}^s}(\|\mu_{k+1}^s - \mu^s\|^2 - \|\mu_k^s - \mu^s\|^2 + \|\mu_k^s\|^2)
\end{aligned}$$

Recall that I_k^s denotes the set of coordinates of $\lambda_k^s + \beta_{k+1}^s(A_s x_{k+1}^{\mathcal{N}_s} - b_s)$ whose values are nonnegative. Then, by the update of λ_k^s in Algorithm 2, we have

$$\begin{aligned}
&\frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x_{k+1}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2 = \sum_{j \in I_k^s} \frac{\beta_{k+1}^s}{2}\|(A_s x_{k+1}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s})_j\|^2 \\
&= \sum_{j \in I_k^s}\left((\lambda_k^s)_j (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j + \frac{\beta_{k+1}^s}{2}((A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j)^2 + \frac{((\lambda_k^s)_j)^2}{2\beta_{k+1}^s}\right) \\
&= \sum_{j \in I_k^s}\left((\lambda_{k+1}^s)_j (A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j - \frac{\beta_{k+1}^s}{2}((A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j)^2 + \frac{((\lambda_k^s)_j)^2}{2\beta_{k+1}^s}\right) \\
&= (\lambda_{k+1}^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s) - \sum_{j \in I_k^s}\frac{\beta_{k+1}^s}{2}((A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j)^2 + \sum_{j \in I_k^s}\frac{((\lambda_k^s)_j)^2}{2\beta_{k+1}^s}
\end{aligned}$$

and

$$\frac{1}{2\beta_{k+1}^s}\|\lambda_{k+1}^s - \lambda_k^s\|^2 = \sum_{j \in I_k^s}\frac{\beta_{k+1}^s}{2}((A_s x_{k+1}^{\mathcal{N}_s} - b_s)_j)^2 + \sum_{j \in [m^s]\setminus I_k^s}\frac{((\lambda_k^s)_j)^2}{2\beta_{k+1}^s}$$

Putting these two equations together yields that

$$\frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x_{k+1}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2 = -\frac{1}{2\beta_{k+1}^s}\|\lambda_{k+1}^s - \lambda_k^s\|^2 + (\lambda_{k+1}^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s) + \frac{\|\lambda_k^s\|^2}{2\beta_{k+1}^s} \tag{30}$$

Combining Eq. (30) with the first inequality in Lemma 9, we have

$$\frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x_{k+1}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2 \geq (\lambda^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s) + \frac{\|\lambda_k^s\|^2}{2\beta_{k+1}^s} + \frac{\|\lambda_{k+1}^s - \lambda^s\|^2 - \|\lambda_k^s - \lambda^s\|^2}{2\beta_{k+1}^s} \tag{31}$$

Plugging Eq. (29) and Eq. (31) into Eq. (28) yields that, for any $(\lambda^s, \mu^s) \in \mathbb{R}_+^{m_s} \times \mathbb{R}^{e_s}$, we have

$$\begin{aligned}
\delta_{k+1} \geq &\left(\sum_{s=1}^S \frac{\|\mu_{k+1}^s - \mu^s\|^2 - \|\mu_k^s - \mu^s\|^2}{2\rho_{k+1}^s} + \frac{\|\lambda_{k+1}^s - \lambda^s\|^2 - \|\lambda_k^s - \lambda^s\|^2}{2\beta_{k+1}^s} + \frac{\|\mu_k^s\|^2}{2\rho_{k+1}^s} + \frac{\|\lambda_k^s\|^2}{2\beta_{k+1}^s}\right) \\
&+ \left(\sum_{s=1}^S (\mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + (\lambda^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s)\right) \\
&- \left(\sum_{s=1}^S \frac{\rho_{k+1}^s}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_{k+1}^s}\|^2 + \frac{\beta_{k+1}^s}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2\right) \\
&+ (x_{k+1} - x)^\top v(x), \text{ for all } x \in \widehat{X}
\end{aligned}$$

Rearranging this inequality with the update of δ_k, ρ_k^s and β_k^s, we have

$$\begin{aligned}
\delta_0 \geq &\left(\sum_{s=1}^S \frac{\|\mu_{k+1}^s - \mu^s\|^2 - \|\mu_k^s - \mu^s\|^2}{2\rho_0^s} + \frac{\|\lambda_{k+1}^s - \lambda^s\|^2 - \|\lambda_k^s - \lambda^s\|^2}{2\beta_0^s} + \frac{\|\mu_k^s\|^2}{2\rho_0^s} + \frac{\|\lambda_k^s\|^2}{2\beta_0^s}\right) \\
&+ \gamma^{k+1}\left(\sum_{s=1}^S (\mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + (\lambda^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s)\right) \\
&- \left(\sum_{s=1}^S \frac{\rho_0^s \gamma^{2k+2}}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_{k+1}^s}\|^2 + \frac{\beta_0^s \gamma^{2k+2}}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_{k+1}^s}\}\|^2\right) \\
&+ \gamma^{k+1}(x_{k+1} - x)^\top v(x)
\end{aligned} \tag{32}$$

By Proposition 3, there exists a triple $(\overline{x} \in \widehat{X}, \overline{\lambda} \geq 0, \overline{\mu})$ such that

$$A_s \overline{x}^{\mathcal{N}_s} \leq b_s, \; E_s \overline{x}^{\mathcal{N}_s} = d_s, \; \langle \overline{\lambda}^s, A_s \overline{x}^{\mathcal{N}_s} - b_s \rangle = 0, \text{ for all } s \in [S] \tag{33}$$

and the following VI holds true for all $\nu \in \mathcal{N}$,

$$(\overline{x}^\nu - x^\nu)^\top \left(v_\nu(\overline{x}) + \sum_{s \in \mathcal{I}_\nu} ((A_s^\nu)^\top \overline{\lambda}^s + (E_s^\nu)^\top \overline{\mu}^s) \right) \leq 0, \quad \text{for all } x^\nu \in \widehat{X}_\nu \tag{34}$$

Note that Eq. (33) implies that

$$\begin{aligned} \frac{\rho_0^s \gamma^{2k+2}}{2} \| E_s \overline{x}^{\mathcal{N}_s} - d_s + \frac{\mu_k^s}{\rho_0^s \gamma^{k+1}} \|^2 - \frac{\|\mu_k^s\|^2}{2\rho_0^s} = 0 \\ \frac{\beta_0^s \gamma^{2k+2}}{2} \| \max\{0, A_s \overline{x}^{\mathcal{N}_s} - b_s + \frac{\lambda_k^s}{\beta_0^s \gamma^{k+1}}\} \|^2 - \frac{\|\lambda_k^s\|^2}{2\beta_0^s} \leq 0 \end{aligned} \tag{35}$$

Bounding Lagrangian multipliers. Plugging Eq. (35) into Eq. (32) with $x = \overline{x}$, we have

$$\begin{aligned} \delta_0 \geq \gamma^{k+1} (x_{k+1} - \overline{x})^\top v(\overline{x}) + \gamma^{k+1} \left(\sum_{s=1}^S (\mu^s)^\top (E_s x_{k+1}^{\mathcal{N}_s} - d_s) + (\lambda^s)^\top (A_s x_{k+1}^{\mathcal{N}_s} - b_s) \right) \\ + \left(\sum_{s=1}^S \frac{\|\mu_{k+1}^s - \mu^s\|^2 - \|\mu_k^s - \mu^s\|^2}{2\rho_0^s} + \frac{\|\lambda_{k+1}^s - \lambda^s\|^2 - \|\lambda_k^s - \lambda^s\|^2}{2\beta_0^s} + \frac{\|\mu_k^s\|^2}{2\rho_0^s} + \frac{\|\lambda_k^s\|^2}{2\beta_0^s} \right) \end{aligned} \tag{36}$$

Summing Eq. (36) over $k = 0, 1, \ldots, t-1$ and rearranging the resulting inequality with $\lambda^s = \overline{\lambda}^s$ and $\mu^s = \overline{\mu}^s$ for all $s \in [S]$ and $\widehat{x}_t = \frac{\sum_{k=1}^t \gamma^k x_k}{\sum_{k=1}^t \gamma^k}$, we have

$$\begin{aligned} \frac{\delta_0 t}{\sum_{k=1}^t \gamma^k} \geq \left(\sum_{s=1}^S (\overline{\mu}^s)^\top (E_s \widehat{x}_t^{\mathcal{N}_s} - d_s) + (\overline{\lambda}^s)^\top (A_s \widehat{x}_t^{\mathcal{N}_s} - b_s) \right) \\ + (\widehat{x}_t - \overline{x})^\top v(\overline{x}) + \frac{1}{\sum_{k=1}^t \gamma^k} \left(\sum_{s=1}^S \frac{\|\mu_t^s - \overline{\mu}^s\|^2 - \|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_t^s - \overline{\lambda}^s\|^2 - \|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s} \right) \end{aligned}$$

Applying a similar argument as in Lemma 6 together with Eq. (34) and $\widehat{x}_t \in \widehat{X}$, we have

$$(\overline{x} - \widehat{x}_t)^\top v(\overline{x}) + \sum_{s=1}^S ((A_s \overline{x}^{\mathcal{N}_s} - A_s \widehat{x}_t^{\mathcal{N}_s})^\top \overline{\lambda}^s + (E_s \overline{x}^{\mathcal{N}_s} - E_s \widehat{x}_t^{\mathcal{N}_s})^\top \overline{\mu}^s) \leq 0$$

Summing the above two inequalities and using the fact that $E_s \overline{x}^{\mathcal{N}_s} = d_s$ and $(\overline{\lambda}^s)^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0$ for all $s \in [S]$, we have

$$\delta_0 t \geq \sum_{s=1}^S \left(\frac{\|\mu_t^s - \overline{\mu}^s\|^2 - \|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_t^s - \overline{\lambda}^s\|^2 - \|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s} \right)$$

Changing the index t back to k for the simplicity, we conclude that

$$\sum_{s=1}^S \frac{\|\mu_k^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_k^s - \overline{\lambda}^s\|^2}{2\beta_0^s} \leq \delta_0 k + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s} \right) \tag{37}$$

Proof of Eq. (19)**.** Considering Eq. (36) with $k = T-1$ and

$$\lambda^s = \overline{\lambda}^s + \frac{\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}}{\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\|}, \quad \mu^s = \overline{\mu}^s + \frac{E_s x_T^{\mathcal{N}_s} - d_s}{\|E_s x_T^{\mathcal{N}_s} - d_s\|}, \quad \text{for all } s \in [S]$$

Then, we have

$$\begin{aligned}
\delta_0 \;\; \geq \;\; & \gamma^T (x_T - \overline{x})^\top v(\overline{x}) - \left(\sum_{s=1}^{S} \frac{\left\| \mu_{T-1}^s - \overline{\mu}^s - \frac{E_s x_T^{\mathcal{N}_s} - d_s}{\|E_s x_T^{\mathcal{N}_s} - d_s\|} \right\|^2}{2\rho_0^s} + \frac{\left\| \lambda_{T-1}^s - \overline{\lambda}^s - \frac{\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}}{\|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\|} \right\|^2}{2\beta_0^s} \right) \\
& + \gamma^T \left(\sum_{s=1}^{S} (\overline{\mu}^s)^\top (E_s x_T^{\mathcal{N}_s} - d_s) + (\overline{\lambda}^s)^\top (A_s x_T^{\mathcal{N}_s} - b_s) \right) \\
& + \gamma^T \left(\sum_{s=1}^{S} \|E_s x_T^{\mathcal{N}_s} - d_s\| + \|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \right) \\
\geq \;\; & \gamma^T (x_T - \overline{x})^\top v(\overline{x}) - \left(\sum_{s=1}^{S} \frac{\|\mu_{T-1}^s - \overline{\mu}^s\|^2 + 1}{\rho_0^s} + \frac{\|\lambda_{T-1}^s - \overline{\lambda}^s\|^2 + 1}{\beta_0^s} \right) \\
& + \gamma^T \left(\sum_{s=1}^{S} (\overline{\mu}^s)^\top (E_s x_T^{\mathcal{N}_s} - d_s) + (\overline{\lambda}^s)^\top (A_s x_T^{\mathcal{N}_s} - b_s) \right) \\
& + \gamma^T \left(\sum_{s=1}^{S} \|E_s x_T^{\mathcal{N}_s} - d_s\| + \|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \right)
\end{aligned}$$

Applying a similar argument as in Lemma 6 together with Eq. (34) and $x_T \in \widehat{X}$, we have

$$(\overline{x} - x_T)^\top v(\overline{x}) + \sum_{s=1}^{S} ((A_s \overline{x}^{\mathcal{N}_s} - A_s x_T^{\mathcal{N}_s})^\top \overline{\lambda}^s + (E_s \overline{x}^{\mathcal{N}_s} - E_s x_T^{\mathcal{N}_s})^\top \overline{\mu}^s) \leq 0$$

Combining the above two inequalities and using that $E_s \overline{x}^{\mathcal{N}_s} = d_s$ and $(\overline{\lambda}^s)^\top (A_s \overline{x}^{\mathcal{N}_s} - b_s) = 0$ for all $s \in [S]$ and $\gamma^T > 0$, we have

$$\delta_0 \geq \gamma^T \left(\sum_{s=1}^{S} \|E_s x_T^{\mathcal{N}_s} - d_s\| + \|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \right) - \left(\sum_{s=1}^{S} \frac{\|\mu_{T-1}^s - \overline{\mu}^s\|^2 + 1}{\rho_0^s} + \frac{\|\lambda_{T-1}^s - \overline{\lambda}^s\|^2 + 1}{\beta_0^s} \right)$$

This together with Eq. (37) yields that

$$\sum_{s=1}^{S} \|E_s x_T^{\mathcal{N}_s} - d_s\| + \|\max\{0, A_s x_T^{\mathcal{N}_s} - b_s\}\| \leq \frac{1}{\gamma^T} \left(\delta_0 T + \left(\sum_{s=1}^{S} \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s} \right) \right)$$

Then, Eq. (19) follows from the definition of $\widehat{\epsilon}_1$.

Proof of Eq. (20). Considering Eq. (32) with $k = T-1$ and $\lambda^s = 0$ and $\mu^s = 0$ for all $s \in [S]$. Then, we have

$$
\begin{aligned}
\delta_0 \geq \gamma^T (x_T - x)^\top v(x) &- \left(\sum_{s=1}^S \frac{\|\mu_{T-1}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_{T-1}^s\|^2}{2\beta_0^s}\right) + \left(\sum_{s=1}^S \frac{\|\mu_{T-1}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_{T-1}^s\|^2}{2\beta_0^s}\right) \\
&- \left(\sum_{s=1}^S \frac{\rho_0^s \gamma^{2T}}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu_{T-1}^s}{\rho_0^s\gamma^T}\|^2 + \frac{\beta_0^s\gamma^{2T}}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda_{T-1}^s}{\beta_0^s\gamma^T}\}\|^2\right)
\end{aligned}
$$

Rearranging this inequality and using Eq. (37) yields that

$$
\begin{aligned}
(x_T - x)^\top v(x) \leq \frac{1}{\gamma^T}&\left(\delta_0 T + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s}\right)\right) - \left(\sum_{s=1}^S \frac{\|\mu_{T-1}^s\|^2}{2\rho_0^s\gamma^T} + \frac{\|\lambda_{T-1}^s\|^2}{2\beta_0^s\gamma^T}\right) \\
&+ \left(\sum_{s=1}^S \frac{\rho_0^s\gamma^T}{2}\|E_s x^{\mathcal{N}_s} - d_s + \frac{\mu_{T-1}^s}{\rho_0^s\gamma^T}\|^2 + \frac{\beta_0^s\gamma^T}{2}\|\max\{0, A_s x^{\mathcal{N}_s} - b_s + \frac{\lambda_{T-1}^s}{\beta_0^s\gamma^T}\}\|^2\right)
\end{aligned}
$$

Fixing $\nu \in \mathcal{N}$, we plug $x^{-\nu} = x_T^{-\nu}$ into this inequality and obtain that

$$
\begin{aligned}
(x_T^\nu - x^\nu)^\top v_\nu(x^\nu, x_T^{-\nu}) \leq \frac{1}{\gamma^T}&\left(\delta_0 T + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s}\right)\right) \\
&+ \underbrace{\sum_{s\in\mathcal{I}_\nu} \frac{\beta_0^s\gamma^T}{2}\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s + \frac{\lambda_{T-1}^s}{\beta_0^s\gamma^T}\}\|^2 - \frac{\|\lambda_{T-1}^s\|^2}{2\beta_0^s\gamma^T}}_{\mathbf{I}} \\
&+ \underbrace{\sum_{s\in\mathcal{I}_\nu} \frac{\rho_0^s\gamma^T}{2}\|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s + \frac{\mu_{T-1}^s}{\rho_0^s\gamma^T}\|^2 - \frac{\|\mu_{T-1}^s\|^2}{2\rho_0^s\gamma^T}}_{\mathbf{II}}, \quad \text{for all } x^\nu \in \widehat{X}_\nu
\end{aligned}
$$

Suppose that $x^\nu \in \widehat{X}_\nu$ satisfies

$$
\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \leq \epsilon, \quad \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \leq \epsilon, \quad \text{for all } s \in \mathcal{I}_\nu
$$

Applying the above conditions to **I** and **II** and using Young's inequality, we have

$$
\begin{aligned}
\mathbf{I} \;\; &\leq \;\; \sum_{s\in\mathcal{I}_\nu} \frac{\beta_0^s\gamma^T}{2}\|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\|^2 \\
&\quad + \sum_{s\in\mathcal{I}_\nu} \|\lambda_{T-1}^s\| \|\max\{0, A_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} A_s^j x_T^j - b_s\}\| \\
&\leq \;\; \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s\in\mathcal{I}_\nu} \beta_0^s\right) + \frac{\epsilon^2}{2}\left(\sum_{s\in\mathcal{I}_\nu} \frac{\|\lambda_{T-1}^s\|^2}{\beta_0^s}\right) \\
&\leq \;\; \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s\in\mathcal{I}_\nu} \beta_0^s\right) + \epsilon^2\left(\sum_{s\in\mathcal{I}_\nu} \frac{\|\lambda_{T-1}^s - \overline{\lambda}^s\|^2 + \|\overline{\lambda}^s\|^2}{\beta_0^s}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{II} \quad &\leq \quad \sum_{s\in\mathcal{I}_\nu} \frac{\rho_0^s\gamma^T}{2}\|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\|^2 + \|\mu_{T-1}^s\| \|E_s^\nu x^\nu + \sum_{j\in\mathcal{N}_s, j\neq\nu} E_s^j x_T^j - d_s\| \\
&\leq \quad \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s\in\mathcal{I}_\nu}\rho_0^s\right) + \frac{\epsilon^2}{2}\left(\sum_{s\in\mathcal{I}_\nu}\frac{\|\mu_{T-1}^s\|^2}{\rho_0^s}\right) \\
&\leq \quad \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s\in\mathcal{I}_\nu}\rho_0^s\right) + \epsilon^2\left(\sum_{s\in\mathcal{I}_\nu}\frac{\|\mu_{T-1}^s - \overline{\mu}^s\|^2 + \|\overline{\mu}^s\|^2}{\rho_0^s}\right)
\end{aligned}
$$

Putting these pieces together with $\mathcal{I}_\nu \subseteq [S]$ and Eq. (37) yields that

$$
\begin{aligned}
&(x_T^\nu - x^\nu)^\top(v_\nu(x^\nu, x_T^{-\nu})) \leq \frac{1}{\gamma^T}\left(\delta_0 T + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2}{2\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2}{2\beta_0^s}\right)\right) \\
&\quad + \frac{\epsilon^2(1+\gamma^T)}{2}\left(\sum_{s=1}^S \beta_0^s + \rho_0^s\right) + \epsilon^2\left(2\delta_0(T-1) + \left(\sum_{s=1}^S \frac{\|\mu_0^s - \overline{\mu}^s\|^2 + \|\overline{\mu}^s\|^2}{\rho_0^s} + \frac{\|\lambda_0^s - \overline{\lambda}^s\|^2 + \|\overline{\lambda}^s\|^2}{\beta_0^s}\right)\right) \\
&\quad \text{for all } x^\nu \in \widehat{X}_\nu
\end{aligned}
$$

Then, Eq. (20) follows from the definition of $\widehat{\epsilon}_2$. □

C Experiments

We conduct experiments using several datasets from [Facchinei and Kanzow(2009)]. All of the algorithms were implemented with MATLAB R2020b on a MacBook Pro with an Intel Core i9 2.4GHz (8 cores and 16 threads) and 16GB memory.

Although some of the GNEPs in [Facchinei and Kanzow(2009)] (i.e., **A1**-**A9**) are more general than the nonlinear GNEPs considered in Definition 1, we can implement our Algorithm 1 and 2 with player-specific parameters: $u_{\max}^\nu = 10^6$ and $\beta_0^\nu = \rho_0^\nu = 1$ for each $\nu \in \mathcal{N}$, where the AMP algorithm is used for solving a general VI.[4] For other GNEPs (i.e., **A11**-**A18**), we use $u_{\max}^s = 10^6$ and $\beta_0^s = \rho_0^s = 1$ for every $s \in [S]$. Following the strategies used in Algorithm 11 (S4) [Kanzow and Steck(2016)], we do not update (β, ρ) at each iteration but only when the certain feasibility conditions are satisfied. The associated parameters are chosen according to the size of the problem: $\gamma = 4$ if $n < 100$ and $\gamma = 2$ otherwise. Compared to [Kanzow and Steck(2016)], this represents a less aggressive penalization and is thus suitable for our algorithm. Indeed, for the subproblem solving, [Kanzow and Steck(2016)] employed a Levenberg-Marquardt algorithm [Fan and Yuan(2005)] that used both first-order and second-order information while our algorithm only exploits first-order information. Moreover, we initialize the multipliers (λ_0, μ_0) using the same nonnegative least-squares approach as in [Kanzow and Steck(2016)] and also set the same stopping criterion but with the tolerance 10^{-4}. The maximum iteration number is set as 50 and the maximum penalty parameter is set as 10^{12}.

Since we perform a full penalization, the subproblems are unconstrained NEPs. Solving these subproblems is equivalent to computing a solution of the nonlinear equation $F(\mathbf{x}) + \nabla G(\mathbf{x}) = 0$. We use the AMP algorithm for subproblem solving and set ℓ_G and ℓ_F properly depending on the problem. The maximum iteration number is set as 2000 and we stop each inner loop when either the iteration number exceeds 2000 or $\|F(\mathbf{x}) + \nabla G(\mathbf{x})\|$ is below 10^{-6}. We summarize the results in Table 1 and make several comments. With the exception of problem A.16d, Algorithm 2 solves every problem given a proper initial point. Although the algorithm compares unfavorably to existing second-order methods for GNEPs in terms of solution accuracy, Algorithm 2 is matrix-free and can scale to large problems. Further, the speed of Algorithm 2 depends on how quickly the subproblems are solved. In this regard, the AMP algorithm benefits from avoiding computationally expensive operations, e.g., matrix inversion. However, as pointed by [Kanzow and Steck(2016)], the subproblems have a semismooth structure in which the Levenberg-Marquardt algorithm can be superlinearly convergent in practice. As such, it is promising to investigate whether or not we can design first-order algorithms that benefit from such structure. It is also important to make good choices of the parameters that determine the multipliers and penalty parameters since they greatly affect the performance of the algorithm. For many application problems, we observe that fine-tuning the parameters can yield a speed improvement.

It is also worth mentioning that we have conducted numerical experiments with both Algorithm 1 and Algorithm 2 and Algorithm 2 dominates Algorithm 1. This is consistent with the observation from [Kanzow and Steck(2016)] so we omit the results for Algorithm 1. It is important to acknowledge that existing second-order methods based on Levenberg-Marquardt or other linear system solvers are faster than our algorithms on small instances. This is because solving linear systems will not be an issue when the size of problem is small. However, these linear system solvers will be prohibitive as the problem size increases. In contrast, our methods do not need a sophisticated matrix factorization and can be thus scaled to large problems.

[4] We set $G = 0$ in which case the AMP algorithm reduces to the extragradient algorithm.

Table 1. Numerical results for Algorithm 2. N denotes the number of players, n is the total number of variables, k is the number of outer iterations, i_{total} is the total number of inner iterations and F denotes a failure. We also include R_f, R_o and R_c which measure the feasibility, optimality and complementary slackness at the solution in terms of KKT condition; see [Kanzow and Steck(2016)] for the definition.

Example	N	n	x_0	k	i_{total}	R_f	R_o	R_c	$\rho_{\max}$
A.1	10	10	0.01	7	393	4.7e-05	1.9e-06	1.3e-05	4
			0.1	7	378	2.9e-05	1.9e-06	7.9e-06	4
			1	9	7296	4.0e-08	1.3e-06	1.1e-08	4096
A.2	10	10	0.01	7	3725	2.9e-05	1.0e-06	8.7e-06	16
			0.1	6	3562	9.6e-05	2.8e-06	2.9e-05	16
			1	F					
A.3	3	7	0	F					
			1	F					
			10	12	5837	4.3e-07	9.9e-07	3.6e-05	1024
A.4	3	7	0	12	618	2.0e-07	8.3e-07	6.1e-05	1024
			1	0	0	0	0	0	1
			10	14	2395	8.2e-08	9.3e-07	1.5e-05	4096
A.5	3	7	0	8	2830	2.3e-05	9.9e-07	5.2e-05	64
			1	9	3289	9.0e-06	9.9e-07	2.2e-05	64
			10	9	5327	3.7e-05	1.1e-06	9.0e-05	256
A.6	3	7	0	12	8636	1.0e-07	1.0e-06	2.1e-05	1024
			1	12	7578	2.1e-08	1.2e-06	2.4e-05	1024
			10	F					
A.7	4	20	0	14	22844	1.2e-07	7.2e-05	5.3e-05	4096
			1	13	22609	4.4e-08	4.9e-05	2.0e-05	4096
			10	17	23796	2.0e-07	9.3e-07	7.8e-05	4096
A.8	3	3	0	F					
			1	1	115	1.7e-06	4.1e-06	1.7e-06	1
			10	3	267	0	8.5e-07	2.1e-07	4
A.9a	7	56	0	9	2353	8.2e-06	1.0e-06	2.2e-05	4
A.9b	7	112	0	16	1793	4.2e-06	8.2e-07	4.1e-05	1
A.11	2	2	0	7	306	8.0e-05	2.3e-06	4.0e-05	4
A.12	2	2	0	1	83	0	2.9e-06	0	1
A.13	3	3	0	4	8000	0	8.9e-05	1.2e-05	1
A.14	10	10	0.01	1	69	0	1.2e-06	0	1
A.15	3	6	0	3	6000	0	1.6e-05	0	1
A.16a	5	5	10	9	1895	1.1e-06	1.0e-06	3.1e-05	4
A.16b	5	5	10	12	1288	1.5e-06	2.4e-06	2.8e-05	1
A.16c	5	5	10	8	1130	7.1e-06	1.7e-06	5.1e-05	1
A.16d	5	5	10	F					
A.17	2	3	0	7	2902	0	1.3e-06	3.5e-05	16
A.18	2	12	0	11	11950	4.5e-06	1e-06	8.5e-05	64
			1	11	11950	4.5e-06	1e-06	8.5e-05	64
			10	11	11945	4.5e-06	1e-06	8.5e-05	64