Efficient Gradient Methods for Distributed Saddle Problems

Ruichen Luo RLUO@IST.AC.AT
Institute of Science and Technology Austria

Anton Rodomanov ANTON.RODOMANOV@CISPA.DE
CISPA Helmholtz Center

Sebastian U. Stich STICH@CISPA.DE
CISPA Helmholtz Center

Abstract

The distributed setting for Saddle Problems (SPs) has recently emerged as a framework for various modern applications in machine learning and multiagent systems. Despite its relevance, the theoretical foundations of this setting have not yet been thoroughly established. In this paper, we advance this research direction by formalizing the distributed setup for SPs and providing rigorous definitions of communication and computational costs. Our main result is a novel decoupled method that achieves optimal communication cost within the gradient-span framework. Our method is based on a multi-stage reduction to the decoupled minimization of residual norms, which matches the lower bound on communication complexity and yields a strict improvement over the long-standing computational complexity of the Extragradient method in the general convex-concave setting. Finally, we study the extension of distributed SP into Variational Inequality Problem (VIP), which generalizes two-player zero-sum games to multiplayer general-sum games. We show that our decoupled method achieves a new state-of-the-art communication complexity for this broader class.

Keywords: convex optimization, distributed optimization, complexity bounds, saddle problems, variational inequalities, extra-gradient method, fast gradient method.

1. Introduction

Motivation. Saddle problems (SPs) and their generalizations, variational inequality problems (VIPs), are fundamental problem classes in optimization and variational analysis. Beyond their theoretical significance, they have a wide array of modern applications, including the training of Generative Adversarial Networks (Goodfellow et al., 2014), robust optimization (Ben-Tal and Nemirovski, 2002), and equilibrium computation in game theory and multiagent systems (von Neumann and Morgenstern, 1947; Rosen, 1965; Hu et al., 1998).

The growing scale of modern problems—driven by massive datasets in machine learning, complex game dynamics, and multiagent protocols—renders reliance on a single central processor increasingly impractical. Beyond scalability, many applications are inherently distributed: agents are often geographically dispersed, driven by their own individual interests, and bound by privacy constraints that prohibit the sharing of raw data or utilities. Consequently, distributed computation has become an essential regime for these problems. This perspective underlies a growing body of work in large-scale learning, game-theoretic models, and multiagent systems (McMahan et al., 2017; Zhang et al., 2024; Conitzer and Sandholm, 2004; Nisan and Segal, 2006; Hart and Mansour, 2010; Yoon et al., 2025).

In this work, we consider a natural setup where the decision variables of the SPs or VIPs are partitioned among different agents. For instance, in a classic saddle problem $\min_{\mathbf{x}} \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$,

we consider one agent controls the minimizing variable $\mathbf{x}$ while another controls the maximizing variable $\mathbf{y}$. This partition naturally models, for instance, the strategic autonomy of players in game theory, the interaction protocol in multiagent systems, and the physical separation of the generator and discriminator in GANs (Conitzer and Sandholm, 2004; Goodfellow et al., 2014). Since the decision variables are coupled within their utilities, these agents must coordinate to reach a mutual equilibrium. To do so, they form a communication network that allows them to exchange certain information, such as their current decision variables. Thus, this provides a natural distributed setup where the decision variables are separated among the different agents.

While distributed optimization is well-established for finite-sum minimization and federated learning (Schmidt et al., 2017; McMahan et al., 2017), the literature on Saddle Problems has primarily focused on extending these data-distributed paradigms (Deng and Mahdavi, 2021; Beznosikov et al., 2025). In contrast, the setting where decision variables and utility functions are partitioned among autonomous agents—essential for the multi-agent applications discussed earlier—remains a relatively new topic.

Although a few recent works have touched upon this direction (Zhang et al., 2024; Zindari et al., 2025; Yoon et al., 2025; Yoon and Loizou, 2025), they predominantly focus on algorithms tailored to specific, favorable scenarios. Consequently, a fundamental gap persists: the lack of a systematic theoretical framework for the general setting of distributed SPs and VIPs. Existing discussions regarding performance often remain at a vague conceptual level, lacking a rigorous formalization of the distributed environment itself. Specifically, there are no standardized definitions for communication and computational complexities in this context. Without such a foundation, it can be difficult to determine the inherent performance limits (lower bounds) or to rigorously compare the efficiency of different protocols.

To enable a rigorous analysis, it is essential to establish metrics that reflect the constraints of distributed multiagent systems, where network latency and bandwidth often dwarf local processing time. In this regime, the primary bottleneck is the communication cost (exchange rounds), while the computational cost (local gradient queries) is a secondary objective. Viewed through this lens, the Extragradient (EG) method (Tseng, 1995; Nemirovski, 2004) serves as the "gold standard" baseline, though the challenges associated with it differ by metric. Regarding computational complexity, strictly improving upon EG for general monotone problems has remained an elusive goal despite over two decades of research. Regarding communication complexity—a metric that has recently come into focus with the rise of distributed systems—EG similarly defines the current state-of-the-art. Surpassing this baseline in the general setting represents a new but critical open problem.

This leads to the following research questions:

- **Formalization and Limits:** How can we rigorously formalize the communication and computational complexities for distributed SPs?
- **Computational Efficiency:** Is it possible to strictly improve upon the long-standing computational complexity of the EG method for general SPs?
- **Communication Efficiency:** Can we design an algorithm that surpasses the state-of-the-art communication bounds for general SPs and VIPs?

Contributions. We answer the aforementioned questions in the affirmative, which advances the current theory of distributed SPs and VIPs.

- In Section 2, we rigorously define the distributed setup, with the information-based notions of distributed algorithms and their communication and computational costs.

- In Section 3, we propose a template communication-efficient method for SPs, whose main idea is to reduce SPs to coordinate-wise subproblems that can be solved in a decoupled way. This method (a) improves the state-of-the-art communication cost, and (b) strictly improves the long-standing computational cost of the classic EG method.
- Then, in Section 4 we explain the technical components, with which we obtain a concrete algorithm in Section 5. Further, Section 6 establishes the *communication optimality* of our algorithm within the gradient-span framework.
- Finally, in Section 7, we discuss extensions to VIPs (with details relegated to Appendix D). We show that an extended variant of our decoupled method improves the state-of-the-art communication complexity of the class.

Notations. Let $[n] \triangleq \{1, \cdots, n\}$. For any finite-dimensional real vector space $\mathcal{E}$ with Euclidean norm $\|\cdot\|_{\mathcal{E}}$, we use $\|\cdot\|_{\mathcal{E}^*}$ to denote its dual norm. Let the finite-dimensional real vector space $\mathcal{E}_x$ be equipped with the Euclidean norm $\|\mathbf{x}\|_x = \langle \mathbf{P}_x \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}$, where $\mathbf{P}_x \colon \mathcal{E}_x \to \mathcal{E}_x^*$ is a self-adjoint positive definite linear operator, and the dual pairing $\langle \phi_x, \mathbf{x} \rangle$ represents $\phi_x(\mathbf{x})$; and let $\|\cdot\|_{x^*}$ be the dual norm of $\|\cdot\|_x$. We consider similar geometries for the spaces of $\mathcal{E}_y$, $\mathcal{E}_i$ ($i \in [K]$), and $\hat{\mathcal{E}}$ considered in this paper, with their respective self-adjoint positive definite linear operators $\mathbf{P}_y$, $\mathbf{P}_i$, and $\hat{\mathbf{P}}$. For any function $\psi \colon \mathcal{E} \to \mathbb{R} \cup \{+\infty\}$, let $\operatorname{dom} \psi$ denote its effective domain, and let $\partial \psi(\mathbf{z})$ denote its subdifferential at point $\mathbf{z} \in \operatorname{dom} \psi$.

2. Convex-concave saddle problems

We are interested in the convex-concave saddle problems (SPs) given by (f, ψ_x, ψ_y) as follows:

$$\min_{\mathbf{x} \in \operatorname{dom} \psi_x} \max_{\mathbf{y} \in \operatorname{dom} \psi_y} [F(\mathbf{x}, \mathbf{y}) \triangleq f(\mathbf{x}, \mathbf{y}) + \psi_x(\mathbf{x}) - \psi_y(\mathbf{y})], \tag{1}$$

where the functions $\psi_x(\cdot) \colon \mathcal{E}_x \to \mathbb{R} \cup \{+\infty\}$ and $\psi_y(\cdot) \colon \mathcal{E}_y \to \mathbb{R} \cup \{+\infty\}$ have simple structures, the function $f(\cdot, \cdot)$ is real-valued and defined on an open set containing $Q \triangleq \operatorname{dom} \psi_x \times \operatorname{dom} \psi_y$. To simplify the notations, we denote $\mathbf{z} \triangleq (\mathbf{x}, \mathbf{y}) \in Q$ in the context of SPs.

2.1. Distributed first-order algorithm: Communication and computational costs

We consider a setting where there are two distributed agents: Agent x controls decision variable $\mathbf{x} \in \operatorname{dom} \psi_x$ and has direct access to the function ψ_x, while Agent y controls $\mathbf{y} \in \operatorname{dom} \psi_y$ and has direct access to the function ψ_y. Moreover, we assume Agents x and y can also interact with the function f, but only through their respective oracles $\mathscr{O}_x$ and $\mathscr{O}_y$. An oracle is a standard notion in black-box optimization (Nemirovskij and Yudin, 1983) that takes an input point and returns certain information about the function at this point. It should be noted that Agent x does not have direct access to oracle $\mathscr{O}_y$, and Agent y does not have direct access to oracle $\mathscr{O}_x$. However, Agents x and y are connected via a global communication channel, where they can exchange certain information from time to time.

More precisely, we consider algorithms defined in the following way.

Distributed algorithm. Given oracles $\mathscr{O}_x$ and $\mathscr{O}_y$, a distributed algorithm for a problem class $\mathscr{P}$ is a procedure that proceeds in a sequence of *communication rounds*. A problem class is the collection of all the saddle problems of form (1) satisfying certain assumptions. (We will introduce a specific problem class of interest in Section 2.2.) We consider an algorithm as a sequence of methods that

prescribe the agents what to do at each communication round. Thus, let us denote such a method as $\mathscr{M} = (\mathscr{M}_x^t, \mathscr{M}_y^t, \bar{\mathscr{M}}^t)_{t\geq 0}$, where the tuple $(\mathscr{M}_x^t, \mathscr{M}_y^t, \bar{\mathscr{M}}^t)$ contains the prescribed methods to the agents at communication round t. At the beginning, the local information sets of the agents, $\mathcal{I}_x^0$ and $\mathcal{I}_y^0$, are both empty sets. At each communication round $t \geq 0$, the agents proceed with three phases:

1. **Local Computation Tasks:** Each agent runs a local method that iteratively queries its oracle a number of times and accumulates information about the problem. In the distributed SP considered here, Agents x and y are prescribed with the methods $\mathscr{M}_x^t$ and $\mathscr{M}_y^t$ for them to run locally. Let us explain the local methods in more detail. Starting with information set $\mathcal{I}_x^t$, Agent x runs the local computation method $\mathscr{M}_x^t$: specifically, Agent x is guided by its local method $\mathscr{M}_x^t$ to **iteratively query oracle** $\mathscr{O}_x$, collect the answers $\mathscr{A}_x^t$ from the oracle, and expand its local information set with these answers $\widetilde{\mathcal{I}}_x^t = (\mathcal{I}_x^t, \mathscr{A}_x^t)$. Similarly, starting with information set $\mathcal{I}_y^t$, Agent y runs the local computation method $\mathscr{M}_y^t$ to **iteratively query oracle** $\mathscr{O}_y$ and expand its local information set $\widetilde{\mathcal{I}}_y^t$.

2. **Waiting:** Upon completing its local task, an agent enters a **wait** state (idles).

3. **Communication:** Communication is triggered only when both agents have entered the wait state. At this moment, the agents start with their respective local information sets $\widetilde{\mathcal{I}}_x^t$ and $\widetilde{\mathcal{I}}_y^t$ and run the global communication method of $\bar{\mathscr{M}}^t$. The global communication method guides the agents to produce certain (to-be-exchanged) information from their local information sets, to communicate with the other agents with their produced information, and then guide them to generate an approximate solution together. Specifically, guided by $\bar{\mathscr{M}}^t$: Agent x produces certain information $\bar{\mathcal{I}}_x^t$ from its local information set $\widetilde{\mathcal{I}}_x^t$ and sends them to Agent y; Agent y produces certain information $\bar{\mathcal{I}}_y^t$ from its local information set $\widetilde{\mathcal{I}}_y^t$ and sends them to Agent x; the agents expand their local information sets (that is, $\mathcal{I}_x^{t+1} = (\widetilde{\mathcal{I}}_x^t, \bar{\mathcal{I}}_y^t)$ and $\mathcal{I}_y^{t+1} = (\widetilde{\mathcal{I}}_y^t, \bar{\mathcal{I}}_x^t)$); and the agents jointly **generate an approximate solution** $\bar{\mathbf{z}}^{t+1}$ from these updated information sets. Formally, we denote this entire communication procedure as

$$(\bar{\mathbf{z}}^{t+1}, \mathcal{I}_x^{t+1}, \mathcal{I}_y^{t+1}) = \bar{\mathscr{M}}^t(\widetilde{\mathcal{I}}_x^t, \widetilde{\mathcal{I}}_y^t).$$

   Upon completion of procedure $\bar{\mathscr{M}}^t$, the agents proceed to the next round with $\mathcal{I}_x^{t+1}$ and $\mathcal{I}_y^{t+1}$.

In the above definitions, we work with the information-based complexity, which does not put any restrictions on the arithmetical or memory complexities of the methods. In practice, however, it is important that each operation can be efficiently implemented and memory is maintained efficiently (e.g., agents need to keep in memory and send to each other only a few vectors). This is typically the case for all algorithms we consider in this paper. We also remark that the above description defines only *deterministic methods*. But every rule can in principle be randomized and then this would naturally define a *randomized* method. However, in this paper, for the sake of simplicity, we only focus on deterministic methods.

Communication and computational costs. Let the problem class $\mathscr{P}$ comprise certain SP instances. For any problem instance $P \in \mathscr{P}$, let $\Delta_P(\cdot)$ be an accuracy measure that maps any solution $\mathbf{z} \in Q$ to a non-negative number indicating its approximation error. For any distributed first-order algorithm $\mathscr{M}$ and any problem instance $P \in \mathscr{P}$, let $\bar{\mathbf{z}}^{t+1}(\mathscr{M}, P)$ denote the approximate solution $\bar{\mathbf{z}}^{t+1}$ obtained when algorithm $\mathscr{M}$ runs on the instance P. For any target accuracy ε, let $T^{\mathscr{M}}(\varepsilon, P)$ denote the number of rounds to reach an ε-approximate solution: that is,

$$T^{\mathscr{M}}(\varepsilon, P) = \min\big\{T \geq 1 \mid \Delta_P\big(\bar{\mathbf{z}}^T(\mathscr{M}, P)\big) \leq \varepsilon\big\}.$$

If the ε-approximate solution is never reached, then $T^{\mathscr{M}}(\varepsilon, P) = +\infty$. Furthermore, if $T^{\mathscr{M}}(\varepsilon, P)$ is finite, let us denote by $N_x^{\mathscr{M}}(\varepsilon, P)$ the number of times Agent x queries oracle $\mathscr{O}_x$ and by $N_y^{\mathscr{M}}(\varepsilon, P)$ the number of times Agent y queries oracle $\mathscr{O}_y$, from round 0 to round $T^{\mathscr{M}}(\varepsilon, P) - 1$. While if $T^{\mathscr{M}}(\varepsilon, P) = +\infty$, let $N_x^{\mathscr{M}}(\varepsilon, P) = N_y^{\mathscr{M}}(\varepsilon, P) = +\infty$.

Now, we define the following complexity metrics of algorithm $\mathscr{M}$ for problem class $\mathscr{P}$: (for any target accuracy $\varepsilon > 0$)

- **Communication cost:** $T_{\mathscr{P}}^{\mathscr{M}}(\varepsilon) = \max_{P \in \mathscr{P}} T^{\mathscr{M}}(\varepsilon, P)$;
- **Computational cost**: $N_{\mathscr{P}}^{\mathscr{M}}(\varepsilon) = \max_{P \in \mathscr{P}} \left[c_x N_x^{\mathscr{M}}(\varepsilon, P) + c_y N_y^{\mathscr{M}}(\varepsilon, P)\right]$, where the weights $c_x, c_y \geq 0$ are constants.

We consider the weighted computational cost for the simplicity of presentation. For any given algorithm: with $(c_x = 1, c_y = 0)$ or $(c_x = 0, c_y = 1)$, our weighted cost is equivalent to the query complexity of oracle $\mathscr{O}_x$ or oracle $\mathscr{O}_y$, respectively; and with $(c_x = \frac{1}{2}, c_y = \frac{1}{2})$, our weighted cost is equivalent to the query complexity of full oracle $(\mathscr{O}_x, \mathscr{O}_y)$ in the classic non-distributed settings up to a constant 2.

Finally, among all the algorithms, we consider the communication cost as the main metric and the computational cost as the secondary metric. Specifically, for any fixed accuracy ε, among the distributed algorithms, we consider the algorithm $\mathscr{M}$ with smaller communication cost $T_{\mathscr{P}}^{\mathscr{M}}(\varepsilon)$ to be more efficient; and if the algorithms have the same communication cost, we consider the algorithm $\mathscr{M}$ with smaller computational cost $N_{\mathscr{P}}^{\mathscr{M}}(\varepsilon)$ to be more efficient.

2.2. Problem class, oracles, and accuracy measure

Now, we introduce the problem class, oracles, and accuracy measures for the class of SPs studied in this paper. They are indeed standard descriptions in the literature of convex-concave SPs (Nemirovski, 2004; Wang and Li, 2020; Zhang et al., 2022).

Problem class $\mathscr{P}_{SP}$. We focus on the problem class $\mathscr{P}_{SP}$ that comprises all the SP instances of form (1) under Assumptions (A1) to (A3):

(A1) The function $f(\cdot, \mathbf{y})$ is convex in $\mathbf{x} \in \operatorname{dom} \psi_x$ for any fixed $\mathbf{y} \in \operatorname{dom} \psi_y$. The function $f(\mathbf{x}, \cdot)$ is concave in $\mathbf{y} \in \operatorname{dom} \psi_y$ for any fixed $\mathbf{x} \in \operatorname{dom} \psi_x$. The functions ψ_x and ψ_y are proper closed convex functions.

(A2) Problem (1) has a saddle point $(\mathbf{x}^*, \mathbf{y}^*) \in Q$ such that $\mathbf{x}^* \in \mathcal{B}_x$ and $\mathbf{y}^* \in \mathcal{B}_y$, where $\mathcal{B}_x \triangleq \{\mathbf{x} \in \mathcal{E}_x \mid \|\mathbf{x}^0 - \mathbf{x}\|_x \leq D_x\}$, $\mathcal{B}_y \triangleq \{\mathbf{y} \in \mathcal{E}_y \mid \|\mathbf{y}^0 - \mathbf{y}\|_y \leq D_y\}$, $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) \in Q$ is a given point, and $D_x, D_y > 0$ are given distances.

(A3) The function $f(\cdot, \cdot)$ is continuously differentiable over Q. Moreover, there exists $L_x, L_{xy}, L_y > 0$ such that for all $\mathbf{x}, \mathbf{x}' \in \operatorname{dom} \psi_x$ and $\mathbf{y}, \mathbf{y}' \in \operatorname{dom} \psi_y$, we have

$$\|\nabla_x f(\mathbf{x}', \mathbf{y}') - \nabla_x f(\mathbf{x}, \mathbf{y})\|_{x^*} \leq L_x \|\mathbf{x}' - \mathbf{x}\|_x + L_{xy} \|\mathbf{y}' - \mathbf{y}\|_y,$$
$$\|\nabla_y f(\mathbf{x}', \mathbf{y}') - \nabla_y f(\mathbf{x}, \mathbf{y})\|_{y^*} \leq L_{xy} \|\mathbf{x}' - \mathbf{x}\|_x + L_y \|\mathbf{y}' - \mathbf{y}\|_y.$$

Oracles. This paper considers the following deterministic partial gradient oracles: for any input point $\mathbf{z} \in Q$, the oracle $\mathscr{O}_x$ returns $\mathscr{O}_x(\mathbf{z}) = \nabla_x f(\mathbf{z})$; and the oracle $\mathscr{O}_y$ returns $\mathscr{O}_y(\mathbf{z}) = -\nabla_y f(\mathbf{z})$. These oracles can be queried separately by their respective agents, by different number of times, and at different input points.

Let us provide a concrete example for the abstract oracle settings. Consider $f(\mathbf{z}) = g(\mathbf{z}) + f_x(\mathbf{x}) - f_y(\mathbf{y})$, where the function g is the global coupled utility, and the functions f_x and f_y are the

private utilities of Agents x and y, respectively. Then, the deterministic partial gradient oracles are given by $\mathscr{O}_x(\mathbf{z}) = \nabla_x g(\mathbf{z}) + \nabla f_x(\mathbf{x})$ and $\mathscr{O}_y(\mathbf{z}) = -\nabla_y g(\mathbf{z}) + \nabla f_y(\mathbf{y})$, for all $\mathbf{z} \in Q$. The agents are distributed in the sense that an agent cannot query the other agent's oracle. In this example, Agent x does not have access to Agent y's private utility f_y, and vice versa.

Accuracy measure. For a pair of decisions $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in Q$, we use the following accuracy measure:

$$\Delta(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \triangleq \max_{\mathbf{z} \in \mathcal{B} \cap Q} [F(\bar{\mathbf{x}}, \mathbf{y}) - F(\mathbf{x}, \bar{\mathbf{y}})],$$

where $\mathcal{B} \triangleq \mathcal{B}_x \times \mathcal{B}_y$. We say that $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in Q$ is an *ε-approximate saddle point* of Problem (1) if $\Delta(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq \varepsilon$. Our goal is to find such an ε-approximate saddle point for any $\varepsilon > 0$. For the classic problem of constrained optimization with bounded domains, one can enclose the constrained sets in the balls $\mathcal{B}_x$ and $\mathcal{B}_y$ with sufficiently large radius (e.g., the diameter of the constrained sets), then the restricted saddle problem in form (1) coincides with the original one.

2.3. Communication-efficient algorithmic results from existing literature

The classic EG method has convergence guarantee in the number of iterations (Nemirovski, 2004; Juditsky et al., 2011); while in our distributed setup, each iteration of EG takes two communication rounds and two queries to $\mathscr{O}_x$ and $\mathscr{O}_y$, respectively. Therefore, we state the communication and computational costs of EG implied from its existing analysis.

Proposition 1 (Juditsky et al. (2011, Eq. (6.21))) *For the* EG *method for SPs, for any $\varepsilon > 0$,*

$$T^{\text{EG}}_{\mathscr{P}_{SP}}(\varepsilon) = \frac{L_{xy} D_x D_y}{\varepsilon} + \frac{L_x D_x^2}{\varepsilon} + \frac{L_y D_y^2}{\varepsilon},$$
$$N^{\text{EG}}_{\mathscr{P}_{SP}}(\varepsilon) = (c_x + c_y) \cdot \left(\frac{L_{xy} D_x D_y}{\varepsilon} + \frac{L_x D_x^2}{\varepsilon} + \frac{L_y D_y^2}{\varepsilon} \right).$$

In the distributed setup, some recent papers (Zhang et al., 2024; Yoon et al., 2025; Yoon and Loizou, 2025) propose different communication-efficient approaches using stochastic gradient oracles; however, when applied to standard deterministic oracles, these methods fail to outperform the communication cost of EG. While Zindari et al. (2025) achieves logarithmic communication rounds, their result is limited to the special case of weakly-coupled problems. Consequently, the classic EG method remains a formidable baseline, and *improving its communication complexity for general distributed SPs remains a significant challenge*.

3. (Template) Decoupled method for saddle problems

In this section, we provide a (template) communication-efficient algorithm, which reduces an SP into a sequence of coordinate-wise computation tasks that can be solved in a fully decoupled way.

Assembled norm. Let $\alpha_x, \alpha_y > 0$ (to be fixed later). Let the block diagonal linear operator $\mathbf{P} = \alpha_x \mathbf{P}_x \oplus \alpha_y \mathbf{P}_y$.[1] We consider the space $\mathcal{E}$ to be measured by the following assembled norm: $\|\mathbf{z}\|_{\mathcal{E}} = \langle \mathbf{P}\mathbf{z}, \mathbf{z} \rangle^{\frac{1}{2}} \equiv \sqrt{\alpha_x \|\mathbf{x}\|_x^2 + \alpha_y \|\mathbf{y}\|_y^2}$, for all $\mathbf{z} \in \mathcal{E}$.

1. That is, for all $\mathbf{z} \in \mathcal{E}_x \times \mathcal{E}_y$, $\mathbf{P}\mathbf{z} = \alpha_x \mathbf{P}_x \mathbf{x} + \alpha_y \mathbf{P}_y \mathbf{y}$.

DM-SP. Now, we introduce the Decoupled Method for Saddle Problems (DM-SP) in Algorithm 1. This algorithm adopts the Reduced-Operator Method (ROM) framework proposed by Nesterov (2023). While the original ROM defines the midpoint $(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})$ via an abstract "Essential Step", we tailor this step for the distributed setting and show that it allows agents to compute their respective coordinates in a fully decoupled way. Consequently, each iteration requires only two communication rounds: first, to exchange the computed midpoints (Line 4); and second, to exchange the "extragradient" vectors evaluated at these midpoints (Line 6).

Algorithm 1 DM-SP$(f, (\psi_x, \psi_y), \mathbf{z}^0, (\lambda_t)_{t\geq 1}, (\alpha_x, \alpha_y))$

1: $\mathbf{v}^0 = (\mathbf{v}_x^0, \mathbf{v}_y^0) = \mathbf{z}^0$.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: Let $f^t(\mathbf{z}) \triangleq f(\mathbf{z}) + \frac{\alpha_x \lambda_{t+1}}{2}\|\mathbf{x} - \mathbf{v}_x^t\|_x^2 - \frac{\alpha_y \lambda_{t+1}}{2}\|\mathbf{y} - \mathbf{v}_y^t\|_y^2$.
4: Agent x finds $\mathbf{x}^{t+1} \approx \operatorname{argmin}_{\mathbf{x}\in \operatorname{dom}\psi_x} \left[f^t(\mathbf{x}, \mathbf{v}_y^t) + \psi_x(\mathbf{x})\right]$ and $\psi_x'(\mathbf{x}^{t+1}) \in \partial\psi_x(\mathbf{x}^{t+1})$ s.t.

$$\|\nabla_x f^t(\mathbf{x}^{t+1}, \mathbf{v}_y^t) + \psi_x'(\mathbf{x}^{t+1})\|_{x^*} \leq \delta_x^{(t+1)}\|\mathbf{x}^{t+1} - \mathbf{v}_x^t\|_x,$$

and Agent y finds $\mathbf{y}^{t+1} \approx \operatorname{argmin}_{\mathbf{y}\in \operatorname{dom}\psi_y} \left[-f^t(\mathbf{v}_x^t, \mathbf{y}) + \psi_y(\mathbf{y})\right]$ and $\psi_y'(\mathbf{y}^{t+1}) \in \partial\psi_y(\mathbf{y}^{t+1})$ s.t.

$$\|-\nabla_y f^t(\mathbf{v}_x^t, \mathbf{y}^{t+1}) + \psi_y'(\mathbf{y}^{t+1})\|_{y^*} \leq \delta_y^{(t+1)}\|\mathbf{y}^{t+1} - \mathbf{v}_y^t\|_y,$$

where $\delta_x^{(t+1)} = \frac{\alpha_x \lambda_{t+1}}{2}$ and $\delta_y^{(t+1)} = \frac{\alpha_y \lambda_{t+1}}{2}$; then exchange $\mathbf{z}^{t+1} = (\mathbf{x}^{t+1}, \mathbf{y}^{t+1})$.
5: Generate $\bar{\mathbf{z}}^{t+1} = (\bar{\mathbf{x}}^{t+1}, \bar{\mathbf{y}}^{t+1}) = \left(\sum_{t=1}^T a_t\right)^{-1} \sum_{t=1}^T a_t \mathbf{z}^t$.
6: Agents x and y compute $\nabla_x f(\mathbf{z}^{t+1})$ and $-\nabla_y f(\mathbf{z}^{t+1})$ respectively; then exchange

$$V_\psi(\mathbf{z}^{t+1}) \triangleq \left(\nabla_x f(\mathbf{z}^{t+1}) + \psi_x'(\mathbf{x}^{t+1}), -\nabla_y f(\mathbf{z}^{t+1}) + \psi_y'(\mathbf{y}^{t+1})\right).$$

7: $a_{t+1} = \frac{2\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}^t - \mathbf{z}^{t+1}\rangle}{\|V_\psi(\mathbf{z}^{t+1})\|_{\mathcal{E}^*}^2}$.
8: $\mathbf{v}^{t+1} = (\mathbf{v}_x^{t+1}, \mathbf{v}_y^{t+1}) = \operatorname{argmin}_{\mathbf{v}\in Q} \left[a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}\rangle + \frac{1}{2}\|\mathbf{v} - \mathbf{v}^t\|_{\mathcal{E}}^2\right]$.
9: **end for**

We say Algorithm 1 is a template method because we have not yet specified the implementation of its Line 4. We defer the detailed implementation to Section 5. For now, let us proceed with the main convergence result of Algorithm 1. Our result also considers the case that D_x and D_y are not known, and we use the inexact estimates $\hat{D}_x$ and $\hat{D}_y$ instead. Further, let us denote $\theta \triangleq \sqrt{\frac{D_x \hat{D}_y}{\hat{D}_x D_y}}$.

Theorem 2 *For the problem class $\mathscr{P}_{SP}$, with the choices of $\alpha_x = \frac{L_{xy}\hat{D}_y}{\hat{D}_x}$, $\alpha_y = \frac{L_{xy}\hat{D}_x}{\hat{D}_y}$, and $\lambda_t \equiv \lambda = 2$, the solution $\bar{\mathbf{z}}^T$ in* DM-SP *(Algorithm 1) is an ε-approximate saddle point, where*

$$T = 2 + \frac{L_{xy} D_x D_y}{\varepsilon}(\theta^2 + \theta^{-2}).$$

Corollary 3 *Following from Theorem 2, the communication cost in* DM-SP *(Algorithm 1) is minimized when $\theta = 1$ (that is, when the distance estimates $\hat{D}_x$ and $\hat{D}_y$ satisfy $\frac{\hat{D}_x}{D_x} = \frac{\hat{D}_y}{D_y}$). This results*

in the following communication cost:

$$T^{\text{DM-SP}}_{\mathscr{P}_{SP}}(\varepsilon) = \mathcal{O}\Big(\frac{L_{xy}D_xD_y}{\varepsilon}\Big). \tag{2}$$

Our communication result shows clear improvement over the existing state-of-the-art methods in distributed setup. Moreover, it also conveys the clear intuition: the communication cost is decided by the *coupled conditioning* of $L_{xy}D_xD_y$ (which measures how much a player's decision variable impacts upon *the other player's* objective), and is independent of the *diagonal conditioning* of $L_xD_x^2 + L_yD_y^2$. As we will show later in Section 6, this result matches the communication lower bound, and cannot be further improved by any algorithm under "gradient-span framework".

Furthermore, with certain concrete implementation of Line 4, this decoupled method would simultaneously lead to strict improvement over the long-standing EG method. We defer these detailed discussions to in Section 5.

4. Technical components

We first introduce the general notion of composite variational inequality problem (VIP) in Section 4.1, which is the backbone of our problems. Then, in Sections 4.2 to 4.4, we introduce the three technical components of our algorithm: (i) Reduced-Operator Method (ROM) that reduces an SP to an MS subproblem; (ii) Fully Decoupled Solver (FDS) that further reduces a (weakly coupled) MS subproblem to the coordinate-wise Minimization of Residual Norms (MRNs); and finally, (iii) the efficient solver for MRNs.

4.1. Preliminary: Variational inequality problems

Let the finite-dimensional real vector space $\mathcal{E}$ be equipped with the norm $\|\mathbf{z}\|_{\mathcal{E}} = \langle \mathbf{P}\mathbf{z}, \mathbf{z}\rangle^{\frac{1}{2}}$, where $\mathbf{P}\colon \mathcal{E} \to \mathcal{E}^*$ is a self-adjoint positive definite linear operator. For any operator $V(\cdot)\colon \operatorname{dom}\psi \to \mathcal{E}^*$ and any function $\psi(\cdot)\colon \mathcal{E} \to \mathbb{R}\cup\{+\infty\}$, we say that $\mathbf{z}^* \in \mathcal{E}$ is a *solution* of the VIP of (V,ψ) if

$$\langle V(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^*\rangle + \psi(\mathbf{z}) \geq \psi(\mathbf{z}^*), \text{ for all } \mathbf{z} \in \operatorname{dom}\psi. \tag{3}$$

Assumption for VIPs. Let us introduce the following assumption:

(A1') The function ψ is a (simple) proper closed convex function. The operator V is monotone: $\langle V(\mathbf{z}') - V(\mathbf{z}), \mathbf{z}' - \mathbf{z}\rangle \geq 0$, for all $\mathbf{z}', \mathbf{z} \in \operatorname{dom}\psi$.

Under (A1'), $\mathbf{z}^* \in \mathcal{E}$ is the solution of (3) if and only if $\mathbf{0} \in V(\mathbf{z}^*) + \partial\psi(\mathbf{z}^*)$.

Indeed, under (A1), the saddle points of (1) coincide with the solutions to the VIP given by

$$(V(\mathbf{z}) = (\nabla_x f(\mathbf{z}), -\nabla_y f(\mathbf{z})), \psi(\mathbf{z}) = \psi_x(\mathbf{x}) + \psi_y(\mathbf{y})).$$

Moreover, the above VIP satisfies (A1').

4.2. Reduced-operator method for composite variational inequality problems

Now, we introduce the Reduced-Operator Method (ROM) recently introduced in Nesterov (2023), with which we reduce the VIP to a sequence of Monteiro-Svaiter subproblems (Monteiro and Svaiter, 2013).[2] Let us first define the Monteiro-Svaiter (MS) subproblem, which asks to find an approximate solution for the regularized problem with small subgradient norm.

2. Appendix A details the nuances of our stepsize choice compared to prior literature.

MS subproblem. Given an operator $V\colon \operatorname{dom}\psi \to \mathcal{E}^*$, a function $\psi\colon \mathcal{E} \to \mathbb{R}\cup\{+\infty\}$, a reference point $\mathbf{v}\in\operatorname{dom}\psi$, and a real number $\lambda>0$, we say $(\mathbf{z}^+, \psi'(\mathbf{z}^+))$ is a *solution* of the MS subproblem if $\mathbf{z}^+\in\operatorname{dom}\psi$, $\psi'(\mathbf{z}^+)\in\partial\psi(\mathbf{z}^+)$, and

$$\|V(\mathbf{z}^+)+\psi'(\mathbf{z}^+)+\lambda\mathbf{P}(\mathbf{z}^+-\mathbf{v})\|_{\mathcal{E}^*}\le\lambda\|\mathbf{z}^+-\mathbf{v}\|_{\mathcal{E}}. \tag{4}$$

We will discuss how to solve the MS subproblems later in Section 4.3. But for now, let us assume there exists a solver $\mathcal{M}^{\mathrm{MS}}(V,\psi,v,\lambda)$ for the MS subproblems. Built upon such a solver $\mathcal{M}^{\mathrm{MS}}$, we now introduce ROM in Algorithm 2. At each iteration t: the solver $\mathcal{M}^{\mathrm{MS}}$ returns a solution $(\mathbf{z}^{t+1},\psi'(\mathbf{z}^{t+1}))$ for the MS subproblem built at reference point $\mathbf{v}^t$; this solution is used as a midpoint to compute subgradient $V_\psi(\mathbf{z}^{t+1})$; then the 'extra' step is taken with the stepsize a_{t+1}.

Algorithm 2 $\mathrm{ROM}_{\|\cdot\|_{\mathcal{E}}}(V,\psi,\mathbf{z}^0,(\lambda_t)_{t\ge1}\mid\mathcal{M}^{\mathrm{MS}})$

Require: A solver $\mathcal{M}^{\mathrm{MS}}$ for the MS subproblems.

$\mathbf{v}^0=\mathbf{z}^0$.
for $t=0,1,\cdots$ **do**
  $(\mathbf{z}^{t+1},\psi'(\mathbf{z}^{t+1}))=\mathcal{M}^{\mathrm{MS}}(V,\psi,\mathbf{v}^t,\lambda_{t+1})$.
  $V_\psi(\mathbf{z}^{t+1})=V(\mathbf{z}^{t+1})+\psi'(\mathbf{z}^{t+1})$.
  $a_{t+1}=\frac{2\langle V_\psi(\mathbf{z}^{t+1}),\mathbf{v}^t-\mathbf{z}^{t+1}\rangle}{\|V_\psi(\mathbf{z}^{t+1})\|^2_{\mathcal{E}^*}}$.
  $\mathbf{v}^{t+1}=\operatorname{argmin}_{\mathbf{v}\in\operatorname{dom}\psi}\big[a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}),\mathbf{v}\rangle+\frac12\|\mathbf{v}-\mathbf{v}^t\|^2_{\mathcal{E}}\big]$.
end for

Next, let us present the bound for ROM.

Lemma 4 ROM *(Algorithm 2) ensures for all* $\mathbf{v}\in\operatorname{dom}\psi$ *and for all* $T\ge1$,

$$\sum_{t=0}^{T-1}a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}),\mathbf{z}^{t+1}-\mathbf{v}\rangle\le\frac12\|\mathbf{v}^0-\mathbf{v}\|^2_{\mathcal{E}}-\frac12\|\mathbf{v}^T-\mathbf{v}\|^2_{\mathcal{E}}.$$

Moreover, we have $a_{t+1}\ge\frac{1}{\lambda_{t+1}}$, *for all* $t\ge0$.

4.3. Fully decoupled solver for MS subproblems with weak couplings

We now address the MS subproblems introduced in Section 4.2. Specifically, we focus on the subproblems arising from the application of ROM to SPs, given by $(V,\psi,\mathbf{v},\lambda)$ where

$$V=(\nabla_x f,-\nabla_y f),\quad \psi(\mathbf{z})=\psi_x(\mathbf{x})+\psi_y(\mathbf{y}),\quad \mathbf{v}=(\mathbf{v}_x,\mathbf{v}_y), \tag{5}$$

and the assembled norm $\|\cdot\|_{\mathcal{E}}$ with parameters α_x and α_y. In this section, we will introduce a Fully Decoupled Solver (FDS), which reduces the MS subproblems to coordinate-wise Minimization of Residual Norms (MRNs).

Let us first define the problem of MRN. The MRN asks to find an approximate solution $\hat{\mathbf{z}}^+$ of the VIP of $(\hat V,\hat\psi)$ with the accuracy specified as follows:[3]

3. To avoid the notation conflict, we add a hat to the parameter (that looks like $\hat{\cdot}$) to indicate that it corresponds to the coordinate-wise problem.

Minimization of residual norm. Given an operator $\hat{V}\colon \operatorname{dom}\hat{\psi} \to \hat{\mathcal{E}}^*$, a function $\hat{\psi}\colon \hat{\mathcal{E}} \to \mathbb{R}\cup\{+\infty\}$, a reference point $\hat{\mathbf{v}} \in \operatorname{dom}\hat{\psi}$, and an accuracy $\hat{\delta} > 0$, we say $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ minimizes the residual norm to $\hat{\delta}$-*relative distance accuracy*, if $\hat{\mathbf{z}}^+ \in \operatorname{dom}\hat{\psi}$, $\hat{\psi}'(\hat{\mathbf{z}}^+) \in \partial\hat{\psi}(\hat{\mathbf{z}}^+)$, and

$$\|\hat{V}(\hat{\mathbf{z}}^+) + \hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*} \le \hat{\delta}\|\hat{\mathbf{z}}^+ - \hat{\mathbf{v}}\|_{\hat{\mathcal{E}}}.$$

Let us, again, defer the discussion of solving MRNs to Section 4.4. But for now, let us assume there exist solvers $\mathcal{M}_x^{\mathrm{MRN}}$ and $\mathcal{M}_y^{\mathrm{MRN}}$ for the coordinate-wise MRNs. We say that an MS subproblem has a *weak coupling* if $\lambda \ge 2\bar{L}_{\mathrm{c}} \triangleq \frac{2L_{xy}}{\sqrt{\alpha_x\alpha_y}}$. In particular, for an MS subproblem with a weak coupling, we apply Algorithm 3—Fully Decoupled Solver (FDS)—that solves for each decision variable separately. We show in Lemma 5 that FDS returns the correct solution *in one round*.

Algorithm 3 $\mathrm{FDS}_{\|\cdot\|_{\mathcal{E}}}((\nabla_x f, -\nabla_y f), (\psi_x, \psi_y), \mathbf{v}, \lambda \mid (\mathcal{M}_x^{\mathrm{MRN}}, \mathcal{M}_y^{\mathrm{MRN}}))$

Require: Solvers $\mathcal{M}_x^{\mathrm{MRN}}$ and $\mathcal{M}_y^{\mathrm{MRN}}$ for the coordinate-wise MRNs.

1: Agent x and Agent y respectively compute

$$(\mathbf{x}^+, \psi_x'(\mathbf{x}^+)) = \mathcal{M}_x^{\mathrm{MRN}}\big(\nabla_x f(\cdot, \mathbf{v}_y), \psi_x + \frac{\alpha_x\lambda}{2}\|\cdot - \mathbf{v}_x\|_x^2, \mathbf{v}_x, \delta_x\big) \text{ and}$$

$$(\mathbf{y}^+, \psi_y'(\mathbf{y}^+)) = \mathcal{M}_y^{\mathrm{MRN}}\big(-\nabla_y f(\mathbf{v}_x, \cdot), \psi_y + \frac{\alpha_y\lambda}{2}\|\cdot - \mathbf{v}_y\|_y^2, \mathbf{v}_y, \delta_y\big),$$

where $\delta_x = \frac{\alpha_x\lambda}{2}$ and $\delta_y = \frac{\alpha_y\lambda}{2}$.

2: **return** $(\mathbf{z}^+, \psi'(\mathbf{z}^+))$, where $\mathbf{z}^+ = (\mathbf{x}^+, \mathbf{y}^+)$ and $\psi'(\mathbf{z}^+) = (\psi_x'(\mathbf{x}^+), \psi_y'(\mathbf{y}^+))$.

Lemma 5 *Under (A3), for $\lambda \ge 2\bar{L}_{\mathrm{c}}$,* FDS *(Algorithm 3) returns a solution of the MS subproblem introduced in Equation (5).*

4.4. Minimization of residual norms

We arrive at the last building component, the Minimization of Residual Norms (MRNs).

Assumptions for MRNs. Let us introduce the following assumptions:

(Â1) The function $\hat{\psi}$ is a (simple) proper closed convex function. The operator $\hat{V}$ is monotone over $\operatorname{dom}\hat{\psi}$.

(Â2) The set-valued operator $\hat{V} + \partial\hat{\psi}$ is $\hat{\mu}$-strongly maximally monotone over $\operatorname{dom}\hat{\psi}$.

(Â3) The operator $\hat{V}(\hat{\mathbf{z}})$ is $\hat{L}$-Lipschitz continuous in $\hat{\mathbf{z}} \in \operatorname{dom}\hat{\psi}$.

We will leverage efficient existing solvers for MRNs. In particular, for the SPs of interest, the corresponding coordinate-wise residuals are gradients of convex functions. Therefore, we can use, for instance, the Accelerated Gradient Method (AGM) from the literature (Lan et al., 2023).

The theoretical guarantee provided in the literature is usually based on the distance-to-solution accuracy (cf. Definition 6). We show in Lemma 7 that, under strong maximal monotonicity, the relative distance accuracy required in this paper can be implied from distance-to-solution accuracy.

Definition 6 (Distance-to-solution accuracy) *We say that $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ satisfies $\hat{\varepsilon}$-*distance-to-solution accuracy *if $\hat{\mathbf{z}}^+ \in \operatorname{dom}\hat{\psi}$, $\hat{\psi}'(\hat{\mathbf{z}}^+) \in \partial\hat{\psi}(\hat{\mathbf{z}}^+)$, and $\|\hat{V}(\hat{\mathbf{z}}^+) + \hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*} \le \hat{\varepsilon}\|\hat{\mathbf{v}} - \hat{\widetilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}}$ for some $\hat{\widetilde{\mathbf{z}}}$ in the solution set of the VIP of $(\hat{V}, \hat{\psi})$.*

Lemma 7 *Let $\hat{\varepsilon} < \hat{\mu}$. Under (Â2), if $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ satisfies $\hat{\varepsilon}$-distance-to-solution accuracy, then $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ minimizes the residual norm to $\frac{\hat{\mu}\hat{\varepsilon}}{\hat{\mu}-\hat{\varepsilon}}$-relative distance accuracy.*

Now, we state the gradient query complexity with distance-to-solution accuracy. The original result of Lan et al. (2023) is given in projected gradient norm, which can be converted to the subgradient norm considered in this paper.

Lemma 8 (Lan et al. (2023)) *Assume (Â1), (Â3), and that the solution set of the VIP of $(\hat{V}, \hat{\psi})$ is non-empty. Assume $\hat{V} = \nabla \hat{f}$, where $\hat{f}$ is a continuously differentiable function defined on an open set containing $\operatorname{dom} \hat{\psi}$. Then, there exists an algorithm, denoted by*

$$(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+)) = \text{AGM}\ (\hat{f}, \hat{\psi}, \hat{\mathbf{v}}, \hat{\varepsilon} \mid \hat{L}),$$

that takes no more than $34 \cdot \sqrt{\frac{3\hat{L}}{2\hat{\varepsilon}}}$ queries to $\nabla \hat{f}(\cdot)$, and ensures $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ satisfies $\hat{\varepsilon}$-distance-to-solution accuracy.

5. Decoupled method for saddle problems: Concrete implementation

We are now back to considering the SPs in Equation (1). Let us combine the technical components in Section 4 and present the final, implementable algorithm.

Implementation of DM-SP. We use the AGM in Lemma 8 for Minimization of Residual Norms:

$$\mathcal{M}_x^{\text{MRN}}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \hat{\delta}) \triangleq \text{AGM}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \frac{2\hat{\delta}}{3} \mid L_x),$$

$$\mathcal{M}_y^{\text{MRN}}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \hat{\delta}) \triangleq \text{AGM}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \frac{2\hat{\delta}}{3} \mid L_y).$$

Let us denote $(\psi_x, \psi_y)(\mathbf{z}) \triangleq \psi_x(\mathbf{x}) + \psi_y(\mathbf{y})$. For the assembled norm $\|\cdot\|_{\mathcal{E}}$ with parameters α_x and α_y, we implement the DM-SP as follows:

$$\text{ROM}_{\|\cdot\|_{\mathcal{E}}}\big((\nabla_x f, -\nabla_y f), (\psi_x, \psi_y), \mathbf{z}^0, (\lambda_t)_{t\geq 1} \mid \text{FDS}_{\|\cdot\|_{\mathcal{E}}}(\cdot,\cdot,\cdot,\cdot \mid (\mathcal{M}_x^{\text{MRN}}, \mathcal{M}_y^{\text{MRN}}))\big). \quad (6)$$

Combining Lemmas 4, 5, 7 and 8, we obtain the following result on the computational cost.

Theorem 9 *For the problem class $\mathscr{P}_{SP}$, with the same choices of α_x, α_y, and $(\lambda_t)_{t\geq 0}$ as in Theorem 2,* DM-SP *with the implementation in Equation (6) generates an ε-approximate solution with the computational cost bounded by*

$$(c_x + c_y)\frac{L_{xy}D_xD_y}{\varepsilon}(\theta^2 + \theta^{-2}) + 51\sqrt{2}\Big(\frac{L_{xy}D_xD_y}{\varepsilon}\Big)^{\frac{1}{2}}\Big(c_x\Big(\frac{L_xD_x^2}{\varepsilon}\Big)^{\frac{1}{2}} + c_y\Big(\frac{L_yD_y^2}{\varepsilon}\Big)^{\frac{1}{2}}\Big)(\theta + \theta^{-1}).$$

Following from Theorem 9, the computational cost in DM-SP (Algorithm 1) is minimized when $\theta = 1$ (that is, when the distance estimates $\hat{D}_x$ and $\hat{D}_y$ satisfy $\frac{\hat{D}_x}{D_x} = \frac{\hat{D}_y}{D_y}$). This results in the following communication cost:

$$N_{\mathscr{P}_{SP}}^{\text{DM-SP}}(\varepsilon) = \mathcal{O}\Bigg((c_x + c_y)\frac{L_{xy}D_xD_y}{\varepsilon} + \Big(\frac{L_{xy}D_xD_y}{\varepsilon}\Big)^{\frac{1}{2}}\Big(c_x\Big(\frac{L_xD_x^2}{\varepsilon}\Big)^{\frac{1}{2}} + c_y\Big(\frac{L_yD_y^2}{\varepsilon}\Big)^{\frac{1}{2}}\Big)\Bigg). \quad (7)$$

Now, let us compare the above computational cost of DM-SP with the long-standing EG method, given by $\mathcal{O}\big((c_x + c_y) \cdot \frac{L_{xy}D_xD_y + L_xD_x^2 + L_yD_y^2}{\varepsilon}\big)$, in Proposition 1. Our computational cost is consistently no worse than EG, and is substantially better when

$$L_xD_x^2 + L_yD_y^2 \gg L_{xy}D_xD_y + \sqrt{L_{xy}D_xD_y} \cdot \Big(\frac{c_x}{c_x + c_y}\sqrt{L_xD_x^2} + \frac{c_y}{c_x + c_y}\sqrt{L_yD_y^2}\Big).$$

For instance, for $c_x = c_y$, our computational cost is substantially better when the *diagonal conditioning dominates*—that is, when

$$L_xD_x^2 + L_yD_y^2 \gg L_{xy}D_xD_y.$$

To our knowledge, DM-SP is *the first method that strictly improves over the classic* EG *method* in computational cost, without redundant logarithm term.

We include a more detailed discussion on the technical difference between DM-SP and the existing methods in Appendix B.

6. Lower complexity bounds under gradient-span framework

In this section, as a first step toward characterizing the complexity of the distributed setting, we present lower complexity bounds for the distributed algorithms within the gradient-span framework. Within this framework, at any communication round, an agent's next query point stays in the linear span of: (a) all local oracle answers accumulated by that agent thus far; and (b) all oracle answers from the other agent received up to the previous communication round. We include the detailed definitions in Section C.3.

This gradient-span framework is common in the literature (Nesterov, 2004; Woodworth, 2021), and it encompasses our proposed method as well as the majority of existing deterministic first-order algorithms for SPs (Nemirovski, 2004; Chambolle and Pock, 2011; Lin et al., 2020; Yoon et al., 2025; Zindari et al., 2025).

Now, we present the lower complexity bounds implied from the existing worst-case construction for SPs. In the original paper of Zhang et al. (2022), they present the lower bound for the complexity of full gradient oracle. Indeed, we show that the analysis in Zhang et al. (2022) can be adapted to the distributed setup, and implies the lower bounds on the communication and computational costs considered in our paper.

Proposition 10 *Consider any distributed first-order algorithm $\mathscr{M}$ within the gradient-span framework. For any $\varepsilon > 0$,*

$$T^{\mathscr{M}}_{\mathscr{P}_{SP}}(\varepsilon) = \Omega\Big(\frac{L_{xy}D_xD_y}{\varepsilon}\Big),$$

$$N^{\mathscr{M}}_{\mathscr{P}_{SP}}(\varepsilon) = \Omega\Bigg((c_x + c_y)\frac{L_{xy}D_xD_y}{\varepsilon} + c_x\sqrt{\frac{L_xD_x^2}{\varepsilon}} + c_y\sqrt{\frac{L_yD_y^2}{\varepsilon}}\Bigg).$$

Proposition 10 confirms that DM-SP is *communication-optimal* within the gradient-span framework. Its complexity of $\mathcal{O}\big(\frac{L_{xy}D_xD_y}{\varepsilon}\big)$ matches the lower bound exactly, proving that communication cost depends solely on the "coupled conditioning" $L_{xy}D_xD_y$, independent of the "diagonal conditioning". We, however, notice the gap between our computational cost and the lower bound in Proposition 10. This is indeed an open question for the class of non-distributed SPs as well.

7. Discussions and conclusions

Thus far, we have studied the SPs that correspond to two-player zero-sum games. Finally, we extend our algorithmic results to the broader class of monotone Variational Inequality Problems (VIPs) with separable proximal terms. Our main result for VIPs generalizes our decoupled method from two-player zero-sum games to multiplayer general-sum games, and improves the state-of-the-art communication complexity for the class by dropping the dependency on "diagonal conditioning" (cf. Remark 18). See Appendix D for the detailed presentation.

This paper studies communication and computational complexity in convex-concave SPs and monotone VIPs. For the class of SPs, we settle the communication complexity in the distributed setup within gradient-span framework, and strictly improve the long-standing gradient query complexity of EG method, while it remains an important open question to further close the gap towards the lower bound of gradient query complexity. For the class of distributed VIPs, we improve the state-of-the-art communication complexity, while it remains open whether a non-trivial lower bound of communication complexity can be proven for the class.

Acknowledgments

The authors thank Ali Zindari and Krishnendu Chatterjee for their helpful discussions and suggestions on this paper. RL acknowledges the support of ERC CoG 863818 (ForM-SMArt) and Austrian Science Fund (FWF) 10.55776/COE12.

References

Aharon Ben-Tal and Arkadi Nemirovski. Robust optimization–methodology and applications. *Mathematical programming*, 92(3):453–480, 2002.

Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle point problems: lower bounds, near-optimal and robust algorithms. *Optimization Methods and Software*, pages 1–18, 2025.

Radu Ioan Boţ and Enis Chenchene. Extra-gradient method with flexible anchoring: Strong convergence and fast residual decay. *arXiv preprint arXiv:2410.14369*, 2024.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Vincent Conitzer and Tuomas Sandholm. Communication complexity as a lower bound for learning in games. In *Proceedings of the twenty-first international conference on Machine learning*, page 24, 2004.

Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2021.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Sergiu Hart and Yishay Mansour. How long to equilibrium? the communication complexity of uncoupled equilibrium procedures. *Games and Economic Behavior*, 69(1):107–126, 2010.

Junling Hu, Michael P Wellman, et al. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, volume 98, pages 242–250, 1998.

Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

Guanghui Lan, Yuyuan Ouyang, and Zhe Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *arXiv preprint arXiv:2310.12139*, 2023.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pages 2738–2779. PMLR, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Arkadi 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.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. High-order reduced-gradient methods for composite variational inequalities. *arXiv preprint arXiv:2311.15154*, 2023.

Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004. ISBN 978-1-4613-4691-3. doi: 10.1007/978-1-4419-8853-9. URL https://doi.org/10.1007/978-1-4419-8853-9.

Noam Nisan and Ilya Segal. The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory*, 129(1):192–224, 2006.

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

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

Mikhail V Solodov and Benar F Svaiter. A hybrid projection-proximal point algorithm. *Journal of convex analysis*, 6(1):59–70, 1999.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1947.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.

Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020.

TaeHo Yoon and Nicolas Loizou. Pearl-prox: Proximal algorithm for resolving player drift in multiplayer federated learning. In *OPT 2025: Optimization for Machine Learning*, 2025.

TaeHo Yoon, Sayantan Choudhury, and Nicolas Loizou. Multiplayer federated learning: Reaching equilibrium with less communication. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9JX8XrTVEz.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022.

Siqi Zhang, Sayantan Choudhury, Sebastian U Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. In *The Twelfth International Conference on Learning Representations*, 2024.

Ali Zindari, Parham Yazdkhasti, Anton Rodomanov, Tatjana Chavdarova, and Sebastian U Stich. Decoupled SGDA for games with intermittent strategy communication. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZYkFTSEZ6k.

Appendix A. Stepsize choices in Algorithm 2

We remark on our specific stepsize choice in Line 6 of Algorithm 2 as well as the nuanced difference from the choices in the similar methods of Solodov and Svaiter (1999) and Nesterov (2023).

In the original work of Solodov and Svaiter (1999), they choose the classic Extragradient stepsize of $\frac{1}{\lambda_{t+1}}$; Nesterov (2023) shows that it is theoretically safe to set this stepsize to $\frac{\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}^t - \mathbf{z}^{t+1}\rangle}{\|V_\psi(\mathbf{z}^{t+1})\|^2_{\mathcal{E}^*}}$, considering the MS subproblem might be solved to a higher accuracy than the requirement in Equation (4); and in this paper, we follow the refined stepsize choices of Nesterov (2023), but we increase the stepsizes by two times as in Line 5 of Algorithm 2 since we are only dealing with first-order algorithms.

Appendix B. Technical difference from relevant methods

Our approach to convex-concave SPs is built on a multi-stage reduction: from the original problem to a Monteiro-Svaiter (MS) subproblem, then to coordinate-wise Minimization of Residual Norms (MRN), and finally to the deployment of efficient local solvers. This development distinguishes DM-SP from existing methods in several key aspects:

- **EG and HPE framework.** While the EG method (Nemirovski, 2004) relies on a single gradient step to find a midpoint, DM-SP utilizes an inexact local solver to compute a more accurate proximal point. This allows for significantly larger stepsizes for the 'extra' step and fewer total iterations. Furthermore, unlike the general HYBRID-PROXIMAL EXTRAGRADIENT (HPE) framework (Solodov and Svaiter, 1999), DM-SP provides a concrete, decoupled implementation that achieves provable acceleration over the EG baseline.

- **CATALYST variants.** Compared to the CATALYST variants (Lin et al., 2020; Yang et al., 2020; Wang and Li, 2020), which rely on black-box reductions from general convex-concave functions to the strongly convex-(strongly) concave class, our reduction is direct and structural. Consequently, DM-SP avoids the multiplicative squared-logarithmic overhead, $\mathcal{O}(\log^2(1/\varepsilon))$, and eliminates the need for prior knowledge of upper bounds on D_x and D_y.

- **Communication-focused methods.** The PROXSKIP-VI, PEARL-SGD, and PEARL-PROX methods (Zhang et al., 2024; Yoon et al., 2025; Yoon and Loizou, 2025) also address the communication complexity. Their methods do not take the extra step and simply use $\mathbf{x}^t$ and $\mathbf{y}^t$ as reference points, while DM-SP leverages the extra step to ensure faster convergence. Specifically, in the standard deterministic first-order oracle settings, PROXSKIP-VI is randomized and provides no speedup over EG, while PEARL-SGD and PEARL-PROX requires $\mathcal{O}(\varepsilon^{-3/2})$ iterations, making them strictly slower than both EG and our proposed method.

- **Weak-coupling approach.** While DECOUPLED-SGDA (Zindari et al., 2025) achieves logarithmic communication rounds, its results are restricted to the special case of weakly-coupled problems. DM-SP, by contrast, applies to the general class of convex-concave SPs and monotone VIPs without such structural restrictions.

Appendix C. Missing proofs

C.1. Proofs for technical components

We first present a three-point descent lemma in Lemma 11, which will be used repeatedly in the analysis. The following lemma follows directly from strong convexity, and can also be found in classic literature (Chen and Teboulle, 1993).

Lemma 11 *Let Q be a closed convex set. Let $a > 0$, $\mathbf{z} \in Q$, and $\mathbf{g} \in \mathcal{E}^*$. If*

$$\mathbf{z}^+ = \operatorname*{argmin}_{\mathbf{x} \in Q} \left[a\langle \mathbf{g}, \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_{\mathcal{E}}^2 \right], \tag{8}$$

then for all $\mathbf{v} \in Q$, we have

$$a\langle \mathbf{g}, \mathbf{v} - \mathbf{z}^+ \rangle + \frac{1}{2}\|\mathbf{z} - \mathbf{v}\|_{\mathcal{E}}^2 \geq \frac{1}{2}\|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 + \frac{1}{2}\|\mathbf{z} - \mathbf{z}^+\|_{\mathcal{E}}^2.$$

Now, we include the proof of Lemma 4—the regret bound for ROM.

Proof [Proof of Lemma 4] By Lemma 11 and by the optimality of $\mathbf{v}^{t+1}$, we have for all $\mathbf{v} \in \operatorname{dom} \psi$,

$$a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v} - \mathbf{v}^{t+1} \rangle + \frac{1}{2}\|\mathbf{v}^t - \mathbf{v}\|_{\mathcal{E}}^2 \geq \frac{1}{2}\|\mathbf{v}^{t+1} - \mathbf{v}\|_{\mathcal{E}}^2 + \frac{1}{2}\|\mathbf{v}^t - \mathbf{v}^{t+1}\|_{\mathcal{E}}^2,$$

and therefore,

$$\begin{aligned}
&a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v} - \mathbf{z}^{t+1} \rangle + \frac{1}{2}\|\mathbf{v}^t - \mathbf{v}\|_{\mathcal{E}}^2 \\
&\quad \geq a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}^{t+1} - \mathbf{z}^{t+1} \rangle + \frac{1}{2}\|\mathbf{v}^{t+1} - \mathbf{v}\|_{\mathcal{E}}^2 + \frac{1}{2}\|\mathbf{v}^t - \mathbf{v}^{t+1}\|_{\mathcal{E}}^2 \\
&\quad = a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}_t - \mathbf{z}^{t+1} \rangle + \frac{1}{2}\|\mathbf{v}^{t+1} - \mathbf{v}\|_{\mathcal{E}}^2 + a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}^{t+1} - \mathbf{v}_t \rangle + \frac{1}{2}\|\mathbf{v}^t - \mathbf{v}^{t+1}\|_{\mathcal{E}}^2 \\
&\quad \geq a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}_t - \mathbf{z}^{t+1} \rangle + \frac{1}{2}\|\mathbf{v}^{t+1} - \mathbf{v}\|_{\mathcal{E}}^2 - \frac{a_{t+1}^2}{2}\|V_\psi(\mathbf{z}^{t+1})\|_{\mathcal{E}^*}^2 \\
&\quad \geq \frac{1}{2}\|\mathbf{v}^{t+1} - \mathbf{v}\|_{\mathcal{E}}^2,
\end{aligned}$$

where the last inequality follows from the assignment of a_{t+1} in Algorithm 2 of Algorithm 2. Then, the desired bound follows from summing the above inequality over t from 0 to $T - 1$.

Next, we show the lower bound for a_t. For all $t \geq 1$, we have

$$\begin{aligned}
&\langle V_\psi(\mathbf{z}^t), \mathbf{v}^{t-1} - \mathbf{z}^t \rangle - \frac{1}{2\lambda_t}\|V_\psi(\mathbf{z}^t)\|_{\mathcal{E}^*}^2 \\
&\equiv \frac{\lambda_t}{2}\|\mathbf{z}^t - \mathbf{v}^{t-1}\|_{\mathcal{E}}^2 - \frac{1}{2\lambda_t}\|V_\psi(\mathbf{z}^t) + \lambda_t \mathbf{P}(\mathbf{z}^t - \mathbf{v}^{t-1})\|_{\mathcal{E}^*}^2 \\
&\geq 0,
\end{aligned}$$

where the last inequality follows from Equation (4). Therefore, we have

$$a_t = \frac{2\langle V_\psi(\mathbf{z}^t), \mathbf{v}^t - \mathbf{z}^t \rangle}{\|V_\psi(\mathbf{z}^t)\|_{\mathcal{E}^*}^2} \geq \frac{1}{\lambda_t}.$$

■

The correctness of the FDS for SPs can be implied as a direct consequence of the correctness of FDS for VIPs proven in Appendix D. We will get to this result later in Lemma 19.

Now, let us show the gradient query complexity for minimization of residual norms. We first prove a technical lemma.

Proof [Proof of Lemma 7] In view of the triangle inequality and the $\hat{\mu}$-strong maximal monotonicity of $\hat{V}+\hat{\psi}$, we have

$$\begin{aligned}\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}} &\le \|\hat{\mathbf{z}}^+-\hat{\mathbf{v}}\|_{\hat{\mathcal{E}}}+\|\hat{\mathbf{z}}^+-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}}\\ &\le \|\hat{\mathbf{z}}^+-\hat{\mathbf{v}}\|_{\hat{\mathcal{E}}}+\frac{1}{\hat{\mu}}\|\hat{V}(\hat{\mathbf{z}}^+)+\hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*}\\ &\le \|\hat{\mathbf{z}}^+-\hat{\mathbf{v}}\|_{\hat{\mathcal{E}}}+\frac{\hat{\varepsilon}}{\hat{\mu}}\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}}.\end{aligned}$$

Then, we have

$$\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}} \le \frac{\hat{\mu}}{\hat{\mu}-\hat{\varepsilon}}\|\hat{\mathbf{z}}^+-\hat{\mathbf{v}}\|_{\hat{\mathcal{E}}}.$$

Therefore, we have

$$\|V(\hat{\mathbf{z}}^+)+\hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*} \le \hat{\varepsilon}\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}} \le \frac{\hat{\mu}\hat{\varepsilon}}{\hat{\mu}-\hat{\varepsilon}}\|\hat{\mathbf{z}}^+-\hat{\mathbf{v}}\|_{\mathcal{E}}.$$

■

Lemma 8 considers the case of gradient of a convex function, and we include its proof below.

Proof [Proof of Lemma 8] It was originally shown in (Lan et al., 2023, Theorem 3.1) that AGM takes no more than $34\sqrt{\frac{3\hat{L}}{2\hat{\varepsilon}}}$ gradient queries and obtains $\hat{\mathbf{z}}' \in \operatorname{dom}\hat{\psi}$, such that there exists $\hat{\tilde{\mathbf{z}}} \in \operatorname{dom}\hat{\psi}$ with $\nabla\hat{f}(\hat{\tilde{\mathbf{z}}}) \in -\partial\hat{\psi}(\hat{\tilde{\mathbf{z}}})$, and

$$\|2\hat{L}(\hat{\mathbf{z}}^+-\hat{\mathbf{z}}')\|_{\hat{\mathcal{E}}} \le \frac{2}{3}\hat{\varepsilon}\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}}, \text{ where } \hat{\mathbf{z}}^+ = \operatorname*{argmin}_{\hat{\mathbf{z}}\in\operatorname{dom}\hat{\psi}}\left[\langle\nabla\hat{f}(\hat{\mathbf{z}}'),\hat{\mathbf{z}}\rangle+\hat{\psi}(\hat{\mathbf{z}})+\hat{L}\|\hat{\mathbf{z}}-\hat{\mathbf{z}}'\|_{\hat{\mathcal{E}}}^2\right].$$

By the optimality of $\hat{\mathbf{z}}^+$, there exists $\hat{\psi}'(\hat{\mathbf{z}}^+) \in \partial\hat{\psi}(\hat{\mathbf{z}}^+)$ such that

$$\nabla\hat{f}(\hat{\mathbf{z}}')+\hat{\psi}'(\hat{\mathbf{z}}^+)+2\hat{L}\hat{\mathbf{P}}(\hat{\mathbf{z}}^+-\hat{\mathbf{z}}')=\mathbf{0}.$$

Then, we have

$$\begin{aligned}&\|\nabla\hat{f}(\hat{\mathbf{z}}^+)+\hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*}\\ &\le \|\nabla\hat{f}(\hat{\mathbf{z}}')+\hat{\psi}'(\hat{\mathbf{z}}^+)\|_{\hat{\mathcal{E}}^*}+\|\nabla\hat{f}(\mathbf{z}^+)-\nabla\hat{f}(\hat{\mathbf{z}}')\|_{\hat{\mathcal{E}}}\\ &= \|2\hat{L}\hat{\mathbf{P}}(\hat{\mathbf{z}}^+-\hat{\mathbf{z}}')\|_{\hat{\mathcal{E}}^*}+\|\nabla\hat{f}(\hat{\mathbf{z}}^+)-\nabla\hat{f}(\hat{\mathbf{z}}')\|_{\hat{\mathcal{E}}}\\ &\le 3\hat{L}\|\hat{\mathbf{z}}^+-\hat{\mathbf{z}}'\|_{\hat{\mathcal{E}}}\\ &\le \hat{\varepsilon}\|\hat{\mathbf{v}}-\hat{\tilde{\mathbf{z}}}\|_{\hat{\mathcal{E}}}.\end{aligned}$$

■

C.2. Proofs for the main theorems

First, we prove the main convergence lemma for SPs in Lemma 12.

Lemma 12 *Under (A1) to (A3), for* $\lambda_{t+1} \equiv \lambda \geq \frac{2L_{xy}}{\sqrt{\alpha_x \alpha_y}}$, DM-SP *with the implementation in Equation (6) takes no more than* $2T$ *communication rounds, no more than*

$$T \cdot \big(1 + 34\sqrt{\frac{9L_x}{2\alpha_x \lambda}}\big)$$

queries to $\mathscr{O}_x$*, and no more than*

$$T \cdot \big(1 + 34\sqrt{\frac{9L_y}{2\alpha_y \lambda}}\big)$$

queries to $\mathscr{O}_y$*, and obtains an* ε*-approximate saddle point* $\bar{\mathbf{z}}^T$*, where*

$$T = \left\lceil \frac{\alpha_x \lambda D_x^2 + \alpha_y \lambda D_y^2}{2\varepsilon} \right\rceil.$$

Proof [Proof of Lemma 12] By (A1), we have

$$\Delta(\bar{\mathbf{z}}^T) \leq \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \max_{\mathbf{z} \in \mathcal{B} \cap Q} \left[\sum_{t=0}^{T-1} a_{t+1} \langle V_\psi(\mathbf{z}^{t+1}), \mathbf{z}^{t+1} - \mathbf{z} \rangle\right].$$

Further, with $\lambda \geq 2\bar{L}_{\mathrm{c}}$, by Lemmas 4 and 5, we have

$$\begin{aligned}
\Delta(\bar{\mathbf{z}}^T) &\leq \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \max_{\mathbf{z} \in \mathcal{B} \cap Q} \left[\sum_{t=0}^{T-1} a_{t+1} \langle V_\psi(\mathbf{z}^{t+1}), \mathbf{z}^{t+1} - \mathbf{z} \rangle\right] \\
&\leq \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \max_{\mathbf{z} \in \mathcal{B} \cap Q} \left[\frac{\alpha_x}{2} \|\mathbf{x}^0 - \mathbf{x}\|_x^2 + \frac{\alpha_y}{2} \|\mathbf{y}^0 - \mathbf{y}\|_y^2\right] \\
&\leq \Big(\sum_{t=0}^{T-1} \frac{1}{\lambda_{t+1}}\Big)^{-1} \cdot \frac{1}{2}(\alpha_x D_x^2 + \alpha_y D_y^2) \leq \varepsilon,
\end{aligned}$$

where the last inequality follows from the assignments of $(\lambda_t)_{t \geq 1}$ and T. Therefore, the number of communication rounds is bounded by $2T$.

Now we count the number of gradient queries. By Lemma 7, AGM always returns the solution with the required relative distance accuracy; and in view of Lemma 8, it takes no more than $34\sqrt{\frac{3L_x}{\alpha_x \lambda}}$ gradient queries to $\nabla_x f$ and no more than $34\sqrt{\frac{3L_y}{\alpha_y \lambda}}$ gradient queries to $\nabla_y f$. Therefore, the numbers of gradient queries to $\nabla_x f$ and $\nabla_y f$ are bounded by $T \cdot \big(1 + 34\sqrt{\frac{3L_x}{\alpha_x \lambda}}\big)$ and $T \cdot \big(1 + 34\sqrt{\frac{3L_y}{\alpha_y \lambda}}\big)$, respectively. ■

The main convergence result for SPs in Theorems 2 and 9 can be directly implied from Lemma 12.

C.3. Lower complexity bounds

In this section, we include the details for the lower complexity bounds.

Gradient-span framework. We first introduce the formal notions for the gradient-span framework (Nesterov, 2004; Zhang et al., 2022). Given a distributed first-order algorithm $\mathscr{M}$ and an SP instance P, at any communication round $t \geq 0$: suppose Agent x iteratively queries oracle $\mathscr{O}_x$ for $\tau_x(t)$ times, and we denote the input points as $(\mathbf{x}^{(t,1)}, \hat{\mathbf{y}}^{(t,1)}), \cdots, (\mathbf{x}^{(t,\tau_x(t))}, \hat{\mathbf{y}}^{(t,\tau_x(t))})$; and suppose

Agent y iterative queries oracle $\mathscr{O}_y$ for $\tau_y(t)$ times, and we denote the input points as $(\hat{\mathbf{x}}^{(t,1)}, \mathbf{y}^{(t,1)})$, $\cdots$, $(\hat{\mathbf{x}}^{(t,\tau_y(t))}, \mathbf{y}^{(t,\tau_y(t))})$. For Agent x, for all $t \geq 0$ and $0 \leq l \leq \tau_x(t)$, denote the index sets

$$\mathscr{I}_x^{t,l} = \big\{(k,i) \mid 0 \leq k < t \text{ and } 1 \leq i \leq \tau_x(k);\ \text{or } k = t \text{ and } 1 \leq i < l\big\},$$
$$\hat{\mathscr{I}}_y^{t} = \big\{(k,i) \mid 0 \leq k < t \text{ and } 1 \leq i \leq \tau_y(k)\big\},$$

the partial gradient spans

$$G_x^{t,l} = \operatorname{span}\{\mathscr{O}_x(\mathbf{x}^{(k,i)}, \hat{\mathbf{y}}^{(k,i)}) \mid (k,i) \in \mathscr{I}_x^{t,l}\},$$
$$\hat{G}_y^{t} = \operatorname{span}\{\mathscr{O}_y(\hat{\mathbf{x}}^{(k,i)}, \mathbf{y}^{(k,i)}) \mid (k,i) \in \hat{\mathscr{I}}_y^{t}\},$$

and the proximal subgradient spans

$$\Psi_x^{0,0} = \{\mathbf{0}\},$$
$$\Psi_x^{t,0} = \Psi_x^{t-1,\tau_x(t-1)},$$
$$\Psi_x^{t,l} = \operatorname{span}\Bigg(\Psi_x^{t,l-1} \bigcup \Bigg(\bigcup_{\mathbf{x} \in \mathbf{x}^0 - G_x^{t,l} - \Psi_x^{t,l-1}} \partial\psi_x(\mathbf{x})\Bigg)\Bigg) \quad (\text{where } l \geq 1),$$
$$\hat{\Psi}_y^{t} = \Psi_y^{t,0};$$

and we adopt the symmetric notations for Agent y, which we omitted here to simplify the presentation. If $\mathscr{M}$ is an algorithm within the *gradient-span* framework:

- For Agent x, for all $t \geq 0$ and $1 \leq l \leq \tau_x(t)$, the next query point should satisfy:
$$\mathbf{x}^{t,l} \in \mathbf{x}^0 - \mathbf{P}_x^{-1}(G_x^{t,l} + \Psi_x^{t,l}),$$
$$\hat{\mathbf{y}}^{t,l} \in \mathbf{y}^0 - \mathbf{P}_y^{-1}(\hat{G}_y^{t} + \hat{\Psi}_y^{t}).$$
The symmetric argument is made on Agent y as well.
- Moreover, for all $t \geq 0$, the approximate solution generated by the algorithm should satisfy:
$$\bar{\mathbf{x}}^{t+1} \in \mathbf{x}^0 + \mathbf{P}_x^{-1}(G_x^{t+1,0} + \Psi_x^{t+1,0}),$$
$$\bar{\mathbf{y}}^{t+1} \in \mathbf{y}^0 + \mathbf{P}_y^{-1}(G_y^{t+1,0} + \Psi_y^{t+1,0}).$$

Worst-case construction. Following the worst-case construction in Zhang et al. (2022, Eq. (12)), we consider the following SP in n-dimensional Euclidean spaces: $\mathbf{x}^0 = \mathbf{0}$, $\mathbf{y}^0 = \mathbf{0}$, and for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$f(\mathbf{x},\mathbf{y}) = \frac{L_{xy}}{2}\mathbf{x}^\top \mathbf{A}\mathbf{y} - \mathbf{b}^\top \mathbf{y}, \quad \psi_x(\mathbf{x}) = \frac{\mu_x}{2}\mathbf{x}^\top\mathbf{x}, \quad \psi_y(\mathbf{y}) = \frac{\mu_y}{2}\mathbf{y}^\top\mathbf{y}, \tag{9}$$

where μ_x, μ_y, and $\mathbf{b}$ are to be fixed later, and

$$A = \begin{pmatrix} & & & & 1 \\ & & & 1 & -1 \\ & & 1 & -1 & \\ & \iddots & \iddots & & \\ 1 & -1 & & & \end{pmatrix}.$$

Moreover, let us denote

$$\mathcal{H}_x^k \subseteq \begin{cases} \{0\}, & k = 1, \\ \operatorname{span}\big\{A^{2i}(Ab) \colon 0 \leq i \leq \lfloor \frac{k}{2} \rfloor - 1\big\}, & k \geq 2, \end{cases} \quad \textit{and} \quad \mathcal{H}_y^k \subseteq \operatorname{span}\big\{A^{2i}b \colon 0 \leq i \leq \lceil \frac{k}{2} \rceil - 1\big\}.$$

Let $\mathbf{e}_1$ be the unit vector with 1 in the first coordinate and 0 elsewhere.

Lemma 13 (Zhang et al. (2022, Theorem 3.5)) *Consider the SP in (9). For* $n \geq \max\{2\log_q(\frac{\mu_x\mu_y}{\sqrt{2}L_{xy}^2}), 4k\}$ *and* $\mathbf{b} = -\frac{L_{xy}^2}{4\mu_x}\mathbf{e}_1$, *if* $(\bar{\mathbf{x}}^t, \bar{\mathbf{y}}^t) \in \mathcal{H}_x^t \times \mathcal{H}_y^t$, *then*

$$\Delta(\bar{\mathbf{x}}^t, \bar{\mathbf{y}}^t) \geq q^t \cdot \frac{\mu_y \|\mathbf{y}^* - \mathbf{y}^0\|^2}{32},$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ *is the saddle point, and* $q = 1 + \frac{2\mu_x\mu_y}{L^2} - 2\sqrt{\left(\frac{\mu_x\mu_y}{L^2}\right)^2 + \frac{\mu_x\mu_y}{L^2}}$.

Now, we consider the distributed setup in this paper.

Proposition 14 *Consider any algorithm* $\mathscr{M}$ *under the gradient-span framework. Let* $\bar{\mathbf{z}}^{2t}$ *be the approximate solution returned by* $\mathscr{M}$ *after* $2t$ *communication rounds on the SP in (9). At this moment, let* n_x *be the number of times* $\mathcal{O}_x$ *gets queried thus far, and* n_y *be the number of times* $\mathcal{O}_y$ *gets queried thus far. Then, we have*

$$\bar{\mathbf{z}}^{2t} \in \mathcal{H}_x^k \times \mathcal{H}_y^k,$$

where $k = \min\{n_x, n_y, t\}$.

Then, following the scaling reduction in Zhang et al. (2022), we combine Theorems 13 and 14 with the parameter $\mu_x = \frac{64\varepsilon}{D_x^2}$ and $\mu_y = \frac{64\varepsilon}{D_y^2}$, and obtain the following lower bounds

$$T_{\mathscr{P}_{SP}}^{\mathscr{M}}(\varepsilon) = \Omega\Big(\frac{L_{xy}D_xD_y}{\varepsilon}\Big),$$
$$N_{\mathscr{P}_{SP}}^{\mathscr{M}}(\varepsilon) = \Omega\Big((c_x + c_y)\frac{L_{xy}D_xD_y}{\varepsilon}\Big).$$

Moreover, by the classic lower bounds of convex minimization (Nemirovskij and Yudin, 1983; Nesterov, 2004), we have

$$N_{\mathscr{P}_{SP}}^{\mathscr{M}}(\varepsilon) \geq \Omega\Big(c_x\sqrt{\frac{L_xD_x^2}{\varepsilon}} + c_y\sqrt{\frac{L_yD_y^2}{\varepsilon}}\Big).$$

Therefore, we have

$$N_{\mathscr{P}_{SP}}^{\mathscr{M}}(\varepsilon) = \Omega\Bigg((c_x + c_y)\frac{L_{xy}D_xD_y}{\varepsilon} + c_x\sqrt{\frac{L_xD_x^2}{\varepsilon}} + c_y\sqrt{\frac{L_yD_y^2}{\varepsilon}}\Bigg).$$

Appendix D. Monotone composite variational inequality problems

In this section, we study variational inequality problems (VIPs) (Nemirovski, 2004; Juditsky et al., 2011), a generalization of SPs that captures, for instance, multiplayer general-sum games.

D.1. Problem formulation

VIPs (with separable composite terms). Let us consider the VIP in Equation (3), where $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_K$ is the direct product of K finite-dimensional real vector spaces. For all $i \in [K]$: let the mapping $V_i \colon \operatorname{dom}\psi \to \mathcal{E}_i^*$, and let the function $\psi_i \colon \mathcal{E}_i \to \mathbb{R} \cup \{+\infty\}$. We consider the decomposition of $V(\mathbf{z}) = \big(V_1(\mathbf{z}), \cdots, V_K(\mathbf{z})\big)$ and $\psi(\mathbf{z}) = \psi_1(\mathbf{z}_1) + \cdots + \psi_K(\mathbf{z}_K)$, for all $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_K) \in \mathcal{E}$. Moreover, we denote $\operatorname{dom}\psi = \operatorname{dom}\psi_1 \times \cdots \times \operatorname{dom}\psi_K \triangleq Q$.

Assumptions for VIPs. Let us make the following assumptions:

(A2') Let $\mathbf{z}^0 = (\mathbf{z}_1^0, \cdots, \mathbf{z}_K^0) \in Q$ be a given point. There exists $\mathbf{z}^* = (\mathbf{z}_1^*, \cdots, \mathbf{z}_K^*) \in Q$ in the solution set of the VIP of (V, ψ), such that for all $i \in [K]$: $\mathbf{z}_i^* \in \mathcal{B}_i$, where $\mathcal{B}_i \triangleq \{\mathbf{z}_i \in \mathcal{E}_i \mid \|\mathbf{z}_i^0 - \mathbf{z}_i\|_i \leq D_i\}$ and $D_i > 0$ is a given distance.

(A3') The operator $V_i(\mathbf{z}_j; \mathbf{z}_{-j})$ is L_{ij}-Lipschitz continuous in $\mathbf{z}_j \in \operatorname{dom} \psi_j$ for any fixed $\mathbf{z}_{-j} \in \operatorname{dom} \psi_1 \times \cdots \times \operatorname{dom} \psi_{j-1} \times \operatorname{dom} \psi_{j+1} \times \cdots \times \operatorname{dom} \psi_K$. [4]

Notations. To simplify the notations, let us denote $\mathbf{z} \triangleq (\mathbf{z}_1, \cdots, \mathbf{z}_K) \in Q$ in the context of VIPs. Let us denote

$$\bar{L}_{ij} \triangleq \max\{L_{ij}, L_{ji}\},\ A_i \triangleq D_i\Big(\sum_{j \in [K] \setminus \{i\}} \bar{L}_{ij} D_j\Big),\ \text{and } B_i \triangleq \bar{L}_{ii} D_i^2,\ \text{for all } i, j \in [K].$$

We refer to $\sum_i A_i$ as the diagonal conditioning, and we say that the diagonal conditioning dominates when $\sum_i A_i \gg \sum_i B_i$.

D.2. Communication and computational costs

Accuracy measure. We consider the following accuracy measure for VIPs:

$$\Delta(\bar{\mathbf{z}}) \triangleq \sup_{\mathbf{z} \in \mathcal{B} \cap Q} \langle V(\mathbf{z}), \bar{\mathbf{z}} - \mathbf{z} \rangle + \psi(\bar{\mathbf{z}}) - \psi(\mathbf{z}),\ \text{for all } \mathbf{z} \in Q,$$

where $\mathcal{B} \triangleq \mathcal{B}_1 \times \cdots \times \mathcal{B}_K$. We say that a point $\bar{\mathbf{z}} \in Q$ is an *ε-approximate solution* of the VIP if $\Delta(\bar{\mathbf{z}}) \leq \varepsilon$. Our goal is to find such an ε-approximate solution for any $\varepsilon > 0$.

Distributed first-order algorithm, communication and computational costs. We consider a distributed setting with K agents very similar to the one in Section 2.1. For all $i \in [K]$: Agent i controls decision variable $\mathbf{z}_i \in \operatorname{dom} \psi_i$, has direct access to the function ψ_i, and has access to the oracle $\mathscr{O}_i(\mathbf{z}) = V_i(\mathbf{z})$ for $\mathbf{z} \in Q$. For any distributed first-order algorithm $\mathscr{M}$ and target accuracy ε, the communication cost is defined as the number of communication rounds required for $\mathscr{M}$ to generate an ε-approximate solution; and the computational cost is defined by a weighted sum of the oracle queries across all agents, where per query to $\mathscr{O}_i$ costs $c_i \geq 0$, $i \in [K]$.

We first state the classic results of the EG method in Proposition 15, which remains the state-of-the-art communication complexity bound.

Proposition 15 (Juditsky et al. (2011, Eq. (6.21))) *For any target accuracy $\varepsilon > 0$, the communication cost of* EG *is bounded by*

$$\mathcal{O}\Big(\sum_{i \in [K]} \frac{A_i}{\varepsilon}\Big),$$

and the computational cost of EG *is bounded by*

$$\mathcal{O}\Big(\Big(\sum_{i \in [K]} c_i\Big)\Big(\sum_{i \in [K]} \frac{A_i}{\varepsilon}\Big) + \Big(\sum_{i \in [K]} c_i\Big)\Big(\sum_{i \in [K]} \frac{B_i}{\varepsilon}\Big)\Big).$$

4. For all $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_K) \in Q$, we use the following notations for simplicity: $(\mathbf{z}_j; \mathbf{z}_{-j}) \triangleq \mathbf{z}$ and $\mathbf{z}_{-j} \triangleq (\mathbf{z}_1, \cdots, \mathbf{z}_{j-1}, \mathbf{z}_{j+1}, \cdots, \mathbf{z}_K)$.

D.3. Decoupled method for variational inequality problems

Assembled norm. Let $\alpha_i > 0$ (to be fixed later), for all $i \in [K]$. Let the block diagonal linear operator $\mathbf{P} = \alpha_1 \mathbf{P}_1 \oplus \cdots \oplus \alpha_K \mathbf{P}_K$. We consider the space $\mathcal{E}$ to be measured by the following assembled norm: $\|\mathbf{z}\|_{\mathcal{E}} = \sqrt{\langle \mathbf{P}\mathbf{z}, \mathbf{z}\rangle}$, and its dual space $\mathcal{E}^* = \mathcal{E}_1^* \times \cdots \times \mathcal{E}_K^*$ to be measured by $\|\mathbf{g}\|_{\mathcal{E}^*} = \sqrt{\langle \mathbf{g}, \mathbf{P}^{-1}\mathbf{g}\rangle}$.

DM-VIP. Let us consider the case that D_i is not known, and we use the inexact estimates $\hat{D}_i$ instead, for all $i \in [K]$. Let us denote

$$\bar{L}_{\mathrm{c}} \triangleq \sqrt{\max_{j\in[K]}\left[(\alpha_j \hat{D}_j)^{-1} \sum_{i\in[K]\setminus\{j\}} \frac{\bar{L}_{ij}\left(\sum_{l\in[K]\setminus\{i\}} \bar{L}_{il}\hat{D}_l\right)}{\alpha_i}\right]}. \tag{10}$$

Now, we introduce the (template) Decoupled Reduced-Operator Method for block composite variational inequality problems (DM-VIP), as an extended variant of DM-SP to VIPs. The pseudocode is presented in Algorithm 4, provided with a solver for the minimization of residual norms.

Algorithm 4 DM-VIP$(K, (V_i)_{i\in[K]}, (\psi_i)_{i\in[K]}, \mathbf{z}^0, (\lambda_t)_{t\geq 1}, (\alpha_i)_{i\in[K]} \mid (\mathcal{M}_i^{\mathrm{MRN}})_{i\in[K]})$

Require: A solver $\mathcal{M}^{\mathrm{MRN}}$ for the minimization of residual norms.

$\mathbf{v}^0 = (\mathbf{v}_1^0, \cdots, \mathbf{v}_K^0) = \mathbf{z}^0$.

for $t = 0, 1, \cdots, T-1$ **do**

For all $i \in [K]$, Agent i computes

$$(\mathbf{z}_i^{t+1}, \psi_i'(\mathbf{z}_i^{t+1})) = \mathcal{M}_i^{\mathrm{MRN}}(V_i(\cdot; \mathbf{v}_{-i}^t), \psi_i + \frac{\alpha_i\lambda_{t+1}}{2}\|\cdot - \mathbf{v}_i^t\|_i^2, \mathbf{v}_i^t, \frac{\alpha_i\lambda_{t+1}}{2});$$

and then, all agents communicate $\mathbf{z}^{t+1} = (\mathbf{z}_1^{t+1}, \cdots, \mathbf{z}_K^{t+1})$.

The agents compute $V(\mathbf{z}^{t+1})$; and then communicate

$$V_\psi(\mathbf{z}^{t+1}) \triangleq V(\mathbf{z}^{t+1}) + \psi'(\mathbf{z}^{t+1}) \equiv V(\mathbf{z}^{t+1}) + (\psi_1'(\mathbf{z}_1^{t+1}), \cdots, \psi_K'(\mathbf{z}_K^{t+1})).$$

$a_{t+1} = \frac{2\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}^t - \mathbf{z}^{t+1}\rangle}{\|V_\psi(\mathbf{z}^{t+1})\|_{\mathcal{E}^*}^2}$.

$\mathbf{v}^{t+1} = (\mathbf{v}_1^{t+1}, \cdots, \mathbf{v}_K^{t+1}) = \operatorname{argmin}_{\mathbf{v}\in Q}\left[a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{v}\rangle + \frac{1}{2}\|\mathbf{v} - \mathbf{v}^t\|_{\mathcal{E}}^2\right]$.

end for

We say Algorithm 4 is a template method because we have not yet specified the solver. We defer the detailed implementation to Equation (17) to the end of this section.

Theorem 16 *Under (A2') and (A3'), for* $\lambda_{t+1} \equiv \lambda \geq 2\bar{L}_{\mathrm{c}}$, DM-VIP *(Algorithm 4) can be implemented with no more than*

$$2T$$

communication rounds and no more than

$$T \cdot \left(1 + C_0 \cdot \frac{3L_{ii}}{\alpha_i\lambda}\right)$$

queries to V_i, for all $i \in [K]$, and obtains an ε-approximate solution

$$\bar{\mathbf{z}}^T = (\bar{\mathbf{z}}_1^T, \cdots, \bar{\mathbf{z}}_K^T) = \Big(\sum_{t=1}^T a_t\Big)^{-1} \sum_{t=1}^T a_t \mathbf{z}^t,$$

where

$$T = \Big\lceil \frac{\sum_i \alpha_i \lambda D_i^2}{2\varepsilon} \Big\rceil$$

and $C_0 > 0$ is some fixed constant.

Moreover, assume that the target accuracy $\varepsilon \le \sum_{i,j\in[K],\, i\neq j} \bar{L}_{ij} D_i D_j$. For the choices of $\alpha_i = \frac{\sum_{j\in[K]\setminus\{i\}} \bar{L}_{ij}\hat{D}_j}{\hat{D}_i}$ for all $i \in [K]$, and $\lambda = 2$, the communication cost is bounded by

$$2 + \sum_{i,j\in[K],\, i\neq j} \frac{\bar{L}_{ij} D_i D_j}{\varepsilon} \Big(\frac{D_i \hat{D}_j}{\hat{D}_i D_j} + \frac{\hat{D}_i D_j}{D_i \hat{D}_j}\Big) \triangleq T^{\text{DM-VIP}}\big((\hat{D}_i)_{i\in[K]}\big), \tag{11}$$

and the number of queries to V_i is bounded by

$$\Big(1 + 3C_0 \frac{\bar{L}_{ii}\hat{D}_i}{\sum_{j\in[K]\setminus\{i\}} \bar{L}_{ij}\hat{D}_j}\Big)\Big[\sum_{j,l\in[K],\, j\neq l} \frac{\bar{L}_{jl} D_j D_l}{\varepsilon}\Big(\frac{D_j \hat{D}_l}{\hat{D}_j D_l} + \frac{\hat{D}_j D_l}{D_j \hat{D}_l}\Big)\Big] \triangleq k_i^{\text{DM-VIP}}\big((\hat{D}_j)_{j\in[K]}\big). \tag{12}$$

Corollary 17 *Equation (11) is minimized when the distance estimates $(\hat{D}_i)_{i\in[K]}$ satisfy $\frac{\hat{D}_i}{D_i} = \frac{\hat{D}_y}{D_y}$, for all $i, j \in [K]$. This results in*

$$\min_{(\hat{D}_i)_{i\in[K]}} T^{\text{DM-VIP}}\big((\hat{D}_i)_{i\in[K]}\big) = 2 + \frac{2}{\varepsilon} \sum_{i\in[K]} A_i, \tag{13}$$

and in the meantime, for all $i \in [K]$,

$$k_i^{\text{DM-VIP}}\big((\hat{D}_j)_{j\in[K]}\big) = \frac{2}{\varepsilon}\Big(1 + 3C_0 \cdot \frac{B_i}{A_i}\Big)\Big(\sum_{j\in[K]} A_j\Big). \tag{14}$$

Remark 18 (Improved communication cost under comparable computational costs) *The classic* EG *method takes*

$$\frac{1}{\varepsilon} \sum_{i\in[K]} (A_i + B_i)$$

communication rounds, and the same number of queries to V_i for all $i \in [K]$ (Juditsky et al., 2011, Eq. (6.21)). When the gradient estimates $(\hat{D}_i)_{i\in[K]}$ satisfy $\frac{\hat{D}_i D_j}{D_i \hat{D}_j} = \Theta(1)$ for all $i, j \in [K]$, our communication complexity in Equation (13) is consistently no worse compared to the communication complexity of EG*, and is substantially faster when the "diagonal conditioning" dominates—that is,*

$$\sum_{i\in[K]} B_i \gg \sum_{i\in[K]} A_i.$$

Moreover, our computational cost (under the same choice of parameters) is bounded by

$$\frac{2}{\varepsilon}\Big(\sum_{i\in[K]} c_i\Big)\Big(\sum_{i\in[K]} A_i\Big) + \frac{3C_0}{\varepsilon}\Big(\sum_{i\in[K]} \frac{B_i c_i}{A_i}\Big)\Big(\sum_{i\in[K]} A_i\Big). \tag{15}$$

Compared to the computational cost of EG*, given by*

$$\frac{1}{\varepsilon}\Big(\sum_{i\in[K]} c_i\Big)\Big(\sum_{i\in[K]} A_i\Big) + \frac{1}{\varepsilon}\Big(\sum_{i\in[K]} c_i\Big)\Big(\sum_{i\in[K]} B_i\Big),$$

our computational cost in Equation (15) *differs primarily in the second term, and consequently, may offer an advantage or disadvantage depending on the relative conditioning of* A_i*,* B_i*, and* c_i *for* $i \in [K]$*.*

D.4. Detailed proofs

Correctness of FDS. Let us provide the detailed pseudocode of FDS for VIPs in Algorithm 5. Then, we prove the correctness of the solution returned by FDS.

Algorithm 5 $\mathrm{FDS}((V_i)_{i\in[K]}, (\psi_i)_{i\in[K]}, \mathbf{v}, \lambda \mid \mathcal{M}^{\mathrm{MRN}}, K, (\alpha_i)_{i\in[K]})$

Require: Solver $\mathcal{M}_i^{\mathrm{MRN}}$ for the minimization of residual norms, for all $i \in [K]$.

1: $\delta_i = \frac{\alpha_i\lambda}{2}$, for all $i \in [K]$.

2: $(\mathbf{z}_i^+, \psi_i'(\mathbf{z}_i^+)) = \mathcal{M}_i^{\mathrm{MRN}}\big(V_i(\cdot;\mathbf{v}_{-i}), \psi_i + \frac{\alpha_i\lambda}{2}\|\cdot - \mathbf{v}_i\|_i^2, \mathbf{v}_i, \delta_i\big)$, for all $i \in [K]$.

3: **return** $(\mathbf{z}^+, \psi'(\mathbf{z}^+))$, where $\mathbf{z}^+ = (\mathbf{z}_1^+, \cdots, \mathbf{z}_K^+)$ and $\psi'(\mathbf{z}^+) = (\psi_1'(\mathbf{z}_1^+), \cdots, \psi_K'(\mathbf{z}_K^+))$.

Lemma 19 *Under* (A3')*, for* $\lambda \geq 2\bar{L}_c$*,* FDS *(Algorithm 5) returns the correct solution of the MS subproblem given by* $(V, \psi, \mathbf{v}, \lambda)$*.*

Proof [Proof of Lemma 19] For all $i \in [K]$, by (A3') and then by the relative distance accuracy, we have

$$\begin{aligned}
&\|V_i(\mathbf{z}^+) + \psi_i'(\mathbf{z}_i^+) + \alpha_i\lambda\mathbf{P}_i(\mathbf{z}_i^+ - \mathbf{v}_i)\|_{i^*} \\
&\quad\leq \|V_i(\mathbf{z}_i^+;\mathbf{v}_{-i}) + \psi_i'(\mathbf{z}^+) + \alpha_i\lambda\mathbf{P}_i(\mathbf{z}_i^+ - \mathbf{v}_i)\|_{i^*} + \sum_{j\in[K]\setminus\{i\}} L_{ij}\|\mathbf{z}_j^+ - \mathbf{v}_j\|_j \\
&\quad\leq \delta_i\|\mathbf{z}_i^+ - \mathbf{v}_i\|_i + \sum_{j\in[K]\setminus\{i\}} L_{ij}\|\mathbf{z}_j^+ - \mathbf{v}_j\|_j.
\end{aligned} \tag{16}$$

Finally, we assemble the norms:

$$
\begin{aligned}
&\|V(\mathbf{z}^+) + \psi'(\mathbf{z}_i^+) + \lambda \mathbf{P}(\mathbf{z}_i^+ - \mathbf{v}_i)\|_{\mathcal{E}^*}^2 \\
&\quad = \sum_{i\in[K]} \alpha_i^{-1} \|V_i(\mathbf{z}^+) + \psi_i'(\mathbf{z}_i^+) + \alpha_i \lambda \mathbf{P}_i(\mathbf{z}_i^+ - \mathbf{v}_i)\|_{i^*}^2 \\
&\quad \overset{(16)}{\leq} \sum_{i\in[K]} \alpha_i^{-1} \Big(\delta_i \|\mathbf{z}_i^+ - \mathbf{v}_i\|_i + \sum_{j\in[K]\setminus\{i\}} L_{ij} \|\mathbf{z}_j^+ - \mathbf{v}_j\|_j\Big)^2 \\
&\quad \leq \sum_{i\in[K]} 2\alpha_i^{-1} \Big[\delta_i^2 \|\mathbf{z}_i^+ - \mathbf{v}_i\|_i^2 + \Big(\sum_{j\in[K]\setminus\{i\}} L_{ij} \|\mathbf{z}_j^+ - \mathbf{v}_j\|_j\Big)^2\Big] \\
&\quad = \frac{\lambda^2}{2} \sum_{i\in[K]} \Big(\alpha_i \|\mathbf{z}_i^+ - \mathbf{v}_i\|_i^2\Big) + 2 \sum_{i\in[K]} \Big[\alpha_i^{-1} \Big(\sum_{j\in[K]\setminus\{i\}} L_{ij} \|\mathbf{z}_j^+ - \mathbf{v}_j\|_j\Big)^2\Big] \\
&\quad \leq \frac{\lambda^2}{2} \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 + 2 \sum_{i\in[K]} \Big[\alpha_i^{-1} \Big(\sum_{l\in[K]\setminus\{i\}} L_{il} \hat{D}_l\Big) \Big(\sum_{j\in[K]\setminus\{i\}} \frac{L_{ij}}{\hat{D}_j} \|\mathbf{z}_j^+ - \mathbf{v}_j\|_j^2\Big)\Big] \\
&\quad = \frac{\lambda^2}{2} \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 + 2 \sum_{j\in[K]} \Big[\frac{\|\mathbf{z}_j^+ - \mathbf{v}_j\|_j^2}{\hat{D}_j} \Big(\sum_{i\in[K]\setminus\{j\}} \frac{L_{ij} \big(\sum_{l\in[K]\setminus\{i\}} L_{il} \hat{D}_l\big)}{\alpha_i}\Big)\Big] \\
&\quad \overset{(10)}{\leq} \frac{\lambda^2}{2} \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 + 2\|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 \bar{L}_{\mathrm{c}}^2 \\
&\quad \leq \frac{\lambda^2}{2} \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 + \frac{\lambda^2}{2} \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2 = \lambda^2 \|\mathbf{z}^+ - \mathbf{v}\|_{\mathcal{E}}^2.
\end{aligned}
$$

■

Convergence for VIPs. We are now back to considering the VIPs. Our algorithm is built upon Lemma 20, the proof of which can be found in (Boţ and Chenchene, 2024, Corollary 2.4).

Lemma 20 *Assume (Â1), (Â3), and that the solution set of the VIP of $(\hat{V}, \hat{\psi})$ is non-empty. Then, there exists an algorithm, denoted by $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+)) = \mathrm{FEGM}\,(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \hat{\varepsilon} \mid \hat{L})$, which takes no more than $C_0 \cdot \frac{\hat{L}}{\hat{\varepsilon}}$ operator queries and returns $(\hat{\mathbf{z}}^+, \hat{\psi}'(\hat{\mathbf{z}}^+))$ that satisfies $\hat{\varepsilon}$-distance-to-solution accuracy, where $C_0 > 0$ is some fixed constant.*

Let us use FEGM in Lemma 20 for the minimization of residual norms:

$$\mathcal{M}_i^{\mathrm{MRN}}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \hat{\delta}) \triangleq \mathrm{FEGM}(\hat{V}, \hat{\psi}, \hat{\mathbf{v}}, \frac{2\hat{\delta}}{3} \mid L_{ii}).$$

We obtain the final algorithm for VIPs:

$$\mathrm{ROM}_{\|\cdot\|_{\mathcal{E}}}\Big((V_i)_{i\in[K]}, (\psi_i)_{i\in[K]}, \mathbf{z}^0, (\lambda_t)_{t\geq 1} \mid \mathrm{FDS}_{\|\cdot\|_{\mathcal{E}}}\big(\cdot,\cdot,\cdot,\cdot \mid (\mathcal{M}_i^{\mathrm{MRN}})_{i\in[K]}\big)\Big). \tag{17}$$

Combining Lemmas 4, 5, 7 and 20, with the implementation in Equation (17), we conclude that Theorem 16 holds for the constant C_0 from Lemma 20. We include the complete proof below.

Proof [Proof of Theorem 16] By (A1'), we have

$$\Delta(\bar{\mathbf{z}}^T) \le \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \max_{\mathbf{z}\in\mathcal{B}\cap Q} \Big[\sum_{t=0}^{T-1} a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{z}^{t+1} - \mathbf{z}\rangle\Big].$$

Further, with $\lambda \ge 2\bar{L}_{\mathrm{c}}$, by Lemmas 4 and 5, we have

$$\begin{aligned}
\Delta(\bar{\mathbf{z}}^T) &\le \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \max_{\mathbf{z}\in\mathcal{B}\cap Q} \Big[\sum_{t=0}^{T-1} a_{t+1}\langle V_\psi(\mathbf{z}^{t+1}), \mathbf{z}^{t+1} - \mathbf{z}\rangle\Big] \\
&\le \Big(\sum_{t=0}^{T-1} a_{t+1}\Big)^{-1} \Big[\sum_{i\in[K]} \big(\frac{\alpha_i}{2} \max_{\mathbf{z}_i\in\mathcal{B}_i\cap\operatorname{dom}\psi_i} \|\mathbf{z}_i^0 - \mathbf{z}_i\|_i^2\big)\Big] \\
&\le \Big(\sum_{t=0}^{T-1} \frac{1}{\lambda_{t+1}}\Big)^{-1} \cdot \frac{1}{2}\sum_{i\in[K]} \alpha_i D_i^2 \le \varepsilon,
\end{aligned}$$

where the last inequality follows from the assignments of $(\lambda_t)_{t\ge1}$ and T. Therefore, the number of communication rounds is bounded by $2T$.

Now we count the number of gradient queries. By Lemma 7, FEGM always returns the solution with the required relative distance accuracy; and in view of Lemma 20, it takes no more than $C_0\cdot\frac{3L_{ii}}{\alpha_i\lambda}$ gradient queries to V_i, for all $i \in [K]$. Therefore, the numbers of queries to V_i are bounded by $T\cdot\big(1 + C_0\cdot\frac{3L_{ii}}{\alpha_i\lambda}\big)$, for all $i\in[K]$. ■