Randomized Lagrangian Stochastic Approximation for Large-Scale Constrained Stochastic Nash Games

Zeinab Alizadeh*   Afrooz Jalilzadeh*   Farzad Yousefian†

Abstract

In this paper, we consider stochastic monotone Nash games where each player's strategy set is characterized by possibly a large number of explicit convex constraint inequalities. Notably, the functional constraints of each player may depend on the strategies of other players, allowing for capturing a subclass of generalized Nash equilibrium problems (GNEP). While there is limited work that provide guarantees for this class of stochastic GNEPs, even when the functional constraints of the players are independent of each other, the majority of the existing methods rely on employing projected stochastic approximation (SA) methods. However, the projected SA methods perform poorly when the constraint set is afflicted by the presence of a large number of possibly nonlinear functional inequalities. Motivated by the absence of performance guarantees for computing the Nash equilibrium in constrained stochastic monotone Nash games, we develop a single timescale randomized Lagrangian multiplier stochastic approximation method where in the primal space, we employ an SA scheme, and in the dual space, we employ a randomized block-coordinate scheme where only a randomly selected Lagrangian multiplier is updated. We show that our method achieves a convergence rate of $\mathcal{O}\left(\frac{\log(k)}{\sqrt{k}}\right)$ for suitably defined suboptimality and infeasibility metrics in a mean sense.

1   Introduction

Noncooperative game theory provides a mathematical framework to study multi-agent decision making problems that have emerged in a wide range of applications including electricity markets [19], transportation networks [12], and signal processing [7], among many others. While the multidisciplinary field of game theory finds its origin in the work by von Neumann and Morgenstern [44], the notion of a Nash equilibrium (NE) was introduced and its existence was provably shown by John Nash [36]. Noncooperative Nash game is a modeling framework where a finite collection of selfish agents compete with each other and seek to optimize their own individual objectives. Such a competition is often subject to limited resources characterized by functional constraints. In this work, our primary focus lies in computing an NE for large-scale constrained Nash game formulations afflicted by the presence of uncertainty in the objectives of the agents. More precisely, we consider stochastic monotone Nash games with a large number of (possibly nonlinear) functional constraints described as follows. Let $N \geq 1$ denote the number of players. For all $i = 1, \ldots, N$, the ith player is associated with the following constrained stochastic optimization problem.

---

*Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona 85721, USA. zalizadeh@arizona.edu and afrooz@arizona.edu

†Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA. farzad.yousefian@rutgers.edu

$$\min_{x_i \in \mathcal{X}_i} \quad H_i(x) \triangleq \mathbb{E}[h_i(x_i, x_{-i}, \xi)] \qquad\qquad (\mathrm{P}_i(x_{-i}))$$

$$\text{where} \quad \mathcal{X}_i \triangleq \{x_i \in X_i \subseteq \mathbb{R}^{n_i} \mid g_{i,\ell}(x_i, x_{-i}) \leq 0, \quad \text{for all } \ell = 1, \ldots, J_i\}$$

where $x_i \in \mathbb{R}^{n_i}$ denotes the strategy of the ith player, $x_{-i} \in \mathbb{R}^{n-n_i}$ is the collection of the strategies of the other players, $n \triangleq \sum_{i=1}^N n_i$, $h_i : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ denotes the stochastic cost function associated with the ith player. The uncertainty in the game is characterized by the random variable $\xi : \Omega \to \mathbb{R}^d$ associated with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The constraint set of the ith player is expressed in terms of explicit convex constraint inequalities in terms of the jointly convex functions $g_{i,\ell} : \mathbb{R}^n \to \mathbb{R}$, for all $\ell = 1, \ldots, J_i$. The ith player's strategy is a subset of a nonempty convex set denoted by $X_i \subseteq \mathbb{R}^{n_i}$. While we will provide the detailed description of our assumptions in subsequent sections, it is worth emphasizing that throughout, we assume that all the aforementioned functions are merely convex.

Problem $(\mathrm{P}_i(x_{-i}))$ is a subclass of the generalized Nash equilibrium problems (GNEP) that have been extensively employed in the literature in formulating applications arising in economics and operations research, among others [10, 31]. Recall that in GNEPs, players seek the NE by *simultaneously* satisfying the constraints. This is different from other classes of games where players make decisions in a specific order, e.g., in Stackelberg games.

Note that a popular subclass of the problem $(\mathrm{P}_i(x_{-i}))$ is the stochastic minimax problem. Consider the following stochastic merely-convex-merely-concave minimax optimization problem with possibly many functional constraints.

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} \quad H(u, v) \triangleq \mathbb{E}[h(u, v, \xi)] \qquad\qquad (1)$$

$$\text{where} \quad \mathcal{U} \triangleq \{u \in U \mid g_{1,\ell}(u) \leq 0, \quad \text{for all } \ell = 1, \ldots, J_1\} \quad \text{and} \quad U \subseteq \mathbb{R}^{n_1},$$

$$\mathcal{V} \triangleq \{v \in V \mid g_{2,\ell}(v) \leq 0, \quad \text{for all } \ell = 1, \ldots, J_2\} \quad \text{and} \quad V \subseteq \mathbb{R}^{n_2}.$$

Minimax optimization can indeed be viewed as a subclass of two-person zero-sum games. The existence of equilibrium in such a game is established by the celebrated von Neumann's minimax theorem in 1928 [43] that appears amongst the most fundamental results in game theory. The research on the development of gradient-type methods for solving minimax problems, also known as the problem of finding *saddle points*, dates back to as early as 1970s, including the work by Korpelevich [29] and Golshtein [14], followed by efforts on on the development gradient descent ascent as well as primal-dual methods (e.g., see [6, 38, 37, 49, 16] and [11, Chp. 1]). More recently, minimax problems have drawn an increasing attention in areas including adversarial learning [40, 15, 42], fairness in machine learning [48, 23], and distributionally robust federated learning [8], to name a few.

Existing methods and research gap. In addressing deterministic games, iterative methods for approximating an equilibrium find their origin in 1960s in the seminal work by Scarf [41] (see [11, Chapter 12] for a detailed review of deterministic methods). The prior algorithmic efforts in addressing stochastic Nash games, however, find their roots in the work by Jiang and Xu [22] in 2008, where a stochastic approximation (SA) method was developed for addressing stochastic variational inequality (VI) problems with strongly monotone and Lipschitzian mappings. Recall that given a set $\mathcal{X}$ and a single-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, vector $x \in X$ solves $\mathrm{VI}(\mathcal{X}, F)$ if $F(x)^T(y - x) \geq 0$ for all $y \in \mathcal{X}$. Under some mild convexity and differentiability assumptions, it can be shown that [11,

Table 1: Solution methods with rate statements for variational inequality problems

Ref.	Problem	Rate	Nonlinear const.
Proximal Extra-Gradient[33]	VI	$\mathcal{O}(1/\epsilon)$	✗
SMP[24]	SVI	$\mathcal{O}(1/\epsilon^2)$	✗
DS-SA [21]	SVI	$\mathcal{O}(\log(1/\epsilon)/\epsilon^2)$	✗
RLSA (This paper)	SVI	$\mathcal{O}(\log(1/\epsilon)/\epsilon^2)$	✓

Table 2: A subset of methods with guarantees for saddle point problems

Ref.	Stoch.	Non-bilinear	Convex	Nonlinear const.
PDHG [5],Acc-SP-HPE[18]	✗	✗	$\mathcal{O}(1/\epsilon)$	✗
Acc-HPE-type [28],SMP[24]	✗	✓	$\mathcal{O}(1/\epsilon)$	✗
Acc- BD[17]	✓	✗	$\mathcal{O}(1/\epsilon)$	✗
SAA[39],SADMM[50]	✓	✓	$\mathcal{O}(1/\epsilon^2)$	✗
RLSA(This paper)	✓	✓	$\mathcal{O}(\log(1/\epsilon)/\epsilon^2)$	✓

Chapter 1] the set of equilibria of the stochastic game $(\mathrm{P}_i(x_{-i}))$, for $i = 1, \ldots, N$, is characterized by the solution set of $\mathrm{VI}(\mathcal{X}, F)$ where $\mathcal{X} \triangleq \prod_{i=1}^{N} \mathcal{X}_i$ and $F(x) \triangleq (\nabla_{x_1}\mathbb{E}[h_1(x,\xi)]; \ldots; \nabla_{x_N}\mathbb{E}[h_N(x,\xi)])$. In view of this result, seeking a Nash equilibrium of a stochastic game is equivalent to solving the aforementioned stochastic VI. The convergence and rate analysis of SA schemes for solving VIs under weaker monotonicity and smoothness assumptions were studied more recently in works including [25, 30, 46]. Also, stochastic extragradient methods and their variance-reduced variants were studied in [47, 32, 20]. Despite these advances, it is often assumed in the above-mentioned methods that the sets $\mathcal{X}_i$ is easy-to-project on and accordingly, the algorithmic framework in these works relies on projected schemes. However, in the following cases, $\mathcal{X}_i$ may become difficult-to-project on: (i) When the dimensionality of the solution space, i.e., n, is large; (ii) When the number of the constraints is large. For example, in the game setting $\sum_{i=1}^{N} J_i$ could be large; (iii) The constraint set may be characterized by nonlinear constraints. In fact, we are unaware of any iterative methods with provable complexity guarantees for the resolution of even deterministic variants of constrained monotone Nash games. Our research in this paper is precisely motivated by this shortcoming in the literature.

Main contributions. In Table 1 and Table 2, we provide a summary of the main results in our work and we compare them with some of the existing methods for addressing monotone VIs and minimax problems. To highlight our contributions, we first provide a brief review of some of the existing avenues for addressing monotone Nash games and VIs with explicit constraints. The duality theory for VIs and the notion of the dual VI has been studied by Mosco [35] in 1972 which was later improved in [13, 9]. Extending the duality framework devised in [1], Auslender and Teboulle [3] developed a Lagrangian duality scheme for solving multi-valued variational inequality problems with maximal monotone operators and explicit convex constraint inequalities. Leveraging entropic proximal terms, interior proximal point methods were developed for solving constrained VIs in works including [2, 4]. Although the aforementioned dual-based methods are endowed with asymptotic convergence guarantees, convergence speed of Lagrangian dual methods for solving constrained VIs is not known. In particular, we are interested in investigating whether it is possible to devise suitable Lagrangian dual methods that can be guaranteed with convergence speeds of similar order of magnitude to those of primal-dual methods developed for standard constrained optimization methods [45]. We show that this is indeed possible. We summarize our main contributions in the

following.

(i) *A single timescale randomized primal-dual stochastic approximation method.* Leveraging the primal-dual framework for addressing constrained stochastic optimization problems, we devise a randomized primal-dual stochastic approximation method for solving VIs with merely monotone and stochastic mappings with explicit constraint inequalities. To capture large-scale constrained stochastic Nash games, we employ a randomized block scheme for updating the Lagrange multipliers. Importantly, this scheme is single timescale and efficient to implement.

(ii) *New convergence rate statements.* In contrast with standard optimization problems, one of the main challenges in addressing VIs lies in the lack of availability of suitable error metrics that rely on objective function values. In particular, this challenge introduces some difficulty in the convergence rate analysis of monotone VIs, an issue that is exacerbated in the presence of explicit constraint inequalities. Motivated by earlier efforts [46, 27], leveraging the notion of dual gap functions, we analyze the convergence of the proposed method and derive convergence rates of $\mathcal{O}\left(\frac{\log(k)}{\sqrt{k}}\right)$ for both suboptimality and infeasibility metrics in a mean sense.

Outline of the paper. The remainder of the paper is organized as follows. In Section 2, we provide the main assumptions and review some preliminary results that are employed in the analysis. In Section 3, we present the outline of the proposed algorithm along with some definitions. In Section 4 we establish convergence properties of the method and derive explicit performance guarantees. We present some concluding remarks in Section 5. Lastly, Section 6 includes the proofs for some of the results in the paper.

2 Preliminaries

To address the stochastic game $\mathrm{P}_i(x_{-i})$ for $i \in [N]$, we consider the stochastic VI problem described as follows.

> Find $x \in \mathcal{X}$ such that $\quad \mathbb{E}[F(x,\xi)]^T(y-x) \geq 0, \quad$ for all $y \in \mathcal{X}$ $\qquad$ **(cSVI)**
>
> where $\quad \mathcal{X} \triangleq \{x \in X \mid f_j(x) \leq 0, \text{ for all } j = 1, \ldots, J\} \text{ and } X \triangleq \prod_{i=1}^{N} X_i.$

The details of our assumptions on the mapping F, functions f_j, and sets X_i are provided as follows.

Assumption 1 (Problem properties). Consider problem (**cSVI**). Let the following holds.
(i) Mapping $F(\bullet) : \mathbb{R}^n \to \mathbb{R}^n$ is real-valued, continuous, and merely monotone on its domain, i.e. $\langle F(x) - F(y), x - y \rangle \geq 0$, for all $x, y \in X$.
(ii) Function $f_j(\bullet) : \mathbb{R}^n \to \mathbb{R}$ is real-valued, merely convex on its domain for all $j = 1, \ldots, J$.
(iii) Set $X \subseteq \mathrm{int}\left(\mathrm{dom}(F) \cap (\cap_{j=1}^{J}\mathrm{dom}(f_j))\right)$ is nonempty, compact, and convex.
(iv) The Slater condition holds, i.e., there exists $\hat{x} \in X$ such that $f_j(\hat{x}) < 0$ for all $j = 1, \ldots, J$.

Remark 1. Note that problem (**cSVI**) captures the stochastic game $\mathrm{P}_i(x_{-i})$. In fact, given the objective functions $h_i(\bullet, \xi)$ and constraint functions $g_{i,\ell}$ in $\mathrm{P}_i(x_{-i})$, x is an NE if and only if x solves (**cSVI**) where $f_j(x) \triangleq g_{i,\ell}(x)$ where $j := \ell + \sum_{t=1}^{i-1} J_t$ for $\ell \in [J_i]$.

4

Definition 1 (Augmented-Lagrangian function)**.** *Given $x, y \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^J$, and $\rho > 0$, we define*

$$\mathcal{L}_\rho(x, y, \lambda) \triangleq F(y)^T(x - y) + \Phi_\rho(x, \lambda),$$

$$where \quad \Phi_\rho(x, \lambda) \triangleq \frac{1}{J} \sum_{j=1}^{J} \phi_\rho(f_j(x), \lambda^{(j)}) \quad and \quad \phi_\rho(u, v) \triangleq \begin{cases} uv + \frac{\rho}{2}u^2, & if \quad \rho u + v \geq 0, \\ -\frac{v^2}{2\rho}, & otherwise. \end{cases}$$

Similar to the traditional constrained optimization techniques, the nonlinear constrains in problem (**cSVI**) can be combined with the objective function using some multipliers. Using this technique we can characterize the optimality condition of problem (**cSVI**) in the following result.

Proposition 1 (Karush–Kuhn–Tucker (KKT) conditions)**.** Consider problem (**cSVI**) and suppose Assumption 1 holds. Let $f(x) \triangleq (f_1(x), \ldots, f_J(x))^T$ and the gradient matrix $\nabla f(x) \triangleq (\nabla f_1(x), \ldots, \nabla f_J(x))^T \in \mathbb{R}^{n \times J}$. There exists $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^J$ satisfying the following KKT conditions:
(i) $0 \in F(x^*) + J^{-1}\nabla f(x^*)^T \lambda^* + \mathcal{N}_X(x^*)$.
(ii) $0 \leq \lambda^* \perp -f(x^*) \geq 0$.
(iii) $x^* \in X$.

Proof. Note that any solution x of (**cSVI**) is also a solution of the following optimization problem:

$$\min_{y \in X} \ y^T F(x) \tag{2}$$
$$\text{s.t. } f_j(y) \leq 0 \quad \forall j.$$

Since the Slater condition holds, the first-order KKT condition for (2) implies that there exists $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^J$ satisfying conditions (i)-(iii). ∎

We will utilize the following definition in the convergence and rate analysis.

Definition 2. Consider the VI$(\mathcal{X}, F)$ where X is a closed convex set and F is a real-valued monotone map. The dual gap function $\text{Gap}^* : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is defined for any $x \in \mathbb{R}^n$ as

$$\text{Gap}^*(x) \triangleq \sup_{y \in \mathcal{X}} F(y)^T(x - y). \tag{3}$$

Remark 2. Note that by the definition, $\text{Gap}^*(x) \geq 0$ for all $x \in \mathcal{X}$. Also, under some mild conditions, $\text{Gap}^*(x) = 0$ implies that x is a solution to VI$(\mathcal{X}, F)$. This is formally stated below.

Remark 3. Karamardian [26] showed that under continuity and pseudomonotonicity of the operator F, solving (**cSVI**) problem is equivalent to solving Minty stochastic variational inequality (MSVI) [34] problem. Such a problem requires an $x^* \in X$ such that

$$(x^* - x)^T F(x) \leq 0, \qquad \text{for all } x \in \mathcal{X}. \tag{MSVI}$$

Therefore, to obtain the convergence rate we adopt the dual gap function. Note that $\text{Gap}^*(\bullet)$ is well-defined when $\mathcal{X}$ is a compact set, that follows from Assumption 1 (iii).

By invoking Proposition 1 and Assumption 1, we can establish the following two results for problem (**cSVI**). These results will be employed later to demonstrate the boundedness of dual iterates and to obtain convergence rate results. We have provided the proofs of the following lemmas in the appendix.

Lemma 1. Consider problem (**cSVI**) under Assumption 1. Then for any primal-dual solution pair (x^*, λ^*), the following holds

$$F(x^*)^T(x - x^*) + J^{-1}f(x)^T\lambda^* \geq 0, \qquad \text{for all } x \in X.$$

Lemma 2. Consider problem (**cSVI**). Let Assumption 1 holds. Assume that for any $x \in X$ and $\lambda \in \mathbb{R}_+^J$ we have

$$F(x)^T(\hat{x} - x) + J^{-1}f(\hat{x})^T\lambda \leq \Phi_\rho(x, \hat{\lambda}) + C(x, \lambda), \tag{4}$$

where $\hat{x} \in X$ and $\hat{\lambda} \in \mathbb{R}_+^J$ are arbitrary vectors. Then for any primal-dual solution pair (x^*, λ^*), the following holds.
(i) $J^{-1}\mathbf{1}^T[f(\hat{x})]_+ \leq C(x^*, \tilde{\lambda})$, where for all $j \in [J]$ we define

$$\tilde{\lambda}_j \triangleq \begin{cases} 1 + \lambda_j^*, & \text{if } f_j(\hat{x}) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) $\sup_{x \in \mathcal{X}}\{F(x)^T(\hat{x} - x)\} \leq \sup_{x \in \mathcal{X}}\{C(x, 0)\}$.

3   Algorithm outline

The outline of the proposed method is presented by Algorithm 1. The sequence of the primal iterates is denoted by $\{x_k\}$ and the sequence of the dual iterates is denoted by $\{\lambda_k\}$. This is a single timescale Lagrangian stochastic approximation scheme that includes two main steps. At each iteration, in the dual step in equation (5), a randomly selected dual variable $\lambda^{(j)}$ is updated, while in the primal step in equation (6), the primal variables are updated. The stepsize sequence is denoted by $\{\gamma_k\}$ and the penalty sequence is denoted by $\{\rho_k\}$. In addition to the primal and dual variables that are updated at each iteration, both the stepsize and penalty parameter are updated iteratively. Our goal in this work lies in proving that Algorithm 1 can be employed for solving the stochastic VI problem (**cSVI**) where the constraint set is characterized by explicit functional constraints. This result will be presented in the next section by Theorem 1 where we provide specific update rules for both γ_k and ρ_k such that the convergence of the proposed method can be guaranteed and non-asymptotic convergence rates can be derived. Before we proceed with the analysis of the method, we provide some definitions that will be utilized.

Remark 4. *Note that the Augmented Lagrangian function introduced in Definition 1 can be viewed as a relaxed variant of the following standard Augmented Lagrangian function of the form*

$$\mathcal{L}_\rho(x, \lambda) \triangleq \sup_{y \in \mathcal{X}} \mathcal{L}_\rho(x, y, \lambda) = sup_{y \in \mathcal{X}}\{F(y)^T(x - y)\} + \Phi_\rho(x, \lambda)$$
$$= Gap^*(x) + \Phi_\rho(x, \lambda).$$

Indeed, one of the key challenges in employing the Augmented Lagrangian function $\mathcal{L}_\rho(x, \lambda)$ is the presence of the supremum and nondifferentiability of the dual gap function. Further, even when the samples $F(\bullet, \xi_k)$ are unbiased, the standard Augmented Lagrangian function above may be biased, due to the presence of the supremum which again, renders an issue in utilizing this Augmented Lagrangian function. To circumvent these challenges, we employ the relaxed variant of the Augmented Lagrangian function introduced in Definition 1. Importantly, as it will be shown in Theorem 1, utilizing the relaxed variant of the Augmented Lagrangian function allows us to derive the rate statements. This is indeed a key novelty in the design of the proposed method in this work.

Algorithm 1 Randomized Lagrangian stochastic approximation method (RLSA)

1: **input**: Choose $x_0 \in X$, $\lambda_0 := 0_J$, and $\rho_0 > 0$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3:     Generate a random variable j_k uniformly drawn from $\{1, \ldots, J\}$
4:     Generate a random realization of ξ denoted by ξ_k and evaluate $F(x_k, \xi_k)$
5:     Update the dual variable λ_k for all $j = 1, \ldots, J$ as follows.

$$\lambda_{k+1}^{(j)} := \begin{cases} \left[\rho_k f_j(x_k) + \lambda_k^{(j)} \right]_+, & \text{if} \quad j = j_k \\ \lambda_k^{(j)}, & \text{otherwise} \end{cases} \tag{5}$$

6:     Evaluate $\tilde{\nabla} f_{j_k}(x_k) \in \partial f_{j_k}(x_k)$
7:     Update the primal variable x_k as follows.

$$x_{k+1} := \Pi_X \left[x_k - \gamma_k \left(F(x_k, \xi_k) + \lambda_{k+1}^{(j_k)} \tilde{\nabla} f_{j_k}(x_k) \right) \right] \tag{6}$$

8: **end for**

Throughout, we let the history of the method be denoted by $\mathcal{F}_k \triangleq \cup_{t=0}^{k-1} \{\xi_t, j_t\}$ for any $k \geq 1$, and $\mathcal{F}_0 \triangleq \{\xi_0, j_0\}$.

Assumption 2 (Random samples). Let the following holds.
(i) Samples ξ_k are generated independently from the probability distribution of ξ for $k \geq 0$.
(ii) Samples j_k, for $k \geq 0$, are generated independently from a uniform probability distribution such that $\text{Prob}(j_k = j) = J^{-1}$ for all $j = 1, \ldots, J$.
(iii) Samples ξ_k and j_k are generated independently from each other.
(iv) $\mathbb{E}[F(x, \xi_k) - F(x) \mid x] = 0$ for all $x \in X$ and all $k \geq 0$.
(v) There is some $\nu > 0$ such that $\mathbb{E}[\|F(x, \xi_k) - F(x)\|^2 \mid x] \leq \nu^2$ for all $x \in X$ and all $k \geq 0$.

Remark 5. In view of Assumption 1, the subdifferential set $\partial f_j(x)$ is nonempty for all $x \in \text{int}(\text{dom}(f_j))$ and all $j = 1, \ldots, J$. Also, f_j has bounded subgradients over X. Throughout, we let scalars D_X and D_f be defined as $D_X \triangleq \sup_{x \in X} \|x\|$ and $D_f \triangleq \max_{j \in [J]} \sup_{x \in X} |f_j(x)|$, respectively. Also, we let $C_F > 0$ and $C_f > 0$ be scalars such that $\|F(x)\| \leq C_F$ and $\|\tilde{\nabla} f_j(x)\| \leq C_f$ for all $\tilde{\nabla} f_j(x) \in \partial f_j(x)$, for all $x \in X$.

Definition 3 (Stochastic errors). Let us define the following stochastic terms for $k \geq 0$.
(i) $w_k \triangleq F(x_k, \xi_k) - F(x_k)$.
(ii) $\delta_k \triangleq \left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)} \right]_+ \tilde{\nabla} f_{j_k}(x_k) - \frac{1}{J} \sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)} \right]_+ \tilde{\nabla} f_j(x_k)$.

In the next lemma, we show that the stochastic errors defined above are unbiased and have bounded variance. The proof is provided in the appendix.

Lemma 3 (Properties of stochastic errors). Consider Definition 3. Let Assumption 2 holds. Then:
(i) $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu^2$.
(ii) $\mathbb{E}[\delta_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\delta_k\|^2 \mid \mathcal{F}_k] \leq 2C_f^2 \left(\rho_k^2 D_f^2 + \frac{\|\lambda_k\|^2}{J} \right)$.

Remark 6. Note that we have $\frac{1}{J}\sum_{j=1}^{J}\left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k) \in \partial_x \Phi_{\rho_k}(x_k, \lambda_k)$. Throughout, we use the following notation

$$\tilde{\nabla}_x \Phi_{\rho_k}(x_k, \lambda_k) \triangleq \frac{1}{J}\sum_{j=1}^{J}\left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k).$$

Therefore, one can conclude that

$$\|\tilde{\nabla}_x \Phi_{\rho_k}(x_k, \lambda_k)\|^2 \le \frac{1}{J}\sum_{j=1}^{J}\left\|\left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right\|^2 \le 2\rho_k^2 D_f^2 C_f^2 + \frac{2C_f^2}{J}\|\lambda_k\|^2.$$

4 Convergence and rate analysis

To obtain the main results of this paper, we use the following technical lemmas. All related proofs are provided in the appendix.

Lemma 4. Given an arbitrary sequences $\{\sigma_k\}_{k\ge 0} \subset \mathbb{R}^n$ and $\{\tau_k\}_{k\ge 0} \subset \mathbb{R}^{++}$, let $\{v_k\}_{k\ge 0}$ be a sequence such that $v_0 \in \mathbb{R}^n$ and $v_{k+1} = v_k + \tau_k \sigma_k$. Then, for all $k \ge 0$ and $x \in \mathbb{R}^n$,

$$\sigma_k^T(x - v_k) \le \frac{1}{2\tau_k}\|x - v_k\|^2 - \frac{1}{2\tau_k}\|x - v_{k+1}\| + \frac{\tau_k}{2}\|\sigma_k\|^2.$$

Lemma 5. Consider Algorithm 1. Let $J_k^+ = \{j \in [J] \mid \rho_k f_j(x_k) + \lambda_k^{(j)} \ge 0\}$ and $J_k^- = [J]\backslash J_k^+$. Then, for any $\lambda \in \mathbb{R}_+^J$, the following holds:

$$- \Phi_{\rho_k}(x_k, \lambda_k) + \frac{1}{J}\sum_{j=1}^{J}\lambda^{(j)}f_j(x_k) + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda\|^2$$

$$\le \frac{1}{2\rho_k}\|\lambda_k - \lambda\|^2 + (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)) + \Delta_k,$$

where $\Delta_k \triangleq -\frac{1}{J}\sum_{j\in J_k^+}\frac{\rho_k}{2}(f_j(x_k))^2 - \frac{1}{J}\sum_{j\in J_k^-}\frac{(\lambda_k^{(j)})^2}{2\rho_k} + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda_k\|^2$.

Lemma 6. Suppose Assumption 1 holds. Then, the following holds:

(a) $\|Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)\|^2 \le D_f^2$.

(b) Let $\bar{\sigma}_k = Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)$ and $\{\bar{v}_k\}_{k\ge 0}$ be a sequence such that $\bar{v}_0 \in \mathbb{R}^n$ and $\bar{v}_{k+1} = v_k + \bar{\tau}_k \bar{\sigma}_k$ for some $\{\bar{\tau}_k\}_{k\ge 0}$. Then, the following holds.

$$(\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$

$$\le (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + \frac{1}{2\bar{\tau}_k}\|\lambda - \bar{v}_k\|^2 - \frac{1}{2\bar{\tau}_k}\|\lambda - \bar{v}_{k+1}\|^2 + \frac{\bar{\tau}_k\|\bar{\sigma}_k\|^2}{2}.$$

Next, using Lemma 5 and 6 we provide one-step analysis of our method by providing an upper bound on the reduction of the gap function in terms of the consecutive iterates.

Proposition 2. Consider Algorithm 1. Let Assumptions 1 and 2 hold. Then, for any $x \in X$ and $\lambda \in \mathbb{R}_+^J$ the following inequality holds.

$$(x_k - x)^T F(x) + J^{-1} f(x_k)^T \lambda - \Phi_{\rho_k}(x, \lambda_k)$$

$$\leq \frac{1}{2\gamma_k} \left(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \right) + \frac{1}{4\gamma_k} \left(\|x - v_k\|^2 - \|x - v_{k+1}\|^2 \right)$$

$$+ \frac{1}{2\rho_k} \left(\|\lambda_k - \lambda\|^2 - \|\lambda_{k+1} - \lambda\|^2 \right) + \frac{1}{2\bar{\tau}_k} \left(\|\lambda - \bar{v}_k\|^2 - \|\lambda - \bar{v}_{k+1}\|^2 \right) + 2\gamma_k C_F^2$$

$$+ 4\gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \frac{1}{J} \|\lambda_k\|^2 \right) + (v_k - x_k)^T(w_k + \delta_k) + 2\gamma_k \|w_k + \delta_k\|^2$$

$$+ (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + \frac{\bar{\tau}_k \|\bar{\sigma}_k\|^2}{2} - \frac{1}{J} \sum_{j \in J_k^+} \frac{\rho_k}{2} (f_j(x_k))^2 - \frac{1}{J} \sum_{j \in J_k^-} \frac{(\lambda_k^{(j)})^2}{2\rho_k}, \tag{7}$$

where J_k^+ and J_k^- are given by Lemma 5.

Proof. Let $x \in X$ and $\lambda \geq 0$ be arbitrary vectors. From (6) we have

$$(x_{k+1} - x)^T \left(x_{k+1} - x_k + \gamma_k \left(F(x_k, \xi_k) + \left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)} \right]_+ \tilde{\nabla} f_{j_k}(x_k) \right) \right) \leq 0. \tag{8}$$

Using monotonicity of $F(\bullet)$ and Young's inequality, one can obtain

$$(x_{k+1} - x)^T F(x_k, \xi_k)$$

$$= (x_{k+1} - x_k)^T F(x_k) + (x_k - x)^T F(x_k) + (x_{k+1} - x)^T w_k$$

$$\geq -\frac{1}{8\gamma_k} \|x_{k+1} - x_k\|^2 - 2\gamma_k \|F(x_k)\|^2 + (x_k - x)^T F(x) + (x_{k+1} - x)^T w_k$$

$$\geq -\frac{1}{8\gamma_k} \|x_{k+1} - x_k\|^2 - 2\gamma_k C_F^2 + (x_k - x)^T F(x) + (x_{k+1} - x)^T w_k. \tag{9}$$

Similarly from Remark 6 and convexity of $\Phi_{\rho_k}(\bullet, \lambda_k)$, then we have

$$(x_{k+1} - x)^T \left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)} \right]_+ \tilde{\nabla} f_{j_k}(x_k)$$

$$= (x_{k+1} - x_k)^T \tilde{\nabla}_x \Phi_{\rho_k}(x_k, \lambda_k) + (x_k - x)^T \tilde{\nabla}_x \Phi_{\rho_k}(x_k, \lambda_k) + (x_{k+1} - x)^T \delta_k$$

$$\geq -\frac{1}{8\gamma_k} \|x_{k+1} - x_k\|^2 - 2\gamma_k \|\tilde{\nabla}_x \Phi_{\rho_k}(x_k, \lambda_k)\|^2 + \Phi_{\rho_k}(x_k, \lambda_k) - \Phi_{\rho_k}(x, \lambda_k)$$

$$\quad + (x_{k+1} - x)^T \delta_k$$

$$\geq -\frac{1}{8\gamma_k} \|x_{k+1} - x_k\|^2 - 4\gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \frac{1}{J} \|\lambda_k\|^2 \right) + \Phi_{\rho_k}(x_k, \lambda_k) - \Phi_{\rho_k}(x, \lambda_k)$$

$$\quad + (x_{k+1} - x)^T \delta_k. \tag{10}$$

We can also write

$$(x_{k+1} - x)^T (x_{k+1} - x_k) = \frac{1}{2} \left(\|x_{k+1} - x\|^2 - \|x_k - x\|^2 + \|x_{k+1} - x_k\|^2 \right). \tag{11}$$

Using (9),(10) and (11) in (8), we have

$$(x_k - x)^T F(x) + \Phi_{\rho_k}(x_k, \lambda_k) - \Phi_{\rho_k}(x, \lambda_k)$$

$$\leq \frac{1}{2\gamma_k} \left(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 - \frac{1}{2} \|x_{k+1} - x_k\|^2 \right) + 2\gamma_k C_F^2$$

$$+ 4\gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \frac{1}{J} \|\lambda_k\|^2 \right) + \underbrace{(x - x_{k+1})^T(w_k + \delta_k)}_{\text{term (a)}}. \tag{12}$$

9

Now we obtain an upper bound for term (a) in (12).

$$(x - x_{k+1})^T (w_k + \delta_k)$$
$$= (x - x_k)^T (w_k + \delta_k) + (x_k - x_{k+1})^T (w_k + \delta_k)$$
$$\leq (x - v_k)^T (w_k + \delta_k) + (v_k - x_k)^T (w_k + \delta_k) + \frac{1}{4\gamma_k} \|x_k - x_{k+1}\|^2 + \gamma_k \|w_k + \delta_k\|^2.$$

From Lemma 4 we have that $(x - v_k)^T (w_k + \delta_k) \leq \frac{1}{4\gamma_k} \|x - v_k\|^2 - \frac{1}{4\gamma_k} \|x - v_{k+1}\|^2 + \gamma_k \|w_k + \delta_k\|^2$, hence the above inequality can be written as

$$(x - x_{k+1})^T (w_k + \delta_k) \leq \frac{1}{4\gamma_k} \|x - v_k\|^2 - \frac{1}{4\gamma_k} \|x - v_{k+1}\|^2 + \frac{1}{4\gamma_k} \|x_k - x_{k+1}\|^2$$
$$+ (v_k - x_k)^T (w_k + \delta_k) + 2\gamma_k \|w_k + \delta_k\|^2.$$

Using the above inequality in (12), we get

$$(x_k - x)^T F(x) + \Phi_{\rho_k}(x_k, \lambda_k) - \Phi_{\rho_k}(x, \lambda_k) \leq \frac{1}{2\gamma_k} \left(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \right)$$
$$+ 2\gamma_k C_F^2 + 4\gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \frac{1}{J} \|\lambda_k\|^2 \right) + \frac{1}{4\gamma_k} \left(\|x - v_k\|^2 - \|x - v_{k+1}\|^2 \right)$$
$$+ (v_k - x_k)^T (w_k + \delta_k) + 2\gamma_k \|w_k + \delta_k\|^2. \tag{13}$$

Using Lemmas 5 and 6, we can bound the left hand side of (13) from below and one can obtain the following.

$$(x_k - x)^T F(x) + \frac{1}{J} \sum_{j=1}^{J} \lambda^{(j)} f_j(x_k) - \Phi_{\rho_k}(x, \lambda_k)$$
$$\leq \frac{1}{2\gamma_k} \left(\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \right) + \frac{1}{4\gamma_k} \left(\|x - v_k\|^2 - \|x - v_{k+1}\|^2 \right)$$
$$+ \frac{1}{2\rho_k} \left(\|\lambda_k - \lambda\|^2 - \|\lambda_{k+1} - \lambda\|^2 \right) + \frac{1}{2\bar{\tau}_k} \left(\|\lambda - \bar{v}_k\|^2 - \|\lambda - \bar{v}_{k+1}\|^2 \right) + 2\gamma_k C_F^2$$
$$+ 4\gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \frac{1}{J} \|\lambda_k\|^2 \right) + (v_k - x_k)^T (w_k + \delta_k) + 2\gamma_k \|w_k + \delta_k\|^2$$
$$+ (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + \frac{\bar{\tau}_k \|\bar{\sigma}_k\|^2}{2} + \Delta_k,$$

where Δ_k is defined in Lemma 5. ∎

Now we show that the sequence of dual iterates generated by the proposed method is bounded.

Lemma 7. Consider Algorithm 1. Let Assumptions 1 and 2 hold. Let $\rho_k = \frac{\rho}{\sqrt{(k+1)\log(k+1)}}$, $\gamma_k = \frac{\gamma}{\sqrt{(k+1)\log(k+1)}}$, $t_k = \bar{\tau}_k = \frac{1}{\sqrt{(k+1)\log(k+1)}}$ for any $k \geq 1$, where $\rho\gamma \leq \frac{1}{120\rho\gamma C_f^2/J}$. Moreover, we define $\rho_0 = \rho$, $\gamma_0 = \gamma$ and $t_0 = \bar{\tau}_0 = 1$. Then, there exists $B \geq 0$ such that $\mathbb{E}[\|\lambda_K\|^2] \leq B$ for any $K \geq 0$.

Proof. From Lemma 1 we have $(x_k - x^*)^T F(x^*) + J^{-1} f(x_k)^T \lambda^* \geq 0$. Also, since $f_j(x^*) \leq 0$ for all $j \in J$, we have $\Phi_\rho(x^*, \lambda_k) \leq 0$. In view of these relations, from Proposition 2, for $x := x^*$ and

$\lambda := \lambda^*$ we obtain

$$
\begin{aligned}
0 \quad &\le \tfrac{1}{2\gamma_k}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) + \tfrac{1}{4\gamma_k}\left(\|x^* - v_k\|^2 - \|x^* - v_{k+1}\|^2\right) \\
&\quad + \tfrac{1}{2\rho_k}\left(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right) + \tfrac{1}{2\bar{\tau}_k}\left(\|\lambda^* - \bar{v}_k\|^2 - \|\lambda^* - \bar{v}_{k+1}\|^2\right) + 2\gamma_k C_F^2 \\
&\quad + 4\gamma_k C_f^2\left(\rho_k^2 D_f^2 + \tfrac{1}{J}\|\lambda_k\|^2\right) + (v_k - x_k)^T(w_k + \delta_k) + 2\gamma_k\|w_k + \delta_k\|^2 \\
&\quad + (\bar{v}_k - \lambda_k)^T\bar{\sigma}_k + \tfrac{\bar{\tau}_k\|\bar{\sigma}_k\|^2}{2} - \tfrac{1}{J}\sum_{j \in J_k^+}\tfrac{\rho_k}{2}(f_j(x_k))^2 - \tfrac{1}{J}\sum_{j \in J_k^-}\tfrac{(\lambda_k^{(j)})^2}{2\rho_k}.
\end{aligned}
$$

Multiplying both sides by t_k and using the fact that $\tfrac{t_k}{\rho_k} \ge \tfrac{t_{k+1}}{\rho_{k+1}}$, $\tfrac{t_k}{\gamma_k} \ge \tfrac{t_{k+1}}{\gamma_{k+1}}$, $t_k \ge t_{k+1}$, $\rho_k \ge \rho_{k+1}$, $\tfrac{t_k}{\bar{\tau}_k} \ge \tfrac{t_{k+1}}{\bar{\tau}_{k+1}}$, summing over $k = 0, \ldots, T$, where $T \le K$, and from $\|\lambda_{T+1}\|^2 \le 2\|\lambda_{T+1} - \lambda^*\|^2 + 2\|\lambda^*\|^2$ we obtain the following relation.

$$
\begin{aligned}
&\tfrac{t_{T+1}}{4\rho_{T+1}}\|\lambda_{T+1}\|^2 \\
&\le \tfrac{t_0}{2\gamma_0}\|x_0 - x^*\|^2 + \tfrac{t_0}{4\gamma_0}\|x^* - v_0\|^2 + \tfrac{1}{2\rho_0}\|\lambda_0 - \lambda^*\|^2 + \tfrac{t_0}{2\bar{\tau}_0}\|\lambda^* - \bar{v}_0\|^2 + \tfrac{t_{T+1}}{2\rho_{T+1}}\|\lambda^*\|^2 \\
&\quad + \sum_{k=0}^{T} t_k\gamma_k\left(2C_F^2 + 4C_f^2\left(\rho_k^2 D_f^2 + \tfrac{1}{J}\|\lambda_k\|^2\right)\right) + \sum_{k=0}^{T} t_k(v_k - x_k)^T(w_k + \delta_k) \\
&\quad + 2\sum_{k=0}^{T} t_k\gamma_k\|w_k + \delta_k\|^2 + \sum_{k=0}^{T} t_k(\bar{v}_k - \lambda_k)^T\bar{\sigma}_k + \sum_{k=0}^{T} t_k\tfrac{\bar{\tau}_k\|\bar{\sigma}_k\|^2}{2} - \sum_{k=0}^{T} t_k\Delta_k.
\end{aligned}
$$

Taking expectation on the both sides and using Assumption 2(iv-v), Lemma 3 and the fact that $\mathbb{E}[(\bar{v}_k - \lambda_k)^T\bar{\sigma}_k] = \mathbb{E}[\Delta_k] = 0$, we get

$$
\begin{aligned}
&\tfrac{t_{T+1}}{4\rho_{T+1}}\mathbb{E}[\|\lambda_{T+1}\|^2] \\
&\le \tfrac{t_0}{2\gamma_0}\|x_0 - x^*\|^2 + \tfrac{t_0}{4\gamma_0}\|x^* - v_0\|^2 + \tfrac{1}{2\rho_0}\|\lambda_0 - \lambda^*\|^2 + \tfrac{t_0}{2\bar{\tau}_0}\|\lambda^* - \bar{v}_0\|^2 + \tfrac{t_{T+1}}{2\rho_{T+1}}\|\lambda^*\|^2 \\
&\quad + \sum_{k=0}^{T} t_k\gamma_k\left(2C_F^2 + 4C_f^2\left(\rho_k^2 D_f^2 + \tfrac{1}{J}\mathbb{E}[\|\lambda_k\|^2]\right)\right) \\
&\quad + 2\sum_{k=0}^{T} t_k\gamma_k(2\nu^2 + 4C_f^2\rho_k^2 D_f^2 + 4C_f^2\mathbb{E}[\tfrac{\|\lambda_k\|^2}{J}]) + \sum_{k=0}^{T} t_k\tfrac{\bar{\tau}_k D_f^2}{2},
\end{aligned}
$$

where we used part (a) of Lemma (6) and the definition of $\bar{\sigma}_k$, i.e., $\mathbb{E}[\|\bar{\sigma}_k\|^2] \le D_f^2$. Define $A_1 = \tfrac{t_0}{2\gamma_0}\|x_0 - x^*\|^2 + \tfrac{t_0}{4\gamma_0}\|x^* - v_0\|^2 + \tfrac{1}{2\rho_0}\|\lambda_0 - \lambda^*\|^2 + \tfrac{t_0}{2\bar{\tau}_0}\|\lambda^* - \bar{v}_0\|^2$, $A_2 = 2C_F^2 + 4C_f^2\rho^2 D_f^2$, $A_3 = 2\nu^2 + 4C_f^2\rho^2 D_f^2$, then the above inequality can be written as follows, where we used the fact that $\rho_k = \tfrac{\rho}{(k+1)\log(k+1)} \le \rho$.

$$
\begin{aligned}
\tfrac{t_{T+1}}{4\rho_{T+1}}\mathbb{E}[\|\lambda_{T+1}\|^2] &\le A_1 + \tfrac{t_{T+1}}{2\rho_{T+1}}\|\lambda^*\|^2 + \sum_{k=0}^{T} t_k\gamma_k\left(A_2 + 4C_f^2\tfrac{1}{J}\mathbb{E}[\|\lambda_k\|^2]\right) \\
&\quad + 2\sum_{k=0}^{T} t_k\gamma_k(A_3 + \tfrac{4}{J}C_f^2\mathbb{E}[\|\lambda_k\|^2]) + \sum_{k=0}^{T} t_k\tfrac{\bar{\tau}_k D_f^2}{2}. \qquad (14)
\end{aligned}
$$

Letting $T = -1$, one can easily show that $\mathbb{E}[\|\lambda_0\|^2] \le B$. Now suppose $\mathbb{E}[\|\lambda_{T+1}\|^2] \le B$ holds for all $T \in \{-1, 0, \ldots, K-2\}$. We show that $\mathbb{E}[\|\lambda_{T+1}\|^2] \le B$ for $T = K-1$. Multiplying both sides

of (14) by $\frac{4\rho_{T+1}}{t_{T+1}}$ and letting $T = K - 1$, we get

$$\mathbb{E}[\|\lambda_K\|^2] \leq \frac{4\rho_K}{t_K} A_1 + 2\|\lambda^*\|^2 + \frac{4\rho_K}{t_K} \sum_{k=0}^{K-1} t_k \gamma_k \left(A_2 + 4C_f^2 B \tfrac{1}{J}\right)$$

$$+ \frac{8\rho_K}{t_K} \sum_{k=0}^{K-1} t_k \gamma_k (A_3 + \tfrac{4}{J} C_f^2 B) + \frac{4\rho_K}{t_K} \sum_{k=0}^{K-1} t_k \frac{\bar{\tau}_k D_f^2}{2}.$$

From the fact that $\rho_k = \frac{\rho}{\sqrt{(k+1)\log(k+1)}}$, $\gamma_k = \frac{\gamma}{\sqrt{(k+1)\log(k+1)}}$, $t_k = \bar{\tau}_k = \frac{1}{\sqrt{(k+1)\log(k+1)}}$, one can show that $\frac{\rho_k}{t_k} = \rho$ and $\sum_{k=0}^{K-1} t_k \gamma_k \leq 3\gamma$. Therefore, we obtain

$$\mathbb{E}[\|\lambda_K\|^2] \leq 4\rho A_1 + 2\|\lambda^*\|^2 + 12\rho\gamma \left(A_2 + 4C_f^2 B \tfrac{1}{J}\right) + 24\rho\gamma(A_3 + \tfrac{4}{J} C_f^2 B) + 12\rho \frac{D_f^2}{2} \leq B,$$

where in the last inequality we used the fact that $B = \max\left\{\|\lambda_0\|^2, \frac{4\rho A_1 + 2\|\lambda^*\|^2 + 12\rho\gamma A_2 + 24\rho\gamma A_3 + 12\rho D_f^2}{1 - 144\rho\gamma C_f^2/J}\right\}$ and $\rho\gamma \leq \frac{1}{144\rho\gamma C_f^2/J}$. ∎

Now we are ready to state the convergence rates of Algorithm 1.

Theorem 1 (Convergence rate statements for Algorithm 1). Consider Algorithm 1. Let Assumptions 1 and 2 hold. Let $\rho_k = \frac{\rho}{\sqrt{(k+1)\log(k+1)}}$, $\gamma_k = \frac{\gamma}{\sqrt{(k+1)\log(k+1)}}$, $t_k = \bar{\tau}_k = \frac{1}{\sqrt{(k+1)\log(k+1)}}$ for all $k \geq 1$, where $\rho\gamma \leq \frac{1}{144\rho\gamma C_f^2/J}$. Moreover, we define $\rho_0 = \rho$, $\gamma_0 = \gamma$ and $t_0 = \bar{\tau}_0 = 1$. Let us define $\bar{x}_K \triangleq \frac{\sum_{k=0}^{K} t_k x_k}{\sum_{k=0}^{K} t_k}$ for $K \geq 0$. Then, for any $K \geq 0$, we have

$$\mathbb{E}\left[\sup_{x \in \mathcal{X}} \{F(x)^T (\bar{x}_K - x)\}\right] \leq \mathcal{O}\left(\log(K+1)/\sqrt{K+2}\right)$$
$$\mathbb{E}\left[J^{-1} \mathbf{1}^T [f(\bar{x}_K)]_+\right] \leq \mathcal{O}\left(\log(K+1)/\sqrt{K+2}\right).$$

Proof. Multiplying both sides of (7) by t_k, using the fact that $\frac{t_k}{\rho_k} \geq \frac{t_{k+1}}{\rho_{k+1}}$, $\frac{t_k}{\gamma_k} \geq \frac{t_{k+1}}{\gamma_{k+1}}$, $t_k \geq t_{k+1}$, $\rho_k \geq \rho_{k+1}$, $\frac{t_k}{\bar{\tau}_k} \geq \frac{t_{k+1}}{\bar{\tau}_{k+1}}$, and summing $k = 0$ to K, we get

$$\sum_{k=0}^{K} t_k \left((x_k - x)^T F(x) + \frac{1}{J} \sum_{j=1}^{J} \lambda^{(j)} f_j(x_k) - \Phi_\rho(x, \lambda_k)\right)$$

$$\leq \underbrace{\frac{t_0}{2\gamma_0}\|x_0 - x\|^2 + \frac{t_0}{4\gamma_0}\|v_0 - x\|^2 + \frac{t_0}{2\rho_0}\|\lambda_0 - \lambda\|^2}_{\text{term (a)}}$$

$$+ \frac{t_0}{2\bar{\tau}_0} + \sum_{k=0}^{K} 2t_k \gamma_k C_F^2 + \sum_{k=0}^{K} 4t_k \gamma_k C_f^2 \left(\rho_k^2 D_f^2 + \tfrac{1}{J}\|\lambda_k\|^2\right)$$

$$+ \underbrace{\sum_{k=0}^{K} t_k \left((v_k - x_k)^T (w_k + \delta_k) + 2\gamma_k \|w_k + \delta_k\|^2 + (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + \frac{\bar{\tau}_k \|\bar{\sigma}_k\|^2}{2} + \Delta_k\right)}_{\text{term (b)}}. \quad (15)$$

Let right-hand side of (15) denoted by $C(x, \lambda)$. Dividing both sides of the above inequality by $\sum_{k=0}^{K} t_k$ and invoking the definition of $\bar{x}_k$, we get

$$(\bar{x}_K - x)^T F(x) + \frac{1}{J} \sum_{j=1}^{J} \lambda^{(j)} f_j(\bar{x}_K) - \Phi_\rho(x, \bar{\lambda}_k) \leq \frac{1}{\sum_{k=0}^{K} t_k} C(x, \lambda),$$

where in the left-hand side we used Jensen's inequality and the fact that Φ_ρ is concave with respect to λ.

Since $\bar{x}_K \in X$, from Lemma 2 (i) we have $J^{-1}\mathbf{1}^T[f(\bar{x}_K)]_+ \leq C(x^*, \tilde{\lambda})$, where $\tilde{\lambda}$ is defined in Lemma 2. taking expectation on both side and using definition of $C(x, \lambda)$ in (15), Lemmas 3 7, Assumption 2 (iv-v), the fact that $\mathbb{E}[(\bar{v}_k - \lambda_k)^T \bar{\sigma}_k] = \mathbb{E}[\Delta_k] = 0$ and $\mathbb{E}[\|\bar{\sigma}_k\|^2] \leq D_f^2$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[J^{-1}\mathbf{1}^T[f(\bar{x}_K)]_+\right] \leq & \frac{1}{\sum_{k=0}^K t_k}\Big[\tfrac{t_0}{2\gamma_0}\|x_0 - x^*\|^2 + \tfrac{t_0}{4\gamma_0}\|v_0 - x^*\|^2 + \tfrac{t_0}{2\rho_0}\|\lambda_0 - \tilde{\lambda}\|^2 \\
& + \tfrac{t_0}{2\bar{\tau}_0} + \sum_{k=0}^K 2t_k\gamma_k C_F^2 + \sum_{k=0}^K 4t_k\gamma_k C_f^2 \rho_k^2 D_f^2 + \sum_{k=0}^K 4t_k\gamma_k C_f^2 \tfrac{1}{J}B \\
& + \sum_{k=0}^K t_k\left(2\gamma_k(2\nu^2 + 4C_f^2\rho_k^2 D_f^2 + \tfrac{4}{J}C_f^2 B) + \tfrac{\bar{\tau}_k D_f^2}{2}\right)\Big].
\end{aligned}
\tag{16}
$$

Moreover, from Lemma 2 (ii) we have $\sup_{x\in\mathcal{X}}\{F(x)^T(\bar{x}_K - x)\} \leq \sup_{x\in\mathcal{X}}\{C(x, 0)\}$. By taking conditional expectation and then, unconditional expectation on both sides and using the fact that term (a) and term (b) in the definition of $C(x, \lambda)$ do not depend on x, we obtain

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{x\in\mathcal{X}}&\{F(x)^T(\bar{x}_K - x)\}\Big] \\
\leq & \frac{1}{\sum_{k=0}^K t_k}\Big[\sup_{x\in\mathcal{X}}\left\{\tfrac{t_0}{2\gamma_0}\|x_0 - x\|^2 + \tfrac{t_0}{4\gamma_0}\|v_0 - x\|^2 + \tfrac{t_0}{2\rho_0}\|\lambda_0\|^2\right\} + \tfrac{t_0}{2\bar{\tau}_0} \\
& + \sum_{k=0}^K 2t_k\gamma_k C_F^2 + \sum_{k=0}^K 4t_k\gamma_k C_f^2 \rho_k^2 D_f^2 + \sum_{k=0}^K 4t_k\gamma_k C_f^2 \tfrac{1}{J}B \\
& + \sum_{k=0}^K t_k\left(2\gamma_k(2\nu^2 + 4C_f^2\rho_k^2 D_f^2 + \tfrac{4}{J}C_f^2 B) + \tfrac{\bar{\tau}_k D_f^2}{2}\right)\Big].
\end{aligned}
\tag{17}
$$

From $\rho_k = \frac{\rho}{\sqrt{(k+1)\log(k+1)}}$, $\gamma_k = \frac{\gamma}{\sqrt{(k+1)\log(k+1)}}$, $t_k = \bar{\tau}_k = \frac{1}{\sqrt{(k+1)\log(k+1)}}$, and the facts that $\rho_0 = \rho$, $\gamma_0 = \gamma$, $t_0 = \bar{\tau}_0 = 1$, one can show that $\sum_{k=0}^K t_k\gamma_k \leq 3\gamma$ and similarly $\sum_{k=0}^K t_k\bar{\tau}_k \leq 3$, also $\sum_{k=0}^K t_k \geq \frac{1}{\log(K+1)}\int_1^{K+1}\frac{1}{\sqrt{x+1}}dx = \frac{2(\sqrt{K+2}-\sqrt{2})}{\log(K+1)}$. Therefore, we obtain that $\mathbb{E}\left[J^{-1}\mathbf{1}^T[f(\bar{x}_K)]_+\right] \leq \mathcal{O}(\log(K+1)/\sqrt{(K+2)})$ and similarly $\mathbb{E}\left[\sup_{x\in\mathcal{X}}\{F(x)^T(\bar{x}_K - x)\}\right] \leq \mathcal{O}(\log(K+1)/\sqrt{(K+2)})$. ∎

Notably, the rate statements in Theorem 1 are in a mean sense, for both the dual gap function and the infeasibility metric. The latter quantifies the violation of the explicit functional constraints. A natural question is whether we can guarantee the convergence of the infeasibility metric to zero in an almost sure sense. This is partially addressed in the following result.

Corollary 1. *Consider Theorem 1. There exists a subsequence of $\{\bar{x}_k\}$ along which, the infeasibility metric $\mathbf{1}^T[f(\bar{x}_K)]_+$ converges to zero almost surely.*

Proof. From Theorem 1, we have $\lim_{K\to\infty}\mathbb{E}\left[\mathbf{1}^T[f(\bar{x}_K)]_+\right] = 0$. Invoking Fatou's lemma and noting that $\mathbf{1}^T[f(\bar{x}_K)]_+ \geq 0$, we obtain

$$
\liminf_{K\to\infty} \mathbf{1}^T[f(\bar{x}_K)]_+ = 0 \qquad \text{almost surely.}
$$

13

Further, the sequence $\{\bar{x}_k\}$ is bounded, due to the projection onto the compact set X in Algorithm 1. From the continuity of f, it follows that one of the (random) accumulation points of $\{\bar{x}_k\}$ must be a feasible point with respect to the explicit functional constraints almost surely. ∎

5 Conclusion

In this paper, we consider stochastic variational inequality (VI) problems with a monotone mapping and a set that is characterized in terms of explicit functional constraints. Motivated by the absence of convergence rate statements for solving this class of problems, we develop a randomized Lagrangian stochastic approximation method where at each iteration the primal and dual variables are updated recursively. Our main contribution is to show that the existing convergence rates for nonlinearly constrained stochastic optimization problems can be extended to the stochastic VI regime. This is indeed promising and implies that the Lagrangian duality theory can be employed with provable guarantees for several important classes of problems that can be formulated as a stochastic VI. In particular, this work provides convergence speed guarantees for computing a Nash equilibrium in stochastic Nash games where each player may be associated with many hard-to-project constraints.

6 Appendix

6.1 Proof of Lemma 1

Proof. Invoking Proposition 1 and taking into account that $\mathcal{N}_X(x^*) = \partial \mathcal{I}_X(x^*)$, we have that $x^* \in X$ solves the following augmented variational inequality problem $\mathrm{VI}\left(X, F + J^{-1}\nabla f^T \lambda^*\right)$, that is parameterized by J and λ^*. This implies that

$$\left(F(x^*) + J^{-1}\nabla f(x^*)^T \lambda^*\right)^T (x - x^*) \geq 0, \qquad \text{for all } x \in X. \tag{18}$$

From the convexity of function f_j for all $j \in [J]$ and that $\lambda_j \geq 0$, we have

$$\lambda_j^* \left(f_j(x) - f_j(x^*)\right) \geq \lambda_j^* \nabla f_j(x^*)^T (x - x^*).$$

Summing the preceding relation over $j \in [J]$ and recalling the definition of the mapping $f(x)$, we obtain

$$(f(x) - f(x^*))^T \lambda^* \geq \left(\nabla f(x^*)^T \lambda^*\right)^T (x - x^*).$$

Invoking Proposition 1 (ii) we obtain $f(x)^T \lambda^* \geq \left(\nabla f(x^*)^T \lambda^*\right)^T (x - x^*)$. From the preceding relation and (18) we obtain $F(x^*)^T (x - x^*) + J^{-1} f(x)^T \lambda^* \geq 0$ for all $x \in X$. ∎

6.2 Proof of Lemma 2

Proof. (i) Note that x^* is a feasible point to problem (**cSVI**) with respect to the set $\mathcal{X}$, i.e., $x^* \in \mathcal{X}$. Also, note that $\hat{\lambda} \geq 0$. From the definition of Φ_ρ, we have that $\Phi_\rho(x^*, \hat{\lambda}) \leq 0$. Let $x := x^*$ in (4). Then we have

$$F(x^*)^T (\hat{x} - x^*) + J^{-1} f(\hat{x})^T \lambda \leq C(x^*, \lambda). \tag{19}$$

14

Also, from Lemma 1 and that $\hat{x} \in X$ we have

$$0 \le F(x^*)^T(\hat{x} - x^*) + J^{-1}f(\hat{x})^T\lambda^*.$$

The preceding relation and that $\lambda^* \ge 0$ imply that

$$0 \le F(x^*)^T(\hat{x} - x^*) + J^{-1}[f(\hat{x})]_+^T\lambda^*.$$

Summing the preceding relation and (19) and rearranging the terms, we obtain

$$J^{-1}f(\hat{x})^T\lambda - J^{-1}[f(\hat{x})]_+^T\lambda^* \le C(x^*, \lambda). \tag{20}$$

Let us choose $\lambda_j := 1 + \lambda_j^*$ if $f_j(\hat{x}) > 0$, and $\lambda_j := 0$ otherwise for all $j \in [J]$. Then, we obtain the desired relation in (i).

(ii) Let $\lambda = 0$ in (4) and note that $\Phi_\rho(x, \hat{\lambda}) \le 0$ for all $x \in \mathcal{X}$. We have $F(x)^T(\hat{x} - x) \le C(x, 0)$ for all $x \in \mathcal{X}$. Taking supremum from the both sides, we obtain desired results in (ii). ∎

6.3 Proof of Lemma 3

Proof. The relations in part (i) hold as a consequence of Assumption 2. To show $\mathbb{E}[\delta_k \mid \mathcal{F}_k] = 0$, we can write

$$\mathbb{E}[\delta_k \mid \mathcal{F}_k] = \mathbb{E}\left[\left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)}\right]_+ \tilde{\nabla} f_{j_k}(x_k) - \tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k) \mid \mathcal{F}_k\right]$$

$$= \tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k) - \tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k) = 0,$$

where the last inequality is implied from the assumption that j_k is uniformly drawn from the set $[J]$. Next, we derive the bound on $\mathbb{E}[\|\delta_k\|^2 \mid \mathcal{F}_k]$. We have

$$\mathbb{E}[\|\delta_k\|^2 \mid \mathcal{F}_k] = \mathbb{E}\left[\left\|\left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)}\right]_+ \tilde{\nabla} f_{j_k}(x_k)\right\|^2 \mid \mathcal{F}_k\right] + \left\|\tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right\|^2$$

$$- 2\mathbb{E}\left[\left[\rho_k f_{j_k}(x_k) + \lambda_k^{(j_k)}\right]_+ \tilde{\nabla} f_{j_k}(x_k) \mid \mathcal{F}_k\right]^T \left(\tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right)$$

$$= \tfrac{1}{J}\sum_{j=1}^J \left\|\left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right\|^2 - \left\|\tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right\|^2.$$

Dropping the non-negative term in the preceding relation and invoking Remark 5, we obtain

$$\mathbb{E}[\|\delta_k\|^2 \mid \mathcal{F}_k] \le \tfrac{1}{J}\sum_{j=1}^J \left\|\left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+ \tilde{\nabla} f_j(x_k)\right\|^2 = \tfrac{1}{J}\sum_{j=1}^J \left[\rho_k f_j(x_k) + \lambda_k^{(j)}\right]_+^2 \left\|\tilde{\nabla} f_j(x_k)\right\|^2$$

$$\le \tfrac{C_f^2}{J}\sum_{j=1}^J \left(\rho_k f_j(x_k) + \lambda_k^{(j)}\right)^2 \le \tfrac{2C_f^2}{J}\sum_{j=1}^J \left(\rho_k^2 D_f^2 + \left(\lambda_k^{(j)}\right)^2\right) = 2C_f^2\left(\rho_k^2 D_f^2 + \tfrac{\|\lambda_k\|^2}{J}\right).$$

∎

6.4 Proof of Lemma 4

Proof. From the update rule of v_{k+1}, we know $\sigma_k = \tfrac{1}{\tau_k}(v_{k+1} - v_k)$, hence we have that

$$\sigma_k^T(x - v_k) = \sigma_k^T(x - v_{k+1}) + \sigma_k^T(v_{k+1} - v_k)$$

$$\le \frac{1}{2\tau_k}\|x - v_k\|^2 - \frac{1}{2\tau_k}\|x - v_{k+1}\|^2 - \frac{1}{2\tau_k}\|v_{k+1} - v_k\|^2 + \sigma_k^T(v_{k+1} - v_k)$$

$$\le \frac{1}{2\tau_k}\|x - v_k\|^2 - \frac{1}{2\tau_k}\|x - v_{k+1}\|^2 + \frac{\tau_k}{2}\|\sigma_k\|^2.$$

first inequality is obtain from three points inequality. ∎

15

6.5 Proof of Lemma 5

Proof. From the fact that $\lambda_{k+1} - \lambda_k = J\rho_k e_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)$, one can get the following:

$$\frac{1}{\rho_k}(\lambda_k - \lambda)^T(\lambda_{k+1} - \lambda_k) = (\lambda_k - \lambda)^T(\nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$
$$+ (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)).$$

also by knowing that $\frac{1}{\rho_k}(\lambda_k - \lambda)^T(\lambda_{k+1} - \lambda_k) = \frac{1}{2\rho_k}(\|\lambda_{k+1} - \lambda\|^2 - \|\lambda_k - \lambda\|^2 - \|\lambda_{k+1} - \lambda_k\|^2)$ and using previous equality one can obtain:

$$\frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda\|^2 = \frac{1}{2\rho_k}\|\lambda_k - \lambda\|^2 + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda_k\|^2 + (\lambda_k - \lambda)^T(\nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$
$$+ (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)) \tag{21}$$

Using (21), one can easily show that:

$$- \Phi_{\rho_k}(x_k, \lambda_k) + \frac{1}{J}\sum_{j=1}^{J}\lambda^{(j)}f_j(x_k) + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda\|^2$$

$$= -\Phi_{\rho_k}(x_k, \lambda_k) + \frac{1}{J}\sum_{j=1}^{J}\lambda^{(j)}f_j(x_k) + \frac{1}{2\rho_k}\|\lambda_k - \lambda\|^2 + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda_k\|^2 + (\lambda_k - \lambda)^T(\nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$

$$+ (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)). \tag{22}$$

From definition of $\Phi_{\rho_k}(x_k, \lambda_k)$, J_k^+ and J_k^- we have :

$$\Phi_{\rho_k}(x_k, \lambda_k) = \frac{1}{J}\Big[\sum_{j\in J_k^+}(\frac{\rho_k}{2}(f_j(x_k))^2 + \lambda_k^{(j)}f_j(x_k)) - \sum_{j\in J_k^-}\frac{(\lambda_k^{(j)})^2}{2\rho_k}\Big]. \tag{23}$$

Using (22), (23) and the fact that $\nabla_\lambda \Phi_\rho(x, \lambda) = \frac{1}{J}\Big[\max(\frac{-\lambda^{(j)}}{\rho}, f_j(x))\Big]_{j=1}^{J}$, the following holds:

$$- \Phi_{\rho_k}(x_k, \lambda_k) + \frac{1}{J}\sum_{j=1}^{J}\lambda^{(j)}f_j(x_k) + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda\|^2$$

$$= -\frac{1}{J}\sum_{j\in J_k^+}\frac{\rho_k}{2}(f_j(x_k))^2 + \frac{1}{J}\sum_{j\in J_k^-}\Big[\frac{(\lambda_k^{(j)})^2}{2\rho_k} + \lambda^{(j)}f_j(x_k) + (\lambda_k^{(j)} - \lambda^{(j)})(\frac{-\lambda_k^{(j)}}{\rho_k})\Big]$$

$$+ \frac{1}{2\rho_k}\|\lambda_k - \lambda\|^2 + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda_k\|^2 + (\lambda_k - \lambda)^T(\nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$

$$+ (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$

$$= -\frac{1}{J}\sum_{j\in J_k^+}\frac{\rho_k}{2}(f_j(x_k))^2 - \frac{1}{J}\sum_{j\in J_k^-}(\frac{(\lambda_k^{(j)})^2}{2\rho_k} - \lambda^{(j)}(f_j(x_k) + \frac{\lambda_k^{(j)}}{\rho_k})) \tag{24}$$

$$+ \frac{1}{2\rho_k}\|\lambda_k - \lambda\|^2 + \frac{1}{2\rho_k}\|\lambda_{k+1} - \lambda_k\|^2 + (\lambda_k - \lambda)^T(Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)).$$

Note that $\lambda \geq 0$ and by definition J_k^- it holds that $\lambda^{(j)}(f_j(x_k) + \frac{\lambda_k^{(j)}}{\rho_k}) \leq 0$, so we conclude that

$$-\frac{1}{J}\sum_{j\in J_k^+}\frac{\rho_k}{2}(f_j(x_k))^2 - \frac{1}{J}\sum_{j\in J_k^-}(\frac{(\lambda_k^{(j)})^2}{2\rho_k} - \lambda^{(j)}(f_j(x_k) + \frac{\lambda_k^{(j)}}{\rho_k})) \leq -\frac{1}{J}\sum_{j\in J_k^+}\frac{\rho_k}{2}(f_j(x_k))^2 - \frac{1}{J}\sum_{j\in J_k^-}\frac{(\lambda_k^{(j)})^2}{2\rho_k}. \tag{25}$$

Hence we have the desired result by putting (25) in (24). ∎

16

6.6 Proof of lemma 6

Proof. (a) From definition of $\nabla_\lambda \Phi_{\rho_k}$, using Assumption 1 (ii) and the fact that $\lambda_k^{(j_k)} \geq 0$ for all k and j, we have that $\|Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k)\|^2 = \left| \max\left(\frac{-\lambda_k^{(j_k)}}{\rho_k}, f_{j_k}(x_k) \right) \right|^2 \leq D_f^2$.

(b) By definition of $\bar{\sigma}_k$ and $\bar{v}_k$ and using Lemma 4, one can obtain the following.

$$(\lambda - \lambda_k \pm \bar{v}_k)^T (\nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k) - Je_{j_k} \odot \nabla_\lambda \Phi_{\rho_k}(x_k, \lambda_k))$$
$$= (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + (\lambda - \bar{v}_k)^T \bar{\sigma}_k \leq (\bar{v}_k - \lambda_k)^T \bar{\sigma}_k + \frac{1}{2\bar{\tau}_k}\|\lambda - \bar{v}_k\|^2 - \frac{1}{2\bar{\tau}_k}\|\lambda - \bar{v}_{k+1}\| + \frac{\bar{\tau}_k}{2}\|\bar{\sigma}_k\|^2.$$

∎

7 Acknowledgments

References

[1] A. AUSLENDER, *Optimisation*, Méthodes numériques, (1976).

[2] A. AUSLENDER AND M. HADDOU, *An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities*, Mathematical Programming, 71 (1995), pp. 77–100.

[3] A. AUSLENDER AND M. TEBOULLE, *Lagrangian duality and related multiplier methods for variational inequality problems*, SIAM Journal on Optimization, 10 (2000), pp. 1097–1115.

[4] R. S. BURACHIK AND A. N. IUSEM, *A generalized proximal point algorithm for the variational inequality problem in a hilbert space*, SIAM journal on Optimization, 8 (1998), pp. 197–216.

[5] A. CHAMBOLLE AND T. POCK, *On the ergodic convergence rates of a first-order primal–dual algorithm*, Mathematical Programming, 159 (2016), pp. 253–287.

[6] G. H.-G. CHEN AND R. T. ROCKAFELLAR, *Convergence rates in forward–backward splitting*, SIAM Journal on Optimization, 7 (1997), pp. 421–444.

[7] A. DELIGIANNIS, A. PANOUI, S. LAMBOTHARAN, AND J. A. CHAMBERS, *Game-theoretic power allocation and the nash equilibrium analysis for a multistatic mimo radar network*, IEEE Transactions on Signal Processing, 65 (2017), pp. 6397–6408.

[8] Y. DENG, M. M. KAMANI, AND M. MAHDAVI, *Distributionally robust federated averaging*, Advances in Neural Information Processing Systems, 33 (2020).

[9] J. ECKSTEIN AND M. C. FERRIS, *Smooth methods of multipliers for complementarity problems*, Mathematical Programming, 86 (1999), pp. 65–90.

[10] F. FACCHINEI AND C. KANZOW, *Generalized nash equilibrium problems*, Annals of Operations Research, 175 (2010), pp. 177–211.

[11] F. FACCHINEI AND J.-S. PANG, *Finite-dimensional Variational Inequalities and Complementarity Problems. Vols. I,II*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[12] M. C. FERRIS AND J.-S. PANG, *Engineering and economic applications of complementarity problems*, Siam Review, 39 (1997), pp. 669–713.

[13] D. GABAY, *Applications of the method of multipliers to variational inequalities*, vol. 15, Elsevier, 1983, ch. ix. In: Studies in mathematics and its applications, pp. 299–331.

[14] E. GOLSHTEIN, *Generalized gradient method for finding saddlepoints*, Matekon, 10 (1974), pp. 36–52.

[15] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, Advances in Neural Information Processing Systems, 27 (2014).

[16] E. Y. HAMEDANI AND N. S. AYBAT, *A primal-dual algorithm with line search for general convex-concave saddle point problems*, SIAM Journal on Optimization, 31 (2021), pp. 1299–1329.

[17] Y. HE AND R. D. MONTEIRO, *Accelerating block-decomposition first-order methods for solving composite saddle-point and two-player nash equilibrium problems*, SIAM Journal on Optimization, 25 (2015), pp. 2182–2211.

[18] ——, *An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM Journal on Optimization, 26 (2016), pp. 29–56.

[19] X. HU AND D. RALPH, *Using epecs to model bilevel games in restructured electricity markets with locational prices*, Operations research, 55 (2007), pp. 809–827.

[20] A. N. IUSEM, A. JOFRÉ, R. I. OLIVEIRA, AND P. THOMPSON, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM Journal on Optimization, 27 (2017), pp. 686–724.

[21] A. N. IUSEM, A. JOFRÉ, R. I. OLIVEIRA, AND P. THOMPSON, *Variance-based extragradient methods with line search for stochastic variational inequalities*, SIAM Journal on Optimization, 29 (2019), pp. 175–206.

[22] H. JIANG AND H. XU, *Stochastic approximation approaches to the stochastic variational inequality problem*, IEEE Transactions on Automatic Control, 53 (2008), pp. 1462–1475.

[23] Y. JIN, A. SIDFORD, AND K. TIAN, *Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods*, in Conference on Learning Theory, PMLR, 2022, pp. 4362–4415.

[24] A. JUDITSKY, A. NEMIROVSKI, AND C. TAUVEL, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stochastic Systems, 1 (2011), pp. 17–58.

[25] A. KANNAN AND U. V. SHANBHAG, *Distributed computation of equilibria in monotone Nash games via iterative regularization techniques*, SIAM Journal on Optimization, 22 (2012), pp. 1177–1205.

[26] S. KARAMARDIAN, *An existence theorem for the complementarity problem*, Journal of Optimization Theory and Applications, 19 (1976), pp. 227–232.

[27] H. D. KAUSHIK AND F. YOUSEFIAN, *A method with convergence rates for optimization problems with variational inequality constraints*, SIAM Journal on Optimization, 31 (2021), pp. 2171–2198.

[28] O. KOLOSSOSKI AND R. D. MONTEIRO, *An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems*, Optimization Methods and Software, 32 (2017), pp. 1244–1272.

[29] G. M. KORPELEVICH, *An extragradient method for finding saddle points and for other problems*, Eknomika i Matematicheskie Metody, 12 (1976), pp. 747—-756.

[30] J. KOSHAL, A. NEDIĆ, AND U. V. SHANBHAG, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Transactions on Automatic Control, 58 (2013), pp. 594–609.

[31] S. KRILAŠEVIĆ AND S. GRAMMATICO, *Learning generalized nash equilibria in monotone games: A hybrid adaptive extremum seeking control approach*, Automatica, 151 (2023), p. 110931.

[32] X.-J. LONG AND Y.-H. HE, *A fast stochastic approximation-based subgradient extragradient algorithm with variance reduction for solving stochastic variational inequality problems*, Journal of Computational and Applied Mathematics, 420 (2023), p. 114786.

[33] Y. MALITSKY, *Proximal extrapolated gradient methods for variational inequalities*, Optimization Methods and Software, 33 (2018), pp. 140–164. PMID: 29348705.

[34] G. J. MINTY ET AL., *Monotone (nonlinear) operators in hilbert space*, Duke Mathematical Journal, 29 (1962), pp. 341–346.

[35] U. MOSCO, *Dual variational inequalities*, Journal of Mathematical Analysis and Applications, 40 (1972), pp. 202–206.

[36] J. NASH, *Non-cooperative games*, Annals of mathematics, (1951), pp. 286–295.

[37] A. NEDIĆ AND A. OZDAGLAR, *Subgradient methods for saddle-point problems*, Journal of Optimization Theory and Applications, 142 (2009), pp. 205–228.

[38] 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 (2004), pp. 229–251.

[39] A. NEMIROVSKI, A. JUDITSKY, G. LAN, AND A. SHAPIRO, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.

[40] M. SANJABI, J. BA, M. RAZAVIYAYN, AND J. D. LEE, *On the convergence and robustness of training gans with regularized optimal transport*, Advances in Neural Information Processing Systems, 31 (2018).

[41] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM Journal on Applied Mathematics, 15 (1967), pp. 1328–1343.

[42] A. Sinha, H. Namkoong, and J. Duchi, *Certifiable distributional robustness with principled adversarial training*, in International Conference on Learning Representations, 2018.

[43] v. Neumann, *Zur theorie der gesellschaftsspiele*, Mathematische Annalens, 19 (1928), pp. 295–320.

[44] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior, 2nd rev*, (1947).

[45] Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM Journal on Optimization, 30 (2020), pp. 1664–1692.

[46] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems*, Mathematical Programming, 165 (2017), pp. 391–431.

[47] ———, *On stochastic mirror-prox algorithms for stochastic Cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes*, Set-Valued and Variational Analysis, 26 (2018), pp. 789–819.

[48] L. Zhang, D. Xu, S. Yuan, and X. Wu, *FairGAN: Fairness-aware generative adversarial networks*, in CoRR, 2018.

[49] R. Zhao, *Accelerated stochastic algorithms for convex-concave saddle-point problems*, Mathematics of Operations Research, 47 (2022), pp. 1443–1473.

[50] R. Zhao, W. B. Haskell, and V. Y. Tan, *An optimal algorithm for stochastic three-composite optimization*, in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 428–437.