Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games

Farzad Yousefian Angelia Nedić Uday V. Shanbhag

Abstract

We consider multi-user optimization problems and Nash games with stochastic convex objectives, instances of which arise in decentralized control problems. The associated equilibrium conditions of both problems can be cast as Cartesian stochastic variational inequality problems with mappings that are strongly monotone but not necessarily Lipschitz continuous. Consequently, most of the currently available stochastic approximation schemes cannot address such problems. First, through a user-specific local smoothing, we derive an approximate map that is shown to be Lipschitz continuous with a prescribed constant. Second, motivated by the need for a robust schemes that can be implemented in a distributed fashion, we develop a distributed self-tuned stochastic approximation scheme (DSSA) that adapts to problem parameters. Importantly, this scheme is provably convergent in an almost-sure sense and displays the optimal rate of convergence in mean-squared error, i.e., $\mathcal{O}\left(\frac{1}{k}\right)$. A locally randomized variant is also provided to ensure that the scheme can contend with stochastic non-Lipschitzian multi-user problems. We conclude with numerics derived from a stochastic Nash-Cournot game.

I. Introduction

In this paper, we consider the solution of a Cartesian stochastic variational inequality problem, described as follows. Given a set X and a mapping $\Phi(x, \xi)$, where ξ is a random vector, determine

The second author in in the Dept. of Indust. and Enterprise Sys. Engg., Univ. of Illinois, Urbana, IL 61801, USA, while the first and third authors are with the Department of Indust. and Manuf. Engg., Penn. State University, University Park, PA 16802, USA. They are contactable at {angelia}@illinois.edu and szy5,udaybag@psu.edu. Nedić and Shanbhag gratefully acknowledge the support of the NSF through awards NSF CMMI 0948905 ARRA (Nedić and Shanbhag), CMMI-0742538 (Nedić) and CMMI-1246887 (Shanbhag). A part of this paper has appeared in [1].

a vector $x^* \in X$ such that $(x - x^*)^T \mathsf{E}[\Phi(x^*, \xi)] \geq 0$ for all $x \in X$, where the mathematical expectation is taken with respect to the random vector ξ. Our interest is in the case when the set X is a Cartesian product of some component sets X_i and the stochastic mapping $\Phi(x, \xi)$ has a decomposable structure in ξ. Specifically, our focus is on the case when the mapping $\Phi(x, \xi)$ has the form $\Phi(x, \xi) = \prod_{i=1}^{N} \Phi_i(x, \xi_i)$, where ξ_i is a random vector for all $i = 1, \cdots, N$. The resulting Cartesian stochastic variational inequality problem reduces to determining a vector $x^* = (x_1^*; x_2^*; \ldots; x_N^*) \in X$ such that

$$(x_i - x_i^*)^T \mathsf{E}[\Phi_i(x_i^*, \xi_i)] \geq 0 \qquad \text{for all } x_i \in X_i \text{ and all } i = 1, \ldots, N. \tag{1}$$

The problem will be formalized in detail subsequently.

Our goal lies in developing distributed algorithms for solving such problems by utilizing the structure of the set X and the disturbances ξ_i. The problem (1) arises in many situations, typically modeling multi-user systems with either cooperative or competitive user objectives. In particular, the problems of this form appear across applications in power systems [2], [3], cognitive radio networks [4], communication networks [5]–[8], amongst others. More recently, Nash games are also finding increasing utilization in the context of developing distributed control laws (cf. [9]). In these settings, user objectives are often characterized by uncertainty and the gradient map is not necessarily Lipschitz. The resulting class of problems has seen less attention, motivating the development of distributed algorithms for contending precisely with such challenges.

A broad avenue for contending with stochastic variational inequalities is through stochastic approximation (SA) methods. Starting with the seminal work by Robbins and Monro [10] for root-finding problems and Ermoliev for stochastic programs [11], significant effort has been applied towards the examination of such schemes (cf. [12]–[14]). Amongst the first extensions to the regime of stochastic variational inequalities was provided by Jiang and Xu [15], while regularized variants were investigated in [16]. Prox-type methods were first proposed by Nemirovski [17] for solving VIs with monotone and Lipschitz continuous maps, prox-type methods have been developed to address different problems in convex optimization and variational inequalities (c.f. [18]–[20]). Recently, Juditsky et al. [19] introduced the stochastic Mirror-Prox (SMP) algorithm for solving stochastic VIs with non-Lipschitzian but bounded mappings.

The existing algorithms for stochastic variational inequalities have several limitations. First, to recover almost-sure convergence, all of the aforementioned work is done for maps that are

Lipschitzian. In [19], the SMP algorithm can accommodate non-Lipschitzian maps to show convergence of the gap function in mean. Second, the algorithms mostly employ steplength sequences that *do not adapt to problem parameters* leading to poor practical performance. Third, the convergence statements are often *limited to mean-square convergence*, while it is often desirable to have almost-sure convergence guarantees.

In this paper we address these limitations by developing distributed algorithms for problems of the form (1) arising from a class of convex multi-user optimization problems or a class of convex Nash-games (to be specified later). For such problems, we consider settings where the associated variational maps of the user objectives are not Lipschitz continuous. By smoothing each user-specific objective, we construct an approximate problem that leads to a map that can be shown to be Lipschitz continuous. This is done by building on our prior work [21] where we study a local-smoothing technique for convex non-differentiable functions. This work leverages a randomized smoothing technique. first introduced by Steklov [22] in 1907, later on used in optimization [23], [24], and more recently investigated in [21], [25], [26]. Then, for a class of strongly monotone maps, we use a distributed (i.e., user-specific) local smoothing and distributed adaptive stepsize rules to address the problem in (1). Standard SA schemes provide little guidance regarding the choice of a steplength sequence, while the behavior of SA schemes is closely tied to the choices. The standard implementations use the steplengths of the form θ/k where $\theta > 0$ and k denotes the iteration number. Generally, there have been two avenues traversed in choosing steplengths: (1) *Deterministic steplength sequences:* Spall [14, Ch. 4, pg. 113] considered diverse choices of the form $\gamma_k = \frac{\beta}{(k+1+a)^\alpha}$, where $\beta > 0$, $0.5 < \alpha \leq 1$, and $a \geq 0$ (also see Powell [27]). (2) *Stochastic steplength sequences:* An alternative to a deterministic rule is a stochastic scheme that updates steplengths based on observed data. Of note is recent work by George et al. [28] where an adaptive stepsize rule is proposed that minimizes the mean squared error. In a similar vein, Cicek et al. [29] develop an adaptive Kiefer-Wolfowitz SA algorithm and derive general upper bounds on its mean-squared error. Inspired by our efforts to solve nonsmooth stochastic optimization problems [21], we generalize this centralized scheme for optimization problems to a distributed version that can also cope with non-Lipschitzian Cartesian stochastic variational inequalities (CSVIs). To summarize, the main contributions of this paper are as follows:

(i) Locally randomized approximate map: By introducing a user-specific (distributed) smoothing, we derive an approximation of the original map with a prescribed Lipschitz constant. This

approximation forms the basis for constructing stochastic approximation schemes.

(ii) Distributed local smoothing SA scheme for non-Lipschitzian CSVIs: We generalize the local smoothing technique to address non-Lipschitzian CSVIs through the development of distributed smoothing counterparts of the SA schemes developed in (i). These schemes are shown to generate iterates that converge to solution of the approximate problem in an almost sure sense. We further show that the sequence of smoothed solutions converges to its true counterpart.

(iii) Distributed self-tuned SA schemes: We develop a distributed self-tuned steplength rule that allows for distributed implementation of the stochastic approximation scheme for the Cartesian variational inequality problem with Lipschitzian maps. The proposed steplength rule is a generalization of the stepsize developed in [21] for stochastic optimization to the CSVIs. We prove that this SA scheme is characterized by the optimal rate of convergence.

In contrast with our prior work [21], this work is characterized by several key distinctions. First, in this paper, we develop locally randomized SA schemes for computing approximate solutions of non-Lipschitzian Cartesian stochastic variational inequalities rather than stochastic convex optimization problems considered in [21]. Second, we develop distributed counterparts of the schemes presented in [21] and prove that they admit the optimal rate of $\mathcal{O}(1/k)$ in mean-squared error. Third, we show that the sequence of approximate solutions converge to the true solution as the smoothing parameter is driven to zero.

The paper is organized as follows. In Section II, we provide the formulation of the problem and motivate this formulation through an example. By introducing a user-specific local smoothing, we derive an approximation of the map in Section III and show that it is Lipschitz continuous with a prescribed constant. In Section IV we outline an SA algorithm and state assumptions on the problem. A distributed self-tuned steplength SA (DSSA) scheme for the Lipschitzian CSVI is provided in Section V. A distributed locally randomized SA scheme is provided in Section VI. We report some numerical results in Section VII and conclude in Section VIII.

Notation: Throughout this paper, a vector x is assumed to be a column vector and x^T denotes its transpose. $\|x\|$ denotes the Euclidean vector norm, i.e., $\|x\| = \sqrt{x^T x}$, $\|x\|_1$ denotes the 1-norm, i.e., $\|x\|_1 = \sum_{i=1}^n |x_i|$ for $x \in \mathbb{R}^n$, and $\|x\|_\infty$ to denote the infinity vector norm, i.e., $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$ for $x \in \mathbb{R}^n$. We use $\Pi_X(x)$ to denote the Euclidean projection of a vector x on a set X, i.e., $\Pi_X(x) = \min_{y \in X} \|x - y\|$. We write *a.s.* as the abbreviation for "almost surely". We use $\mathsf{E}[z]$ to denote the expectation of a random variable z. The mapping

$F : X \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is strongly monotone with parameter $\eta > 0$ if for any $x, y \in X$, we have $(F(x) - F(y))^T(x - y) \geq \eta\|x - y\|^2$. The mapping F is Lipschitz continuous with parameter $L > 0$ if for any $x, y \in X$, we have $\|F(x) - F(y)\| \leq L\|x - y\|$. The `Matlab` notation $(u_1; u_2; u_3)$ refers to a column vector obtained by stacking the column vectors u_1, u_2 and u_3 into a single column.

II. Source problems

We provide formal description of the Cartesian stochastic variational inequality (CSVI) problem in (1) and specify source problems of our interest in this paper. We let the underlying set $X \subseteq \mathbb{R}^n$ be given by the Cartesian product of N sets:

$$X \triangleq \prod_{i=1}^{N} X_i, \qquad \text{with } X_i \subseteq \mathbb{R}^{n_i}, \tag{2}$$

where $n = \sum_{i=1}^{N} n_i$. We are interested in determining a vector $x^* = (x_1^*; x_2^*; \ldots; x_N^*) \in X$ that satisfies the following system of inequalities:

$$(x_i - x_i^*)^T \mathsf{E}[\Phi_i(x^*, \xi_i)] \geq 0 \qquad \text{for all } x_i \in X_i \text{ and all } i = 1, \ldots, N, \tag{3}$$

where $\xi_i \in \mathbb{R}^{d_i}$ is a random vector and $\Phi_i(x, \xi_i) : \mathbb{R}^n \times \mathbb{R}^{d_i} \to \mathbb{R}^{n_i}$ for $i = 1, \ldots, N$. We let $x = (x_1; x_2; \ldots; x_N)$, and we assume that the expected maps $\mathsf{E}[\Phi_i(x, \xi_i)] : \mathbb{R}^n \to \mathbb{R}^{n_i}$ are well defined (i.e., the expectations are finite). Denote the resulting composite map by F, i.e.,

$$F(x) \triangleq (\mathsf{E}[\Phi_1(x, \xi_1)]; \ldots; \mathsf{E}[\Phi_N(x, \xi_N)]) \quad \text{for all } x \in X. \tag{4}$$

In this notation, the variational inequality problem in (3) will be abbreviated by $\mathrm{VI}(X, F)$. Our interest is in the variational inequalities arising from either one of the following problems:

a) Stochastic multi-user optimization: Consider N users where user i has its own (stochastic) convex objective function $f_i(x_i, \xi_i)$ and a convex closed constraint set X_i. The function depends on the user's decision variable $x_i \in \mathbb{R}^{n_i}$ and a random vector $\xi_i \in \Omega_i \subseteq \mathbb{R}^{d_i}$. The users are coupled through a convex cost function $c(x)$, giving rise to the following generalized multi-user optimization problem: $\min_{x \in X} \sum_{i=1}^{N} \mathbb{E}[f_i(x_i, \xi_i)] + c(x)$, where $c(x)$ is known by all users. It is assumed that user i can observe all the decisions of the other users, i.e., user i has access to $x_{-i} = \{x_j\}_{i \neq j}$. This problem is distributed in the sense that user i knows only f_i and X_i, and it can modify only its own decision variable x_i. Optimal solutions of this

problem are characterized by a suitably defined Cartesian stochastic VI(X, F), where $F(x) = \prod_{i=1}^{N} (\mathsf{E}[\nabla_{x_i} f_i(x_i, \xi_i)] + \nabla_{x_i} c(x))$.

b) Noncooperative convex Nash games: Consider a noncooperative counterpart of the preceding problem, where a user is referred to as a player. In the corresponding $N-$player noncooperative stochastic Nash game, player i solves the following parameterized stochastic convex problem:

$$\min_{x_i \in X_i} \quad \mathbb{E}[f_i(x, \xi_i)]. \tag{$P_i(x_{-i})$}$$

A solution to this N-player game is a Nash equilibrium point $x^* = \{x_i^*\}_{i=1}^N$, where each x_i^* is a solution to problem $(P_i(x_{-i}^*))$ for $i = 1, \ldots, N$. The ith player is characterized by an expected-value objective $\mathbb{E}[f_i(x; \xi_i)]$ where $f_i(x; \xi_i)$ is convex and continuous over $X_i \subseteq \mathbb{R}^{n_i}$, which is a closed and convex set. The set of Nash equilibria for this game are captured by the solutions of the Cartesian stochastic VI(X, F), where $F(x) = \prod_{i=1}^{N} \mathsf{E}[\nabla_{x_i} f_i(x, \xi_i)]$.

As an example of such a game, consider a networked stochastic Nash-Cournot game (see [30], [31]) where a collection of N firms compete at M different locations wherein the production and sales for firm i at location j are denoted by g_{ij} and s_{ij}, respectively. Let firm i's cost of production at location j be given by an uncertain cost function $c_{ij}(g_{ij}, \xi_{ij})$. Furthermore, let the goods sold by firm i at location j fetch a revenue given by $p_j(\bar{s}_j)s_{ij}$, where $p_j(\cdot)$ is a sale price at location j and $\bar{s}_j = \sum_{i=1}^{N} s_{ij}$ is the sale aggregate at location j. We assume that cost and price functions are merely (continuous) convex functions and may have kinks [32]. Finally, firm i's production at node j is capacitated by cap_{ij} and its optimization problem is given by $\min_{x_i \in X_i} \mathsf{E}[f_i(x, \xi_i)]$, where $x = (x_1; \ldots; x_N)$ with $x_i = (g_i; s_i)$, $g_i = (g_{i1}; \ldots; g_{iM})$, $s_i = (s_{i1}; \ldots; s_{iM})$, $\xi_i = (\xi_{i1}; \ldots; \xi_{iM})$, $f_i(x, \xi_i) \triangleq \sum_{j=1}^{M} (c_{ij}(g_{ij}, \xi) - p_j(\bar{s}_j, \xi)s_{ij})$, and

$$X_i \triangleq \left\{ (g_i, s_i) \mid \sum_{j=1}^{M} g_{ij} = \sum_{j=1}^{M} s_{ij}, \quad g_{ij},\ s_{ij} \geq 0, g_{ij} \leq \mathrm{cap}_{ij},\ j = 1, \ldots, M \right\}.$$

Note that transportation costs are assumed to be zero. Under the validity of the interchange between the expectation and the derivative operator, the resulting equilibrium conditions are compactly captured by VI(X, F), where $X \triangleq \prod_{i=1}^{N} X_i$ and $F(x) = \prod_{i=1}^{N} \mathsf{E}[\nabla_{x_i} f_i(x, \xi_i)]$.

Next, we present two instances of Nash–Cournot competition in practical settings.

i) Imperfectly competitive power markets: Cournot models have been extensively to model competition between generation firms [30]. Here, agents $j = 1, \ldots, M$ are generation firms

while load is distributed over the network of N nodes. Generator j aims to determine its profit-maximizing generation level at node i, given by g_{ij}, and the level of sales s_{ij} at node j. The constraint $\sum_{j=1}^{M} g_{ij} = \sum_{j=1}^{M} s_{ij}$ imposes an energy balance for generator j while prices at node j are denoted by $p_j(S_j) = a_j - b_j S_j$ where S_j is the aggregate sales at node j.

ii) Imperfect competition in rate control over communication networks: Competitive models have assumed relevance in rate control in communication networks [7]. In such models, player payoffs, characterized by idiosyncratic utility functions, are coupled through a network-specific congestion cost. The associated Nash games are complicated by the coupling of strategy sets through a set of shared network constraints. An extension of this model, presented in [33], imposes a scaled congestion cost, akin to a Cournot-based framework. In a single-link version, there are N players, characterized by utility functions $U_1(x_1), \ldots, U_N(x_N)$. Consequently, player i's payoff function are given by its utility function less scaled congestion cost i.e. $U_i(x_i) - x_i c(X)$ where $X = \sum_{i=1}^{N} x_i$ while $\sum_{i=1}^{N} x_i \leq c$, where c denotes the link capacity.

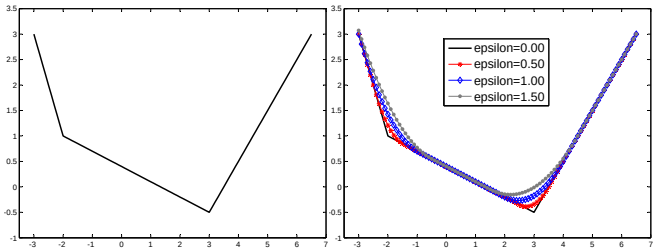
III. A DISTRIBUTED LOCAL SMOOTHING TECHNIQUE

In the development of SA schemes for VI(X, F) with the set X and the map F defined by (2) and (4), respectively, we first approximate the map F with a smoothed map denoted by F^{ϵ}. Then, we use stochastic approximation to solve VI(X, F^{ϵ}) and, thus, obtain an approximate solution to VI(X, F). Our approximation of the map F relies on a "local randomization" technique and its generalizations, as presented in this section. The next proposition (see [23]) presents a smoothing of a nondifferentiable convex function.

Proposition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and consider $\mathsf{E}[f(x-\omega)]$, where ω belongs to the probability space $(\mathbb{R}^n, B_n, P)$, B_n is the $\sigma-$algebra of Borel sets of $\mathbb{R}^n$ and P is a probability measure on B_n. Then, if $\mathsf{E}[f(x-\omega)] < \infty$ for all $x \in \mathbb{R}^n$, the function $\mathsf{E}[f(x-\omega)]$ is everywhere differentiable.*

This technique [21], [25], [34] transforms f into a smooth function. In [25], a Gaussian distribution is used for the smoothing, leading to a smoothed function has Lipschitz gradients, with a prescribed Lipschitz constant, when the original function has bounded subgradients. A challenge in that approach is that, in some situations, the original function f may have a restricted domain and not be defined for some realizations of the Gaussian random variable. In the following

example, we demonstrate how the smoothing technique works for a piecewise linear function.



(a) The original function $f(x)$ (b) Different smooth. parameters

Fig. 1: The smoothing technique

Example 1 (Smoothing of a convex function)**.** Consider the following piecewise linear function $f(x)$ is $-2x-3$ for $x<-2$, $-0.3x+0.4$ for$-2\leq x<3$, and $x-3.5$ for $x\geq 3$. Let z be a uniform random variable defined on $[-\epsilon,\epsilon]$ where $\epsilon>0$ is a given parameter. Consider the approximation $f^{\epsilon}(x)=\mathsf{E}[f(x+z)]$. Prop. 1 implies that f^{ϵ} is a smooth function. A function f that is nonsmooth at $x=-2$ and $x=3$ is shown in Fig. 1a, while Fig. 1b shows the smoothed functions f^{ϵ} for different values of ϵ and illustrates the exactness of the approximation as $\epsilon\to 0$.

Motivated by this smoothing technique, we introduce a distributed smoothing scheme where we simultaneously perturb the value of vectors x_i with a random vector z_i for all i. In this scheme, referred to as *a multi-cubic randomized (MCR) scheme*, we let $z_i\in\mathbb{R}^{n_i}$ be uniformly distributed on the n_i-dimensional cube centered at the origin with an edge length of $2\epsilon_i$.

Now, consider a map F from the set X to $\mathbb{R}^n$. Define an approximation $F^{\epsilon}:X\to\mathbb{R}^n$ as the expectation of $F(x+z)$ where x is perturbed by a random vector $z=(z_1;\ldots;z_N)$. Specifically, when $F_1,\ldots,F_N$ are coordinate-maps of F, F^{ϵ} is given by

$$F^{\epsilon}(x)\triangleq(\mathsf{E}[F_1(x+z)]\,;\ldots;\mathsf{E}[F_N(x+z)])\qquad\text{for all } x\in X. \tag{5}$$

Let us define $C_n(x,\rho)\subset\mathbb{R}^n$ as a cube centered at a point x with the edge length $2\rho>0$ where the edges are along the coordinate axes. More precisely, $C_n(x,\rho)\triangleq\{y\in\mathbb{R}^n\mid\|y-x\|_\infty\leq\rho\}$. In the MCR scheme, we assume that for any $i=1,\ldots,N$, the random vector z_i is uniformly distributed on the set $C_{n_i}(0,\epsilon_i)$ and is independent of the other random vectors z_j for $j\neq i$. For the mapping F^{ϵ} to be well defined, we will assume that F is defined over the set X^{ϵ} given by $X^{\epsilon}\triangleq X+\prod_{i=1}^{N}C_{n_i}(0,\epsilon_i)$, where $\epsilon_i>0$ and $\epsilon\triangleq(\epsilon_1,\ldots,\epsilon_N)$. The following Lemma provides a relation used in establishing the main property of the density function of the MCR scheme.

Lemma 1. *Let the vector $p \in \mathbb{R}^m$ be such that $0 \leq p_i \leq 1$ for all $i = 1, \ldots, m$. Then, we have $1 - \prod_{i=1}^m (1 - p_i) \leq \|p\|_1$.*

Proof. We use induction over m to prove this result. For $m = 1$, we have $1-(1-p_1) = p_1 = \|p\|_1$, implying that the result holds for $m = 1$. Assume $1 - \prod_{i=1}^m (1 - p_i) \leq \|p\|_1$ holds for some $m \geq 1$, i.e., $\prod_{i=1}^m (1 - p_i) \geq 1 - \sum_{i=1}^m p_i$. Multiplying both sides of the preceding relation by $(1 - p_{m+1})$, we obtain

$$\prod_{i=1}^{m+1} (1 - p_i) \geq (1 - \sum_{i=1}^{m} p_i)(1 - p_{m+1}) = 1 - \sum_{i=1}^{m+1} p_i + p_{m+1} \sum_{i=1}^{m} p_i \geq 1 - \sum_{i=1}^{m+1} p_i.$$

Hence, $\prod_{i=1}^{m+1} (1 - p_i) \geq 1 - \sum_{i=1}^{m+1} p_i$ which implies that the result holds for $m + 1$. □

The following result is crucial for establishing the properties of the approximation F^ϵ.

Lemma 2. *Let $z \in \mathbb{R}^n$ be a random vector with the uniform density over an n-dimensional cube $\prod_{i=1}^N C_{n_i}(0, \epsilon_i)$ with $\epsilon_i > 0$ for all i. Let $p_c : \mathbb{R}^n \to \mathbb{R}$ be the probability density function of the random vector z. Then, the following relation holds for all $x, y \in \mathbb{R}^n$:*

$$\int_{\mathbb{R}^n} |p_c(u - x) - p_c(u - y)| du \leq \frac{\sqrt{n}}{\min\limits_{1 \leq i \leq N} \{\epsilon_i\}} \|x - y\|.$$

Proof. The proof is omitted and can be found in the extended version of this paper [35]. □

We make use of the fooling assumption in our analysis.

Assumption 1. *Let the map $F = (F_1, \ldots, F_N)$ have bounded coordinate maps F_i on X^ϵ for an $\epsilon = (\epsilon_1, \ldots, \epsilon_N) > 0$, i.e., $\|F_i(x)\| \leq C_i$ for all $x \in X^\epsilon$, and define $C \triangleq (C_1, \ldots, C_N)$.*

The next proposition derives continuity and monotonicity properties of the approximation F^ϵ.

Proposition 2 (Properties of F^ϵ under the MCR scheme)**.** *Let Assumption 1 hold. Then, the approximate map F^ϵ defined in (5) has the following properties:*

(a) *F^ϵ is bounded on the set X, i.e., $\|F^\epsilon(x)\| \leq \|C\|$ for all $x \in X$.*

(b) *F^ϵ is Lipschitz continuous over the set X, i.e., for any $x, y \in X$,*

$$\|F^\epsilon(x) - F^\epsilon(y)\| \leq \frac{\sqrt{n}\|C\|}{\min_{j=1,\ldots,N}\{\epsilon_j\}} \|x - y\|. \tag{6}$$

(c) *Suppose that mapping F is strongly monotone over X^ϵ with parameter $\eta > 0$. Then F^ϵ is strongly monotone over the set X with constant η.*

Proof. (a) Note that $\|F^\epsilon(x)\| \leq \sqrt{\sum_{i=1}^N \mathsf{E}[\|F_i(x+z)\|^2]} \leq \sqrt{\sum_{i=1}^N C_i^2}$, where the first inequality follows from Jensen's inequality and the result is due to the boundedness property imposed on F by Assumption 1.

(b) Since the random vector z_i is uniformly distributed on the set $C_{n_i}(0,\epsilon_i)$ for each i, the vector $z=(z_1;\ldots;z_N)$ is uniformly distributed on the set $\prod_{i=1}^N C_{n_i}(0,\epsilon_i)$. By the definition of the approximation F^ϵ in (5), it follows that for any $x,y\in X$, $\|F^\epsilon(x)-F^\epsilon(y)\|$ equals to

$$\left\|\int F(x+z)p_c(z)dz - \int F(y+z)p_c(z)dz\right\| = \left\|\int F(u)p_c(u-x)du - \int F(v)p_c(v-y)dv\right\|$$
$$= \left\|\int_{\mathbb{R}^n} F(u)(p_c(u-x)-p_c(u-y))du\right\| \leq \int_{\mathbb{R}^n} \|F(u)\||p_c(u-x)-p_c(u-y)|du,$$

where in the second equality $u=x+z$ and $v=y+z$, while the third inequality follows from the triangle inequality. Invoking Assumption 1 we obtain

$$\|F^\epsilon(x)-F^\epsilon(y)\| \leq \|C\| \int_{\mathbb{R}^n} |p_c(u-x)-p_c(u-y)|du.$$

The desired relation follows from invoking Lemma 2.

(c) It is easily verifiable using the definitions of strong monotonicity and the smoothed map. □

IV. Algorithm outline and assumptions

The focus of this paper is on the development of SA schemes for VI(X,F) with the set X and the map F defined by (2) and (4), respectively. When F is Lipschitz, we consider the distributed SA scheme given by the following:

$$\begin{aligned} x_{k+1,i} &:= \Pi_{X_i}\left(x_{k,i} - \gamma_{k,i}(F_i(x_k)+w_{k,i})\right), \\ w_{k,i} &\triangleq \Phi_i(x_k,\xi_{k,i}) - F_i(x_k), \end{aligned} \tag{7}$$

for all $k\geq 0$ and $i=1,\ldots,N$, where $F_i(x)\triangleq \mathsf{E}[\Phi_i(x,\xi_i)]$ for $i=1,\ldots,N$, $\gamma_{k,i}>0$ is the stepsize for the ith user at iteration k, $x_{k,i}$ denotes the solution for the i-th user at iteration k, and $x_k=(x_{k,1};\ x_{k,2};\ \ldots;\ x_{k,N})$. Moreover, $x_0\in X$ is a random initial vector independent of any other random variables in the scheme and such that $\mathsf{E}[\|x_0\|^2]<\infty$. Regarding (7), we let $\mathcal{F}_k$ denote the history of the method up to time k, i.e., $\mathcal{F}_k=\{x_0,\xi_0,\xi_1,\ldots,\xi_{k-1}\}$ for $k\geq 1$ and $\mathcal{F}_0=\{x_0\}$, where $\xi_k=(\xi_{k,1}\ ;\xi_{k,2}\ ;\ldots\ ;\xi_{k,N})$. Algorithm (7) is distributed in the sense that at any iteration, user i knows F_i and the set X_i, and may observe x, while controlling its own decision variable x_i and stepsize $\gamma_{k,i}$. Note that while algorithm (7) may be rewritten as

$x_{k+1,i} = \Pi_{X_i}\left(x_{k,i} - \gamma_{k,i}\Phi_i(x_k, \xi_{k,i})\right)$, we use (7) to aid in the analysis. We use the following lemma in establishing the convergence of method (7) and its extensions. This result may be found in [36] (cf. Lemma 10, page 49).

Lemma 3. *Let $\{v_k\}$ be a sequence of nonnegative random variables, where $\mathsf{E}[v_0] < \infty$, and let $\{\alpha_k\}$ and $\{\mu_k\}$ be deterministic scalar sequences such that for all $k \geq 0$, we have $0 \leq \alpha_k \leq 1, \mu_k \geq 0$ and $\mathsf{E}[v_{k+1}|v_0, \ldots, v_k] \leq (1-\alpha_k)v_k + \mu_k$ almost surely, and $\sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \mu_k < \infty, \lim_{k\to\infty} \frac{\mu_k}{\alpha_k} = 0$. Then, $v_k \to 0$ almost surely.*

We make the following main assumptions.

Assumption 2. *The sets X_i and the map F satisfy the following:*

(a) *The set $X_i \subseteq \mathbb{R}^{n_i}$ is closed and convex for $i = 1, \ldots, N$.*

(b) *$F(x)$ is strongly monotone with a constant $\eta > 0$.*

Remark 1. *Note that strong monotonicity is observed to hold in a range of practical problems including Nash-Cournot games [30], [31], and congestion control problems in communication networks [7]. Moreover, the strong monotonicity assumption can be weakened by using a regularization [37]. It is known that if mapping $F : X \to \mathbb{R}^n$, defined on a closed and convex set $X \subset \mathbb{R}^n$, is monotone, then for any scalar $\eta > 0$, the mapping $F + \eta \boldsymbol{I}$ is a strongly monotone mapping with the parameter η (cf. [38], Chapter 12). However, given our interest in resolving the challenges arising from the absence of Lipschitz continuity in the map and the need for providing distributed adaptive steplength rules, we impose the strong monotonicity assumption.*

V. Distributed self-tuned SA (DSSA) schemes for Lipschitzian mappings

In this section, we restrict our attention to settings where the mapping $F(x)$ is a Lipschitzian map with constant $L > 0$. If not, we may apply the smoothing scheme described in Section III and construct an approximate problem where the mapping is Lipschitz with the constant given by Lemma 2. A key challenge in practical implementations of stochastic approximation lies in choosing an appropriate diminishing steplength sequence $\{\gamma_k\}$. In [21], we developed a stepsize sequence rule in a convex stochastic optimization regime by leveraging three parameters: (i) Lipschitz constant of the gradients; (ii) strong convexity constant; and (ii) diameter of the set X. Note that a function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is said to be strongly convex with parameter

$\eta > 0$, if $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y) - \frac{\eta}{2}t(1-t)\|x-y\|^2$, for any x, y in the domain and $t \in [0,1]$. Along similar directions, such a rule can be constructed for strongly monotone stochastic variational inequality problem. Unfortunately, in distributed regimes, such a rule requires prescription by a central coordinator, a relatively challenging task in multi-agent regimes. This motivates the development of a distributed counterpart of the aforementioned adaptive rule. We present such a generalization with convergence theory in Section V-A and prove that such a scheme displays the optimal rate of convergence in Section V-B.

A. A distributed self-tuned steplength SA (DSSA) scheme

Given that the set X is the Cartesian product of closed and convex sets $X_1, \ldots, X_N$, our interest lies in developing steplength update rules in the context of method (7) where the i-th user chooses steplength sequence $\{\gamma_{k,i}\}$, assumed to satisfy the following requirements.

Assumption 3. *The steplength sequences $\{\gamma_{k,i}\}$ satisfy the following:*

(a) *The stepsize sequences $\{\gamma_{k,i}\}$, $i = 1, \ldots, N$, are such that $\gamma_{k,i} > 0$ for all k and i, with $\sum_{k=0}^{\infty} \gamma_{k,i} = \infty$ and $\sum_{k=0}^{\infty} \gamma_{k,i}^2 < \infty$ for all i.*

(b) *If $\{\delta_k\}$ and $\{\Gamma_k\}$ are positive sequences such that $\delta_k \leq \min_{1\leq i\leq N} \gamma_{k,i}$ and $\Gamma_k \geq \max_{1\leq i\leq N} \gamma_{k,i}$ for all $k \geq 0$, then $\frac{\Gamma_k - \delta_k}{\delta_k} \leq \beta$ for all $k \geq 0$, where β is a scalar satisfying $0 \leq \beta < \frac{\eta}{L}$.*

Remark 2. *Assumption 3a is a standard requirement on steplength sequences while Assumption (3b) provides an additional condition on the discrepancy between the stepsize values $\gamma_{k,i}$ at each k. This condition is satisfied, for instance, when $\gamma_{k,1} = \ldots = \gamma_{k,N}$, in which case $\beta = 0$. We impose some further conditions on the stochastic errors $w_{k,i}$, as follows.*

Assumption 4. *The errors $w_k = (w_{k,1}; w_{k,2}; \ldots; w_{k,N})$ are such that $\mathsf{E}[w_k \mid \mathcal{F}_k] = 0$ almost surely for all k and for some $\nu > 0$, $\mathsf{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu^2$ for all $k \geq 0$ almost surely.*

We use the following result in deriving an adaptive rule.

Lemma 4. *Let Assumptions 2 and 4 hold.*

(a) *If $\{\delta_k\}$ and $\{\Gamma_k\}$ are defined by Assumption 3, then The following holds a.s. for all $k \geq 0$:*

$$\mathsf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \leq \Gamma_k^2 \nu^2 + (1 - 2(\eta + L)\delta_k + 2L\Gamma_k + L^2\Gamma_k^2)\|x_k - x^*\|^2.$$

(b) *If Assumption (3b) holds, then for all $k \geq 0$:*

$$\mathsf{E}\left[\|x_{k+1} - x^*\|^2\right] \leq (1+\beta)^2\delta_k^2\nu^2 + (1 - 2(\eta - \beta L)\delta_k + (1+\beta)^2L^2\delta_k^2)\mathsf{E}\left[\|x_k - x^*\|^2\right].$$

Proof. (a) From the properties of the projection operator, we know that a vector x^* solves VI(X, F) problem if and only if x^* satisfies $x^* = \Pi_X(x^* - \gamma F(x^*))$ for any $\gamma > 0$. From (7) and the non-expansiveness property of the projection operator, we have for all $k \geq 0$ and all i,

$$\begin{aligned}\|x_{k+1,i} - x_i^*\|^2 &= \|\Pi_{X_i}(x_{k,i} - \gamma_{k,i}(F_i(x_k) + w_{k,i})) - \Pi_{X_i}(x_i^* - \gamma_{k,i}F_i(x^*))\|^2 \\ &\leq \|x_{k,i} - x_i^* - \gamma_{k,i}(F_i(x_k) + w_{k,i} - F_i(x^*))\|^2.\end{aligned}$$

Taking expectations conditioned on the past, and using $\mathsf{E}[w_{k,i} \mid \mathcal{F}_k] = 0$, we have

$$\begin{aligned}\mathsf{E}\left[\|x_{k+1,i} - x_i^*\|^2 \mid \mathcal{F}_k\right] \leq \|x_{k,i} - x_i^*\|^2 + \gamma_{k,i}^2\|F_i(x_k) - F_i(x^*)\|^2 + \gamma_{k,i}^2\mathsf{E}\left[\|w_{k,i}\|^2 \mid \mathcal{F}_k\right] \\ - 2\gamma_{k,i}(x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*)).\end{aligned}$$

Now, by summing the preceding relations over i, we have

$$\begin{aligned}\mathsf{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x^*\|^2 + \sum_{i=1}^N \gamma_{k,i}^2\|F_i(x_k) - F_i(x^*)\|^2 + \sum_{i=1}^N \gamma_{k,i}^2\mathsf{E}\left[\|w_{k,i}\|^2 \mid \mathcal{F}_k\right] \\ - 2\sum_{i=1}^N \gamma_{k,i}(x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*)).\end{aligned}$$

Using $\gamma_{k,i} \leq \Gamma_k$ and Assumption 4, we can see that $\sum_{i=1}^N \gamma_{k,i}^2\mathsf{E}[\|w_{k,i}\|^2 \mid \mathcal{F}_k] \leq \Gamma_k^2\nu^2$ almost surely for all $k \geq 0$. Thus, from the preceding relation, we have

$$\begin{aligned}\mathsf{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x^*\|^2 + \Gamma_k^2\nu^2 + \underbrace{\sum_{i=1}^N \gamma_{k,i}^2\|F_i(x_k) - F_i(x^*)\|^2}_{\textbf{Term 1}} \\ \underbrace{-2\sum_{i=1}^N \gamma_{k,i}(x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*))}_{\textbf{Term 2}}. \quad (8)\end{aligned}$$

Using the Lipschitzian property of the mapping F, we obtain

$$\text{Term 1} \leq \Gamma_k^2\|F(x_k) - F(x^*)\|^2 \leq \Gamma_k^2L^2\|x_k - x^*\|^2. \quad (9)$$

By adding and subtracting $-2\sum_{i=1}^N \delta_k(x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*))$ from Term 2, and using $\sum_{i=1}^N (x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*)) = (x_k - x^*)^T(F(x_k) - F(x^*))$, we further obtain

$$\text{Term 2} \leq -2\delta_k(x_k - x^*)^T(F(x_k) - F(x^*)) - 2\sum_{i=1}^N (\gamma_{k,i} - \delta_k)(x_{k,i} - x_i^*)^T(F_i(x_k) - F_i(x^*)).$$

By Cauchy-Schwartz inequality, the preceding relation yields

$$\begin{aligned}\text{Term 2} \leq & -2\delta_k(x_k - x^*)^T(F(x_k) - F(x^*)) + 2(\gamma_{k,i} - \delta_k)\sum_{i=1}^{N}\|x_{k,i} - x_i^*\|\|F_i(x_k) - F_i(x^*)\| \\ \leq & -2\delta_k(x_k - x^*)^T(F(x_k) - F(x^*)) + 2(\Gamma_k - \delta_k)\|x_k - x^*\|\|F(x_k) - F(x^*)\|,\end{aligned}$$

where in the last relation, we use the definition of Γ_k and Hölder's inequality. Invoking strong monotonicity of the mapping F for bounding the first term we have

$$\text{Term 2} \leq -2\eta\delta_k\|x_k - x^*\|^2 + 2(\Gamma_k - \delta_k)L\|x_k - x^*\|^2. \tag{10}$$

The desired inequality is obtained by combining relations (8), (9), and (10).

(b) Assumption 3b implies that $\Gamma_k \leq (1+\beta)\delta_k$. Therefore, from part (a) we obtain a.s.,

$$\mathsf{E}\big[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\big] \leq (1+\beta)^2\delta_k^2\nu^2 + (1 - 2(\eta - \beta L)\delta_k + (1+\beta)^2L^2\delta_k^2)\|x_k - x^*\|^2.$$

Taking expectations in the preceding inequality, we obtain the desired relation. □

The following proposition proves the almost-sure convergence of the distributed SA scheme when the steplength sequences satisfy the bounds prescribed by Assumption 3b.

Proposition 3 (Almost-sure convergence)**.** *Let Assumptions 2, 3, and 4 hold. Then, the sequence* $\{x_k\}$ *generated by algorithm* (7) *converges almost surely to the unique solution of VI*(X, F).

Proof. Consider the relation of Lemma 4(a). For this relation, we show that the conditions of Lemma 3 are satisfied, which will allow us to claim the almost-sure convergence of x_k to x^*. Let us define $v_k \triangleq \|x_k - x^*\|^2$, and

$$\alpha_k \triangleq 2(\eta - \beta L)\delta_k - L^2\delta_k^2(1+\beta)^2, \qquad \mu_k \triangleq (1+\beta)^2\delta_k^2\nu^2. \tag{11}$$

Next, we show that $0 \leq \alpha_k \leq 1$ for k sufficiently large. Since $\gamma_{k,i}$ tends to zero for all $i = 1, \ldots, N$, we may conclude that δ_k goes to zero as k grows. In turn, as δ_k goes to zero, for k large enough, say $k \geq k_1$, we have $1 - \frac{(1+\beta)^2L^2\delta_k}{2(\eta-\beta L)} > 0$. By Assumption 3b we have $\beta < \frac{\eta}{L}$, which implies $\eta - \beta L > 0$. Thus, we have $\alpha_k \geq 0$ for $k \geq k_1$. Also, for k large enough, say $k \geq k_2$, we have $\alpha_k \leq 1$ since $\delta_k \to 0$. Therefore, when $k \geq \max\{k_1, k_2\}$ we have $0 \leq \alpha_k \leq 1$. Obviously, $v_k, \mu_k \geq 0$. From Assumption 3b we have $\delta_k \leq \gamma_k \leq (1+\beta)\delta_k$ for all k. Using these relations and the conditions on $\gamma_{k,i}$ given in Assumption 3a, we can show that $\sum_{k=0}^{\infty}\delta_k = \infty$ and $\sum_{k=0}^{\infty}\delta_k^2 < \infty$. Furthermore, from the preceding properties of the sequence $\{\delta_k\}$, and the

definitions of α_k and μ_k in (11), we can see that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \mu_k < \infty$. Finally, the definitions of α_k and μ_k imply that $\lim_{k\to\infty} \frac{\mu_k}{\alpha_k} = 0$ since $\delta_k \to 0$. Hence, all conditions of Lemma 3 are satisfied and we may conclude that $\|x_k - x^*\|^2 \to 0$ almost surely. □

Our goal in the remainder of this section lies in providing a stepsize rule that aims to minimize a suitably defined error function of the algorithm, while satisfying Assumption 3. Consider the result of Lemma 4b for all $k \geq 0$: When the stepsizes $\gamma_{k,i}$ are further restricted so that $0 < \delta_k \leq \frac{\eta - \beta L}{(1+\beta)^2 L^2}$, we have $1 - 2(\eta - \beta L)\delta_k + (1+\beta)^2 L^2 \delta_k^2 \leq 1 - (\eta - \beta L)\delta_k$. Thus, for $0 < \delta_k \leq \frac{\eta - \beta L}{(1+\beta)^2 L^2}$, we obtain

$$\mathsf{E}\left[\|x_{k+1} - x^*\|^2\right] \leq (1+\beta)^2 \delta_k^2 \nu^2 + (1 - (\eta - \beta L)\delta_k)\mathsf{E}\left[\|x_k - x^*\|^2\right]. \tag{12}$$

Let us view the quantity $\mathsf{E}[\|x_{k+1} - x^*\|^2]$ as an error e_{k+1} of the method arising from the use of the lower bounds $\delta_0, \delta_1, \ldots, \delta_k$ for the stepsize values $\gamma_{0,i}, \gamma_{1,i} \cdots, \gamma_{k,i}$, $i = 1, \ldots, N$. Relation (12) gives us an error estimate for algorithm (7) in terms of these lower bounds. We proceed to derive a rule for generating lower bounds $\delta_0, \delta_1, \ldots, \delta_K$ by minimizing the error e_{K+1}. We define the following real-valued error function by considering (12):

$$e_{k+1}(\delta_0, \ldots, \delta_k) \triangleq (1 - (\eta - \beta L)\delta_k) e_k(\delta_0, \ldots, \delta_{k-1}) + (1+\beta^2)\nu^2 \delta_k^2 \qquad \text{for all } k \geq 0, \tag{13}$$

where e_0 is a positive scalar, $\{\delta_k\}$ is a sequence of positive scalars such that $0 < \delta_k \leq \frac{\eta - \beta L}{(1+\beta)^2 L^2}$, L is the Lipschitz constant of the mapping F, η is the strong monotonicity parameter of F, and ν^2 is the upper bound for the second moment of the error norms $\|w_k\|$ (cf. Assumption 4).

Now let us consider the stepsize sequence $\{\delta_k^*\}$ given by

$$\delta_0^* = \frac{\eta - \beta L}{2(1+\beta)^2 \nu^2} e_0 \tag{14}$$

$$\delta_k^* = \delta_{k-1}^* \left(1 - \left(\frac{\eta - \beta L}{2}\right) \delta_{k-1}^*\right) \qquad \text{for all } k \geq 1, \tag{15}$$

where e_0 is the same initial error as for the errors e_k in (13). In what follows, we often abbreviate $e_k(\delta_0, \ldots, \delta_{k-1})$ by e_k whenever this is unambiguous. The next proposition shows that the lower bound sequence $\{\delta_k^*\}$ for $\gamma_{k,i}$ given by (14)–(15) minimizes the errors e_k over $[0, \frac{\eta - \beta L}{(1+\beta)^2 L^2}]^k$.

Proposition 4 (A self-tuned lower bound steplength SA scheme)**.** *Let $e_k(\delta_0, \ldots, \delta_{k-1})$ be defined as in (13), where e_0 is a given positive scalar, ν is an upper bound defined in Assumption 4, η and L are the strong monotonicity and Lipschitz constants of the mapping F respectively and*

ν is chosen such that $\nu \geq L\sqrt{\frac{e_0}{2}}$. Let β be a scalar such that $0 \leq \beta < \frac{\eta}{L}$, and let the sequence $\{\delta_k^*\}$ be given by (14)–(15). Then, the following hold:

(a) *For all $k \geq 0$, the error e_k satisfies $e_k(\delta_0^*, \ldots, \delta_k^*) = \frac{2(1+\beta)^2\nu^2}{\eta-\beta L}\,\delta_k^*$.*

(b) *For any $k \geq 1$, the vector $(\delta_0^*, \delta_1^*, \ldots, \delta_{k-1}^*)$ is the minimizer of the function $e_k(\delta_0, \ldots, \delta_{k-1})$ over the set $\mathbb{G}_k \triangleq \left\{\alpha \in \mathbb{R}^k : 0 < \alpha_j \leq \frac{\eta-\beta L}{(1+\beta)^2L^2}, j = 1, \ldots, k\right\}$. More precisely, for any $(\delta_0, \ldots, \delta_{k-1}) \in \mathbb{G}_k$, we have*

$$e_k(\delta_0, \ldots, \delta_{k-1}) - e_k(\delta_0^*, \ldots, \delta_{k-1}^*) \geq (1+\beta)^2\nu^2(\delta_{k-1} - \delta_{k-1}^*)^2.$$

Proof. (a) To show the result, we use induction on k. Trivially, it holds for $k = 0$ from (14). Now, suppose we have $e_k(\delta_0^*, \ldots, \delta_{k-1}^*) = \frac{2(1+\beta)^2\nu^2}{\eta-\beta L}\,\delta_k^*$ for some k, and consider the case for $k+1$. From the definition of e_k in (13), we obtain

$$e_{k+1}(\delta_0^*, \ldots, \delta_k^*) = (1 - (\eta - \beta L)\delta_k^*)\frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\,\delta_k^* + (1+\beta)^2\nu^2(\delta_k^*)^2 = \frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\,\delta_{k+1}^*,$$

where the last equality follows by (15). Hence, the result holds for all $k \geq 0$.

(b) First we need to show that $(\delta_0^*, \ldots, \delta_{k-1}^*) \in \mathbb{G}_k$. By our assumption on e_0, we have $0 < e_0 \leq \frac{2\nu^2}{L^2}$, which by the definition of δ_0^* in (14) implies that $0 < \delta_0^* \leq \frac{\eta-\beta L}{(1+\beta)^2L^2}$, i.e., $\delta_0^* \in \mathbb{G}_1$. Using the induction on k, from relations (14)–(15), it can be shown that $0 < \delta_k^* < \delta_{k-1}^*$ for all $k \geq 1$. Thus, $(\delta_0^*, \ldots, \delta_{k-1}^*) \in \mathbb{G}_k$ for all $k \geq 1$. Using the induction on k again, we now show that the vector $(\delta_0^*, \delta_1^*, \ldots, \delta_{k-1}^*)$ minimizes the error $e_k(\delta_0, \ldots, \delta_{k-1})$ for all $k \geq 1$. From the definition of the error e_1 and the relation $e_1(\delta_0^*) = \frac{2(1+\beta)^2\nu^2}{\eta-\beta L}\,\delta_1^*$ shown in part (a), we have

$$e_1(\delta_0) - e_1(\delta_0^*) = (1 - (\eta - \beta L)\delta_0)e_0 + (1+\beta)^2\nu^2\delta_0^2 - \frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\,\delta_1^*.$$

Using $\delta_1^* = \delta_0^*\left(1 - \frac{\eta-\beta L}{2}\,\delta_0^*\right)$ (cf. (15)), and $e_0 = \frac{2(1+\beta)^2\nu^2}{\eta-\beta L}\,\delta_0^*$ (cf. (14)), it follows that

$$\begin{aligned} e_1(\delta_0) - e_1(\delta_0^*) &= -2(1+\beta)^2\nu^2\delta_0\delta_0^* + (1+\beta)^2\nu^2\delta_0^2 + (1+\beta)^2\nu^2(\delta_0^*)^2 \\ &= (1+\beta)^2\nu^2\left(\delta_0 - \delta_0^*\right)^2, \end{aligned}$$

showing that the hypothesis holds for $k = 1$. Now, suppose

$$e_k(\delta_0, \ldots, \delta_{k-1}) - e_k(\delta_0^*, \ldots, \delta_{k-1}^*) \geq (1+\beta)^2\nu^2(\delta_{k-1} - \delta_{k-1}^*)^2. \tag{16}$$

holds for some k and for all $(\delta_0, \ldots, \delta_{k-1}) \in \mathbb{G}_k$. We next show that relation (16) holds for $k+1$ and for all $(\delta_0, \ldots, \delta_k) \in \mathbb{G}_{k+1}$. To simplify the notation, we use e_{k+1}^* to denote the error e_{k+1}

evaluated at $(\delta_0^*, \delta_1^*, \ldots, \delta_k^*)$, and e_{k+1} when evaluating at an arbitrary vector $(\delta_0, \delta_1, \ldots, \delta_k) \in \mathbb{G}_{k+1}$. Using (13) and part (a), we have

$$e_{k+1} - e_{k+1}^* = (1 - (\eta - \beta L)\delta_k)e_k + (1+\beta)^2\nu^2\delta_k^2 - \frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\delta_{k+1}^*.$$

Under the inductive hypothesis, we have $e_k \geq e_k^*$ (cf. (16)). When $(\delta_0, \delta_1, \ldots, \delta_k) \in \mathbb{G}_k$, we have $\delta_k \leq \frac{(\eta-\beta L)}{(1+\beta)^2L^2}$. Next, we show that $\frac{(\eta-\beta L)}{(1+\beta)^2L^2} \leq \frac{1}{\eta-\beta L}$. By the definition of strong monotonicity and Lipschitzian property, we have $\eta \leq L$. Using $\eta \leq L$ and $0 \leq \beta \leq \frac{\eta}{L}$ we obtain

$$\eta \leq (1+\beta)L \Rightarrow \eta - \beta L \leq (1+\beta)L \Rightarrow (\eta - \beta L)^2 \ \leq \frac{1}{\eta - \beta L}.$$

This implies that for $(\delta_0, \delta_1, \ldots, \delta_k) \in \mathbb{G}_k$, we have $\delta_k \leq \frac{1}{\eta-\beta L}$ or equivalently $1-(\eta-\beta L)\delta_k \geq 0$. Using this, $e_k^* = \frac{2(1+\beta)^2\nu^2}{\eta-\beta L}\delta_k^*$, and the definition of δ_{k+1}^*, we obtain

$$e_{k+1} - e_{k+1}^* \geq (1 - (\eta - \beta L)\delta_k)\frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\delta_k^* + (1+\beta)^2\nu^2\delta_k^2 - \frac{2(1+\beta)^2\nu^2}{\eta - \beta L}\delta_k^*\left(1 - \frac{\eta - \beta L}{2}\delta_k^*\right)$$

$$= (1+\beta)^2\nu^2(\delta_k - \delta_k^*)^2.$$

Hence, we have $e_k - e_k^* \geq (1+\beta)^2\nu^2(\delta_{k-1} - \delta_{k-1}^*)^2$ for all $k \geq 1$ and all $(\delta_0, \ldots, \delta_{k-1}) \in \mathbb{G}_k$.

□

We have just provided an analysis in terms of a lower bound sequence $\{\delta_k\}$. We may conduct a similar analysis for an upper bound sequence $\{\Gamma_k\}$. Following a similar analysis as in the proof of Proposition 4, we can show that when $0 < \Gamma_k \leq \frac{\eta-\beta L}{(1+\beta)L^2}$, the optimal choice of the sequence $\{\Gamma_k^*\}$ is given by

$$\Gamma_0^* = \frac{\eta - \beta L}{2(1+\beta)\nu^2}\, e_0, \tag{17}$$

$$\Gamma_k^* = \Gamma_{k-1}^*\left(1 - \frac{\eta - \beta L}{2(1+\beta)}\Gamma_{k-1}^*\right) \qquad \text{for all } k \geq 1. \tag{18}$$

The following lemma is employed in our main convergence result for adaptive stepsizes $\{\gamma_{k,i}\}$.

Lemma 5. *Suppose that sequences $\{\lambda_k\}$ and $\{\gamma_k\}$ are given by the recursive equations $\lambda_{k+1} = \lambda_k(1-\lambda_k)$ and $\gamma_{k+1} = \gamma_k(1-\bar{c}\gamma_k)$ for all $k \geq 0$, where $\bar{c} > 0$ is a given constant and $\lambda_0 = \bar{c}\gamma_0$. Then for all $k \geq 0$, $\lambda_k = \bar{c}\gamma_k$.*

Proof. We use the induction on k. For $k = 0$, the relation holds since $\lambda_0 = \bar{c}\gamma_0$. Suppose that for some $k \geq 0$ the relation holds. Then, we have

$$\gamma_{k+1} = \gamma_k(1-\bar{c}\gamma_k) \Rightarrow\ \bar{c}\gamma_{k+1} = \bar{c}\gamma_k(1-\bar{c}\gamma_k) \Rightarrow\ \bar{c}\gamma_{k+1} = \lambda_k(1-\lambda_k) \Rightarrow\ \bar{c}\gamma_{k+1} = \lambda_{k+1}.$$

□

Note that using Lemma 5, it can be shown that for all $k \geq 0$,

$$\Gamma_k^* = (1+\beta)\delta_k^* \tag{19}$$

where $\{\delta_k^*\}$ is given by (14)–(15). The relations (14)–(15) and (17)–(18), respectively, are rules for determining the best upper and lower bounds for stepsize sequences $\{\gamma_{k,i}\}$, where "best" corresponds to the minimizers of the associated error bounds. Having provided this intermediate result, our main result, stated next, shows the a.s. convergence of the DSSA scheme.

Theorem 1 (A class of distributed adaptive steplength SA rules)**.** *Suppose that Assumptions 2 and 4 hold, and assume that the set X is bounded. Suppose that, for all $i = 1, \ldots, N$, the stepsizes $\{\gamma_{k,i}\}$ in algorithm (7) are given by the following recursive equations:*

$$\gamma_{0,i} = r_i c \frac{D^2}{\left(1 + \dfrac{\eta - 2c}{L}\right)^2 \nu^2}, \tag{20}$$

$$\gamma_{k,i} = \gamma_{k-1,i}\left(1 - \frac{c}{r_i}\gamma_{k-1,i}\right) \quad \textit{for all } k \geq 1. \tag{21}$$

where $D \triangleq \max_{x \in X} \|x - x_0\|$, c is a scalar satisfying $c \in (0, \frac{\eta}{2})$, r_i is a parameter such that $r_i \in [1, 1 + \frac{\eta - 2c}{L}]$, η is the strong monotonicity parameter of the mapping F, L is the Lipschitz constant of F, and ν is the upper bound defined in Assumption 4. We assume that the constant ν is chosen large enough such that $\nu \geq \frac{DL}{\sqrt{2}}$. Then, the following hold:

(a) *For any $i, j = 1, \ldots, N$ and $k \geq 0$, $\dfrac{\gamma_{k,i}}{r_i} = \dfrac{\gamma_{k,j}}{r_j}$.*

(b) *Assumption 3b holds with $\beta = \frac{\eta - 2c}{L}$, $\delta_k = \delta_k^*$, $\Gamma_k = \Gamma_k^*$, and $e_0 = D^2$, where δ_k^* and Γ_k^* are given by (14)–(15) and (17)–(18), respectively.*

(c) *The sequence $\{x_k\}$ generated by (7) converges a.s. to the unique solution of VI(X, F).*

(d) *The results of Proposition 4 hold for δ_k^* when $e_0 = D^2$ and $\beta = \frac{\eta - 2c}{L}$.*

Proof. (a) Consider the sequence $\{\lambda_k\}$ given by $\lambda_0 = \frac{c^2}{(1 + \frac{\eta - 2c}{L})^2 \nu^2} D^2$, and $\lambda_{k+1} = \lambda_k(1 - \lambda_k)$ for all $k \geq 1$. Since for any $i = 1, \ldots, N$, we have $\lambda_0 = (c/r_i)\,\gamma_{0,i}$, using Lemma 5 we obtain $\lambda_k = (c/r_i)\gamma_{k,i}$ for all $i = 1, \ldots, N$ and $k \geq 0$. Hence, the desired relation follows.

(b) First we show that δ_k^* and Γ_k^* are well defined. Consider the relation of part (a). Let $k \geq 0$ be arbitrarily fixed. If $\gamma_{k,i} > \gamma_{k,j}$ for some $i \neq j$, then we have $r_i > r_j$. Therefore, the minimum

possible $\gamma_{k,i}$ is obtained with $r_i = 1$ and the maximum possible $\gamma_{k,i}$ is obtained with $r_i = 1+\frac{\eta-2c}{L}$. Now, consider (20)–(21). If, $r_i = 1$, and D^2 is replaced by e_0, and c by $\frac{\eta-\beta L}{2}$, we get the same recursive sequence defined by (14)–(15). Therefore, since the minimum possible $\gamma_{k,i}$ is achieved when $r_i = 1$, we conclude that $\delta_k^* \leq \min_{i=1,\ldots,N} \gamma_{k,i}$ for any $k \geq 0$. This shows that δ_k^* is well-defined in the context of Assumption 3b. Similarly, it can be shown that Γ_k^* is also well-defined in the context of Assumption 3b. Now, (19) implies that $\Gamma_k^* = (1+\frac{\eta-2c}{L})\delta_k^*$ for any $k \geq 0$, which shows that Assumption 3b is satisfied with $\beta = \frac{\eta-2c}{L}$, where $0 \leq \beta < \frac{\eta}{L}$ since $0 < c \leq \frac{\eta}{2}$.

(c) In view of Proposition 3, to show the almost-sure convergence, it suffices to show that Assumption 3 holds. Part (b) implies that Assumption 3b is satisfied by the given stepsize choices. As seen in Proposition 3 of [21], Assumption 3a holds for any positive recursive sequence $\{\lambda_k\}$ of the form $\lambda_{k+1} = \lambda_k(1 - a\lambda_k)$. Since each sequence $\gamma_{k,i}$ is a recursive sequence of this form, Assumption 3a follows from Proposition 3 in [21].

(d) It suffices to show that the hypotheses of Proposition 4 hold when $e_0 = D^2$ and $\beta = \frac{\eta-2c}{L}$. Relation $\nu \geq \frac{DL}{\sqrt{2}}$ follows from $\nu \geq L\sqrt{\frac{e_0}{2}}$. Also, as mentioned in part (c), since $0 < c \leq \frac{\eta}{2}$, the relation $0 \leq \beta < \frac{\eta}{L}$ holds for any choice of c within that range. Therefore, the conditions of Proposition 4 are satisfied. □

Remark 3. *Theorem 1 provides a class of self-tuned stepsize rules for a distributed SA algorithm (7), i.e., for any choice of parameter c such that $0 < c \leq \frac{\eta}{2}$, relations (20)–(21) correspond to an adaptive stepsize rule for agents $1, \ldots, N$. Note that if $c = \frac{\eta}{2}$, these rules represent the centralized adaptive scheme. In a distributed setting, each agent may choose its corresponding parameter r_i from the specified range $[1, 1 + \frac{\eta-2c}{L}]$ and requires that all agents agree on a fixed system-specified parameter c and have consistent estimates of parameters η and L. A natural question is the optimal choice of c and r_i, a problem that requires stronger assumptions on each agent's mapping F_i. Unfortunately, this question is beyond the scope of the current paper.*

B. Convergence rate analysis

In this section, we establish the convergence rate of the DASA algorithm. First, in the following result, we provide an upper bound for a recursive sequence.

Lemma 6. *Suppose a sequence $\{\gamma_k\}_{k=0}^{\infty}$ is given by $\gamma_{k+1} = \gamma_k(1 - a\gamma_k)$ for $k \geq 0$, where a is a positive parameter such that $0 < \gamma_0 < \frac{1}{a}$. Then, we have $\gamma_k < \frac{1}{ak}$ for any $k \geq 1$.*

Proof. First, by an induction on k, we show that $0 < \gamma_k \leq \frac{1}{a}$ for any $k \geq 0$. For $k = 0$, the statement holds because $0 < \gamma_0 < \frac{1}{a}$. Assume that $0 < \gamma_k < \frac{1}{a}$. This implies $0 < (1 - a\gamma_k) < 1$. Therefore, the relation $\gamma_{k+1} = \gamma_k(1 - a\gamma_k)$ $\gamma_{k+1} > 0$. On the other hand, $(1 - a\gamma_k) < 1$ implies that $\gamma_{k+1} < \gamma_k < \frac{1}{a}$. In conclusion, the statement holds for $k := k + 1$ showing that it holds for any nonnegative integer k. Next, from definition of the sequence $\{\gamma_k\}_{k=0}^{\infty}$ we have for $k \geq 0$, $\frac{1}{\gamma_{k+1}} = \frac{1}{\gamma_k(1-a\gamma_k)} = \frac{1}{\gamma_k} + \frac{a}{1-a\gamma_k}$. Summing up from $k = 0$ to N, we obtain

$$\frac{1}{\gamma_{N+1}} = \frac{1}{\gamma_0} + a\sum_{k=0}^{N} \frac{1}{1 - a\gamma_k} > a\sum_{k=0}^{N} \frac{1}{1 - a\gamma_k}. \tag{22}$$

The inequality of arithmetic and harmonic means states that $\frac{1}{\frac{1}{n}\sum_{k=1}^{n}\frac{1}{a_k}} \leq \frac{1}{n}\sum_{k=1}^{n} a_k$ holds for arbitrary positive numbers $a_1, a_2, \ldots, a_n$. Thus, for the terms $1 - a\gamma_k$ we obtain

$$\frac{1}{\frac{1}{N+1}\sum_{k=0}^{N}\frac{1}{1-a\gamma_k}} \leq \frac{1}{N+1}\sum_{k=0}^{N}(1 - a\gamma_k) < \frac{\sum_{k=0}^{N} 1}{N+1} = 1$$

This indicates that $\sum_{k=0}^{N}\frac{1}{1-a\gamma_k} > N + 1$. Therefore, using relation (22), we obtain $\frac{1}{\gamma_{N+1}} > a(N + 1)$ for any $N \geq 0$. Therefore, $k\gamma_k < \frac{1}{a}$ implying that the desired result holds. □

This result simply states that the recursive sequence $\{\gamma_k\}_{k=0}^{\infty}$ converges to zero at least as fast as the rate $\frac{1}{k}$. The following Corollary shows that the convergence rate for the DASA schemes is $\mathcal{O}(k^{-1})$, which is the optimal convergence rate for the SA scheme (7) (see [36]).

Corollary 1 (Convergence rate of DASA schemes)**.** *Suppose that Assumptions 2 and 4 hold, and assume that the set X is bounded. Consider algorithm (7) where the stepsizes $\{\gamma_{k,i}\}$ are given by relations (20) and (21) and c is defined in Theorem 1. Then for all $k \geq 0$:*

$$\mathsf{E}\big[\|x_k - x^*\|^2\big] \leq \left(\frac{\nu(1 + \frac{\eta - 2c}{L})}{c}\right)^2 \frac{1}{k}.$$

Proof. Using the definition of D in Theorem 1 we obtain $\mathsf{E}[\|x_0 - x^*\|^2] \leq \mathsf{E}[\max_{x\in X}\|x_0 - x\|^2] = D^2$. Consider the definition of error function e_k given by (13) and let $e_0 = D^2$. The inequality (12) implies that $\mathsf{E}[\|x_k - x^*\|^2] \leq e_k(\delta_0, \ldots, \delta_{k-1})$. In addition, since the stepsizes of algorithm (7) are given by relations (20) and (21), Theorem 1b indicates that $\delta_k = \delta_k^*$ given by (14) and (15) and Theorem 1d states that the results of Proposition 4 holds. Therefore, from Proposition 4a and the preceding relation we obtain

$$\mathsf{E}\big[\|x_k - x^*\|^2\big] \leq e_k(\delta_0^*, \ldots, \delta_{k-1}^*) = \frac{2(1 + \beta)^2\nu^2}{\eta - \beta L}\delta_k^*.$$

From Lemma 6 and the definition of δ^*, we conclude that $\delta_k^* \leq (\frac{2}{\eta-\beta L})\frac{1}{k}$. From this relation and the preceding inequality, we have $\mathsf{E}[\|x_k - x^*\|^2] \leq \frac{4(1+\beta)^2\nu^2}{(\eta-\beta L)^2}\frac{1}{k}$. In addition, Theorem 1 implies that $\beta = \frac{\eta-2c}{L}$. Replacing β by this value in the preceding relation obtains the desired result. □

VI. A DISTRIBUTED LOCALLY SMOOTHED SA SCHEME

The local smoothing scheme presented in Section III facilitates the construction of a distributed locally randomized SA scheme. Consider the CSVI problem VI(X, F^ϵ) given in (5) where the mapping F is not necessarily Lipschitz. Given some $x_0 \in X$, let the sequence $\{x_k\}$ be given by

$$x_{k+1,i} := \Pi_{X_i}\left(x_{k,i} - \gamma_{k,i}\Phi_i(x_k + z_k, \xi_k)\right), \tag{23}$$

for all $k \geq 0$ and $i = 1, \ldots, N$, where $\gamma_{k,i} > 0$ denotes the stepsize of the i-th agent at iteration k, $x_k = (x_{k,1}; x_{k,2}; \ldots; x_{k,N})$, and $z_k = (z_{k,1}; z_{k,2}; \ldots; z_{k,N})$. The following proposition proves the a.s. convergence of the iterates generated by (23) to the solution of the approximation VI(X, F^ϵ). In this result, we proceed to show that the approximation does indeed satisfy the assumptions of Proposition 3 and convergence can then be immediately claimed. We define $\mathcal{F}'_k$, the history of the method up to time $k \geq 1$, as $\mathcal{F}'_k \triangleq \{x_0, z_0, \xi_0, z_1, \xi_1, \ldots, z_{k-1}, \xi_{k-1}\}$, for $\mathcal{F}'_0 = \{x_0\}$. We assume that, at any k, the vectors z_k and ξ_k in (23) are independent given the history $\mathcal{F}'_k$.

Proposition 5 (Almost-sure convergence of locally randomized SA scheme)**.** *Let Assumptions 1, 2a, and 3 hold, and suppose the map F is strongly monotone on X^ϵ with a constant $\eta > 0$. Then, the sequence $\{x_k\}$ generated by (23) converges a.s. to the unique solution of VI(X, F^ϵ).*

Proof. Define $\xi' \triangleq (z_1; z_2; \ldots; z_N; \xi)$, allowing us to rewrite algorithm (23) as follows:

$$\begin{aligned} x_{k+1,i} &= \Pi_{X_i}\left(x_{k,i} - \gamma_{k,i}(F_i^\epsilon(x_k) + w'_{k,i})\right), \\ w'_{k,i} &\triangleq \Phi_i(x_k + z_k, \xi_k) - F_i^\epsilon(x_k). \end{aligned} \tag{24}$$

To prove convergence of the iterates produced by (24), it suffices to show that the conditions of Proposition 3 are satisfied for the set X, the mapping F^ϵ, and the stochastic errors $w'_{k,i}$.
(i) Proposition 2b implies that the mapping F^ϵ is Lipschitz over the set X with the constant $\frac{\sqrt{n}\|C\|}{\min_{j=1,\ldots,N}\{\epsilon_j\}}$. Thus, Assumption 2b holds for the mapping F^ϵ.
(ii) The mapping F^ϵ is strongly monotone over the set X^ϵ with a constant $\eta > 0$ (Prop 2 (iii)).
(iii) The last step requires showing that the stochastic errors $w'_k \triangleq (w_{k,1}; w_{k,2}; \ldots; w_{k,N})$ are well-defined, i.e., $\mathsf{E}[w'_k \mid \mathcal{F}'_k] = 0$ and that Assumption 4 holds with respect to the stochastic error w'_k.

Consider the definition of $w'_{k,i}$ in (24). Taking conditional expectations on both sides, we have for all $i = 1, \ldots, N$ $\mathsf{E}\big[w'_{k,i} \mid \mathcal{F}'_k\big] = \mathsf{E}_{z,\xi}[\Phi_i(x_k + z_k, \xi_k)] - F_i^\epsilon(x_k) = \mathsf{E}[F_i(x_k + z_k)] - F_i^\epsilon(x_k) = 0$, where the last equality is obtained using the definition of F^ϵ in (5). Consequently, it suffices to show that the condition of Assumption 4 holds. This may be expressed as follows:

$$\begin{aligned}
&\mathsf{E}\big[\|w'_k\|^2 \mid \mathcal{F}'_k\big] = \mathsf{E}\left[\sum_{i=1}^N \|w'_{k,i}\|^2 \mid \mathcal{F}'_k\right] = \mathsf{E}\left[\sum_{i=1}^N \|\Phi_i(x_k + z_k, \xi_k) - F_i^\epsilon(x_k)\|^2 \mid \mathcal{F}'_k\right] \\
&\leq 2\mathsf{E}\left[\sum_{i=1}^N \|\Phi_i(x_k + z_k, \xi_k)\|^2 \mid \mathcal{F}'_k\right] + 2\sum_{i=1}^N \|F_i^\epsilon(x_k)\|^2 \\
&\leq 2\mathsf{E}\left[\sum_{i=1}^N C_i^2 \mid \mathcal{F}'_k\right] + 2\sum_{i=1}^N \mathsf{E}\big[|\Phi_i(x_k + z_k, \xi_k)\|^2 \mid \mathcal{F}'_k\big] \leq 4\sum_{i=1}^N C_i^2 = 4\|C\|^2,
\end{aligned}$$

where in the fourth relation, we use Jensen's inequality and the assumption that Φ_i is uniformly bounded over the set X^ϵ (cf. Assumption 1). Thus, the stochastic errors $\{w'_k\}$ satisfy Assumption 4. Thus, the conditions of Proposition 3 are satisfied for the set X, the mapping F^ϵ, and the stochastic errors $w'_{k,i}$ and the convergence result follows. □

Next, we state that by employing algorithm (23) coupled with the self-tuned stepsize sequences (20)–(21), the convergence results will hold.

Corollary 2. *Consider algorithm* (23) *where the stepsizes are given by* (20)–(21). *Let Assumptions 1 and 2a hold, and suppose that F is strongly monotone on X^ϵ with a constant $\eta > 0$. Then, the sequence $\{x_k\}$ generated by algorithm* (23) *converges almost surely to the unique solution of VI(X, F^ϵ). Furthermore, the mean squared error converges to zero at the rate of $\frac{1}{k}$.*

Proof. First, note that the proof of Theorem 1(c) implies that Assumption 3 holds for stepsize sequences (20)–(21). Therefore, from Prop. 5, a.s. convergence follows. From Prop. 2, F^ϵ is Lipschitz with parameter $L = \frac{\sqrt{n}\|C\|}{\min_{j=1,\ldots,N}\{\epsilon_j\}}$ and strong monotone with parameter η. Using the framework (24), in a similar fashion to the analysis in Section V, we may build the error function (13) where L is given by $L = \frac{\sqrt{n}\|C\|}{\min_{j=1,\ldots,N}\{\epsilon_j\}}$. Therefore, the result of Corollary 1 holds for such an L and x^* being the unique solution to VI(X, F^ϵ). This implies that $\{x_k\}$ converges to the unique solution in the mean squared sense with the rate of $\frac{1}{k}$. □

The distributed locally randomized SA scheme produces an approximate solution x^ϵ where ϵ denotes the size of the support of the randomization. If x^* denote the solution of VI(X, F^ϵ) and

VI(X, F), the following proposition provides a bound on the error of the smoothing scheme and proves that $x^\epsilon \to x^*$ as $\epsilon \to 0$.

Proposition 6. *Let Assumptions 1 and 2a hold, and suppose F is strongly monotone on X^ϵ with a constant $\eta > 0$. Then, VI(X, F^ϵ) has a unique solution, denoted by x^ϵ and we have:*

(a) $\|x^* - x^\epsilon\| \leq \dfrac{\sup_{\|z\|\leq\epsilon} \|F(x^* + z) - F(x^*)\|}{\eta}$.

(b) *Suppose there exists a neighborhood of x^* in which F is differentiable with bounded Jacobian. If J_{ub} is the upper bound of the Jacobian, for an sufficiently small ϵ, we have:*

$$\|x^* - x^\epsilon\| \leq \frac{J_{ub}\max_{\{1,\ldots,N\}}\sqrt{n_i}\epsilon_i}{\eta},$$

(c) *Suppose F is a continuous mapping over X^ϵ and the set X is compact. Then $x^\epsilon \to x^*$ when $\epsilon \to 0$.*

Proof. (a) Since set X is assumed to be closed and convex, the definition of X^ϵ implies that X^ϵ is also closed and convex. Thus, the existence and uniqueness of the solution to VI(X, F), as well as VI(X, F^ϵ), is guaranteed by Theorem 2.3.3 of [38]. Let $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$ with $\epsilon_i > 0$ for all i be arbitrary, and let x^ϵ denote the solution to VI(X, F^ϵ). Thus, since x^ϵ is the solution to VI(X, F^ϵ), we have $(x^* - x^\epsilon)^T F^\epsilon(x^\epsilon) \geq 0$. Similarly, since x^* is the solution to VI(X, F), we have $(x^\epsilon - x^*)^T F(x^*) \geq 0$. Adding the preceding two inequalities, we obtain for any $k \geq 0$, $(x^* - x^\epsilon)^T (F^\epsilon(x^\epsilon) - F(x^*)) \geq 0$. Adding and subtracting the term $F^\epsilon(x^*)^T(x^* - x^\epsilon)$, we have

$$\begin{aligned}
&(x^* - x^\epsilon)^T (F^\epsilon(x^\epsilon) - F^\epsilon(x^*)) + (x^* - x^\epsilon)^T (F^\epsilon(x^*) - F(x^*)) \geq 0, \\
\Rightarrow &(x^* - x^\epsilon)^T (F^\epsilon(x^*) - F(x^*)) \geq (x^* - x^\epsilon)^T (F^\epsilon(x^*) - F^\epsilon(x^\epsilon)) \geq \eta \|x^* - x^\epsilon\|^2,
\end{aligned}$$

where the last inequality follows by the strong monotonicity of the mapping F^ϵ. By invoking the Cauchy-Schwartz inequality, we obtain

$$\|F^\epsilon(x^*) - F(x^*)\| \geq \eta \|x^* - x^\epsilon\|. \tag{25}$$

It suffices to show that, $\|F^\epsilon(x^*) - F(x^*)\| \leq \sup_{\|z\|\leq\epsilon} \|F(x^* + z) - F(x^*)\|$. By the definition of F^ϵ, we have $\|F^\epsilon(x^*) - F(x^*)\| = \|\mathsf{E}[F(x^* + z) - F(x^*)]\|$. Then, the right-hand side can be expressed as follows:

$$\begin{aligned}
\|\mathsf{E}[F(x^* + z) - F(x^*)]\| &= \left\| \int_{\|z\|\leq\epsilon} (F(x^* + z) - F(x^*)) p_c(z) dz \right\| \\
&= \int_{\|z\|\leq\epsilon} \sup_{\|z\|\leq\epsilon} \|F(x^* + z) - F(x^*)\| p_c(z) dz = \sup_{\|z\|\leq\epsilon} \|F(x^* + z) - F(x^*)\|,
\end{aligned} \tag{26}$$

where the second equality is implied using the Jensen's inequality and convexity of the norm. Therefore, relations (25) and (26) imply the desired result.

(b) Let J_F denote the Jacobian of F. By assumption, there exists a $\rho > 0$ where $\|J_F(x)\| \leq J_{ub}$ for any $x \in B_n(x^*, \rho)$, where $B_n(x^*, \rho)$ is defined as a n-dimensional ball centered at x^* with radius ρ. Using the mean value theorem,

$$F(x+\delta) - F(x) = \left(\int_0^1 J_F(x + t\delta) dt \right) \delta, \quad \text{for any } \|\delta\| \leq \rho. \tag{27}$$

Assume that $\max_{\{1,\ldots,N\}} \epsilon_i < \rho$. From (27) we obtain

$$\|F(x^* + z) - F(x^*)\| \leq J_{ub}\|z\| \leq J_{ub} \max_{\{1,\ldots,N\}} \sqrt{n_i}\epsilon_i.$$

The desired results follows from the preceding relation and the inequality of part (a).

(c) Since the set X is bounded and $x^\epsilon \in X$, the sequence x^ϵ has at least one convergent subsequence. Let x^{ϵ^t} denote that subsequence and define $\lim_{t\to\infty} x^{\epsilon^t} = \bar{x}$. Using the definition of F^{ϵ^t} and we have

$$\lim_{t\to\infty} F^{\epsilon^t}(x^{\epsilon^t}) = \lim_{t\to\infty} \int F(x^{\epsilon^t} + z) p_c(z) dz = \lim_{t\to\infty} \int F(x^{\epsilon^t} + \|\epsilon^t\| y) p_c(z) dy,$$

where we make use of change of variables $y = \frac{z}{\|\epsilon^t\|}$. By Assumption 1 we have that $\|F(x^\epsilon + z)\| \leq C$, implying that $F(x^{\epsilon^t} + \|\epsilon^t\| y)$ is bounded with respect to the distribution defining the random variable z. Thus, using the Lebesgue's dominated convergence theorem, we interchange the limit and the integral leading to the following relations:

$$\lim_{t\to\infty} F^{\epsilon^t}(x^{\epsilon^t}) = \int \lim_{\epsilon\to 0} F(x^{\epsilon^t} + \|\epsilon^t\| y) p_c(z) dy = \int F(\bar{x}) p_c(z) dy = F(\bar{x}),$$

where the second equality follows by the continuity of the mapping F, and $x^{\epsilon^t} \to \bar{x}$. Since x^{ϵ^t} solves VI(X, F^{ϵ^t}), we have $(x - x^{e^t})^T F^{\epsilon^t}(x^{\epsilon^t}) \geq 0$ for all $x \in X$. Thus, by taking the limit along the subsequence $\{\epsilon^t\}$ for any $x \in X$ we obtain

$$(x - \lim_{t\to\infty} x^{e^t})^T \left(\lim_{t\to\infty} F^{\epsilon^t}(x^{\epsilon^t}) \right) \geq 0 \implies (x - \bar{x})^T F(\bar{x}) \geq 0,$$

showing that $\bar{x}$ is a solution to VI(X, F). However, since VI(X, F) has a unique solution, x^*, and that $\bar{x}$ is an arbitrary accumulation point of x^ϵ, the set of all accumulation points of $\{x^\epsilon\}$ is equal to $\{x^*\}$. It is known that the limit superior of a sequence is equal to the supremum of the set of all accumulation points of that sequence. Therefore, $\limsup_{\epsilon\to 0} x^\epsilon = x^*$. Similarly, $\liminf_{\epsilon\to 0} x^\epsilon = x^*$, We conclude that $\{x^\epsilon\}$ is converges to x^*. □

Remark 4 (Choice of the smoothing parameter ϵ)**.** *One question that may arise in the application of algorithm (7) is pertaining to a good choice of the parameter ϵ. In practice, the Proposition 6 can be applied to address such question. We start the SA algorithm (7) using an arbitrary ϵ. After a suitable stopping criteria is met, we drop the smoothing parameter. Following such scheme, Proposition 6 implies that we approach the solution to the original problem. A detailed analysis on the convergence of the SA algorithm where the smoothing and regularization parameters are updated per iteration can be found in [37].*

VII. Numerical results

In this section, we examine the behavior of our schemes on a stochastic Nash-Cournot game, described in Section VII-A. First, in Section VII-B, we compare the performance of the DSSA scheme given by (20)–(21) with that of SA schemes with harmonic stepsizes (HSA) of the form $\gamma_k = \frac{\beta}{(k+a)^\alpha}$ (see [14, Ch. 4, pg. 113]). Next, in Section VII-C, we examine the convergence of the proposed smoothing SA algorithm when the smoothing parameter decays to zero (see Prop. 6.) Finally, our proposed SA algorithm is compared with sample average approximation (SAA) techniques in Section VII-D where Knitro 6.0 [39] is used to solve the SAA problem.

A. Preliminaries

We consider a networked Nash-Cournot game akin to that described in Section II with 6 firms over a network with 4 nodes. Specifically, let firm i's generation and sales decisions at node j be given by g_{ij} and s_{ij}, respectively. The price at node j is denoted by the function p_j, defined as $p_j(\bar{s}_j, a_j, b_j) = a_j - b_j \bar{s}_j^\sigma$, where $\bar{s}_j = \sum_i s_{ij}$, $\sigma \geq 1$ and a_j is a uniformly distributed random variable defined over the interval $[lb_j^a, ub_j^a]$. We assume the generation cost is a piecewise linear convex function given by $c_{ij}(g_{ij}) = 0.9 g_{ij}$ for $0 \leq g_{ij} < 50$, $c_{ij}(g_{ij}) = g_{ij} - 5$ for $50 \leq g_{ij} < 100$, and $c_{ij}(g_{ij}) = 1.1 g_{ij} - 15$ for $g_{ij} \geq 100$. The objective function of the firm i, given by $f_i(x, \xi) = \sum_{j=1}^M \left(c_{ij}(g_{ij}, \xi) - p_j(\bar{s}_j, \xi) s_{ij}\right)$, is a stochastic nonlinear convex and nonsmooth function. As discussed in [31], when $1 < \sigma \leq 3$ and $M \leq \frac{3\sigma - 1}{\sigma - 1}$, the mapping F is strictly monotone. Given our interest in strongly monotone problems, we consider a regularized map F; more precisely, as explained in Remark 1, $H \triangleq F + \eta \mathbf{I}$ is strongly monotone for any arbitrary scalar $\eta > 0$. On the other hand, when $\sigma > 1$, ascertaining the Lipschitzian property of F is challenging,

motivating the use of the distributed locally randomized SA scheme introduced in Sec. VI. Using regularization and the smoothing scheme, we consider the solution of the approximate problem $\text{VI}(X, H^\epsilon)$. From the definition (5) and by noting that the smoothing variable is mean zero, we have that $H^\epsilon = F^\epsilon + \eta\mathbf{I}$, where $\eta > 0$ denotes the regularization parameter. Further, $\text{VI}(X, H^\epsilon)$ admits a unique solution denoted by $x^*_{\eta,\epsilon}$ since H^ϵ is strongly monotone [38, Theorem 2.3.3]. Throughout our experiment, we set $\sigma = 1.9$, $lb_a = [400; 600; 700; 500]$, $ub_a = lb_a + [1; 1; 1; 1]$, $b = 0.001 \times [5; 6; 5; 3]$, and $cap = 100\,[3\ 4\ 5\ 6\ 5\ 4; 0.6\ 5\ 2\ 4\ 6\ 4; 5\ 0.5\ 6\ 7\ 3\ 4; 6\ 4\ 40\ 7\ 0.3\ 5]$.

B. Comparison with harmonic and constant stepsizes

We consider the following SA schemes for the solution of $\text{VI}(X, F^\epsilon + \eta\mathbf{I})$:

DSSA scheme: Here, we employ (23) coupled with self-tuned stepsizes given by (20)–(21) and assume that the random vector z is generated via the MCR scheme. An immediate benefit of this scheme is that the Lipschitzian parameter L can be estimated from Prop. 2b. r_i is randomly chosen for each firm within the prescribed range. The constant c is maintained at $\frac{\eta}{4}$.

HSA schemes: Analogous to the DSSA scheme, this scheme uses harmonic steplength sequence of the form $\gamma_k = \frac{\beta}{(k+a)^\alpha}$ at k-th iteration for any firm. In our experiment, we assign different values to β, α and a. Note that to guarantee almost-sure convergence, γ_k needs to be non-summable but square-summable, implying that $0.5 < \alpha \leq 1$. We further allow α to take on values given by 1, 0.75, and 0.51. Table I shows 18 settings for the parameters α, β and a and are categorized by the value of the initial stepsize γ_1. In 9 of these settings we set $a = 0$ and in the other 9 settings, a positive value is assigned to a. The positive values of a and the values of α are considered such that γ_1 is 1, 0.1 or 0.01.

-	$\gamma_1 = 1$			$\gamma_1 = 0.1$			$\gamma_1 = 0.01$		
S(i)	β	a	α	β	a	α	β	a	α
1	1	0	1	0.1	0	1	0.01	0	1
2	1	0	0.75	0.1	0	0.75	0.01	0	0.75
3	1	0	0.51	0.1	0	0.51	0.01	0	0.51
4	10	9	1	1	9	1	1	99	1
5	10	21	0.75	1	21	0.75	1	463	0.75
6	10	90	0.51	1	90	0.51	1	8347	0.51

TABLE I: Parameters of HSA schemes

-	-	$\gamma_1 = 1$		$\gamma_1 = 0.1$		$\gamma_1 = 0.01$	
ALG.	S(i)	MSE	Std	MSE	Std	MSE	Std
HSA	1	5.29e+1	1.86e−1	6.73e+1	1.16e−2	9.76e+4	3.63e+0
	2	2.06e−3	1.51e−4	1.67e+0	4.36e−4	1.16e+4	8.14e−1
	3	1.19e−3	6.93e−4	2.66e−3	1.44e−4	3.81e+1	2.00e−3
	4	6.22e−1	4.30e−4	1.23e−2	1.23e−4	3.42e−1	1.70e−4
	5	1.63e−3	1.02e−3	2.05e−3	1.54e−4	2.20e−3	1.53e−4
	6	2.87e+4	1.97e+3	1.20e−3	6.83e−4	8.97e−4	3.82e−4
DSSA	7	9.30e−4	3.61e−4	9.43e−4	3.45e−4	1.19e−3	2.49e−4
CSA	8	2.06e+6	1.52e+3	3.56e−2	2.39e−1	9.47e−4	3.97e−4

TABLE II: DSSA vs HSA and CSA

CSA schemes: Here we employ the the SA algorithm (23) coupled with a constant stepsize rule. We run the simulations for CSA algorithm with each of the initial stepsize values 1, 0.1 and 0.01 and compare the performance with the DSSA and HSA schemes. Note that a constant stepsize policy does not lead to asymptotic convergence.

Throughout this section, the regularization parameter η is set at $\eta = 0.1$ and the smoothing parameter is maintained equal to 0.01 for all the agents. Recall from Prop. 4(b) and Theorem 1(d), the self-tuned stepsize minimizes the error function when the stepiszes are smaller than $\frac{2c}{(1+(\eta-2c)/L)^2L^2}$. However, for the chosen $\epsilon = 0.01$ and $\eta = 0.1$, this upper limit may be small. To achieve a fair comparison between DSSA and HSA schemes, it makes sense to allow both schemes to start with the same initial stepsize values. We run different simulations with the the initial stepsizes set at 1, 0.1, and 0.01. Note that since these values are not in the feasible range explained before, we do not expect the DSSA algorithm to result in the least mean-squared error among all SA schemes. However, we would like to study the robustness of such the DASA and HSA schemes to the initial stepsize. To evaluate the performance, we run 100 simulations with $K = 4000$ steps and calculate the empirical mean-squared error (MSE), as given by $\frac{\sum_{i=1}^{100} \|x_{K,i} - x^*_{\eta,\epsilon}\|^2}{100}$. We also report the standard deviation of the term $\|x_{K,i} - x^*_{\eta,\epsilon}\|^2$.

Results and insights: Table II presents the simulation results for the test settings using the DSSA, HSA and CSA schemes. When $\gamma_1 = 1$, we observe that the DSSA schemes perform better than the HSA and CSA counterparts by having the least MSE. In this case, CSA performs poorly since the $\gamma_1 = 1$ is too large. The HSA scheme with setting $\mathrm{S}(3)$ and $\mathrm{S}(5)$ have a smaller MSE among other HSA settings. However, the DASA scheme has the least standard deviation compared with the two HSA schemes showing more robustness to problem randomness. When $\gamma_1 = 0.1$, we observe that again DSSA has the least MSE among all schemes and HSA with $\mathrm{S}(3)$, $\mathrm{S}(5)$ and $\mathrm{S}(6)$ being the best amongst the HSA settings with $\mathrm{S}(3)$ has the minimum deviation. Note that by changing the initial stepsize from 1 to 0.1, the MSE and Std remains in the order of 10^-4 implying the robustness of this scheme. When $\gamma_1 = 0.01$, the CSA scheme has the least MSE while the HSA scheme with $\mathrm{S}(6)$ has the second lowest MSE while the DSSA scheme has the third lowest MSE. Interestingly, the DSSA scheme has the smallest standard deviation, implying more robustness to the stochasticity of the problem. Comparing across the three cases, one immediate observation is the setting $\mathrm{S}(6)$ performs poorly in the first case while it performs well in the other two cases. This implies that the performance of the HSA scheme with nonzero

a depends on the choice of β and α. This experiment shows how the HSA schemes of the form $\gamma_k = \frac{\beta}{(k+a)^\alpha}$ are extremely sensitive to the choice of all the parameters α, β and a. A natural question is how the CSA scheme performs if we decrease γ_1. We performed the simulation for $\gamma_1 = 10^{-3}$ and 10^{-4} and observed that the resulting MSE increases by further decreasing the size of the stepsize. The MSE of all the schemes is demonstrated in Figure 2 for iterations

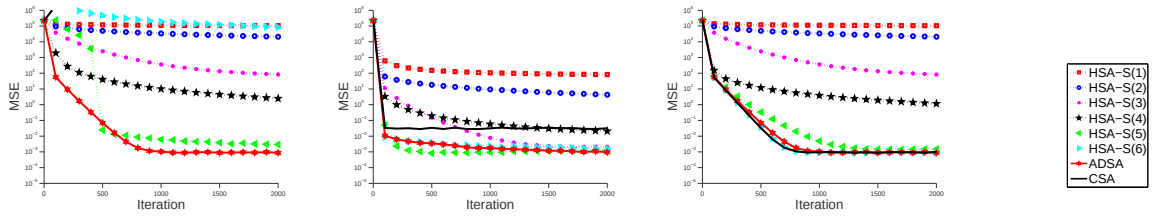


Fig. 2: MSE of the SA schemes: $\gamma_1 = 1, 0.1,$ and 0.01; and Legend

ranging from 1 to 4000. It is observed that the DSSA scheme (assigned red solid line with "*" character) performs well for different settings of the starting stepsize. To study the performance of the DSSA scheme in problems with different number of variables, we performed another set of simulations. For this purpose, we compared the performance of HSA (6) with that of the DSSA scheme as M grows from 5 to 25 while N is fixed at 4. Table III presents such a comparison with $\gamma_1 = 0.1$. The results show that the DSSA scheme displays lower MSE with smaller standard deviations than the HSA scheme S(6) in all cases.

ALG.	-	$M = 5$	$M = 10$	$M = 15$	$M = 20$	$M = 25$
DSSA	MSE	4.26e−4	5.14e−4	5.69e−4	5.97e−4	6.22e−4
	Std	2.89e−4	3.23e−4	3.58e−4	3.56e−4	3.95e−4
HSA(6)	MSE	9.86e−4	1.30e−3	1.59e−3	1.70e−3	1.86e−3
	Std	6.24e−4	9.00e−4	1.03e−3	9.76e−4	9.60e−4

TABLE III: DSSA vs HSA(6) – varying number of agents

C. *Convergence of smoothing scheme*

In this section, we study the convergence of the proposed SA algorithm given by (23) to the solution of the original problem. We apply the DSSA scheme and study convergence of $x_{\eta,\epsilon}$ to the solution of VI$(X, F+\eta\mathbf{I})$ when the smoothing parameter is reduced to zero. Specifically, we generate 50 runs of the DSSA algorithm with $\epsilon = 1, 0.1, 0.01, 0.001$ and report the MSE of the scheme defined by $\frac{\sum_{i=1}^{50} \|x_{K,i} - x_\eta^*\|^2}{50}$. Note that in this section, we keep the problem regularized

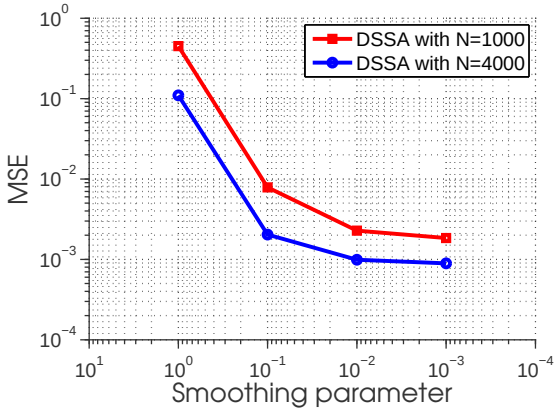


Fig. 3: Varying smoothing parameter

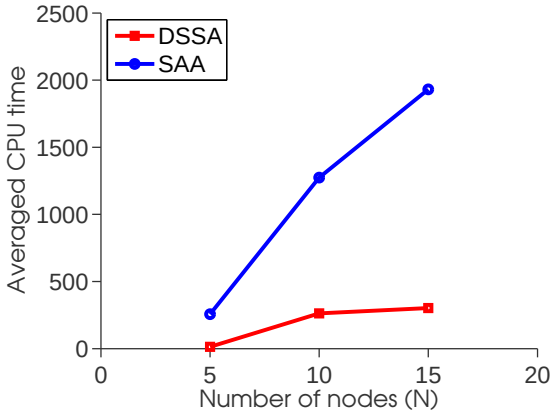


Fig. 4: SAA vs DSSA: varying number of nodes

and focus on reducing the support of the local randomization.

Results and insights: Figure 3 demonstrates the results of our experiment. It is seen that by decreasing the smoothing parameter to zero, the MSE approaches zero, supporting Prop. 6 (Note that the continuity condition of F is not satisfied here due to nonsmooth cost functions. Extension of Prop. 6 to discontinuous cases is a subject of our future research). We also see that for any fixed value of the smoothing parameter, the MSE decreases as the number of iteration increases from 1000 to 4000.

D. Comparison with SAA method

Next, we compare the DSSA scheme with sample-average approximation techniques in terms of CPU time, MSE and the standard deviation, based on 50 simulations.

ALG.	-	$N = 5$	$N = 10$	$N = 15$
DSSA	CPU(s)	1.40e+1	2.63e+2	3.02e+2
	MSE	1.68e−3	2.96e−3	4.55e−3
	Std	1.03e−3	1.12e−3	1.72e−3
SAA	CPU(s)	2.57e+2	1.27e+3	1.93e+3
	MSE	1.03e−5	2.09e−5	3.07e−5
	Std	5.65e−6	9.35e−6	8.95e−6

TABLE IV: DSSA vs SAA – varying number of nodes

Results and insights: The purpose of this section is to study how the DSSA method performs in terms of CPU time comparing with the well-known SAA method. We performed a set of simulations to do this comparison where we change number of nodes N fro 5 to 15 in two steps. Table IV presents the results of such an experiment. Note that SAA is a framework and

in our computation, the commercial solver KNITRO [39] (uses Newton based technique) is used to solve the SAA problem. This explains why the order of MSE for the SAA method is smaller than that of the DSSA scheme. However, the CPU time for the SAA method is significantly larger than that of the DSSA scheme. Moreover, we observe that the required CPU time for SAA method grows much faster than the DSSA method. Figure 4 shows this observation. This experiment supports that the solution time for SA algorithm is significantly smaller than that for SAA. A more detailed comparison between SA and SAA schemes can be found in [18].

VIII. Concluding remarks

We consider the solution of strongly monotone CSVI problems arising from stochastic multi-user systems via stochastic approximation schemes. The resulting maps associated with such problems are possibly non-Lipschitzian and few SA schemes, if any, can cope with such problems to display almost-sure convergence. By introducing a user-specific local smoothing, we derive an approximate map that is shown to be Lipschitz continuous. Given that the majority of SA schemes are reliant on naively chosen steplength sequences with highly varying performance, we develop a distributed multi-user recursive rule based on minimizing a bound on the mean-squared error. The resulting choices of steplength sequences adapt to problem parameters such as the Lipschitz and monotonicity constants of the map. It is shown that the rule produces iterates that converge to the unique solution in an a.s. sense and at the optimal rate. We utilize this technique in developing a distributed locally randomized variant that can cope with non-Lipschitzian stochastic maps. It is shown that this scheme produces iterates converging to a solution of an approximate problem and the sequence of approximate solutions converges to the original solution as the smoothing parameter is driven to zero. Finally, we apply our scheme on a stochastic Nash-Cournot game, for which the DSSA scheme displays far more robustness than the standard implementations that leverage harmonic stepsizes of the form $\frac{\beta}{(k+a)^{\alpha}}$.

References

[1] F. Yousefian, A. Nedić, and U. V. Shanbhag, "A distributed adaptive steplength stochastic approximation method for monotone stochastic Nash games," *IEEE American Control Conference (ACC)*, pp. 4772–4777, 2013.

[2] B. F. Hobbs, "Linear complementarity models of Nash-Cournot competition in bilateral and poolco power markets," *IEEE Transactions on Power Systems*, vol. 16, no. 2, pp. 194–202, 2001.

[3] A. Kannan, U. V. Shanbhag, and H. . M. Kim, "Addressing supply-side risk in uncertain power markets: stochastic Nash models, scalable algorithms and error analysis," *Optimization Methods and Software (online first)*, vol. 0, no. 0, pp. 1–44, 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10556788.2012.676756

[4] J. Wang, G. Scutari, and D. P. Palomar, "Robust mimo cognitive radio via game theory," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1183–1201, 2011.

[5] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness, and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[6] S. Shakkottai and R. Srikant, "Network optimization and control," *Foundations and Trends in Networking*, vol. 2, no. 3, pp. 271–379, 2007.

[7] T. Alpcan and T. Başar, "A game-theoretic framework for congestion control in general topology networks," in *Proceedings of the 41st IEEE Conference on Decision and Control*, December 2002, pp. 1218– – 1224.

[8] H. Yin, U. V. Shanbhag, and P. G. Mehta, "Nash equilibrium problems with scaled congestion costs and shared constraints," *IEEE Transactions of Automatic Control*, vol. 56, no. 7, pp. 1702–1708, 2009.

[9] N. Li and J. R. Marden, "Designing games to handle coupled constraints," in *Proceedings of the IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 250–255.

[10] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statistics*, vol. 22, pp. 400–407, 1951.

[11] Y. M. Ermoliev, *Stochastic Programming Methods*. Moscow: Nauka, 1976.

[12] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[13] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer New York, 2003.

[14] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken, NJ, 2003.

[15] H. Jiang and H. Xu, "Stochastic approximation approaches to the stochastic variational inequality problem," *IEEE Transactions on Automatic Control*, vol. 53, no. 6, pp. 1462–1475, 2008.

[16] J. Koshal, A. Nedić, and U. V. Shanbhag, "Regularized iterative stochastic approximation methods for variational inequality problems," *IEEE Transactions on Automatic Control*, vol. 58(3), pp. 594–609, 2013.

[17] 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*, vol. 15, pp. 229–251, 2004.

[18] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[19] A. Juditsky, A. Nemirovski, and C. Tauvel, "Solving variational inequalities with stochastic mirror-prox algorithm," *Stochastic Systems*, pp. DOI: 10.1214/10–SSY011, 17–58, 2011.

[20] C. D. Dang and G. Lan, "On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators," Dept. of Industrial and Systems Eng., University of Florida, Tech. Rep.

[21] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On stochastic gradient and subgradient methods with adaptive steplength sequences," *Automatica*, vol. 48, no. 1, pp. 56–67, 2012, an extended version of the paper available at: http://arxiv.org/abs/1105.4549.

[22] V. A. Steklov, "Sur les expressions asymptotiques decertaines fonctions dfinies par les quations diffrentielles du second

ordre et leers applications au problme du dvelopement d'une fonction arbitraire en sries procdant suivant les diverses fonctions," *Comm. Charkov Math. Soc.*, vol. 2, no. 10, pp. 97–199, 1907.

[23] D. P. Bertsekas, "Stochastic optimization problems with nondifferentiable functionals with an application in stochastic programming," in *Proceedings of 1972 IEEE Conference on Decision and Control*, 1972, pp. 555–559.

[24] V. I. Norkin, "The analysis and optimization of probability functions," International Institute for Applied Systems Analysis technical report, Tech. Rep., 1993, wP-93-6.

[25] H. Lakshmanan and D. Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 911–940, 2008.

[26] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, "Randomized smoothing for stochastic optimization," *SIAM Journal on Optimization (SIOPT)*, vol. 22, no. 2, pp. 674–701, 2012.

[27] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2010.

[28] A. P. George and W. B. Powell, "Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming," *Machine Learning*, vol. 65, pp. 167–198, 2006.

[29] D. Cicek, M. Broadie, and A. Zeevi, "General bounds and finite-time performance improvement for the kiefer-wolfowitz stochastic approximation algorithm," *Operations Research*, vol. 59, no. 5, pp. 1211–1224, 2011.

[30] C. Metzler, B. F. Hobbs, and J.-S. Pang, "Nash-cournot equilibria in power markets on a linearized dc network with arbitrage: Formulations and properties," *Networks and Spatial Theory*, vol. 3, no. 2, pp. 123–150, 2003.

[31] A. Kannan and U. V. Shanbhag, "Distributed computation of equilibria in monotone Nash games via iterative regularization techniques," *SIAM Journal of Optimization*, vol. 22, no. 4, pp. 1177–1205, 2012.

[32] B. F. Hobbs and J. S. Pang, "Nash-Cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints," *Oper. Res.*, vol. 55, no. 1, pp. 113–127, 2007.

[33] H. Yin, U. V. Shanbhag, and P. G. Mehta, "Nash equilibrium problems with scaled congestion costs and shared constraints," *, IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1702–1708, 2011.

[34] A. M. Gupal, *Stochastic methods for solving nonsmooth extremal problems (Russian)*. Naukova Dumka, 1979.

[35] F. Yousefian, A. Nedić, and U. Shanbhag, "Distributed adaptive steplength stochastic approximation schemes for cartesian stochastic variational inequality problems," *arXiv:1301.1711v1*, 2013.

[36] B. Polyak, *Introduction to optimization*. New York: Optimization Software, Inc., 1987.

[37] F. Yousefian, A. Nedić, and U. V. Shanbhag, "A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems," *Proceedings of the Winter Simulation Conference*, pp. 933–944, 2013.

[38] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vols. I,II*, ser. Springer Series in Operations Research. New York: Springer-Verlag, 2003.

[39] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM Journal on Optimization*, vol. 9, pp. 877–900, 1999.