A Distributed Primal-Dual Algorithm for Bandit Online Convex Optimization with Time-Varying Coupled Inequality Constraints

Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl H. Johansson

Abstract— **This paper considers distributed bandit online optimization with time-varying coupled inequality constraints. The global cost and the coupled constraint functions are the summations of local convex cost and constraint functions, respectively. The local cost and constraint functions are held privately and only at the end of each period the constraint functions are fully revealed, while only the values of cost functions at queried points are revealed, i.e., in a so called bandit manner. A distributed bandit online primal-dual algorithm with two queries for the cost functions per period is proposed. The performance of the algorithm is evaluated using its expected regret, i.e., the expected difference between the outcome of the algorithm and the optimal choice in hindsight, as well as its constraint violation. We show that $\mathcal{O}(T^c)$ expected regret and $\mathcal{O}(T^{1-c/2})$ constraint violation are achieved by the proposed algorithm, where T is the total number of iterations and $c \in [0.5, 1)$ is a user-defined trade-off parameter. Assuming Slater's condition, we show that $\mathcal{O}(\sqrt{T})$ expected regret and $\mathcal{O}(\sqrt{T})$ constraint violation are achieved. The theoretical results are illustrated by numerical simulations.**

I. Introduction

Online convex optimization is a promising methodology for modeling sequential tasks and can be traced back to the 1990s [1]–[4]. It has important applications in machine learning and control, see, e.g., [5]–[10]. Bandit online convex optimization is online convex optimization with so called bandit feedback meaning that in each period only the values of the cost functions at some points are revealed, rather than other information such as the gradient of the cost function. Gradient information may not be available in many applications, such as online source localization, online routing in data networks, and online advertisement placement in web search [7]. Essentially, bandit online convex optimization is a derivative-free method to solve convex optimization problems. Derivative-free methods have an evident advantage since computing a function value is much simpler than computing its gradient [11]. Early works of studying bandit online convex optimization include [12], [13], where the expected regret is used to measure the performance of the algorithms. The expected regret is the expected difference between the outcome of the algorithm and the optimal choice in hindsight.

A key step in bandit online convex optimization is to estimate the gradient of the cost function through querying the cost function. Various algorithms have been developed and can be divided into two categories depending on the number of queries. Algorithms with one query per period have been proposed in [13]–[22]. Algorithms with two or more queries per period have been proposed in [22]–[28] and the expected regret bounds can then be further reduced.

Aforementioned studies did not consider equality or inequality constraints. In the literature, there are only few papers considering bandit online convex optimization with equality or inequality constraints, although such constraints are common in applications. The authors of [29] studied online convex optimization with static inequality constraints and bandit feedback for constraints. They proposed an algorithm with two queries per period and achieved $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(T^{3/4})$ bounds on the expected regret and constraint violation, respectively. The authors of [30] studied online convex optimization with time-varying inequality constraints and bandit feedback for cost functions. Under Slater's condition, they proposed a class of algorithms with one or two queries per period.

Most existing bandit online convex optimization studies are in a centralized setting and only few papers considered distributed bandit online convex optimization. When cost functions are strongly convex, the authors of [31] proposed a consensus-based distributed bandit online algorithm with one query per period and obtained $\mathcal{O}(\sqrt{T}\log(T))$ expected regret. When cost functions are quadratic, the authors of [32] proposed a consensus-based distributed bandit online algorithm with two queries per period and obtained $\mathcal{O}(\sqrt{T})$ expected regret when there are set constraints.

This paper studies distributed bandit online convex optimization with time-varying coupled inequality constraints. The global cost and the coupled constraint functions are the sum of local convex cost and constraint functions, respectively. The local cost and constraint functions are held privately and only at the end of each period the constraint functions are fully revealed, while the values of the cost functions are revealed in a bandit manner. Specifically, per period each agent can sample the value of its local cost function at two points and can observe the value and the exact gradient of its local constraint function at one point. Compared to existing studies, the contributions of this paper are summarized as follows.

This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Research Council, Ministry of Education of Republic of Singapore under Grant MoE Tier 1 RG72/19, and the National Natural Science Foundation of China under grants 61991403, 61991404, and 61991400.

X. Yi and K. H. Johansson are with School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44, Stockholm, Sweden. {xinleiy, kallej}@kth.se.

X. Li and L. Xie are with School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. {xiuxianli, elhxie}@ntu.edu.sg.

T. Yang and T. Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China. {yangtao,tychai}@mail.neu.edu.cn.

1) We develop a distributed bandit online primal-dual algorithm, where per period each agent uses two queries to estimate the gradient of its local cost function. The proposed algorithm uses different non-increasing stepsize sequences for the primal and dual updates and a non-increasing sequence of regularization parameters. Moreover, it also uses non-increasing shrinkage and exploration sequences in the gradient estimation model. These sequences give some freedom in the expected regret and constraint violation bounds, as they allow the trade-off between how fast these two bounds tend to zero. It should be highlighted that the total number of iterations is not used as a variable in the algorithm, which is different from most existing bandit online algorithms.

2) For the proposed algorithm, we show that $\mathcal{O}(T^c)$ expected regret and $\mathcal{O}(T^{1-c/2})$ constraint violation are achieved, where $c \in [0.5, 1)$ is a user-defined trade-off parameter. Compared with the bandit algorithm in [25], which achieved $\mathcal{O}(\sqrt{T})$ expected regret under static set constraints and centralized computations using the total number of iterations in the algorithm, we are relaxing all these assumptions.

3) Assuming Slater's condition, we show that $\mathcal{O}(\sqrt{T})$ expected regret and $\mathcal{O}(\sqrt{T})$ constraint violation can be achieved. Although the two-query bandit algorithm in [30] also achieved the same expected regret and constraint violation bounds, it is a centralized algorithm and uses the total number of iterations. Moreover, a slightly stronger Slater's condition was assumed in [30].

The rest of this paper is organized as follows. Section II introduces the preliminaries. Section III gives the problem formulation. Section IV provides a distributed bandit online algorithm and presents expected regret and constraint violation bounds. Section V gives numerical simulations. Finally, Section VI concludes the paper.

Notations: All inequalities and equalities are understood componentwise. $[n]$ represents the set $\{1, \dots, n\}$ for any $n \in \mathbb{N}_+$. $\mathbb{S}^p$ stands for the unit sphere centered around the origin in $\mathbb{R}^p$ under Euclidean norm. $\mathrm{col}(z_1, \dots, z_k)$ represents the concatenated column vector of vectors $z_i \in \mathbb{R}^{n_i}$, $i \in [k]$. Given two scalar sequences $\{\alpha_t,\ t \in \mathbb{N}_+\}$ and $\{\beta_t > 0,\ t \in \mathbb{N}_+\}$, $\alpha_t = \mathcal{O}(\beta_t)$ means that $\limsup_{t\to\infty}(\alpha_t/\beta_t)$ is bounded. For a set $\mathbb{K} \subseteq \mathbb{R}^p$, $\mathcal{P}_{\mathbb{K}}(\cdot)$ denote the projection operator, i.e., $\mathcal{P}_{\mathbb{K}}(x) = \arg\min_{y\in\mathbb{K}} \|x - y\|^2,\ \forall x \in \mathbb{R}^p$. For simplicity, $[\cdot]_+$ is used to denote $\mathcal{P}_{\mathbb{K}}(\cdot)$ when $\mathbb{K} = \mathbb{R}^p_+$.

II. Preliminaries

In this section, we present some definitions and assumptions related to graph theory and gradient approximation.

A. Graph Theory

Interactions between agents are modeled by a time-varying directed graph. Specifically, at time t, agents communicate with other agents according to a directed graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, where $\mathcal{V} = [n]$ is the agent set and $\mathcal{E}_t \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set. Let $\mathcal{N}_i^{\mathrm{in}}(\mathcal{G}_t) = \{j \in [n] \mid (j, i) \in \mathcal{E}_t\}$ and $\mathcal{N}_i^{\mathrm{out}}(\mathcal{G}_t) = \{j \in [n] \mid (i, j) \in \mathcal{E}_t\}$ be the sets of in- and out-neighbors, respectively, of agent i at time t. The mixing matrix $W_t \in \mathbb{R}^{n\times n}$ at time t fulfills $[W_t]_{ij} > 0$ if $(j, i) \in \mathcal{E}_t$ or $i = j$, and $[W_t]_{ij} = 0$ otherwise.

The following standard assumption is made on the graph, see e.g., [33].

Assumption 1. *For any $t \in \mathbb{N}_+$, the graph $\mathcal{G}_t$ satisfies the following conditions: (i) There exists a constant $w \in (0, 1)$, such that $[W_t]_{ij} \geq w$ if $[W_t]_{ij} > 0$; (ii) The mixing matrix W_t is doubly stochastic, i.e., $\sum_{i=1}^n [W_t]_{ij} = \sum_{j=1}^n [W_t]_{ij} = 1,\ \forall i, j \in [n]$; (ii) There exists an integer $\iota > 0$ such that the graph $(\mathcal{V}, \cup_{l=0,\dots,\iota-1}\mathcal{E}_{t+l})$ is strongly connected.*

B. Gradient Approximation

Let $f : \mathbb{K} \to \mathbb{R}$ be a function with $\mathbb{K} \subset \mathbb{R}^p$. We assume that $\mathbb{K}$ is bounded and has a nonempty interior. Without loss of generality, we assume that $\mathbb{K}$ contains the ball of radius $r(\mathbb{K})$ centered at the origin and is contained in the ball of radius $R(\mathbb{K})$, i.e., $r(\mathbb{K})\mathbb{B}^p \subseteq \mathbb{K} \subseteq R(\mathbb{K})\mathbb{B}^p$. We use the following gradient estimator

$$\hat{\nabla}_2 f(x) = \frac{p}{\delta}(f(x + \delta u) - f(x))u,\ \forall x \in (1 - \xi)\mathbb{K}, \quad (1)$$

where u is a random vector which is uniformly distributed in the unit sphere $\mathbb{S}^p$, $\delta \in (0, r(\mathbb{K})\xi]$ is an exploration parameter, and $\xi \in (0, 1)$ is a shrinkage coefficient. The estimator (1) thus requires to sample the function at two points, so it is a two-query model. The intuition follows from directional derivatives [24]. The properties of $\hat{f}$ can be found in [34].

III. Problem Formulation

Consider a network of n agents indexed by $i \in [n]$. For each i, let the local decision set $\mathbb{X}_i \subseteq \mathbb{R}^{p_i}$ be a closed convex set with p_i being a positive integer. Let $\{f_{i,t} : \mathbb{X}_i \to \mathbb{R}\}$ and $\{g_{i,t} : \mathbb{X}_i \to \mathbb{R}^m\}$ be sequences of local convex cost and constraint functions over time $t = 1, 2, \dots$, respectively, where m is a positive integer. At each t, the network's objective is to solve the constrained convex optimization problem

$$\begin{aligned} \min_{x_t \in \mathbb{X}} \quad & f_t(x_t) \\ \text{s.t.} \quad & g_t(x_t) \leq \mathbf{0}_m, \quad t = 1, 2, \dots \end{aligned} \quad (2)$$

where $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \subseteq \mathbb{R}^p$ with $p = \sum_{i=1}^n p_i$ being the global decision set, $x_t = \mathrm{col}(x_{1,t}, \dots, x_{n,t})$ is the global decision variable, $f_t(x_t) = \sum_{i=1}^n f_{i,t}(x_{i,t})$ is the global cost function, and $g_t(x_t) = \sum_{i=1}^n g_{i,t}(x_{i,t})$ is the coupled constraint function. In order to guarantee that problem (2) is feasible, we assume that for any $T \in \mathbb{N}_+$, the set of all feasible sequences $\mathcal{X}_T = \{(x_1, \dots, x_T) :\ x_t \in \mathbb{X},\ g_t(x_t) \leq \mathbf{0}_m,\ t \in [T]\}$ is non-empty. With this standing assumption, an optimal sequence to (2) always exists.

It is interesting and challenging to consider the problem (2) in a bandit online setting and propose distributed algorithms to solve it. For a distributed bandit online algorithm, at time t, each agent i selects a decision $x_{i,t} \in \mathbb{X}_i$. After the selection, the agent can sample partial information of its cost function $f_{i,t}$ and constraint function $g_{i,t}$ at some points. At the same moment, due to the lack of other agents' cost and constraint function information, the agents exchange data (to be determined in the next section) with their neighbors over a time-varying directed graph. For bandit online algorithms, expected regret and constraint violation are commonly used as performance metrics. The expected regret is the expected accumulation over time of the loss difference between the decision determined by the algorithm and the static optimal decision if the sequences of cost and constraint functions are known in advance. Specifically, the efficacy of a decision sequence $\boldsymbol{x}_T = (x_1, \dots, x_T)$ is characterized by the expected regret

$$\mathrm{Reg}(\boldsymbol{x}_T) = E[\sum_{t=1}^{T} f_t(x_t)] - \sum_{t=1}^{T} f_t(x_T^*),$$

where $x_T^* \in \mathbb{X}$ such that $(x_T^*, \dots, x_T^*) = \arg\min_{\boldsymbol{x}_T \in \mathcal{X}_T} \sum_{t=1}^{T} f_t(x_t)$ with $\mathcal{X}_T = \{(x, \dots, x) : x \in \mathbb{X},\ g_t(x) \le \mathbf{0}_m,\ t = 1, \dots, T\}$ is the set of feasible static sequences. In order to guarantee the existence of x_T^*, we assume that for any $T \in \mathbb{N}_+$, $\mathcal{X}_T$ is non-empty. Moreover, the constraint violation measure is

$$\|[\sum_{t=1}^{T} g_t(x_t)]_+\|.$$

This definition implicitly allows constraint violations at some times to be compensated by strictly feasible decisions at other times. This is appropriate for constraints that have a cumulative nature such as energy budgets enforced through average power constraints.

In this paper, we consider the problem (2) in the bandit setting. Specifically, at each period each agent can sample the value of its local cost function at two points and can observe the value and the exact gradient of its local constraint function at one point. In other words, this paper can be viewed as an extension of the problem considered in [10], from full information feedback of the cost functions to bandit feedback. Bandit feedback is suitable to model many applications, where the gradient information is not available, such as online source localization, online routing in data networks, and online advertisement placement [7].

To end this section, we make the following assumptions on the cost and constraint functions.

Assumption 2. *(i) Each set $\mathbb{X}_i$ is convex and closed. Moreover, there exist $r_i > 0$ and $R_i > 0$ such that $r_i\mathbb{B}^{p_i} \subseteq \mathbb{X}_i \subseteq R_i\mathbb{B}^{p_i}$. (ii) $\{f_{i,t}\}$ and $\{g_{i,t}\}$ are convex and uniformly bounded on $\mathbb{X}_i$, i.e., there exist constants $F_f > 0$ and $F_g > 0$ such that $|f_{i,t}(x)| \le F_f$ and $\|g_{i,t}(x)\| \le F_g$ for all $t \in \mathbb{N}_+,\ i \in [n],\ x \in \mathbb{X}_i$. (iii) $\{\nabla f_{i,t}\}$ and $\{\nabla g_{i,t}\}$ exist and they are uniformly bounded on $\mathbb{X}_i$, i.e., there exist constants $G_f > 0$ and $G_g > 0$ such that $\|\nabla f_{i,t}(x)\| \le G_f$ and $\|\nabla g_{i,t}(x)\| \le G_g$ for all $t \in \mathbb{N}_+,\ i \in [n],\ x \in \mathbb{X}_i$.*

Assumption 2 is common in the literature on bandit online convex optimization, see, e.g., [29], [30].

Assumption 3. *(Slater's condition) There exists a constant $\varepsilon > 0$ and a vector $x_s \in \mathbb{X}$, such that $g_t(x_s) \le -\varepsilon\mathbf{1}_m,\ \forall t \in \mathbb{N}_+$.*

Assumption 3 is slightly weaker than the Assumption (as3) used in [30] since the later requires $x_s \in (1-\xi)\mathbb{X}$ for a small $\xi > 0$.

A. Motivating Example

As a motivating example, consider a power grid with n power generation units. Each unit i has p_i conventional and renewable power generators. The units can communicate through the information infrastructure. At stage t, let $x_{i,t} \in \mathbb{X}_i$ and $\mathbb{X}_i \subset \mathbb{R}^{p_i}$ be the output and the set of feasible outputs of the generators in unit i, respectively. To generate the output, each unit i suffers a cost $f_{i,t}(x_{i,t})$. This local cost $f_{i,t}$ is usually unknown in advance, since fossil fuel price is fluctuating and renewable energy is uncertain and unpredictable. Except the local generator limit constraints $\mathbb{X}_i$, all units need to cooperatively take into account global constraints, such as power balance and emission constraints. The global constraints can be modelled as $\sum_{i=1}^{n} g_{i,t}(x_{i,t}) \le \mathbf{0}_m$, where $g_{i,t}$ is unit i's local constraint function. The goal of the units is to reduce the global cost while satisfying the constraints.

IV. Main Results

In this section, we propose a distributed bandit online primal-dual algorithm to solve the problem (2) and derive expected regret and constraint violation bounds for this algorithm. For space purposes, all proofs are omitted, but can be found in [34].

A. Distributed Bandit Online Primal-Dual Algorithm

The proposed distributed bandit online algorithm is given in pseudo-code as Algorithm 1. Note that each agent i maintains three local sequences: the local primal decision variable sequence $\{x_{i,t}\} \subseteq \mathbb{X}_i$, the local dual variable sequence $\{q_{i,t}\} \subseteq \mathbb{R}_+^m$, and the estimator of the average of local dual variables $\{\tilde{q}_{i,t}\} \subseteq \mathbb{R}_+^m$. They are initialized by an arbitrary $x_{i,1} \in \mathbb{X}_i$ and $q_{i,1} = \tilde{q}_{i,1} = \mathbf{0}_m$, and updated recursively using the update rules (3a)–(3c). In (3b), $a_{i,t}$ is the updating direction information for the local primal variable defined as

$$a_{i,t} = \hat{\nabla}_2 f_{i,t-1}(x_{i,t-1}) + (\nabla g_{i,t-1}(x_{i,t-1}))^\top \tilde{q}_{i,t}. \tag{4}$$

The intuition of the update rules (3a)–(3c) is as follows. The augmented Lagrangian function associated with the constrained optimization problem with cost function f and constraint function g is

$$\mathcal{A}(x, \mu) = f(x) + \mu^\top g(x) - \frac{\beta}{2}\|\mu\|^2, \tag{5}$$

where $\mu \in \mathbb{R}_+^m$ is the Lagrange multiplier and $\beta > 0$ is the regularization parameter. $\mathcal{A}(x, \mu)$ is a convex-concave function. A standard primal-dual algorithm to find its saddle point is

$$x_{k+1} = \mathcal{P}_{\mathbb{X}}(x_k - \alpha(\nabla f(x_k) + (\nabla g(x_k))^\top \mu_k)), \tag{6a}$$

Algorithm 1 Distributed Bandit Online Primal-Dual Descent

```
Input: non-increasing stepsize sequences $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\gamma_t\} \subseteq (0,1]$; non-increasing shrinkage coefficients $\{\xi_{i,t}\} \subseteq (0,1),\ i \in [n]$; exploration parameters $\{\delta_{i,t}\} \subseteq (0, r_i\xi_{i,t}],\ i \in [n]$.
Initialize: $x_{i,1} \in (1-\xi_{i,1})\mathbb{X}_i$ and $q_{i,1} = \mathbf{0}_m,\ \forall i \in [n]$.
for $t = 2, \dots, T$ do
    for $i \in [n]$ in parallel do
        Select vector $u_{i,t-1} \in \mathbb{S}^{p_i}$ independently and uniformly at random.
        Sample $f_{i,t-1}(x_{i,t-1} + \delta_{i,t-1}u_{i,t-1})$ and $f_{i,t-1}(x_{i,t-1})$ and observe $g_{i,t-1}(x_{i,t-1})$ and $\nabla g_{i,t-1}(x_{i,t-1})$.
        Receive $[W_{t-1}]_{ij}q_{j,t-1}$ from $j \in \mathcal{N}_i^{\mathrm{in}}(\mathcal{G}_{t-1})$.
        Update
            $\tilde{q}_{i,t} = \sum_{j=1}^{n} [W_{t-1}]_{ij} q_{j,t-1},$   (3a)
            $x_{i,t} = \mathcal{P}_{(1-\xi_{i,t})\mathbb{X}_i}(x_{i,t-1} - \alpha_t a_{i,t}),$   (3b)
            $q_{i,t} = [\tilde{q}_{i,t} + \gamma_t(g_{i,t-1}(x_{i,t-1}) - \beta_t \tilde{q}_{i,t})]_+.$   (3c)
        Broadcast $q_{i,t}$ to $\mathcal{N}_i^{\mathrm{out}}(\mathcal{G}_t)$.
    end for
end for
Output: $\boldsymbol{x}_T$.
```

$$\mu_{k+1} = [\mu_k + \gamma(g(x_k) - \beta\mu_k)]_+, \tag{6b}$$

where $\alpha > 0$ and $\gamma > 0$ are the stepsizes used in the primal and dual updates, respectively. The update rules (3a)–(3c) are the distributed, online, and gradient-free extensions of (6a) and (6b). The term $-\beta_t\tilde{q}_{i,t}$ in (3c) is derived from penalty term $-\frac{\beta}{2}\|\mu\|^2$ in the augmented Lagrangian function (5) and it has an important role to guarantee that the dual variable sequence is not growing too large as shown in the proof of Lemma 1 given in [34].

In Algorithm 1, the data exchanged between agents is the local dual variable rather than any information related to the local primal variable, so our algorithm is well suited to account for privacy requirements.

B. Expected Regret and Constraint Violation Bounds

The following lemma provides the expected regret and constraint violation bounds for the general case.

Lemma 1. *Suppose Assumptions 1–2 hold. For any $T \in \mathbb{N}_+$, let $\boldsymbol{x}_T$ be the sequence generated by Algorithm 1. Then,*

$$\begin{aligned}
&\mathrm{Reg}(\boldsymbol{x}_T)\\
&\le \sum_{t=1}^{T} E[d_1(t)] + C_1 \sum_{t=1}^{T} \gamma_{t+1} + npG_f^2 \sum_{t=1}^{T} \alpha_{t+1}\\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n} \eta_t E[\|q_{i,t}\|^2] + \frac{2nR_{\max}^2}{\alpha_{T+1}},
\end{aligned} \tag{7a}$$

$$\begin{aligned}
&\|[\sum_{t=1}^{T} g_t(x_t)]_+\|^2\\
&\le d_0(T)\Big\{ \sum_{t=1}^{T} d_2(t) + C_1 \sum_{t=1}^{T} \gamma_{t+1} + \frac{2nR_{\max}^2}{\alpha_{T+1}}\\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n} \eta_t \|q_{i,t} - q^*\|^2 + 2pG_f R_{\max} T \Big\},
\end{aligned} \tag{7b}$$

where $C_1 = \frac{3n(2+n\tau)F_g^2}{1-\lambda} + 2nF_g^2$, $\tau = (1 - w/2n^2)^{-2} > 1$, $\lambda = (1 - w/2n^2)^{1/\iota}$, $\eta_t = \frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} - \beta_{t+1} + 4G_g^2\alpha_{t+1}$, and $R_{\max} = \max_{i\in[n]}\{R_i\}$, and, w and ι are given in Assumption 1, and $\{R_i\}$ are given in Assumption 2. Furthermore,

$$\begin{aligned}
d_0(T) &= 2n(\frac{1}{\gamma_1} + \sum_{t=1}^{T}(4G_g^2\alpha_{t+1} + \beta_{t+1})),\\
d_1(t) &= \sum_{i=1}^{n}(G_f\delta_{i,t} + G_f R_{\max}\xi_{i,t}) + d_2(t),\\
d_2(t) &= \sum_{i=1}^{n}(\frac{2R_{\max}^2(\xi_{i,t} - \xi_{i,t+1})}{\alpha_{t+1}} + G_g R_{\max}\xi_{i,t}\|\tilde{q}_{i,t+1}\|),
\end{aligned}$$

and $q^* = \frac{2[\sum_{t=1}^{T} g_t(x_t)]_+}{d_0(T)} \in \mathbb{R}_+^m$.

In (7a) and (7b), only two terms $d_1(t)$ and $d_2(t)$ depend on the shrinkage coefficients ($\xi_{i,t},\ i \in [n]$) and the exploration parameters ($\delta_{i,t},\ i \in [n]$) which are used to calculate the gradient estimator. Thus, the two terms are zero if the accurate gradient is used. In other words, we can regard $d_1(t)$ and $d_2(t)$ as the error caused by the inaccuracy of the gradient estimator. Note that the dependence on the stepsize sequences (α_t, β_t, and γ_t), shrinkage coefficients ($\xi_{i,t},\ i \in [n]$), exploration parameters ($\delta_{i,t},\ i \in [n]$), the number of agents n, and the network connectivity (w and ι), are all characterized in (7a) and (7b). In order to obtain sublinear expected regret and constraint violation bounds, the stepsize sequences, shrinkage coefficients, and exploration parameters should be properly chosen. Firstly, note that α_t appears in both the denominator and numerator of (7a) and (7b), so we should let $\alpha_t = \mathcal{O}(\frac{1}{t^c})$ with $c \in (0,1)$ because otherwise one of the terms that contained α_t will grow linearly or superlinearly. Note that it is not clear whether the dual sequence is bounded or not, so we should let $\eta_t \le 0$. Finally, note that $\xi_{i,t}$ and $\delta_{i,t}$ only appear in the numerator, so we should let them be as small as possible.

The following theorem characterizes expected regret and constraint violation bounds based on such selected stepsizes, shrinkage coefficients, and exploration parameters.

Theorem 1. *Suppose Assumptions 1–2 hold. For any $T \in \mathbb{N}_+$, let $\boldsymbol{x}_T$ be the sequence generated by Algorithm 1 with*

$$\begin{aligned}
&\alpha_t = \frac{1}{t^c},\ \beta_t = \frac{4G_g^2 + 1}{t^c},\ \gamma_t = \frac{1}{t^{1-c}},\\
&\xi_{i,t} = \frac{1}{t+1},\ \delta_{i,t} = \frac{r_i}{t+1},\ \forall t \in \mathbb{N}_+,
\end{aligned} \tag{8}$$

where $c \in [0.5, 1)$ is a constant. Then,

$$\mathrm{Reg}(\boldsymbol{x}_T) \le C_2 T^c, \tag{9a}$$

$$\|[\sum_{t=1}^{T} g_t(x_t)]_+\| \leq \sqrt{C_3}T^{1-c/2}, \tag{9b}$$

where $C_2 = 4nG_fR_{\max} + C_{2,1}$, $C_{2,1} = \frac{2nF_gG_gR_{\max}}{c(4G_g^2+1)} + \frac{C_1}{c} + \frac{np^2G_f^2}{1-c} + 8nR_{\max}^2$, $C_3 = C_{3,1}(2pG_fR_{\max} + C_{2,1})$, *and* $C_{3,1} = 2n(1 + \frac{8G_g^2+1}{1-c})$ *are constants independent of* T.

Remark 1. *The parameter* c *in Theorem 1 is a user-defined parameter which enables the trade-off between the expected regret bound and the constraint violation bound. For example, setting* $c = 0.5$ *gives* $\mathrm{Reg}(\boldsymbol{x}_T) = \mathcal{O}(\sqrt{T})$ *and* $\|[\sum_{t=1}^{T} g_t(x_t)]_+\| = \mathcal{O}(T^{3/4})$. *These two bounds are the same as the bounds achieved in [8], [9]. So in average sense, Algorithm 1 is as efficient as the algorithms proposed in [8], [9]. However, [8], [9] are in full-information setting and the algorithms proposed in them are centralized. Moreover, the constraint functions considered in [8] are time-invariant.*

From (9b), we see that the constraint violation bound is strictly greater than $\mathcal{O}(\sqrt{T})$ since $c < 1$. In the following theorem we show that $\mathcal{O}(\sqrt{T})$ bound on constraint violation can be achieved if Slater's condition holds.

Theorem 2. *Suppose Assumptions 1–3 hold. For any* $T \in \mathbb{N}_+$, *let* $\boldsymbol{x}_T$ *be the sequence generated by Algorithm 1 with*

$$\alpha_t = \frac{1}{\sqrt{t}},\ \beta_t = \frac{4G_g^2+1}{\sqrt{t}},\ \gamma_t = \frac{1}{\sqrt{t}},$$
$$\xi_{i,t} = \frac{1}{t+1},\ \delta_{i,t} = \frac{r_i}{t+1},\ \forall t \in \mathbb{N}_+. \tag{10}$$

Then,

$$\mathrm{Reg}(\boldsymbol{x}_T) \leq C_4\sqrt{T}, \tag{11a}$$
$$\|[\sum_{t=1}^{T} g_t(x_t)]_+\| \leq 4\sqrt{n}B_1\sqrt{T}, \tag{11b}$$

where $C_4 = 4G_fR_{\max} + 2B_1F_gG_gR_{\max} + 2C_1 + 2np^2G_f^2 + 8nR_{\max}^2$, $B_1 = \max\{\frac{2nB_2}{\varepsilon}, 2\sqrt{4nR_{\max}^2 + 2B_2}\}$, *and* $B_2 = C_1 + 2pG_fR_{\max} + 2nR_{\max}^2 + \frac{F_gG_gR_{\max}}{4G_g^2+1}$ *are constants independent of* T.

Remark 2. *From* (9a) *or* (11a), *we know that, no matter whether Slater's condition holds or not,* $\mathcal{O}(\sqrt{T})$ *expected regret is achieved by Algorithm 1. It was pointed out in [6] that* $\mathcal{O}(\sqrt{T})$ *is a tight bound to regret for online convex optimization problems in full-information setting, so in average sense, Algorithm 1 is as efficient as the optimal online algorithms using full-information. The same expected regret bound and the same expected regret as well as constraint violation bounds were also achieved by the two-query bandit algorithm in [25] and [30], respectively. However, in [25] static set constraints (rather than time-varying inequality constraints) were considered and in [30] a slightly stronger Slater's condition was assumed (see the discussion after Assumption 3 for details). Moreover, in [25], [30] the proposed algorithms are centralized (rather than distributed) and the total number of iterations needs to be known in advance to design the algorithm.*

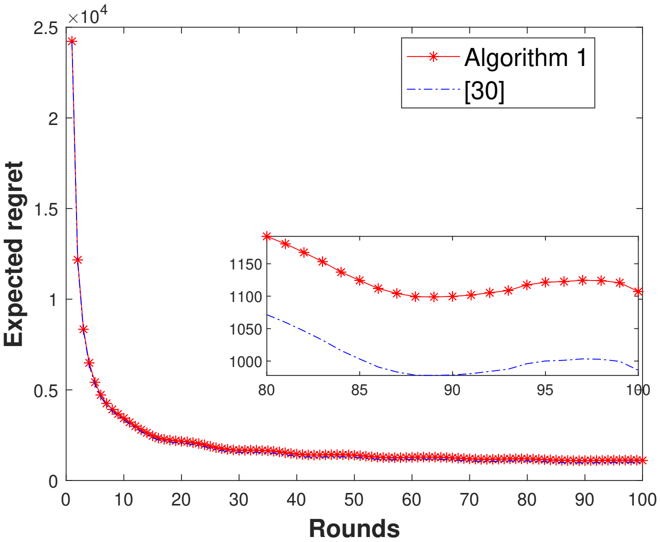

Fig. 1: Comparison of evolutions of the expected regret $\mathrm{Reg}(\boldsymbol{x}_T)/T$.

V. Numerical Simulations

This section evaluates the performance of Algorithm 1 in solving the power generation example introduced in Section III-A. The local cost and constraint functions are denoted

$$f_{i,t}(x_{i,t}) = x_{i,t}^\top \Pi_{i,t}^\top \Pi_{i,t} x_{i,t} + \langle \pi_{i,t}, x_{i,t} \rangle,$$
$$g_{i,t}(x_{i,t}) = x_{i,t}^\top \Phi_{i,t}^\top \Phi_{i,t} x_{i,t} + \langle \phi_{i,t}, x_{i,t} \rangle + c_{i,t},$$

where $\Pi_{i,t} \in \mathbb{R}^{p_i \times p_i}$, $\pi_{i,t} \in \mathbb{R}_+^{p_i}$, $\Phi_{i,t} \in \mathbb{R}^{p_i \times p_i}$, $\phi_{i,t} \in \mathbb{R}^{p_i}$, and $c_{i,t} \in \mathbb{R}$. At each time t, an undirected graph is used as the communication graph. Specifically, connections between vertices are random and the probability of two vertices being connected is $\rho > 0$. Moreover, edges $(i, i+1)$, $i \in [n-1]$ are added and $[W_t]_{ij} = 1/n$ if $(j,i) \in \mathcal{E}_t$ and $[W_t]_{ii} = 1 - \sum_{j \in \mathcal{N}_i^{\mathrm{in}}(\mathcal{G}_t)} [W_t]_{ij}$. The parameters are set as: $n = 50$, $m = 1$, $p_i = 6$, $\mathbb{X}_i = [-10, 10]^{p_i}$, and $\rho = 0.2$. Each element of $\Pi_{i,t}$, $\pi_{i,t}$, $\Phi_{i,t}$, $\phi_{i,t}$, and $c_{i,t}$ are drawn from the discrete uniform distribution in $[-5, 5]$, $[0, 10]$, $[-5, 5]$, $[-5, 5]$, and $[-5, -1]$, respectively. Under above settings, Assumptions 1–2 hold.

Since there are no other distributed bandit online algorithms to solve the problem of online optimization with time-varying coupled inequality constraints, we compare our Algorithm 1 with the centralized two-query bandit algorithm in [30]. Figs. 1 and 2 show the evolutions of $\mathrm{Reg}(\boldsymbol{x}_T)/T$ and $\|[\sum_{t=1}^{T} g_t(x_t)]_+\|/T$, respectively. The average is taken over 100 realizations. From Figs. 1 and 2, we see that our proposed distributed algorithm achieves comparable results as the centralized algorithm proposed in [30].

VI. Conclusions

In this paper, we considered a distributed bandit online convex optimization problem with time-varying coupled inequality constraints. We proposed a distributed bandit online

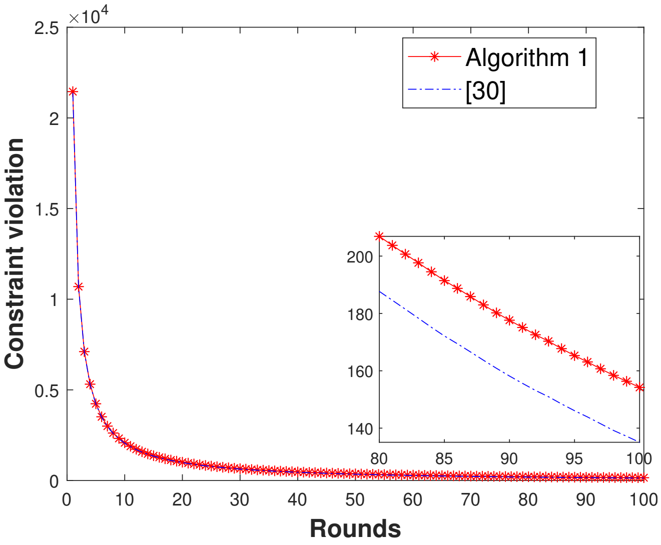

Fig. 2: Comparison of evolutions of the constraint violation $\|[\sum_{t=1}^{T} g_t(x_t)]_+\|/T$.

primal-dual algorithm to solve this problem, where a two-query procedure was used to estimate the gradients of the cost functions. We showed that sublinear expected regret and constraint violation can be achieved. We also showed that the results in this paper can be cast as non-trivial extensions of existing literature on online optimization and bandit feedback. An interesting future research directions is to consider an adaptive choice of the number of queries per period by different agents.

References

[1] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth, "Worst-case quadratic loss bounds for prediction using linear functions and gradient descent," *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 604–619, 1996.

[2] C. Gentile and M. K. Warmuth, "Linear hinge loss and average margin," in *Advances in Neural Information Processing Systems*, 1999, pp. 225–231.

[3] G. J. Gordon, "Regret bounds for prediction problems," in *Conference on Learning Theory*, 1999, pp. 29–40.

[4] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *International Conference on Machine Learning*, 2003, pp. 928–936.

[5] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[6] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2-3, pp. 169–192, 2007.

[7] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[8] R. Jenatton, J. Huang, and C. Archambeau, "Adaptive algorithms for online convex optimization with long-term constraints," in *International Conference on Machine Learning*, 2016, pp. 402–411.

[9] W. Sun, D. Dey, and A. Kapoor, "Safety-aware algorithms for adversarial contextual bandit," in *International Conference on Machine Learning*, 2017, pp. 3280–3288.

[10] X. Yi, X. Li, L. Xie, and K. H. Johansson, "Distributed online convex optimization with time-varying coupled inequality constraints," *IEEE Transactions on Signal Processing*, vol. 68, pp. 731–746, 2020.

[11] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[12] R. D. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Advances in Neural Information Processing Systems*, 2005, pp. 697–704.

[13] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005, pp. 385–394.

[14] V. Dani, S. M. Kakade, and T. P. Hayes, "The price of bandit information for online optimization," in *Advances in Neural Information Processing Systems*, 2008, pp. 345–352.

[15] J. D. Abernethy, E. Hazan, and A. Rakhlin, "Competing in the dark: An efficient algorithm for bandit linear optimization," in *Conference on Learning Theory*, 2008, pp. 263–273.

[16] ——, "Interior-point methods for full-information and bandit online learning," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4164–4175, 2012.

[17] A. Saha and A. Tewari, "Improved regret guarantees for online smooth convex optimization with bandit feedback," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 636–642.

[18] E. Hazan and K. Levy, "Bandit convex optimization: Towards tight bounds," in *Advances in Neural Information Processing Systems*, 2014, pp. 784–792.

[19] S. Bubeck, O. Dekel, T. Koren, and Y. Peres, "Bandit convex optimization: $\sqrt{T}$ regret in one dimension," in *Conference on Learning Theory*, 2015, pp. 266–278.

[20] S. Bubeck and R. Eldan, "Multi-scale exploration of convex functions and bandit convex optimization," in *Conference on Learning Theory*, 2016, pp. 583–589.

[21] E. Hazan and Y. Li, "An optimal algorithm for bandit convex optimization," *arXiv preprint arXiv:1603.04350*, 2016.

[22] A. V. Gasnikov, E. A. Krymova, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko, "Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case," *Automation and Remote Control*, vol. 78, no. 2, pp. 224–234, 2017.

[23] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback." in *Conference on Learning Theory*, 2010, pp. 28–40.

[24] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[25] O. Shamir, "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback," *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[26] T. Yang, L. Zhang, R. Jin, and J. Yi, "Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient," in *International Conference on Machine Learning*, 2016, pp. 449–457.

[27] T. Tatarenko and M. Kamgarpour, "Minimizing regret in bandit online optimization in unconstrained and constrained action spaces," *arXiv preprint arXiv:1806.05069*, 2018.

[28] I. Shames, D. Selvaratnam, and J. H. Manton, "Online optimization using zeroth order oracles," *IEEE Control Systems Letters*, vol. 4, no. 1, pp. 31–36, 2019.

[29] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: online convex optimization with long term constraints," *Journal of Machine Learning Research*, vol. 13, no. Sep, pp. 2503–2528, 2012.

[30] T. Chen and G. B. Giannakis, "Bandit convex optimization for scalable and dynamic iot management," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1276–1286, 2019.

[31] D. Yuan, D. W. Ho, Y. Hong, and G. Jiang, "Online bandit convex optimization over a network," in *Chinese Control Conference*, 2016, pp. 8090–8095.

[32] D. Yuan, A. Proutière, and G. Shi, "Distributed online linear regression," *arXiv preprint arXiv:1902.04774*, 2019.

[33] S. Lee and M. M. Zavlanos, "On the sublinear regret of distributed primal-dual algorithms for online constrained optimization," *arXiv preprint arXiv:1705.11128*, 2017.

[34] X. Yi, X. Li, T. Yang, L. Xie, T. Chai, and K. H. Johansson, "A distributed primal-dual algorithm for bandit online convex optimization with time-varying coupled inequality constraints," *https://people.kth.se/~xinleiy/papers/Bandit_ACC2020_Long.pdf*, 2020.