2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

Minimizing Regret in Unconstrained Online Convex Optimization

Tatiana Tatarenko[1] and Maryam Kamgarpour[2]

***Abstract*— We consider online convex optimizations in the bandit setting. The decision maker does not know the time-varying cost functions, or their gradients. At each time step, she observes the value of the cost function for her chosen action. The objective is to minimize the regret, that is, the difference between the sum of the costs she accumulates and that of the optimal action computable had she known the cost functions *a priori*. We present a novel algorithm in order to minimize the regret in an unconstrained action space. Our algorithm hinges on the idea of introducing randomization to approximate the gradients of the cost functions using only their observed values. We establish an almost sure regret bound for the mean values of actions and an expected regret bound for the actions.**

I. Introduction

Decision making under uncertainty is at the core of control theory. Whereas in classical control the uncertainty enters in the structure of the dynamical system or as disturbances acting on the system, in several complex applications, arising for example in reinforcement learning or learning over a (routing, traffic, power grid) network, the dynamics and cost functions may themselves be unknown. The controller gathers information about the system sequentially, by playing an action and observing the cost corresponding to the played action. The problem of choosing the actions based on the obtained cost information is referred to as online optimization.

The early formulations of online optimization include the multi-armed bandit problem [12], [6]. Here, the decision maker (gambler) at each time step can make a choice from finitely many actions (pulling one of the K arms of a gambling machine). For this choice, she receives a cost. The *regret minimization* problem is how to use the observed costs to minimize the difference between the sum of the costs she acquires and that of the optimal action, had she known the cost functions (winning probability of each machine) *a priori*. Ever since its original formulation, the multi-armed bandit problem has been explored extensively, mainly in the machine learning but also in the control community [14], [2].

A natural extension of multi-armed bandit problems, more relevant to control and reinforcement learning, is considering a continuous action space. In the continuous action space, under the stochastic model setup, the cost functions obey a probability distribution. Here, a common assumption is that only the cost function for the played actions are observed (zero order oracle). Then, algorithms inspired by stochastic approximation procedures can be used to derive regret bounds [3]. In the adversarial setup considered in the celebrated work of Zinkevich [16], the cost functions are convex and time-varying. At each time step the decision maker chooses an action from a compact action space and observes the gradient of the time-varying cost function evaluated at this played action (first order oracle). Within the control community, this formulation has been extended to network online optimization in [5], [13]. In these works, agents communicate over a graph and cooperatively minimize the regret, under the assumption of observing the sub-gradients of cost functions at played actions [5] or having common knowledge of the minimizer's dynamics [13]. Furthermore, online optimization has been used in applications ranging from human decision making [11] to electricity markets [7].

The formulation of zero order oracle is a natural setup for reinforcement learning and multi-agent problems, in which the decision maker(s) can only observe the cost functions for a played action. Algorithms based on zero order information for classical optimization problems over compact sets are presented in [8], [10]. As for online optimization, the work in [4] considers also compact action spaces and proposes a regret minimizing algorithm based purely on zero order information. However, to the best of our knowledge, the zero order online optimization formulations have not dealt with the rather elementary setup of an unconstrained action space.

Our contribution is to propose a regret minimizing algorithm for the unconstrained continuous action space online optimization problems, in the adversarial setup and under the zero order oracle assumption. Our novel algorithm is inspired by the approach of continuous learning automata [15]. In particular, by playing a randomized strategy from a properly chosen Gaussian distribution, we derive samples of the gradients of the cost functions, in a relaxed decision space. We then define a stochastic gradient type update law. By properly choosing the step sizes of the algorithm and the variances of the Gaussian distribution, we provide almost sure bounds on regret for the mean values of actions. This, in turn provides an expected regret bound for the actions.

Our paper is organized as follows. In Section II we formulate the problem. In Section III we propose our algorithm and prove its regret bound. We present a numerical case study in Section IV and end with concluding remarks in V.

Notation
The standard inner product on $\mathbb{R}^n$ is denoted by $(\cdot,\cdot)\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, with the associated norm $\|\boldsymbol{x}\| := \sqrt{(\boldsymbol{x},\boldsymbol{x})}$. We say that a function $f(\boldsymbol{x})$ grows not faster than a function $g(\boldsymbol{x})$ as $\boldsymbol{x} \to \infty$, if there exists a positive constant Q such that $f(\boldsymbol{x}) \leq g(\boldsymbol{x})$ for any $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \geq Q$.

[1] Tatiana Tatarenko is with the Department of Control Theory and Robotics, TU Darmstadt, Germany `tatarenk@rmr.tu-darmstadt.de`. [2] M. Kamgarpour is with the Automatic Control Laboratory, ETH Zürich, Switzerland `maryamk@ethz.ch`. This work is gratefully supported by ERC Starting Grant CONENE.

978-3-9524-2699-9 ©2018 EUCA

II. Online Optimization Problem

A. Problem formulation

An unconstrained online convex optimization problem consists of an infinite sequence $\{c_1, c_2, \ldots\}$, where each $c_t : \mathbb{R}^n \to \mathbb{R}$ is a convex function. At each time step t, an online convex programming algorithm selects a vector $\mathbf{x}_t \in \mathbb{R}^n$. After the vector is selected, it receives the cost function $\hat{c}_t = c_t(\mathbf{x}_t)$. Efficiency of any online optimization is measured with respect to a regret function defined below.

Definition 1: Given an algorithm updating $\{\mathbf{x}_t\}$, and a convex programming problem $\{c_1, c_2, \ldots\}$, if $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ are the vectors selected by this algorithm, then the cost of the algorithm until time T is

$$C(T) = \sum_{t=1}^{T} c_t(\mathbf{x}_t).$$

The cost of a static solution $\mathbf{x} \in \mathbb{R}^n$ until time T is

$$C(\mathbf{x}, T) = \sum_{t=1}^{T} c_t(\mathbf{x}).$$

The regret of the algorithm until time T is

$$R(T) = C(T) - \min_{\mathbf{x} \in \mathbb{R}^n} C(\mathbf{x}, T). \tag{1}$$

The goal of the online optimization is to propose a procedure for the update of $\{\mathbf{x}_t\}$ such that the average regret function approaches zero with time, i.e.

$$\overline{\lim_{T\to\infty}} \frac{R(T)}{T} = \lim_{T\to\infty} \sup \frac{R(T)}{T} \leq 0,$$

with a sufficiently fast sublinear rate.

To address this problem, we need the following standard assumptions [16], [4].

Assumption 1: The convex functions $c_t(\boldsymbol{x})$, $t = 1, 2, \ldots$, are differentiable and the gradients ∇c_t are uniformly bounded on $\mathbb{R}^n$. Moreover, each gradient ∇c_t is Lipschitz continuous on $\mathbb{R}^n$ with some constant L_t and there exists L such that $L_t < L$ for all t.

Remark 1: Since Assumption 1 requires the uniform boundedness of ∇c_t, the functions $c_t(\boldsymbol{x})$, $t = 1, 2, \ldots$, grow not faster than a linear function as $\|\boldsymbol{x}\| \to \infty$. Thus, these functions are Lipschitz continuous with some constant l_t uniformly bounded by some constant l.

Moreover, we require the following behavior of the cost functions on infinity.

Assumption 2: There exists a finite constant $K > 0$ such that for all $t = 1, 2, \ldots$, $(\boldsymbol{x}, \nabla c_t(\boldsymbol{x})) > 0$, $\forall \|\boldsymbol{x}\|^2 > K$.

Remark 2: Note that under Assumption 1, Assumption 2 is equivalent to the assumption that all functions $c_t(\boldsymbol{x})$, $t = 1, 2, \ldots$, achieve their minima in some compact set. Indeed, if the latter assumption holds, due to convexity of c_t on $\mathbb{R}^n$, the function c_t is coercive, namely $\lim_{\|\boldsymbol{x}\|\to\infty} c_t(\boldsymbol{x}) = \infty$. Hence, there exists K such that $c_t(\boldsymbol{x}) > c_t(\mathbf{0})$ for all $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|^2 > K$. Thus,

$$c_t(\mathbf{0}) \geq c_t(\boldsymbol{x}) + (\nabla c_t(\boldsymbol{x}), \mathbf{0} - \boldsymbol{x}) > c_t(\mathbf{0}) - (\nabla c_t(\boldsymbol{x}), \boldsymbol{x})$$

for any such $\boldsymbol{x}$. On the other hand, if $(\boldsymbol{x}, \nabla c_t(\boldsymbol{x})) > 0$, $\forall \|\boldsymbol{x}\|^2 > K$, there is no minima of c_t on the set $\{\boldsymbol{x} : \|\boldsymbol{x}\|^2 > K\}$ (due to the first order optimality condition). Next, the continuous functino c_t iattains its minimum on the compact set $\{\boldsymbol{x} : \|\boldsymbol{x}\|^2 \leq K\}$. However, the value K is not known *a priori* and, thus, cannot be used in the algorithm's design.

B. Supporting Theorems

To prove convergence of the algorithm we will use the results on convergence properties of the Robbins-Monro stochastic approximation procedure analyzed in [9].

We start by introducing some important notation. Let $\{\mathbf{X}(t)\}_t$, $t \in \mathbb{Z}_+$, be a discrete-time Markov process on some state space $E \subseteq \mathbb{R}^d$, namely $\mathbf{X}(t) = \mathbf{X}(t, \omega) : \mathbb{Z}_+ \times \Omega \to E$, where Ω is the sample space of the probability space on which the process $\mathbf{X}(t)$ is defined. The transition function of this chain, namely $\Pr\{\mathbf{X}(t+1) \in \Gamma | \mathbf{X}(t) = \mathbf{X}\}$, is denoted by $P(t, \mathbf{X}, t+1, \Gamma)$, $\Gamma \subseteq E$.

Definition 2: The operator L defined on the set of measurable functions $V : \mathbb{Z}_+ \times E \to \mathbb{R}$, $\mathbf{X} \in E$, by

$$\begin{aligned} LV(t, \mathbf{X}) &= \int P(t, \mathbf{X}, t+1, dy)[V(t+1, y) - V(t, \mathbf{X})] \\ &= E[V(t+1, \mathbf{X}(t+1)) \mid \mathbf{X}(t) = \mathbf{X}] - V(t, \mathbf{X}), \end{aligned}$$

is called a *generating operator* of a Markov process $\{\mathbf{X}(t)\}_t$.

Now, we recall the following theorem for discrete-time Markov processes, which is proven in [9], Theorem 2.5.2.

Theorem 1: Consider a Markov process $\{\mathbf{X}(t)\}_t$ and suppose that there exists a function $V(t, \mathbf{X}) \geq 0$ such that $\inf_{t\geq 0} V(t, \mathbf{X}) \to \infty$ as $\|\mathbf{X}\| \to \infty$ and

$$LV(t, \mathbf{X}) \leq -\alpha(t+1)\psi(t, \mathbf{X}) + f(t)(1 + V(t, \mathbf{X})),$$

where $\psi \geq 0$ on $\mathbb{R} \times \mathbb{R}^d$, $f(t) > 0$, $\sum_{t=0}^{\infty} f(t) < \infty$. Let $\alpha(t)$ be such that $\alpha(t) > 0$, $\sum_{t=0}^{\infty} \alpha(t) = \infty$. Then, almost surely $\sup_{t\geq 0} \|\mathbf{X}(t, \omega)\| = R(\omega) < \infty$.

III. Proposed Online Optimization Algorithm

We propose an algorithm for the online optimization problem that at each time step t uses only the experienced value of the function, $c_t(\mathbf{x}_t)$ for a played action $\mathbf{x}_t$. Then, we derive the regret bounds for our approach.

A. Randomization to minimize regret

The algorithm is as follows. At each time t the decision maker chooses the vector $\mathbf{x}_t$ according to the n-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_t, \sigma_t)$, meaning that the coordinates $x_t^1, \ldots, x_t^n$ of the random vector $\mathbf{x}_t$ are independently distributed with the variance σ_t and the mean values $\mu_t^1 \ldots, \mu_t^n$. The mean value initial condition $\boldsymbol{\mu}_0$, can be chosen arbitrarily. Its iterates $\boldsymbol{\mu}_t$, are updated using the observed costs $\hat{c}_t = c_t(\mathbf{x}_t)$ and the played action $\mathbf{x}_t$ as follows

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}(\boldsymbol{\mu}_t, \sigma_t) \\ \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t - \alpha_t \hat{c}_t \frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\sigma_t^2}. \end{aligned} \tag{2}$$

Before presenting our regret bound on the above procedure, we provide some insights into our proposed choice. In particular, we show that our procedure can be interpreted as a stochastic optimization. By introducing randomization in the algorithm, we obtain samples of the gradients of the cost functions in a relaxed decision space.

First, let $\mathrm{E}_{\mathbf{x}_t}\{\cdot\}$ denote the conditional expectation of some random variable with respect to the σ-algebra $\mathcal{F}_t$ generated by the random variables $\{\boldsymbol{\mu}_k, \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k)\}_{k\le t}$, i.e. $\mathrm{E}_{\mathbf{x}_t}\{\cdot\} = \mathrm{E}\{\cdot|\mathcal{F}_t\}$. Then, introduce $\nabla\tilde{c}_t(\boldsymbol{\mu}_t)$ as

$$\nabla\tilde{c}_t(\boldsymbol{\mu}_t, \sigma_t) = \int_{\mathbb{R}^n} \nabla c_t(\boldsymbol{x}) p(\boldsymbol{\mu}_t, \sigma_t, \boldsymbol{x}) d\boldsymbol{x},$$

where $\boldsymbol{x} = (x^1, \ldots, x^n)$ and

$$p(\boldsymbol{\mu}_t, \sigma_t, \boldsymbol{x}) = \frac{1}{(\sqrt{2\pi}\sigma_t)^n} \exp\left\{-\sum_{k=1}^{n} \frac{(x^k - \mu_t^k)^2}{2\sigma_t^2}\right\}$$

is the density of $\mathcal{N}(\boldsymbol{\mu}_t, \sigma_t)$. Thus, $\nabla\tilde{c}_t(\boldsymbol{\mu}_t, \sigma_t)$ can be considered the expectation $\mathrm{E}_{\mathbf{x}_t}\{\nabla c_t(\mathbf{x}_t)\}$ of the random variable $\nabla c_t(\mathbf{x}_t)$, given that $\mathbf{x}_t$ has the normal distribution $\mathcal{N}(\boldsymbol{\mu}_t, \sigma_t)$.

The iteration (2) can then be rewritten as a stochastic gradient descent

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \alpha_t \nabla c_t(\boldsymbol{\mu}_t) + \alpha_t Q_t(\boldsymbol{\mu}_t, \sigma_t) + \alpha_t \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t), \tag{3}$$

where

$$Q_t(\boldsymbol{\mu}_t, \sigma_t) = \nabla c_t(\boldsymbol{\mu}_t) - \nabla\tilde{c}_t(\boldsymbol{\mu}_t, \sigma_t),$$
$$\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t) = \nabla\tilde{c}_t(\boldsymbol{\mu}_t, \sigma_t) - \hat{c}_t \frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\sigma_t^2}.$$

Note that under Assumption 1, we can show that ξ_t is a Martingale difference

$$\mathrm{E}_{\mathbf{x}_t}\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t) = \nabla\tilde{c}_t(\boldsymbol{\mu}_t, \sigma_t) - \mathrm{E}_{\mathbf{x}_t}\left\{\hat{c}_t \frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\sigma_t^2}\right\} = 0. \tag{4}$$

Indeed the above holds since

$$\mathrm{E}_{\mathbf{x}_t}\left\{\hat{c}_t \frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\sigma_t^2}\right\} = \int_{\mathbb{R}^n} c_t(\boldsymbol{x}) \frac{\boldsymbol{x} - \boldsymbol{\mu}_t}{\sigma_t^2} p(\boldsymbol{\mu}_t, \sigma_t, \boldsymbol{x}) d\boldsymbol{x},$$

and furthermore, for any $k \in [n]$, with $x^{-k} = (x^1, \ldots, x^{k-1}, x^{k+1}, \ldots, x^n)$,

$$\begin{aligned}\int_{\mathbb{R}^n} c_t(\boldsymbol{x}) \frac{x^k - \mu_t^k}{\sigma_t^2} p(\boldsymbol{\mu}_t, \sigma_t, \boldsymbol{x}) d\boldsymbol{x} &= -\frac{1}{(\sqrt{2\pi}\sigma_t)^n} \\ \times \int_{\mathbb{R}^{n-1}} \left[\int_{x^k=-\infty}^{x^k=+\infty} c_t(\boldsymbol{x}) d\left(\exp\left\{-\sum_{k=1}^{n} \frac{(x^k - \mu_t^k)^2}{2\sigma_t^2}\right\}\right)\right] & \\ \times \exp\left\{-\sum_{j\neq k}^{n} \frac{(x^j - \mu_t^j)^2}{2\sigma_t^2}\right\} dx^{-k} & \\ = \int_{\mathbb{R}^n} \frac{\partial c_t(\boldsymbol{x})}{\partial x^k} p(\boldsymbol{\mu}_t, \sigma_t, \boldsymbol{x}) d\boldsymbol{x}, &\end{aligned}$$

where in the above, we use integration by parts and the fact that each c_t grows at most linearly as $\|\mathbf{x}\| \to \infty$ to get to the last equality. Equipped with the parallels of the proposed Algorithm (2) with stochastic gradient procedure (3), we are ready to present the regret bounds.

B. Derivation of regret bounds

Theorem 2: Let (2) define the optimization algorithm for the unconstrained online convex optimization problem $(\mathbb{R}^n, \{c_1, c_2, \ldots\})$. Choose the step sizes and variances according to $\{\alpha_t = \frac{1}{t^a}\}$, $\{\sigma_t = \frac{1}{t^b}\}$, where $0 < a < 1$, $b > 0$, $a + b > 1$, $2a - 2b > 1$. Then, under Assumptions 1-2,

- i) almost surely, the regret of the algorithm (2) estimated with respect to the mean parameters $\{\boldsymbol{\mu}_t\}$ denoted by $R_{\boldsymbol{\mu}}(T)$ (see (1), where $\mathbf{x}_t$ is replaced by $\boldsymbol{\mu}_t$) satisfies
$$\overline{\lim_{T\to\infty}} \frac{R_{\boldsymbol{\mu}}(T)}{T} \le 0.$$
- ii) the regret of algorithm (2) estimated with respect to the actions $\{\mathbf{x}_t\}$ satisfies
$$\overline{\lim_{T\to\infty}} \mathrm{E}\left\{\frac{R_{\mathbf{x}}(T)}{T}\right\} \le 0.$$

To prove the main theorem above, we first show that under the conditions of this theorem, the mean values $\{\boldsymbol{\mu}_t\}$ stay almost surely bounded during the process (2).

Lemma 1: Consider the optimization algorithm (2). Under Assumptions 1-2 and given $\{\alpha_t = \frac{1}{t^a}\}$, $\{\sigma_t = \frac{1}{t^b}\}$, where $0 < a < 1$, $b > 0$, $a + b > 1$, $2a - 2b > 1$,

$$\Pr\{\|\boldsymbol{\mu}_t\|, t = 1, 2, \ldots, \text{ is bounded}\} = 1.$$

In words, in the run of the algorithm $\|\boldsymbol{\mu}_t\|$ is bounded almost surely for any t.

Proof: First, we notice that the conditions on the sequences $\{\alpha_t\}$, $\{\sigma_t\}$ imply that

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \quad \sum_{t=1}^{\infty} \alpha_t\sigma_t < \infty, \quad \sum_{t=1}^{\infty} \frac{\alpha_t^2}{\sigma_t^2} < \infty. \tag{5}$$

Let us consider the function $V(\boldsymbol{\mu}) = W(\|\boldsymbol{\mu}\|^2)$, where $W : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$W(x) = \begin{cases} 0, & \text{if } x < K, \\ (x - K)^2, & \text{if } x \ge K, \end{cases} \tag{6}$$

where K is the constant from Assumption 2. The function W fulfils the following property

$$W(y) - W(x) \le W'(x)(y - x) + (y - x)^2. \tag{7}$$

Thus, taking this inequality into account, we obtain

$$\begin{aligned} V(\boldsymbol{\mu}_{t+1}) - V(\boldsymbol{\mu}_t) \le & W'(\|\boldsymbol{\mu}_t\|^2)(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}_t\|^2) \\ & + (\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}_t\|^2)^2. \end{aligned} \tag{8}$$

According to (3) norm of $\boldsymbol{\mu}_t$ evolves as

$$\begin{aligned} &\|\boldsymbol{\mu}_{t+1}\|^2 \\ &= \|\boldsymbol{\mu}_t - \alpha_t \nabla c_t(\boldsymbol{\mu}_t) + \alpha_t Q_t(\boldsymbol{\mu}_t, \sigma_t) + \alpha_t \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)\|^2 \\ &= \|\boldsymbol{\mu}_t\|^2 \\ &\quad + \alpha_t^2(\|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|^2 + \|\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)\|^2) \\ &\quad + 2\alpha_t(-\nabla c_t(\boldsymbol{\mu}_t) + Q_t(\boldsymbol{\mu}_t, \sigma_t) + \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t), \boldsymbol{\mu}_t) \\ &\quad - 2\alpha_t^2(\nabla c_t(\boldsymbol{\mu}_t), Q_t(\boldsymbol{\mu}_t, \sigma_t) + \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)) \\ &\quad + 2\alpha_t^2(Q_t(\boldsymbol{\mu}_t, \sigma_t), \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)). \end{aligned} \tag{9}$$

Hence, taking into account (4), we obtain

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}_t}\|\boldsymbol{\mu}_{t+1}\|^2 = \|\boldsymbol{\mu}_t\|^2 + \alpha_t^2 \\
&\times (\|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t,\sigma_t)\|^2 + \mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2) \\
&\quad - 2\alpha_t(\nabla c_t(\boldsymbol{\mu}_t),\boldsymbol{\mu}_t) + 2\alpha_t(Q_t(\boldsymbol{\mu}_t,\sigma_t),\boldsymbol{\mu}_t) \\
&\quad - 2\alpha_t^2(\nabla c_t(\boldsymbol{\mu}_t), Q_t(\boldsymbol{\mu}_t,\sigma_t)).
\end{aligned} \tag{10}$$

We proceed with estimation of the terms containing $Q_t(\boldsymbol{\mu}_t,\sigma_t)$ and $\mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2$. Taking into account Assumption 1, we can write:

$$\begin{aligned}
\|Q(\boldsymbol{\mu}_t,\sigma_t)\| &= \|\int_{\mathbb{R}^n}[\nabla c_t(\boldsymbol{\mu}_t) - \nabla c_t(\boldsymbol{x})]p(\boldsymbol{\mu}_t,\sigma_t,\boldsymbol{x})d\boldsymbol{x}\| \\
&\le \int_{\mathbb{R}^n}\|\nabla c_t(\boldsymbol{\mu}_t) - \nabla c_t(\boldsymbol{x})\|p(\boldsymbol{\mu}_t,\sigma_t,\boldsymbol{x})d\boldsymbol{x} \\
&\le \int_{\mathbb{R}^n} L\|\boldsymbol{\mu}_t - \boldsymbol{x}\|p(\boldsymbol{\mu}_t,\sigma_t,\boldsymbol{x})d\boldsymbol{x} \\
&\le \int_{\mathbb{R}^n} L\left(\sum_{i=1}^n|\boldsymbol{\mu}_t^i - x^i|\right)p(\boldsymbol{\mu},\boldsymbol{x})d\boldsymbol{x} = O(\sigma_t).
\end{aligned} \tag{11}$$

Since $\mathrm{E}\|\mathbf{X} - \mathrm{E}\mathbf{X}\|^2 \le \mathrm{E}\|\mathbf{X}\|^2$ and taking into account (4), we have

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}_t}\{\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2\} \\
&\le \sum_{i=1}^n\int_{\mathbb{R}^n}{c_t}^2(\mathbf{x})\frac{(x^i-\mu_t^i)^2}{\sigma_t^4}p(\boldsymbol{\mu}_t,\sigma_t,\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}$$

Thus, we can use Assumption 1 (see Remark 1) to get the next inequality:

$$\mathrm{E}_{\mathbf{x}_t}\{\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2\} \le \frac{f_1(\boldsymbol{\mu}_t,\sigma_t)}{\sigma_t^2}, \tag{12}$$

where $f_1(\boldsymbol{\mu}_t,\sigma_t)$ is a quadratic function of σ_t and μ_t^i, $i\in[n]$. Hence, due to Assumptions 1-2 and due to (8)-(12),

$$\begin{aligned}
LV(\boldsymbol{\mu}) &= \mathrm{E}\{V(\boldsymbol{\mu}_{t+1})|\boldsymbol{\mu}_t = \boldsymbol{\mu}\} - V(\boldsymbol{\mu}) \\
&\le (\mathrm{E}_{\mathbf{x}_t}\{\|\boldsymbol{\mu}_{t+1}\|^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\} - \|\boldsymbol{\mu}\|^2)W'(\|\boldsymbol{\mu}\|) \\
&\qquad + \mathrm{E}_{\mathbf{x}_t}\{(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\} \\
&\le -2\alpha_t(\nabla c_t(\boldsymbol{\mu}),\boldsymbol{\mu})W'(\|\boldsymbol{\mu}\|) \\
&\qquad + W'(\|\boldsymbol{\mu}\|)\alpha_t^2(\|\nabla c_t(\boldsymbol{\mu})\|^2 + \|Q_t(\boldsymbol{\mu},\sigma_t)\|^2 \\
&\qquad + \mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu},\sigma_t)\|^2) \\
&\qquad + 2W'(\|\boldsymbol{\mu}\|)\alpha_t^2\|\nabla c_t(\boldsymbol{\mu})\|\|Q_t(\boldsymbol{\mu},\sigma_t)\| \\
&\qquad + 2W'(\|\boldsymbol{\mu}\|)\alpha_t\|Q_t(\boldsymbol{\mu},\sigma_t)\|\|\boldsymbol{\mu}\| \\
&\qquad + \mathrm{E}_{\mathbf{x}_t}\{(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\} \\
&\le -2\alpha_t(\nabla c_t(\boldsymbol{\mu}),\boldsymbol{\mu})W'(\|\boldsymbol{\mu}\|) + g_1(t)(1 + V(\boldsymbol{\mu})) \\
&\qquad + \mathrm{E}_{\mathbf{x}_t}\{(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\},
\end{aligned} \tag{13}$$

where, due to the choice of the parameters α_t, σ_t, we have $g_1(t) = O\left(\frac{\alpha_t^2}{\sigma_t^2} + \alpha_t\sigma_t\right)$. Thus, according to the condition in (5), $\sum_{t=1}^\infty g_1(t) < \infty$.

Finally, we estimate the term $\mathrm{E}_{\mathbf{x}_t}\{(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\}$. According to (9)

$$\begin{aligned}
&(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2 \\
&= [\alpha_t^2(\|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t,\sigma_t)\|^2 + \|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2) \\
&\qquad + 2\alpha_t(-\nabla c_t(\boldsymbol{\mu}_t) + Q_t(\boldsymbol{\mu}_t,\sigma_t) + \xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t),\boldsymbol{\mu}_t) \\
&\qquad - 2\alpha_t^2(\nabla c_t(\boldsymbol{\mu}_t), Q_t(\boldsymbol{\mu}_t,\sigma_t) + \xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)) \\
&\qquad + 2\alpha_t^2(Q_t(\boldsymbol{\mu}_t,\sigma_t),\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t))]^2 \\
&\le [\alpha_t^2(\|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t,\sigma_t)\|^2 + \|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^2) \\
&\qquad + 2\alpha_t\|\boldsymbol{\mu}_t\|[\|\nabla c_t(\boldsymbol{\mu}_t)\| + \|Q_t(\boldsymbol{\mu}_t,\sigma_t)\| \\
&\qquad\qquad + \|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|] \\
&\qquad + 2\alpha_t^2\|\nabla c_t(\boldsymbol{\mu}_t)\|\|Q_t(\boldsymbol{\mu}_t,\sigma_t)\| \\
&\qquad + 2\alpha_t^2\|\nabla c_t(\boldsymbol{\mu}_t)\|\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\| \\
&\qquad + 2\alpha_t^2\|Q_t(\boldsymbol{\mu}_t,\sigma_t)\|\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|]^2.
\end{aligned} \tag{14}$$

From the equality above we can see that there are additional terms, namely $\mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^3$ and $\mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^4$, to be bounded. Here, analogously to (12) by using Assumption 1, Remark 1, and the properties of the normal distribution, one can demonstrate that

$$\mathrm{E}_{\mathbf{x}_t}\{\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^3\} \le \frac{f_2(\boldsymbol{\mu}_t,\sigma_t)}{\sigma_t^3}, \tag{15}$$

$$\mathrm{E}_{\mathbf{x}_t}\{\|\xi_t(\mathbf{x}_t,\boldsymbol{\mu}_t,\sigma_t)\|^4\} \le \frac{f_3(\boldsymbol{\mu}_t,\sigma_t)}{\sigma_t^4}, \tag{16}$$

where $f_2(\boldsymbol{\mu}_t,\sigma_t)$ and $f_3(\boldsymbol{\mu}_t,\sigma_t)$ are third and forth order polynomials of σ_t and μ_t^i, $i\in[n]$, respectively. Hence, by expanding (14) and taking into account (10)-(16) we get

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}_t}\{(\|\boldsymbol{\mu}_{t+1}\|^2 - \|\boldsymbol{\mu}\|^2)^2|\boldsymbol{\mu}_t = \boldsymbol{\mu}\} \\
&\le g_2(t)(1 + W(\|\boldsymbol{\mu}\|^2)) = g_2(t)(1 + V(\boldsymbol{\mu})),
\end{aligned} \tag{17}$$

where $g_2(t) = O\left(\frac{\alpha_t^2}{\sigma_t^2}\right)$. It follows from (5) that $\sum_{t=1}^\infty g_2(t) < \infty$. Thus, we obtain from (13) and (17) that

$$\begin{aligned}
LV(\boldsymbol{\mu}) \le &- 4\alpha_t(\nabla c_t(\boldsymbol{\mu}),\boldsymbol{\mu})W'(\|\boldsymbol{\mu}\|) \\
&+ (g_1(t) + g_2(t))(1 + V(\boldsymbol{\mu})).
\end{aligned}$$

Due to Assumption 2, $(\nabla c_t(\boldsymbol{\mu}),\boldsymbol{\mu})W'(\|\boldsymbol{\mu}\|) \ge 0$ for any $t = 1, 2, \ldots$ and $\boldsymbol{\mu}\in\mathbb{R}^d$, and Theorem 1 implies that $\Pr\{\|\boldsymbol{\mu}_t\|, t = 1, 2, \ldots, \text{ is bounded}\} = 1$.

■

With this lemma in place, we can prove the main result.

Proof: [Proof of Theorem 2]

i) We will follow the idea of the proof in [16] and instead of $R_{\boldsymbol{\mu}}(T)$, we estimate $\sum_{t=1}^T(\nabla c_t(\boldsymbol{\mu}_t),\boldsymbol{\mu}_t - \boldsymbol{x}^*)$, where $\boldsymbol{x}^*$ is any point from $\mathbb{R}^n$ such that $\|\boldsymbol{x}^*\|$ is bounded. Notice that by convexity of c_t, for any $\{\boldsymbol{\mu}_t\},\boldsymbol{x}^*\in\mathbb{R}^n$

$$\begin{aligned}
&\sum_{t=1}^T c_t(\boldsymbol{\mu}_t) - \sum_{t=1}^T c_t(\boldsymbol{x}^*) \\
&\qquad \le \sum_{t=1}^T(\nabla c_t(\boldsymbol{\mu}_t),\boldsymbol{\mu}_t) - \sum_{t=1}^T(\nabla c_t(\boldsymbol{\mu}_t),\boldsymbol{x}^*) \\
&\qquad = \sum_{t=1}^T(\nabla c_t(\boldsymbol{\mu}_t),\boldsymbol{\mu}_t - \boldsymbol{x}^*).
\end{aligned} \tag{18}$$

The above indicates that $R(T)$ is at least as much as the regret calculated for the function $h(\boldsymbol{\mu}) = (\nabla c_t(\boldsymbol{\mu}_t), \boldsymbol{\mu})$. To bound this term, analogously to equality (9), by evaluating $\|\boldsymbol{\mu}_{t+1} - \boldsymbol{x}^*\|$ instead of $\|\boldsymbol{\mu}_{t+1}\|$ we can write

$$\begin{aligned}
(\nabla c_t(\boldsymbol{\mu}_t), \boldsymbol{\mu}_t - \boldsymbol{x}^*) &= \frac{1}{2\alpha_t}(\|\boldsymbol{\mu}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{x}^*\|^2) \\
&+ \frac{\alpha_t}{2}(\|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|^2 + \|\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)\|^2) \\
&- (Q_t(\boldsymbol{\mu}_t, \sigma_t) + \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t), \boldsymbol{\mu}_t - \boldsymbol{x}^*) \\
&- \alpha_t(\nabla c_t(\boldsymbol{\mu}_t), Q_t(\boldsymbol{\mu}_t, \sigma_t) + \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)) \\
&+ \alpha_t(Q_t(\boldsymbol{\mu}_t, \sigma_t), \xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)).
\end{aligned} \tag{19}$$

By summing up equality (19) from $t = 1$ to $t = T$ and by taking the conditional expectation with respect to $\mathcal{F}_T$ of the both sides in the resulting equality, we obtain that $\forall T > 0$, almost surely

$$\begin{aligned}
&\sum_{t=1}^{T}(\nabla c_t(\boldsymbol{\mu}_t), \boldsymbol{\mu}_t - \boldsymbol{x}^*) \\
&\quad \le \sum_{t=1}^{T-1} \frac{1}{2\alpha_t}(\|\boldsymbol{\mu}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{x}^*\|^2) \\
&\quad + \frac{1}{2\alpha_T}(\|\boldsymbol{\mu}_T - \boldsymbol{x}^*\|^2 - \mathrm{E}\{\|\boldsymbol{\mu}_{T+1} - \boldsymbol{x}^*\|^2 | \mathcal{F}_T\} \\
&\quad + \sum_{t=1}^{T} \frac{\alpha_t}{2}(\mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)\|^2 + \|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|^2 \\
&\qquad + \|\nabla c_t(\boldsymbol{\mu}_t)\|^2 + 2\|\nabla c_t(\boldsymbol{\mu}_t)\|\|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|) \\
&\qquad + \sum_{t=1}^{T}\|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|\|\boldsymbol{\mu}_t - \boldsymbol{x}^*\|.
\end{aligned} \tag{20}$$

Above, we used the property of the conditional expectation, namely $\mathrm{E}\{\boldsymbol{\mu}_{t_1} | \mathcal{F}_{t_2}\} = \boldsymbol{\mu}_{t_1}$ almost surely for any $t_1 \le t_2$, as well as the fact that $\mathrm{E}\{\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t) | \mathcal{F}_T\} = \mathrm{E}_{\mathbf{x}_t}\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t) = 0$ for all $t \le T$, which is implied by (4). Further, according to Lemma 1 and Assumption 2 (see Remark 2), there exists M such that $\|\boldsymbol{\mu}_t - \boldsymbol{x}^*\| \le M$ almost surely for all t. Moreover, by taking into account (11), (12), and Assumption 1, we conclude that almost surely the terms $\|Q_t(\boldsymbol{\mu}_t, \sigma_t)\|$, and $\|\nabla c_t(\boldsymbol{\mu}_t)\|$ are bounded and $\mathrm{E}_{\mathbf{x}_t}\|\xi_t(\mathbf{x}_t, \boldsymbol{\mu}_t, \sigma_t)\|^2 \le O(1/\sigma_t^2)$. Hence, almost surely

$$\begin{aligned}
\sum_{t=1}^{T}(\nabla c_t(\boldsymbol{\mu}_t), \boldsymbol{\mu}_t - \boldsymbol{x}^*) &\le \frac{1}{2\alpha_1}\|\boldsymbol{\mu}_1 - \boldsymbol{x}^*\|^2 \\
&+ \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right)\|\boldsymbol{\mu}_t - \boldsymbol{x}^*\|^2 \\
&+ O\left(\sum_{t=1}^{T}\frac{\alpha_t}{\sigma_t^2}\right) + O\left(\sum_{t=1}^{T}\sigma_t\right) \\
&\le \frac{M}{2\alpha_T} + O\left(\sum_{t=1}^{T}\frac{\alpha_t}{\sigma_t^2}\right) + O\left(\sum_{t=1}^{T}\sigma_t\right).
\end{aligned} \tag{21}$$

Next, taking into account settings for α_t and σ_t, we get

$$\begin{aligned}
\sum_{t=1}^{T}\frac{\alpha_t}{\sigma_t^2} &= \sum_{t=1}^{T}\frac{1}{t^{a-2b}} \le 1 + \int_1^T \frac{dt}{t^{a-2b}} \\
&= \frac{T^{1-a+2b}}{1-a+2b} - \frac{a-2b}{1-a+2b}, \\
\sum_{t=1}^{T}\sigma_t &= \sum_{t=1}^{T}\frac{1}{t^b} \le 1 + \int_1^T \frac{dt}{t^b} = \frac{T^{1-b}}{1-b} - \frac{b}{1-b}.
\end{aligned}$$

Thus, almost surely

$$\begin{aligned}
\sum_{t=1}^{T}(\nabla c_t(\boldsymbol{\mu}_t), \boldsymbol{\mu}_t - \boldsymbol{x}^*) \le &\frac{MT^a}{2} + O\left(T^{1-a+2b}\right) \\
&+ O\left(T^{1-b}\right) + c,
\end{aligned} \tag{22}$$

where $c = -\frac{a-2b}{1-a+2b} - \frac{b}{1-b}$. Hence, from (18), almost surely

$$\frac{R_{\boldsymbol{\mu}}(T)}{T} \le O\left(\frac{1}{T^{1-a}}\right) + O\left(\frac{1}{T^{a-2b}}\right) + O\left(\frac{1}{T^b}\right). \tag{23}$$

As $0 < a < 1$, $b > 0$, $a - 2b > 0$, the inequality above implies $\overline{\lim}_{T\to\infty}\frac{R_{\boldsymbol{\mu}}(T)}{T} \le 0$ almost surely.

ii) Notice that

$$\begin{aligned}
R_{\boldsymbol{\mu}}(T) &= C_{\boldsymbol{\mu}}(T) - \min_{\mathbf{x}\in\mathbb{R}^n} C(\mathbf{x}, T) \\
&= C_{\boldsymbol{\mu}}(T) - C_{\mathbf{x}}(T) + C_{\mathbf{x}}(T) - \min_{\mathbf{x}\in\mathbb{R}^n} C(\mathbf{x}, T).
\end{aligned}$$

Here, $C_{\boldsymbol{\mu}}(T)$ and $C_{\mathbf{x}}(T)$ emphasize that the cost $C(T)$ is considered for the sequences $\{\boldsymbol{\mu}_t\}$ and $\{\mathbf{x}_t\}$, respectively. Hence, by convexity of the functions $\{c_t\}$ and their bounded gradients (see Assumption 1), we obtain that almost surely

$$\begin{aligned}
R_{\mathbf{x}}(T) &= R_{\boldsymbol{\mu}}(T) + C_{\mathbf{x}}(T) - C_{\boldsymbol{\mu}}(T) \\
&= R_{\boldsymbol{\mu}}(T) + \sum_{t=1}^{T}(c_t(\mathbf{x}_t) - c_t(\boldsymbol{\mu}_t)) \\
&\le R_{\boldsymbol{\mu}}(T) + \sum_{t=1}^{T} l\|\mathbf{x}_t - \boldsymbol{\mu}_t\|,
\end{aligned}$$

where l is some positive bounded constant. Then by taking expectation conditioned on $\mathcal{F}_T$, we get almost surely

$$\begin{aligned}
\mathrm{E}\{R_{\mathbf{x}}(T)|\mathcal{F}_T\} &\le R_{\boldsymbol{\mu}}(T) + \sum_{t=1}^{T} l\mathrm{E}\{\|\mathbf{x}_t - \boldsymbol{\mu}_t\||\mathcal{F}_T\} \\
&\le R_{\boldsymbol{\mu}}(T) + nl\sum_{t=1}^{T}\sigma_t.
\end{aligned}$$

In the inequality above we used the fact that $\mathrm{E}\{\|\mathbf{x}_t - \boldsymbol{\mu}_t\||\mathcal{F}_T\} \le n\sigma_t$ for all $t \le T$. Now by taking the full expectation and using the inequality $\sum_{t=1}^{T}\sigma_t \le \frac{T^{1-b}}{1-b} - \frac{b}{1-b}$, we conclude that

$$\mathrm{E}\left\{\frac{R_{\mathbf{x}}(T)}{T}\right\} \le O\left(\frac{1}{T^{1-a}}\right) + O\left(\frac{1}{T^{a-2b}}\right) + O\left(\frac{1}{T^b}\right)$$

and, thus, $\overline{\lim}_{T\to\infty}\mathrm{E}\left\{\frac{R_{\mathbf{x}}(T)}{T}\right\} \le 0$ as desired. ■

We next state our result on the rate of the regret bound.

Corollary 1: The expected regret satisfies the bound:

$$\mathrm{E}\left\{\frac{R_{\mathbf{x}}(T)}{T}\right\} = O\left(\frac{1}{T^{\frac{1}{4}}}\right).$$

Proof: This result follows from the rate derived in (23) in the proof and by optimizing this rate with respect to a, b subject to the constraints $0 < a < 1$, $2a - 2b > 1$, $a + b > 1$. ■

Note that this rate is consistent with that derived in [4]. The work in [4] considered similar assumptions on the functions c_t but the case of compact action spaces. Indeed, the derivation of the convergence proof and the rate provided depended on the diameter of the action space and hence, the method is not applicable to the unconstrained setting here.

IV. Numerical Example

We illustrate the proposed algorithm with a simple case study. Consider the functions $c_t : \mathbb{R} \to \mathbb{R}$ defined as

$$c_t(\mathbf{x}) = \begin{cases} \frac{1}{2}m(\mathbf{x} - x_o)^2, & |\mathbf{x} - x_o| \leq z_t \\ mz_t\mathbf{x} - \frac{1}{2}mz_t^2 - mz_tx_o, & \mathbf{x} - x_o > z_t \\ mz_t\mathbf{x} - \frac{1}{2}mz_t^2 + mz_tx_o, & \mathbf{x} - x_o < -z_t \end{cases}$$

where $m \in \mathbb{R}_{>0}$ and $x_o \in \mathbb{R}$ are constants and $z_t \in \mathbb{R}_{>0}$ is time-dependent. One can verify that c_t's satisfy Assumptions 1-3. The above functions c_t are chosen for this numerical study since the optimizer in $\min_{\mathbf{x} \in \mathbb{R}} \sum_{t=0}^{T} c_t(\mathbf{x})$ is $x^\star = x_o$, $\forall T > 0$. Hence, we can evaluate empirically the regret of our proposed algorithm varying the time horizon T.

We assume at each time t, we can only observe $c_t(\mathbf{x}_t)$ and have no other knowledge on the functions. Our goal is to choose $\mathbf{x}_t$ in order to minimize the regret. We use algorithm (2). The parameters for the following numerical simulations were $z_t \sim \mathcal{U}[2.9, 3.1]$ where $\mathcal{U}[u_1, u_2]$ denotes the uniform distribution on the interval $[u_1, u_2]$ and $x_o = 1$, $m = 2$. The variance and step sizes were chosen according to sequences $\{\alpha_t = \frac{1}{t^a}\}$, $\{\sigma_t = \frac{1}{t^b}\}$ with $a = 10/11$, $b = 2/11$. These sequences satisfy conditions in Theorem 2.

The evolution of $\boldsymbol{\mu}$ and $\mathbf{x}$ for an instance of the problem corresponding to $T = 20000$ is shown in Fig. IV. The initial condition of $\boldsymbol{\mu}$ was chosen randomly from $\mathcal{U}[0, 1]$. For varying horizons of $T = 2 \times 10^2, 2 \times 10^3, 2 \times 10^4, 2 \times 10^5$, the empirical regret defined in (1) was $R_T = 0.2900, 0.1081, 0.0433, 0.0225$, respectively. These rates are consistent with sublinear convergence of regret.

V. Conclusion

We provided a novel algorithm for an online learning problem in which the cost functions are unknown and time-varying. We considered the case of the unconstrained action space. Our algorithm only required information on the values of the cost functions for the played actions. We established regret bounds of the algorithm using results on stochastic processes. Currently, we are developing an extension of the presented procedure to handle both the unconstrained setup and the compact action setup. Furthermore, we aim to establish lower bounds on regret for our settings and match them with the corresponding results presented in the literature so far [1], [10].

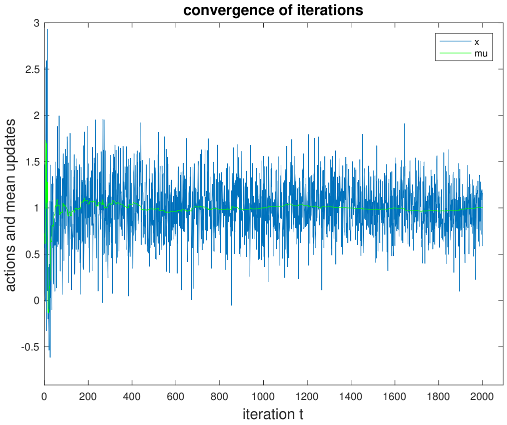

Fig. 1. Iterations $\mathbf{x}$ and $\boldsymbol{\mu}$ of the algorithm corresponding to $T = 2000$.

References

[1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, May.

[2] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. *IEEE Transactions on Automatic Control*, 32(11):977–982, 1987.

[3] E. W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

[4] 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*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[5] S. Hosseini, A. Chapman, and M. Mesbahi. Online distributed convex optimization on dynamic networks. *IEEE Transactions on Automatic Control*, 61(11):3545–3550, 2016.

[6] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[7] W.-J. Ma, V. Gupta, and U. Topcu. Distributed charging control of electric vehicles using online learning. *IEEE Transactions on Automatic Control*, 62(10):5289–5295, 2017.

[8] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, Apr. 2017.

[9] M. B. Nevelson and R. Z. Khasminskii. *Stochastic approximation and recursive estimation [translated from the Russian by Israel Program for Scientific Translations ; translation edited by B. Silver]*. American Mathematical Society, 1973.

[10] B. Recht, K. G. Jamieson, and R. Nowak. Query complexity of derivative-free optimization. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2681–2689. 2012.

[11] P. B. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.

[12] H. Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[13] S. Shahrampour and A. Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 2017.

[14] S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[15] A. L. Thathachar and P. S. Sastry. *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Springer US, 2003.

[16] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 928–936, 2003.