Diminishing Regret for Online Nonconvex Optimization

SangWoo Park, Julie Mulvaney-Kemp, Ming Jin, and Javad Lavaei
University of California, Berkeley

Abstract—**A single nonconvex optimization is NP-hard in the worst case, and so is a sequence of nonconvex problems viewed separately. For online nonconvex optimization (ONO) problems, the widely used local search algorithms are only guaranteed to track a sequence of local optima and not necessarily global optima. In this paper, we introduce the concept of nonconvexity regret that measures the performance of a local search method against a global optimization solver for ONO. We show that memory and random explorations drive the nonconvexity regret to zero if the ONO problem has robustness in its global minima throughout the time horizon. We prove probabilistic guarantees on the regret bound that depend on the evolution of the landscapes of the time-varying objective functions. Then, based on the notions of missing mass and 1-occupancy set, we develop a practical algorithm that works even when there is no information on the landscape of ONO. The theoretical results imply that the existence of a low-complexity optimization at any arbitrary time instance of ONO can nullify the NP-hardness of the entire ONO problem. The results are verified through numerical simulations.**

I. Introduction

Nonconvex optimization is ubiquitous in real-world applications, such as the training of deep neural nets [1], matrix sensing/completion [2], [3], state estimation of dynamic systems [4], and the optimal power flow problem [5]. Moreover, most of these practical problems are solved sequentially over time with time-varying input data, leading to online (real-time) versions of the aforementioned examples [4], [6], [7].

In this paper, we study an online optimization problem whose objective function changes over discrete time periods, namely,

$$\underset{x\in\mathbb{S}}{\text{minimize}} \quad f_t(x) \tag{1}$$

where $t \in \mathbb{Z}^+$ denotes the time and $\mathbb{S} \subset \mathbb{R}^n$ is the time-invariant feasible region. At each time t, the objective function f_t is differentiable but could potentially be nonconvex in x with non-unique local minima. In general, nonconvex optimization problems are NP-hard and the commonly used local search algorithms such as gradient descent method or Newton-Raphson method may converge to a spurious local minimum (i.e., a local minimum that is not globally optimal). In other words, at each instance of time, the sub-optimality gap incurred between the obtained solution and the globally optimal solution, hereafter called *nonconvexity regret*, could be nonzero. The main goal of this paper is to analyze how this nonconvexity regret evolves over time in the ONO setting. More specifically, we study algorithms with memory and random exploration and connect the complexity of ONO with the evolution of the landscapes of f_t over time.

Our first main result shows that the proposed algorithm with random explorations will stop accumulating nonconvexity regret with a high probability if the global minimum is sufficiently robust in the sense to be defined later. This probability depends on the volume ratio of the region of attraction of the global minimum over the problem sequence. The results imply that the existence of a single low-complexity problem (among the sequence of nonconvex problems) can lower the complexity of the entire ONO problem, which is also verified through simulations. The second main result extends the earlier idea to the case when the volume ratio of the region of attraction of the global solution is unknown. This contributes to designing a practical algorithm that stops taking random samples when the missing mass is below a certain threshold with high probability. By analyzing the 1-occupancy set, we draw a connection between the diminishing rate of the nonconvexity regret and the landscape of ONO.

A. Problem Setup

In the ONO framework, the decision maker at each time t chooses $x_t \in \mathbb{S}$ in order to minimize the nonconvex objective function $f_t : \mathbb{S} \to \mathbb{R}$. Unlike in some online learning settings where an adversary also chooses the function f_t at time t, we assume that the sequence of objective functions $f_1, f_2, \ldots$ is fixed ahead of the decision making process. However, a decision maker at time t does not have information about the future objective functions. At time t, the system accrues (instantaneous) regret of the following form:

$$[\textit{Nonconvexity}\text{ regret}] \qquad f_t(x_t) - \inf_{x\in\mathbb{S}} f_t(x) \tag{2}$$

This regret is mainly due to the fact that f_t is non-convex and the decision maker may fail to find the global optimum at each time step. It is straightforward to verify that the nonconvexity regret can be arbitrarily high in a general setting. Therefore, existing works in the literature have derived regret bounds in terms of various quantities such as the regularity of the comparator sequence [8] and the temporal variability of the objective functions [9]. We take a slightly different approach and make the following two main assumptions:

- (*Memory*) The global optimum cannot change significantly over single time intervals: $\|x^*_{t+1} - x^*_t\| < r$.

The authors are with the Department of Industrial Engineering and Operations Research at the University of California Berkeley. Emails: {`spark111`, `julie_mulvaney-kemp`, `jinming`, `lavaei`}`@berkeley.edu`.

This work was supported by grants from ARO, ONR, AFOSR, and NSF.

- (*Robust Global*) The region of attraction of the global solution includes a ball of radius at least r around the global solution: $\exists r > 0$ such that $\mathcal{RA}(x_t^*) \supseteq B(x_t^*, r)$ at any time t.

Under the robust global condition alone, there does not always exist an efficient algorithm to find and track the global minima of a sequence of nonconvex optimization problems. Therefore, we consider random explorations, which we describe in Section III. It turns out that random explorations help with finding the global solution at some point in time, and memory enables the tracking of the robust global solution once it is found.

B. Related Work

Parametric optimization and homotopy methods provide one of the tools for analyzing ONO problems. In the classic work of [10], the authors lay out the theories behind the structure and singularity of the Karush-Kuhn-Tucker (KKT) trajectories for time-varying optimization problems. Taking a different type of approach, [11] presents conditions under which the solution of some ordinary differential equation (ODE) is close to the KKT trajectory (of ONO) and presents a predictor-corrector method to tract the ODE solution. Using similar approaches, [12] studied a gradient flow system with inertia, as a continuous-time limit of the proximal algorithm and developed sufficient conditions under which the solution trajectory would escape spurious local solutions and begin tracking time-varying global solutions. Recently, [13] and [14] explored how variation in the input data can help the ONO solution trajectories escape non-global local solutions.

In the machine learning community, the performance of online optimization schemes is often analyzed through the notion of stationary (or static) regret [15], which is the comparison to a single best action in hindsight. In this paper, we analyze the regret against a more stringent comparator, which we call the nonconvexity regret. A related notion named *dynamic regret* is studied in the literature [9], [15], [16]. Most of the existing works on online optimization with provable guarantees have been focused on the convex setting, where the objective is to minimize a sequence of convex functions over a convex domain [17]–[21]. Unlike convex optimization for which there is no distinction between global and local minima, it is impossible to design an efficient algorithm that always converges to a global minimum even in hindsight under the nonconvex setting. Therefore, the papers [22]–[24] utilize an alternate concept of local regret that is based on stationary points. Contrary to this line of research, we study the *global regret* and establish probabilistic guarantees using memory and random explorations.

Randomization is a useful tool for solving nonconvex optimization problems. It can be employed within the algorithm itself or when initializing a local search algorithm. In both cases, the goal is to facilitate exploration of the solution space and avoid poor local minima. Simulated annealing [25] is one such approach in which a random move is chosen at each iteration. The move is executed if it represents an improvement. However, even if the move results in a worse solution, it is still executed with some nonzero probability. Multi-start methods address algorithm initialization by repeatedly constructing an initial point, applying a local search method to obtain a solution based on said point, and comparing with past solutions, until specified stopping criterion are satisfied. Using the terminology of [26], multi-start methods rely on a combination of three key elements: memory (using knowledge of previous good solutions), randomization (degree to which initial points are generated in a random versus deterministic way), and the degree of rebuild (whether or not some elements are fixed for a number of iterations). This paper adapts the multi-start techniques to the online setting, focusing on the history and randomization elements. [27]–[29] are a few key works on this topic, and we refer the reader to [26], [30] for a more extensive review of multi-start methods.

In optimization, the connection between the convergence of an algorithm and the stability of a dynamical system has been long known, where the region of attraction around an equilibrium point is related to the convergence region for a local optimum [31], [32]; recently, there are also renewed interests in optimizing the hyper-parameters for convergence rate [33]–[35]. However, these results are only for the optimization of a single problem, rather than a sequence of time-varying problems as considered in this study. Existing works often assume that there exists a reasonably large region of attraction around the global optimum in order to guarantee the successful convergence of the iterative methods like gradient descent [36]. The assumption of *robust global* made in this paper is similar in essence but weaker.

C. Organization

The remainder of this paper is organized as follows. In Section II, we provide some notations and preliminaries. In Section III, we develop an online optimization algorithm and derive regret bounds under the assumption that the volumes of the regions of attraction are known for the local minima. In Section IV, we extend the results to the case where the volumes are unknown. The theoretical results are supported by numerical simulations and analyses in Section V. Finally, we conclude the paper in Section VI.

II. Notations and Preliminaries

Let $||\cdot||$ indicate the 2-norm of a vector and $|\cdot|$ represent the cardinality of a set. The symbols $\mathbb{R}^n$ and $\mathbb{Z}^+$ denote the space of n-dimensional real vectors and the set of positive integers, respectively. Define $B(x,r)$ to be a ball centered at x with radius r. In this paper, we assume that $S = \{x \mid g(x) \leq 0\}$ is a compact set of dimension n, where all entries of $g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are convex functions. The global optimum of the optimization problem at time t is denoted by

$$x_t^* = \operatorname*{arginf}_{x \in \mathbb{S}} f_t(x) \tag{3}$$

If the global minimum is not unique, x_t^* denotes a particular global solution for which the memory and robust global conditions stated before hold.

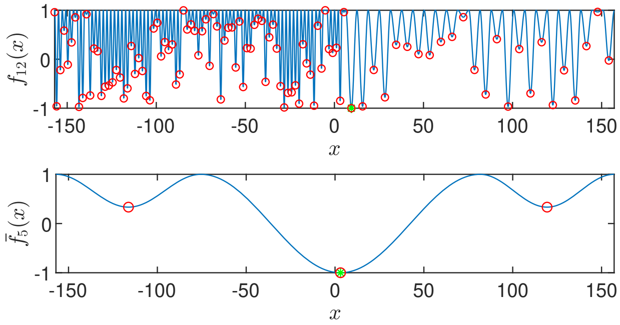

Fig. 1. Top and bottom plots show examples of a problem with high and low complexity, respectively. Local minima are marked by red circles; the global minimum is further marked by a green star.

Definition 1. *Given an arbitrary natural number $T \in \mathbb{Z}^+$, define the cumulative nonconvexity regret up to time T as follows:*

$$\gamma(T) = \sum_{t=1}^{T} \Big(f_t(x_t) - \inf_{x \in \mathbb{S}} f_t(x) \Big). \tag{4}$$

where x_t is the solution obtained by a given local search method. In addition, $\frac{\gamma(T)}{T}$ is regarded as the average nonconvex regret over time period $[1, T]$. It is said that the nonconvexity regret is a (ϵ, δ)-regret if, for any $\epsilon \geq 0$ and $\delta \in [0, 1]$, the following relation holds:

$$\mathbb{P}\Big\{ \frac{\gamma(T)}{T} \leq \epsilon \Big\} \geq 1 - \delta \tag{5}$$

where $\mathbb{P}\{\cdot\}$ represents the probability.

Note that the notion of nonconvexity regret is algorithmic dependent. In this work, we equip local search with memory and randomization, and study their effect on the regret. To proceed with the paper, it is necessary to define a region of attraction (RoA) for each local minimum of $f_t(x)$, which includes the set of all initial points that can be used to find that particular local minimum. Since RoA would be a chaotic set if the step sizes of a gradient-based method are allowed to vary arbitrarily, we circumvent the issue by using sufficiently small step sizes. This allows us to define RoA via ordinary differential equations.

Definition 2. *(Region of attraction) Let $\text{proj}_{\mathbb{S}}(x, -\nabla f_t(x))$ be the solution to the following quadratic program:*

$$\min_{x \in \mathbb{R}^n} \|w + \nabla f_t(x)\|^2 \quad \text{s.t. } \nabla_{\mathcal{J}} g(x)^T w \leq 0$$

where $\nabla_{\mathcal{J}} g(x)$ is the Jacobian of the active constraints at x. Given a local minimum or a saddle point $\bar{x}$ of the optimization problem $\min_{x \in \mathbb{S}} f_t(x)$, define the region of attraction of $\bar{x}$ as follows:

$$\begin{aligned} \mathcal{RA}(\bar{x}) = \Big\{ x_0 : & \lim_{k \to \infty} x(k) = \bar{x}, \ \textit{where} \\ & \frac{dx(k)}{dk} = \text{proj}_{\mathbb{S}}(x, -\nabla f_t(x(k)), \ x(0) = x_0 \Big\} \end{aligned} \tag{6}$$

We have defined the RoA based on a continuous version of the projected gradient method, but one can use other first-order methods to refine this. Next, we make a standard assumption that the optimization problem (1) has a finite number of stationary points. The volume of the set $\mathcal{RA}(\bar{x})$ is defined as

$$V(\mathcal{RA}(\bar{x})) = \int \cdots \int_{\mathcal{RA}(\bar{x})} 1 \ d^n x \tag{7}$$

where $V(\cdot)$ indicates the volume of a set. At each time t, we make the mild assumption that the sum of all RoA is equal to the volume of the entire domain (this means that the algorithm must always converge to a stationary point).

III. Online Algorithm with Random Explorations

In this section, we introduce an online algorithm with random explorations and prove bounds on the average regret, which are dependent on the volumes of the regions of attraction for the global minima at different times. We also derive a hitting time for the time-horizon length such that the desired precision level for the probabilistic guarantee is achieved.

Let $y_t^1, \ldots, y_t^m$ be independent and identically distributed samples from a fixed distribution $\mathcal{P}$ at time t and define the set $Y_t = \{y_t^1, \ldots, y_t^m\}$. Let $h_t : \mathbb{S}^{m+1} \to \mathbb{S}$ be an operator that takes in the current solution x_t along with the m random initial points and outputs the best local minimum or saddle point resulting from these points. In other words,

$$x_{t+1} = h_t(x_t, Y_t) = \operatorname*{argmin}_{x \in x_t \cup Y_t} l_t(x) \tag{8}$$

where l_t represents a continuous-time projected gradient algorithm that takes in x as the initial point and outputs a stationary point of the optimization problem (1) (note that if the initial point is a local maximum, the algorithm will stay at that point). At times, we will simply use $h(\cdot)$ instead of $h_t(\cdot)$. For each time t, let $\mathbb{X}_t$ define the set of local minima and saddle points of the optimization problem (1). Then, since we utilize the solution found at the previous time step as the initial point of the current time step in the proposed algorithm, the above mapping generates a sequence of stationary points over time. We denote this sequence by

$$\phi(x_0) = \{(x_0, x_1, \ldots, x_T) \mid x_{i+1} = h_i(x_i, Y_i)\} \tag{9}$$

More generally, we can define a *forward solution sequence* at time t given x_0:

$$\phi^t(x_0) = \{(x_t, x_{t+1}, \ldots, x_T) \mid x_{i+1} = h_i(x_i, Y_i)\} \tag{10}$$

Note that $\phi(x_0) = \phi^0(x_0)$. For a set of initial conditions $\mathbb{X}_0$, we define a *set of forward solution sequences* at time t as $\Phi^t(\mathbb{X}_0) = \{\phi^t(x_0) \mid x_0 \in \mathbb{X}_0\}$.

The above procedure for solving the online optimization (1) is summarized in Algorithm 1. This algorithm is the natural counterpart of the classical gradient descent method for time-invariant (static) optimization, but is enhanced by incorporating memory and random exploration, meaning that: (i) the solution at each time is used as a memory to guide the algorithm in solving the problem at the next time instance, (ii) the landscape of the problem at each time instance is explored via m random points. In this section, the objective is to understand how the memory and the number of random samples affect the nonconvexity regret as a function of time.

Algorithm 1 Online Local Search with Random Exploration

```
Given: $x^0 \in \mathbb{S}$ and $\{f_t\}_{t=0}^{\infty}$
for $t = 1, 2, \dots$ do
    • Create $Y_t = \{y_t^1, \dots, y_t^m\}$ by sampling $m$ random points from $\mathbb{S}$ using the probability distribution $\mathcal{P}$
    • Set $x_{t+1} = h_t(x_t, Y_t)$
end for
```

Definition 3. *(Volume ratio of global optimum) Let ρ_t symbolize the fraction of the entire solution space belonging to the RoA of the global minimum x_t^* at time t. That is,*

$$\rho_t = \frac{V(\mathcal{RA}(x_t^*))}{V(\mathbb{S})} \tag{11}$$

Note that $V(\mathbb{S}) < \infty$ since $\mathbb{S}$ is assumed to be compact. In this paper, any time t for which the optimization problem has a large ρ_t is regarded as the time with a low-complexity optimization.

An example showing low-complexity and high-complexity problems is shown in Fig. 1. The first main result of this paper provides a probabilistic guarantee that the sequence of online optimization solutions generated by Algorithm 1 will find and keep tracking the global optimum over time. The probability of achieving this property depends on the values of ρ_t. This will be formalized below.

Theorem 1. *Consider two arbitrary natural numbers T and $\bar{T}$ with the property that $\bar{T} \leq T$. Suppose that there exists a positive number r such that*

$$\mathcal{RA}(x_t^*) \supseteq B(x_t^*, r) \tag{12a}$$

$$\|x_{t+1}^* - x_t^*\| < r \tag{12b}$$

for $t = 1, 2, \dots, T$. Let the set Y_t be generated by sampling m points according to a uniform distribution on the set $\mathbb{S}$. Then, for any $\phi^0(x_0) \in \Phi^0(\mathbb{S})$, the following statement holds true when the online optimization is solved via Algorithm 1:

$$\mathbb{P}\Big\{\frac{\gamma(T)}{T} = \frac{\gamma(\bar{T})}{T}\Big\} \geq 1 - \prod_{t=1}^{\bar{T}} \big(1 - \rho_t\big)^m \tag{13}$$

Proof. First, we show that the event of finding the global minimum at some time $t \leq \bar{T}$ via Algorithm 1 leads to the event of finding the global minimum at all time $t \geq \bar{T}$. Suppose that the algorithm has found the global minimum $x_{\bar{t}}^*$ at time $\bar{t} \leq \bar{T}$. It follows from (12b) that $x_{\bar{t}+1}^* \in B(x_{\bar{t}}^*, r)$, which implies that $x_{\bar{t}}^* \in B(x_{\bar{t}+1}^*, r)$. Therefore, it can be concluded that $h(x_{\bar{t}}^*, Y_{\bar{T}}) = x_{\bar{t}+1}^*$. Using an induction process, it holds that all solutions returned by Algorithm 1 after time $\bar{t}$ are also globally optimal:

$$x_{\bar{t}} = x_{\bar{t}}^* \implies x_t = x_t^* \quad \forall t \in \{\bar{t}, \dots, T\}$$

This implies that the regret will stop accumulating after $\bar{T}$:

$$f_t(x_t) - \inf_{x \in \mathbb{S}} f_t(x) = 0 \quad \forall t \in \{\bar{t}, \dots, T\}$$

Accordingly,

$$\frac{1}{T} \sum_{t=1}^{T} \Big(f_t(x_t) - \inf_{x \in \mathbb{S}} f_t(x)\Big) = \frac{1}{T} \sum_{t=1}^{\bar{t}} \Big(f_t(x_t) - \inf_{x \in \mathbb{S}} f_t(x)\Big)$$

Now, it is desirable to calculate the probability that Algorithm 1 will find the global minimum of one of the functions $f_1, \dots, f_{\bar{T}}$. Recall that at each time t, the previous solution x_{t-1} and m random points represented by the set $Y_{t-1} = \{y_{t-1}^1, \dots, y_{t-1}^m\}$ are fed into Algorithm 1. The probability of failing to find the global minimum $x_{\bar{T}}^*$ at time $\bar{T}$ is then less than or equal to the probability that none of the random points up until time $\bar{T}$ belongs to the regions of attraction of the respective solutions $x_1^*, \dots, x_{\bar{T}}^*$:

$$\begin{aligned} &\mathbb{P}\{h(x_{\bar{T}-1}, Y_{\bar{T}-1}) \neq x_{\bar{T}}^*\} \\ &\leq \prod_{t=1}^{\bar{T}} \mathbb{P}\{y_{t-1}^i \notin \mathcal{RA}(x_t^*),\ 1 \leq i \leq m\} = \prod_{t=1}^{\bar{T}} \big(1 - \rho_t\big)^m \end{aligned}$$

Therefore, the probability of arriving at the global minimum up until time $\bar{T}$ is

$$\begin{aligned} \mathbb{P}\{h(x_{\bar{T}-1}, Y_{\bar{T}-1}) = x_{\bar{T}}^*\} &= 1 - \mathbb{P}\{h(x_{\bar{T}-1}, Y_{\bar{T}-1}) \neq x_{\bar{T}}^*\} \\ &\geq 1 - \prod_{t=1}^{\bar{T}} \big(1 - \rho_t\big)^m \end{aligned}$$

It can be inferred from the above arguments that

$$\begin{aligned} &\mathbb{P}\Big\{\frac{\gamma(T)}{T} = \frac{\gamma(\bar{T})}{T}\Big\} \\ &\geq \mathbb{P}\Big\{\exists \bar{t} \leq \bar{T} \text{ s.t. } x_{\bar{t}} = x_{\bar{t}}^*\Big\} \geq 1 - \prod_{t=1}^{\bar{T}} \big(1 - \rho_t\big)^m \quad \square \end{aligned}$$

Theorem 1 states that the regret stops accumulating after any arbitrary time $\bar{T} \leq T$ with a probability that depends on the volume of $\mathcal{RA}(x_t^*)$ for $t = 1, \dots, \bar{T}$. In the special case when $\rho_{\bar{T}} = 1$ (e.g. when $f_{\bar{T}}(x)$ is convex), inequality (13) holds with probability one. This implies that the existence of a single convex problem, in between the sequence of numerous nonconvex problems, is enough to break down the NP-hardness of solving a nonconvex problem for all future times, under the memory and robust global conditions. Notice that $\gamma(\bar{T})$ itself is a random variable because it is based on the random sampling of initial points. To refine the result of this theorem, the next corollary provides a deterministic upper-bound that is based on the maximum distance between any stationary point and the best global optimum over time.

Corollary 1. *Under the assumptions of Theorem 1, define*

$$C_{\bar{T}} := \max_{t \leq \bar{T}} \Big\{ \sup_{x \in \mathbb{X}_t} \big[f_t(x) - \inf_{x \in \mathbb{S}} f_t(x)\big] \Big\} \tag{14}$$

The following statement holds true:

$$\mathbb{P}\Big\{\frac{\gamma(T)}{T} \leq \frac{C_{\bar{T}} \bar{T}}{T}\Big\} \geq 1 - \prod_{t=1}^{\bar{T}} \big(1 - \rho_t\big)^m \tag{15}$$

Proof. Due to Theorem 1, it is enough to prove that $\gamma(\bar{T}) \leq C_{\bar{T}} \bar{T}$. This follows by definition:

$$\gamma(\bar{T}) \leq \bar{T} \max_{t \leq \bar{T}} \Big\{ \sup_{x \in \mathbb{S}} \big[f_t(x) - \inf_{x \in \mathbb{S}} f_t(x)\big] \Big\} = C_{\bar{T}} \bar{T} \quad \square$$

Given a pair (ϵ, δ), it is essential to estimate the earliest time at which the *average nonconvexity regret* associated with

Algorithm 1 is a (ϵ, δ)-regret. Such time is called the (ϵ, δ)-hitting time and is denoted by T^h.

In what follows, we will study the case where there is a time $\bar{T}$ at which the global solution $x^*_{\bar{T}}$ dominates other local and saddle points in terms of the volumes of their RoA. A special case of this scenario corresponds to having at least one convex function in the sequence $f_1(x), f_2(x), \ldots$

Corollary 2. *Suppose that the two conditions in Theorem 1 hold. Given an arbitrary constant $q > 0$, let $\bar{T}$ be the first time such that $|\mathbb{X}_{\bar{T}}| \leq q$ and*

$$V(\mathcal{RA}(x^*_{\bar{T}})) \geq V(\mathcal{RA}(x)), \quad \forall x \in \mathbb{X}_{\bar{T}} \tag{16}$$

(we set $\bar{T}$ to infinity if such number does not exist). Then, the (ϵ, δ)-hitting time T^h, for Algorithm 1, is upper bounded by

$$\max\left(\bar{T}, \frac{C_{\bar{T}}\bar{T}}{\epsilon}\right) \tag{17}$$

if m is chosen to be greater than $\frac{\ln(\delta)}{\ln(1-1/q)}$.

Proof. It follows from (16) that

$$\rho_{\bar{T}} = \frac{V(\mathcal{RA}(x^*_{\bar{T}}))}{V(\mathbb{S})} = \frac{V(\mathcal{RA}(x^*_{\bar{T}}))}{\sum_{x \in \mathbb{X}_{\bar{T}}} V(\mathcal{RA}(x))}$$
$$\geq \frac{V(\mathcal{RA}(x^*_{\bar{T}}))}{V(\mathcal{RA}(x^*_{\bar{T}})) \cdot |\mathbb{X}_{\bar{T}}|} = \frac{1}{|\mathbb{X}_{\bar{T}}|} \geq \frac{1}{q}$$

Furthermore, since we have $m \geq \frac{\ln(\delta)}{\ln(1-1/q)}$, one can write:

$$1 - \prod_{t=1}^{\bar{T}} \left(1 - \rho_t\right)^m \geq 1 - \left(1 - \rho_{\bar{T}}\right)^m \geq 1 - \left(1 - \frac{1}{q}\right)^m \geq 1 - \delta$$

As a result,

$$1 - \prod_{t=1}^{T'} \left(1 - \rho_t\right)^m \geq 1 - \prod_{t=1}^{\bar{T}} \left(1 - \rho_t\right)^m \geq 1 - \delta, \quad \forall T' \geq \bar{T}$$

Now, it results from Theorem 1 that

$$\mathbb{P}\Big\{\frac{1}{T'} \sum_{t=1}^{T'} \Big(f_t(x_t) - \inf_{x \in \mathbb{S}} f_t(x)\Big) \leq \epsilon\Big\} \geq 1 - \delta,$$

as long as $\epsilon \geq \frac{C_{\bar{T}}\bar{T}}{T'}$, or equivalently $T' \geq \frac{C_{\bar{T}}\bar{T}}{\epsilon}$. Therefore, $\max(\bar{T}, \frac{C_{\bar{T}}\bar{T}}{\epsilon})$ is an upper bound on the $(\epsilon, \delta)-$hitting time. □

From Corollary 2, one can analyze the role that a "low-complexity problem" at some time $\bar{T}$ plays in determining the complexity of the entire online nonconvex optimization. As an extreme but important case, suppose that there is a finite time $\bar{T}$ such that $|\mathbb{X}_{\bar{T}}| \leq q = 1$, and let $\bar{T}$ denote the smallest number with this property. Then, Algorithm 1 at any time after $T \simeq O(1/\epsilon)$ provides the desired level of confidence on nonconvexity regret. Note that there is no need for using random initial points in this scenario (i.e., m = 0).

IV. Missing Mass and Dynamic Stopping Rule

In practice, information on $V(\mathcal{RA}(x^*_{\bar{T}}))$ may be limited and difficult to estimate. Therefore, we modify Algorithm 1 to account for this fact and present a heuristic algorithm that does not require knowledge of the volumes beforehand. In essence, these results rely on the notion of missing mass.

Algorithm 2 Online Local Search with Random Exploration and Dynamic Stopping Rule

```
Given: $x^0 \in \mathbb{S}$, $\bar{m} \in \mathbb{Z}^+$, $\alpha, \delta \in [0,1]$ and $\{f_t\}_{t=0}^{\infty}$
for $t = 1, 2, \ldots$ do
    Set: $m = 1$ and $|\mathbb{W}_1^0| = 0$
    while $\frac{|\mathbb{W}_1^{m-1}|}{m} + 5\sqrt{\frac{\ln(3/\alpha)}{m}} > \delta$ and $m \leq \bar{m}$ do
        • Sample $y_t^m \in \mathbb{S}$ using $\mathcal{P}$
        • Update $|\mathbb{W}_1^m|$ based on $l_t(y_t^1), \ldots, l_t(y_t^m)$
        • Set $m = m + 1$
    end while
    Set: $m_t = m$ and $x_{t+1} = h(x_t, \{y_t^1, \ldots, y_t^{m_t}\})$
end for
```

Definition 4. *Given m random points at time t and any arbitrary point $x \in \mathbb{S}$, define $c_t(x)$ to be the number of samples that lie in the RoA of x:*

$$c_t(x) = \sum_{i=1}^{m} \mathbb{I}[y_t^i \in \mathcal{RA}(x)] \tag{18}$$

where $\mathbb{I}[\cdot]$ is the indicator function. For any integer $k \geq 0$, let $\mathbb{W}_k^t$ denote the set of local minima $x \in \mathbb{X}_t$ with the property that $c_t(x) = k$. Finally, define M_k^t to be the probability of reaching a local minimum in $\mathbb{W}_k^t$:

$$M_k^t = \sum_{x \in \mathbb{W}_k^t} V(\mathcal{RA}(x))/V(\mathbb{S}) \tag{19}$$

Note that M_k^t depends on the random samples and therefore is a random variable. The quantity M_0^t is the missing mass at time t and signifies the volume of the RoA of all the stationary points that have not yet been found. In [37], the authors provide an upper-bound on the missing mass using the Good-Turing estimator. We reformulate the result for our purpose and state it without reiterating the proof.

Lemma 1. *([37], Theorem 9) For all $\alpha \in (0, 1]$, the following inequality holds with probability at least $1 - \alpha$:*

$$M_0^t \leq \frac{|\mathbb{W}_1^t|}{m} + (2\sqrt{2} + \sqrt{3})\sqrt{\frac{\ln(3/\alpha)}{m}} \tag{20}$$

From hereon, we will simplify the above inequality by the approximation $2\sqrt{2} + \sqrt{3} \simeq 5$. As mentioned before, the volumes of the regions of attraction cannot be estimated in general. Therefore, an alternative implementation of Algorithm 1 would be to adaptively change the number of samples and yet keep it below a user-defined threshold $\bar{m}$. At each time, Algorithm 2 continues taking one sample at a time as long as two conditions are satisfied:

- The missing mass is not below the desired threshold;
- The maximum sample number is not exhausted.

In doing so, the algorithm attempts to explore until the missing mass is small enough but also guards against taking many samples when facing a high-complexity problem (e.g. when the global minimum is sharp and has a small RoA [36]). Let m_t denote the number of randomly chosen points at each time t. It may seem plausible that the missing mass provides an upper-bound of the probability of not being able to find the global minimum. However, this is not true in general because m_t, W_1^t and M_0^t are correlated variables. To illustrate this, consider the probability of finding the global minimum conditional on a given upper-bound of the missing mass, i.e., $\mathbb{P}\{x_t = x_t^* \mid M_0^t(Y_t) \leq \delta\}$. Here, we use $M_0^t(Y_t)$ to clarify the dependence of the missing mass on random samples $Y_t = \{y_t^1, \ldots, y_t^{m_t}\}$. Then, the following always holds:

$$\begin{aligned}
&\mathbb{P}\{x_t = x_t^* \mid M_0^t(Y_t) \leq \delta\} \\
=&\frac{\mathbb{P}\{x_t = x_t^*,\ \sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}}{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}} \\
=&\frac{\mathbb{P}\{\exists\ i \text{ s.t. } y_t^i \in \mathcal{RA}(x_t^*),\ \sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}}{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}} \\
\geq&\frac{\mathbb{P}\{\exists!\ i \text{ s.t. } y_t^i \in \mathcal{RA}(x_t^*),\ \sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}}{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}} \\
=&\frac{\mathbb{P}\{\exists!\ i \text{ s.t. } y_t^i \in \mathcal{RA}(x_t^*), \sum_{k=1}^{\infty} M_k^t(Y_t \setminus y_t^i) \geq 1-\delta-\rho_t\}}{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}}
\end{aligned}$$

Let Ω denote the event of that y_t^i is the only sample that belongs to the RoA of the global minimum. Then, the final equation is equal to

$$\frac{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t \setminus y_t^i) \geq 1-\delta-\rho_t \mid \Omega\} \cdot \mathbb{P}\{\Omega\}}{\mathbb{P}\{\sum_{k=1}^{\infty} M_k^t(Y_t) \geq 1-\delta\}} \tag{21}$$

which provides a lower-bound on the probability of finding the global minimum given that we have observed a total mass greater than or equal to $1-\delta$. We evaluate the above probability for two scenarios: (i) ρ_t is very small (e.g., exponentially small in the dimension for a high-dimensional problem), and (ii) $V(\mathcal{RA}(x))$ is equal for every $x \in \mathbb{X}_t$. In case (i), (21) can be approximated by $\mathbb{P}\{\Omega\} = \rho_t(1-\rho_t)^{m_t-1}$, since the other two probability terms are very close due to the infinitesimal effect of y_t^i and ρ_t. Thus, the missing mass reveals little about the probability of finding a global minimum. In case (ii), let δ take on a value such that $q = (1-\delta)\cdot|\mathbb{X}_t|$ is an integer. Then, (21) can be approximated by

$$\binom{|\mathbb{X}_t|-1}{q-1} \Big/ \binom{|\mathbb{X}_t|}{q} = \frac{q}{|\mathbb{X}_t|} \simeq 1-\delta.$$

Thus, the missing mass is indeed informative on the probability of finding a global minimum. Note that Algorithm 2 is a heuristic in the sense that stopping when the missing mass is small enough does not always guarantee that the global has been found with high probability. However, it is speculated that the connection holds as long as $V(\mathcal{RA}(x_T^*))$ is sufficiently large as is in case (ii).

In the regime where the missing mass is closely related to the probability of finding the global solution, the value of $|\mathbb{W}_1^t|$ plays an important role, especially because the number of samples that we can take is limited. If $|\mathbb{W}_1^t|$ decays fast with the number of samples taken, then the algorithm performs better. This motivates a more careful analysis on the behavior of $|\mathbb{W}_1^t|$ with respect to the landscape of ONO.

A. Analysis using 1-occupancy Set

For each local minima $x \in \mathbb{X}_t$, the parameter $c_t(x)$ defined in (18) is sometimes called the occupancy count of x. We call $\mathbb{W}_k^t$ the *k-occupancy set*. The analysis of the 1-occupancy set can be performed under the setting of *multinomial allocations* [38]. Let x_i denote an arbitrary element in $\mathbb{X}_t$. Denote p_i as the proportion of the entire space that is included in the region of attraction of the point x_i, namely $p_i = V(\mathcal{RA}(x_i))/V(\mathbb{S})$. It holds that $\sum_i p_i = 1$. The first and second moments of $|\mathbb{W}_1^t|$ can be obtained as

$$\mathbb{E}[|\mathbb{W}_1^t|] = m_t \sum_{i=1}^{|\mathbb{X}_t|} p_i(1-p_i)^{m_t-1} \tag{22}$$

$$\begin{aligned}
\mathbb{E}[|\mathbb{W}_1^t|^2] = m_t &\sum_{i=1}^{|\mathbb{X}_t|} p_i(1-p_i)^{m_t-1} \\
&+ m_t(m_t-1)\sum_{i\neq j} p_i p_j (1-p_i-p_j)^{m_t-2}
\end{aligned} \tag{23}$$

The values of these moments depend on the distribution of the p_i's and the magnitude of m_t. To illustrate this correlation, consider an example where the set of $\{p_i\}$, ordered from largest to smallest, constitutes a *probability mass function (PMF)* that is discretized from an exponential distribution with parameter μ. The *probability density function (PDF)* of the exponential distribution is given by $d(x) = \mu e^{-\mu x}$. For each value of μ, the number of random initial points m_t is varied from 1 to 20, and then the corresponding moments are calculated and plotted as heat maps in Fig. 2. It can be observed that as $\{p_i\}$ move away from heavy tail (uniform) to light tail, both the first and second moments decrease.

A more accurate analysis would be based on the asymptotic behavior of $|\mathbb{W}_1^t|$ when m_t and $|\mathbb{X}_t|$ grow towards infinity. If the $\{p_i\}$ change in an arbitrary fashion as $|\mathbb{X}_t|$ increases, there will be numerous types of asymptotic behavior. Therefore, we impose some natural restrictions on the $\{p_i\}$ and analyze them case by case.

Definition 5. *Parameters $m_t \to \infty$, $|\mathbb{X}_t| \to \infty$ and $\{p_i\}$ are said to vary in the central domain if there exist positive constants c and $\beta_0 < \beta_1$ such that*

$$|\mathbb{X}_t| p_i \leq c, \quad \beta_0 \leq \beta = \frac{m_t}{|\mathbb{X}_t|} \leq \beta_1 \tag{24}$$

Define $\bar{p} = \max_i p_i$ and $\underline{p} = \min_i p_i$. It is said that $m_t \to \infty$, $|\mathbb{X}_t| \to \infty$ and $\{p_i\}$ vary in the left-hand 1-domain if

$$m_t \bar{p} \to 0, \quad \mathbb{E}[|\mathbb{W}_1^t|] \to \lambda < \infty \tag{25}$$

Finally, we shall say that $m_t \to \infty$, $|\mathbb{X}_t| \to \infty$ and $\{p_i\}$ vary in the right-hand 1-domain if

$$m_t \underline{p} \to \infty, \quad \mathbb{E}[|\mathbb{W}_1^t|] \to \lambda < \infty \tag{26}$$

By adapting theories laid out in [38], the asymptotic behavior of 1-occupancy set is described in the following lemma.

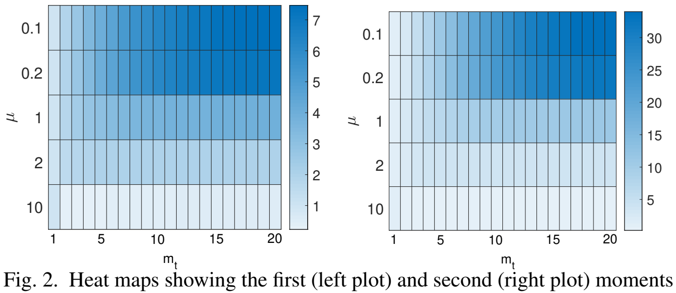

Fig. 2. Heat maps showing the first (left plot) and second (right plot) moments of $|\mathbb{W}_1^t|$. The x-axis represents the number of random initial points and the y-axis represents the exponential distribution parameter.

Lemma 2. *In the central domain, the distribution of $|\mathbb{W}_1^t|$ approaches a Normal distribution with mean (μ_t) and variance (σ_t^2) as follows:*

$$\mu_t = \sum_i^{|\mathbb{X}_t|} \nu(p_i) e^{-\nu(p_i)} \tag{27}$$

$$\sigma_t^2 = \sum_i^{|\mathbb{X}_t|} \nu(p_i) e^{-\nu(p_i)} - \sum_i^{|\mathbb{X}_t|} \left(\nu(p_i) e^{-\nu(p_i)}\right)^2 \tag{28}$$

$$-\frac{1}{\beta|\mathbb{X}_t|}\left(\sum_i^{|\mathbb{X}_t|} \nu(p_i) e^{-\nu(p_i)}(\nu(p_i)-1)\right)^2$$

where $\nu(p_i) = \beta|\mathbb{X}_t|p_i$. In the left-hand 1-domain, the distribution of $(m_t - |\mathbb{W}_1^t|)/2$ approaches a Poisson distribution

$$\lim \mathbb{P}\left\{\frac{m_t - |\mathbb{W}_1^t|}{2} = k\right\} = \frac{\lambda^k}{k!} e^{-\lambda} \tag{29}$$

where the parameter λ is given by

$$\lambda = \lim \frac{m_t^2}{2} \sum_i^{|\mathbb{X}_t|} p_i^2 \tag{30}$$

In the right-hand 1-domain, the distribution of $|\mathbb{W}_1^t|$ approaches a Poisson distribution with parameter λ given by

$$\lambda = \lim m_t \sum_{i=1}^{|\mathbb{X}_t|} p_i (1-p_i)^{m_t - 1} \tag{31}$$

In all of the above, the probability mass of $|\mathbb{W}_1^t|$ is concentrated on some small values if the distribution of $\{p_i\}$ is non-uniform. This explains the experimental analysis provided in Fig. 2. In the context of Algorithm 2, this implies that a low-complexity problem will once again drastically drive the nonconvexity regret to zero, similar to what we observed for Algorithm 1. This phenomenon is demonstrated in Section V.

V. Numerical Results

The objective of this section is to support the results of this paper through numerical analysis and demonstrate the role that a single comparatively low-complexity problem can play in a sequence of nonconvex problems. First, we will illustrate the performance of Algorithm 1 with respect to the parameters ρ_t and m, i.e., the fraction of the solution space belonging to the region of attraction of the global minimum at time t and the number of random initial points, respectively. This analysis

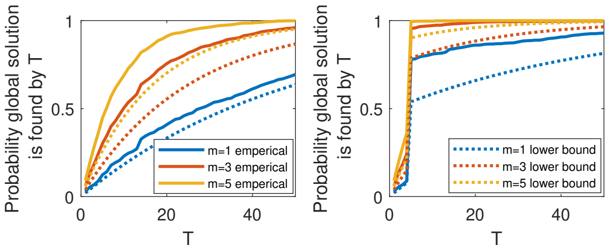

Fig. 3. Empirical validation of Theorem 1.

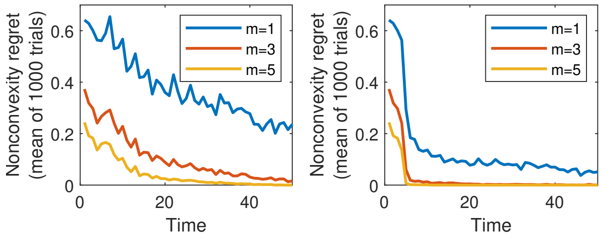

Fig. 4. Nonconvexity regret over time resulting from applying Algorithm 1 to Case 1 (left) and Case 2 (right).

considers two cases satisfying the conditions of Theorem 1 with $r = \pi$ and $T = 50$:

1) *"No low-complexity problem":* In this case, $\{f_t\}_{t=1}^{50}$ are deterministic nonconvex functions bounded between -1 and 1 with the number of local minima ranging from 51 to 247. While the number of local minima varies, ρ_t is kept at 0.02 for $t = 1, ..., 50$. A representative function from this sequence is shown on the top plot of Fig. 1.
2) *"Low-complexity problem at time 5":* This case is identical to Case 1 at every time period except for time 5. The bottom plot of Fig. 1 shows $\bar{f}_5$, which has three local minima and $\bar{\rho}_5 = 0.5$.

We conducted 1000 trials of Algorithm 1 on Case 1 and Case 2 in three scenarios of $m = 1$, $m = 3$, and $m = 5$. Fig. 3 plots the empirical probability that $\bar{t} \leq T$ versus the theoretical lower bound provided in Theorem 1. Note that for the same value of m, the two cases are identical until time 5. The results support Theorem 1. The distribution of $\bar{t}$ is a key driver of the nonconvex regret over time, which is shown in Fig. 4. In Case 2 (right plot), the nonconvexity regret falls sharply at time 5 because the majority of trials achieve zero nonconvexity regret at this time, if they had not before. However, even without such "low-complexity problem" at time 5, the nonconvexity regret still trends downward over time on average. Further, Fig. 4 highlights the role of increasing m in improving regret across cases.

Now consider Algorithm 2. Again, we consider a "no low-complexity problem" case (Case 1-b) and a "low-complexity problem at time 5" (Case 2-b) case. However, the function sequence $\{f_t\}_{t=1}^{50}$ is now more complex with the number of local minima between 250 and 1250 and ρ_t kept at 0.004 for $t = 1, ..., 50$. The low-complexity problem $\bar{f}_5$ is unchanged. Fig. 5 plots the observed nonconvexity regret over time for these two cases under different parameter values.

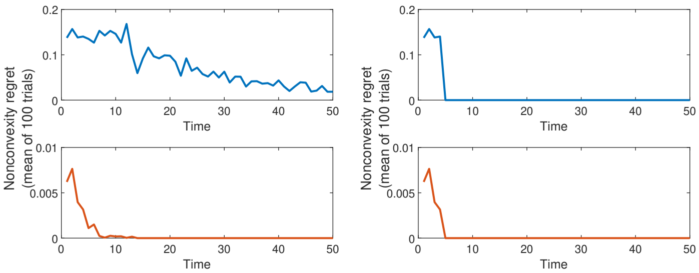

Fig. 5. Nonconvexity regret over time resulting from applying Algorithm 2 to Case 1-b (left) and Case 2-b (right), with parameters $\alpha = 0.1, \delta = 0.1, \bar{m} = 10$ (top) and $\alpha = 0.1, \delta = 0.9, \bar{m} = 125$ (bottom).

VI. Conclusion

In this paper, we studied how the regret attributed to the nonconvexity evolves over time in an online nonconvex optimization (ONO) setting. We showed that the probability of finding and tracking the global solution over time via a local search method that uses memory and random initialization at each time instance depends on the robustness of the global minimum. The results imply that the existence of a single low-complexity problem (among the sequence of nonconvex problems) can lower the complexity of the entire ONO problem. We developed various bounds to quantify the nonconvexity regret and its asymptotic behavior.

References

[1] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” *Advances in Neural Information Processing Systems*, pp. 6389–6399, 2018.

[2] E. J. Candes and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[3] E. J. Candes and Y. Plan, “Matrix completion with noise,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[4] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, “Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations,” *IEEE Transactions on Automatic Control*, vol. 48, no. 2, 2003.

[5] F. Zohrizadeh, C. Josz, M. Jin, R. Madani, J. Lavaei, and S. Sojoudi, “Conic relaxations of power system optimization: Theory and algorithms,” *to appear in European Journal of Operational Research*, 2020.

[6] M. S. Asif and J. Romberg, “Sparse recovery of streaming signals using l1-homotopy,” *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4209–4223, 2014.

[7] Y. T. ad Krishnamurthy Dvijotham, , and S. Low, “Real-time optimal power flow,” *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2963–2973, 2017.

[8] E. C. Hall and R. M. Willett, “Online convex optimization in dynamic environments,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 647–662, 2015.

[9] O. Besbes, Y. Gur, and A. Zeevi, “Non-stationary stochastic optimization,” *Operations Research*, vol. 63, no. 5, pp. 1227–1244, 2015.

[10] J. Guddat, D. Nowack, and F. Guerra, *Parametric Optimization: Embeddings, Path Following and Singularities*. Springer, 1990.

[11] O. Massicot and J. Marecek, “On-line non-convex constrained optimization,” 2019, https://arxiv.org/pdf/1909.07492.pdf.

[12] Y. Ding, J. Lavaei, and M. Arcak, “Escaping spurious local minimum trajectories in online time-varying nonconvex optimization,” 2019, https://lavaei.ieor.berkeley.edu/Online_opt_2019_2.pdf.

[13] S. Fattahi, C. Josz, R. Mohammadi, J. Lavaei, and S. Sojoudi, “On the absence of spurious local trajectories in online nonconvex optimization,” 2019, https://lavaei.ieor.berkeley.edu/Time_Varing_2019_1.pdf.

[14] J. Mulvaney-Kemp, S. Fattahi, and J. Lavaei, “Smoothing property of load variation promotes finding global solutions of time-varying optimal power flow,” 2020, https://lavaei.ieor.berkeley.edu/DOPF_2020_2.pdf.

[15] M. Zinkevich, “Online convex programming and generalized infinitesimal gradient ascent,” *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

[16] A. Jadbabaie, S. S. Alexander Rakhlin, , and K. Sridharan, “Online optimization: Competing with dynamic comparators,” *Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, vol. 38, pp. 398–406, 2015.

[17] E. Hazan, *Introduction to Online Convex Optimization*. Independently published, 2019, https://arxiv.org/pdf/1909.05207.pdf.

[18] A. Simonetto, A. K. Aryan Mokhtari, G. Leus, and A. Ribeiro, “A class of prediction-correction methods for time-varying convex optimization,” *IEEE Transactions on Signal Processing*, vol. 64, no. 17, pp. 4576–4591, 2016.

[19] M. Fazlyab, C. Nowzari, G. J. Pappas, A. Ribeiro, and V. M. Preciado, “Self-triggered time-varying convex optimization,” *Conference on Decision and Control (CDC)*, 2016.

[20] A. Bernstein, E. Dall’Anese, and A. Simonetto, “Online primal-dual methods with measurement feedback for time-varying convex optimization,” *Conference on Decision and Control (CDC)*, 2018, https://arxiv.org/abs/1804.05159.

[21] A. Simonetto, “Time-varying convex optimization via time-varying averaged operators,” 2017, https://arxiv.org/abs/1704.07338v3.

[22] E. Hazan, K. Singh, and C. Zhang, “Efficient regret minimization in nonconvex games,” *International Conference on Machine Learning (ICML)*, vol. 3, pp. 2278–2288, 2017.

[23] Y. Tang, E. Dall’anese, A. Bernstein, and S. Low, “Running primal-dual gradient method for time-varying nonconvex problems,” 2017, https://arxiv.org/pdf/1812.00613.pdf.

[24] L. Yang, L. Deng, M. H. Hajiesmaili, C. Tan, and W. S. Wong, “An optimal algorithm for online non-convex learning,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 25, 2018.

[25] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[26] R. Martí, *Multi-Start Methods*. Boston, MA: Springer US, 2003, pp. 355–368. [Online]. Available: https://doi.org/10.1007/0-306-48056-5_12

[27] C. Boender and A. R. Kan, “Bayesian stopping rules for multistart global optimization methods,” *Mathematical Programming*, vol. 37, no. 1, pp. 59–80, 1987.

[28] T. A. Feo and M. G. Resende, “Greedy randomized adaptive search procedures,” *Journal of global optimization*, vol. 6, no. 2, pp. 109–133, 1995.

[29] W. E. Hart, “Sequential stopping rules for random optimization methods with applications to multistart local search,” *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 270–290, 1998.

[30] T. Dick, E. Wong, and C. Dann, “How many random restarts are enough?” 2014. [Online]. Available: https://www.cs.cmu.edu/~epxing/Class/10715-14f/project-reports/DannDickWong.pdf

[31] B. T. Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[32] P. T. Boggs, “The solution of nonlinear systems of equations by a-stable integration techniques,” *SIAM Journal on Numerical Analysis*, vol. 8, no. 4, pp. 767–785, 1971.

[33] L. Lessard, B. Recht, and A. Packard, “Analysis and design of optimization algorithms via integral quadratic constraints,” *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[34] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *International Conference on Machine Learning*, 2016, pp. 1225–1234.

[35] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, “On the convergence rate of incremental aggregated gradient algorithms,” *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1035–1048, 2017.

[36] Y. Chi, Y. M. Lu, and Y. Chen, “Nonconvex optimization meets low-rank matrix factorization: An overview,” *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[37] D. A. McAllester and R. E. Schapire, “On the convergence rate of good-turing estimators,” *Conference on Learning Theory (COLT)*, pp. 1–6, 2000.

[38] V. F. Kolchin, B. A. Sevast’yanov, V. P. Chistyakov, and A. V. Balakrishnan, *Random Allocations*. Washington, D.C. Winston, 1978.