A Hitting Time Analysis of Non-convex Optimization with Time-Varying Revelations

Han Feng, Ali Yekkehkhany, and Javad Lavaei

Industrial Engineering and Operations Research, University of California, Berkeley

***Abstract*— Time-varying optimization is an integral part of online learning, where we minimize a function whose values are revealed over time. To understand the time-varying phenomenon of optimizing non-convex functions with changing local or global minima, we propose to study two models of time variation. In the first model, additive noisy evaluations of functions are revealed over time, for which we obtain bounds for the local minima of the revealed function to coincide with those of the true function. Moreover, we generalize the additive noisy model to a linear variation model, and define the notion of shape dominant operator. In this second model, the noisy revealed function is able to reach the true function quickly — both the bias and noise decrease exponentially fast. Under the further assumption of sub-Gaussian perturbation, we bound the hitting time for reaching a neighborhood of the target function. Even though the theoretical analysis is performed for discrete functions, we discuss the implications of our results on optimization over continuous domains.**

I. Introduction

In many practical applications of optimization, such as those in the training of neural networks [1], [2], in online advertising [3], in decision-making process of power systems [4], [5], and in the real-time state estimation of nonlinear systems [6], the parameters of the problem are often uncertain and change over time. Furthermore, the decision made in the current round can affect the problem to be optimized in the future. In its most general form, a time-varying optimization problem aims to find the solution trajectories determined by

$$x_t^* = \arg\min_{x\in\mathcal{X}} \left\{ f_t(x) = \mathbb{E}F_t(x,\xi) \right\}, \quad t = 1, 2, \ldots, \tag{1}$$

where the random variable ξ models the uncertainty in the objective that comes from disturbance, inexactness of model, use of small batches, or injected noise. The expectation $\mathbb{E}$ over the ξ can only be evaluated approximately since the nature of the noise is often unknown. Furthermore, due to the limitation of numerical solvers, the operator $\arg\min$ in (1) often returns a local solution satisfying the first- and second-order necessary optimality conditions, while it is desirable to find a global solution. In practice, the series of problems in (1) are normally solved in an online fashion, where the current solution x_t^* is obtained based on the previous solutions and the observations of the objective functions [7]. Our work attempts to address the theoretical gap introduced by the online optimization of a non-convex function over time to a local minimum.

Email: han_feng@berkeley.edu, aliyek@berkeley.edu, lavaei@berkeley.edu
This work was supported by grants from ARO, ONR, AFOSR, and NSF.

When $f_t(x)$ is convex, there is no difference between local and global minima, and efficient algorithms have been proposed to track the solution trajectories [8]. In particular, the field of online convex optimization has studied a wide range of algorithms with regret guarantees [7], [9]. When the objective $f_t(x)$ is non-convex and changes over time, the notion of regret can be generalized to local regret [10]. Despite the use of various numerical algorithms in the non-convex setting [11] , non-trivial definitions of local and global solution trajectories [12], the concept of no-spurious solutions and the singularity issue [13] all stand in the way of a clear picture of online optimization algorithms. The existing results on benign landscape of optimization [14]–[16] and the escape of saddle points [17], [18] do not seem to have a natural counterpart.

To further motivate the study and the generality of the framework (1), we first explain several examples and discuss related works along the way.

A. Bandit Optimization

Consider the classical multi-armed bandit problem, where a set $\mathcal{X}$ of arms is to be pulled in each round. After a decision is made, a stochastic loss $l_t(x_t, \xi_t)$ is incurred. The goal is to find a strategy that generates a series of arms $x_1, \ldots, x_t$ such that a particular notion of regret is sub-linear in t. One solution strategy is the online mirror descent [19], which can be written in the proximal form

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \left\{ \eta\langle \nabla l_t(x_t,\xi), x\rangle + D_R\langle x, x_t\rangle \right\},$$

where $D_R(u,v) = R(u) - R(v) - \langle \nabla R(v), u - v\rangle$ is the Bregman divergence induced by a strictly convex function $R(\cdot)$. This is an instance of optimizing a time-varying function with uncertainty. Regularization in the form of $D_R(u,v)$ is widely used in online optimization and the recent work [20] has provided a theoretical analysis of its effect on eliminating spurious local minima.

B. Model Predictive Control

Model predictive control (MPC) has found widespread applications in control systems due to its ability in handling hard constraints on inputs and states [21]. One instance of

MPC with horizon T is given by

$$\begin{aligned} V_t(x_t) = &\min_{\bar{u}_t,\dots,\bar{u}_{t+T-1}} c_{t+T}(\bar{x}_{t+T}) + \sum_{k=t}^{t+T-1} c_k(\bar{x}_k, \bar{u}_k) \\ s.t. \quad & \bar{x}_{k+1} = g_k(\bar{x}_k, \bar{u}_k), \text{ for } t \le k \le t+T-1 \\ & \bar{x}_t = x_t \\ & \bar{x}_k \in X_k, \bar{u}_k \in U_k, \text{ for } t \le k \le t+T-1, \end{aligned}$$

where X_k and U_k are the sets of allowable states and actions; $g_k(\cdot)$ and $c_k(\cdot)$ represent the dynamics and cost, respectively. After the solutions $\bar{u}_t^*, \dots, \bar{u}_{t+T-1}^*$ are obtained, MPC applies the control action $u_t = \bar{u}_t^*$, makes the transition to x_{n+1} and re-solves the problem $V_{t+1}(x_{t+1})$ to obtain u_{t+1}. The series of actions u_t comes from the optimization of a time-varying function. When the costs $c_k(\cdot,\cdot)$ are non-convex or come from simulations, the function to be optimized is noisy and non-convex.

C. Empirical Risk Minimization

Many machine learning applications involve solving a regression problem whose goal is to find the best parameter x that minimizes the empirical loss. Formally, given the samples $\xi_1, \dots, \xi_n$, we solve

$$\min \frac{1}{n} \sum_{i=1}^{n} l(x, \xi_i).$$

The loss function $l(x, \xi)$ is often non-convex in x, which may arise from the use of deep neural networks [1]. In practice, the optimization is often solved iteratively. At iteration t, we obtain a small set of samples $S_t \subseteq \{1, 2, \dots, n\}$ and then run one-step stochastic gradient descent (SGD) [22]:

$$\begin{aligned} x_{t+1} &= x_t - \frac{\eta}{|S_t|} \sum_{i \in S_t} \nabla l(x, \xi_i) \\ &= \arg\min_x \left\{ \frac{1}{|S_t|} \sum_{i \in S_t} \langle \nabla l(x_t, \xi_i), x \rangle + \frac{1}{2\eta} \|x_t - x\|^2 \right\}, \end{aligned}$$

where $\eta > 0$ is the step size. A large body of work addresses the issue of convergence and generalization properties of SGD [18], [23]–[25]. We observe that the iterates of SGD can be modeled as the solutions to a time-varying optimization problem, and the functions to be optimized over time are noisy.

D. Contribution

We investigate the optimization of time-varying functions and ask whether optimization of functions that reveal their values over time can provide useful information about the optimization of the ground truth. This study requires assumptions on the noise and on the manner of time revelation of functions. In Section II, we study our first model, where the functions revealed over time have additive noise. We develop a bound on the hitting time for the local minima of the sequential optimization problem to coincide with the local minima of the true model; we further specialize the model to the optimization of a unimodal function. Even though the two sections above consider functions over the discrete domain and the noise is additive, they provide insight into a much broader phenomenon. We describe the connections between the local minima of continuous and discrete optimization problems in Section III. Furthermore, in Section IV, we explain how nonlinear variations of functions over time can be described by a linear model. The generality of the linear model allows us to describe hitting time with the new notion of shape-dominant operator that drives the function towards a particular target. We characterize the approximate shape of the function in finite time and bound the hitting time for reaching a neighborhood of the target function. It is shown that an eigenvector of the operator modeling the time-variation plays a key role in shaping the time-varying function.

II. Optimization of Functions with Additive Noise

Consider an unknown discrete function $f : \mathcal{X} \to \mathcal{R}$, where $\mathcal{X} \subseteq \mathbb{Z}^d$ is a bounded subset of d integer tuples and $\mathcal{R} \subseteq \mathbb{R}$ is a subset of real numbers ($\mathbb{Z}$ and $\mathbb{R}$ denote the sets of integer and real numbers). Denote the strict local minima and maxima, known collectively as strict local extrema, of the unknown function f by $\mathcal{X}^* = \{x^* \in \mathcal{X} : f(x^*) < f(x), \forall x \in \mathcal{B}(x^*)\} \cup \{x^* \in \mathcal{X} : f(x^*) > f(x), \forall x \in \mathcal{B}(x^*)\}$, where $\mathcal{B}(x^*) = \cup_{j=1}^{d} \{x^* \pm e_j\} \cap \mathcal{X}$ with $e_1, \dots, e_d$ being the standard basis of $\mathbb{Z}^d$. The goal is to find $\mathcal{X}^*$, the set of strict local extrema of the unknown function f. Although the function f is unknown, inquiries of the function values at given input points can be made in consecutive rounds, which are evaluated with added noise signals that are mean zero, independent and identically distributed (i.i.d.) over time and over $\mathcal{X}$. Formally speaking, the revealed values of the target function f at round $t \in \{1, 2, \dots\}$ are

$$f_t(x) = f(x) + N_t(x), \ \forall x \in \mathcal{X}, \tag{2}$$

where $N_t(x)$ are i.i.d. random variables that are strictly bounded by an interval with length L and $\mathbb{E}[N_t(x)] = 0$ for all $t \in \{1, 2, \dots\}$ and $x \in \mathcal{X}$. In order to simplify the analysis, function values at adjacent points are considered to be different so that their noisy values become distinguishable after enough observations. Note that if the noise is disruptive enough, a single set of observed noisy function values $f_t(x)$ for all $x \in \mathcal{X}$ may not represent the unknown target function accurately, making it impossible to find local extrema of the function. Putting this into perspective, the estimate of the target function f at round $t-1$ can be updated with the new observation at round $t \in \{2, 3, \dots\}$ as

$$\widehat{f}_t(x) = \frac{t-1}{t} \cdot \widehat{f}_{t-1}(x) + \frac{1}{t} \cdot f_t(x), \ \forall x \in \mathcal{X}. \tag{3}$$

Note that the estimated function $\widehat{f}_t(x)$ changes over time and may not be well-behaved from an optimization perspective when t is small. However, there is a point of time, called hitting time T, at which optimizing the estimated function $\widehat{f}_t$ determines the local extrema of the target function f with an associated confidence level $1 - a$, where $0 < a \le 1$. As a result, the complexity of finding local extrema of the target function f may be irrelevant to the complexity of finding the local extrema of the function $\widehat{f}_t$ before the hitting time T. Consequently, the complexity of finding the local extrema of the target function f is related to the hitting time T and the computational complexity of the function $\widehat{f}_T$. By denoting the set of strict local extrema of $\widehat{f}_t$ by $\widehat{\mathcal{X}}_t^* = \{\widehat{x}^* \in \mathcal{X} : \widehat{f}_t(\widehat{x}^*) < \widehat{f}_t(x), \forall x \in \mathcal{B}(\widehat{x}^*)\} \cup \{\widehat{x}^* \in \mathcal{X} : \widehat{f}_t(\widehat{x}^*) > \widehat{f}_t(x), \forall x \in \mathcal{B}(\widehat{x}^*)\}$, the hitting time $T(a)$ for $0 < a \leq 1$ can be defined as

$$T(a) = \min\left\{T : \mathbb{P}\left(\widehat{\mathcal{X}}_T^* = \mathcal{X}^*\right) \geq 1 - a\right\}, \tag{4}$$

where $\mathbb{P}$ takes the probability of the event.

The minimum distance of the function values of f at point $x \in \mathcal{X}$ from the function values at its neighbor points is denoted by $\delta(x)$, which is defined as

$$\delta(x) = \min_{x' \in \mathcal{B}(x)} \|f(x) - f(x')\|, \tag{5}$$

where $\|\cdot\|$ is the L^2-norm throughout this article. It is assumed that $\delta(x) > 0$ for all $x \in \mathcal{X}$, which implies that $\delta_m = \min_{x \in \mathcal{X}} \delta(x) > 0$.

Theorem 1: Consider the time-varying function $\hat{f}_t$ in (3). The hitting time $T(a)$, defined in Equation (4), is bounded by

$$T(a) \leq \frac{2L^2}{\delta_m^2} \cdot \ln\left(\frac{2|\mathcal{X}|}{a}\right), \tag{6}$$

where $|\mathcal{X}|$ denotes the number of elements in the set $\mathcal{X}$ and depends on the dimension d.

Proof: In order to find an upper bound on the hitting time $T(a)$, note that the hitting event used in Equation (4) satisfies the condition

$$\left\{\frac{1}{T} \cdot \left\|\sum_{t=1}^{T} N_t(x)\right\| < \frac{\delta(x)}{2},\ \forall x \in \mathcal{X}\right\} \subseteq \left\{\widehat{\mathcal{X}}_T^* = \mathcal{X}^*\right\}. \tag{7}$$

The above equation holds because Equations (2) and (3) result in $\widehat{f}_T(x) = f(x) + \frac{1}{T} \cdot \sum_{t=1}^{T} N_t(x)$, and if the magnitude of the added noise to the true value of function f at point x is less than half of $\delta(x)$ for all $x \in \mathcal{X}$, the set of the local extrema of the function $\widehat{f}_T$ coincides with set $\mathcal{X}^*$, the local extrema of function f. The probability of the event on the left-hand side of Equation (7) can be lower bounded as

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{T} \cdot \left\|\sum_{t=1}^{T} N_t(x)\right\| < \frac{\delta(x)}{2},\ \forall x \in \mathcal{X}\right\} \\
&\overset{(a)}{=} \prod_{i=1}^{|\mathcal{X}|} \mathbb{P}\left\{\frac{1}{T} \cdot \left\|\sum_{t=1}^{T} N_t(x)\right\| < \frac{\delta(x)}{2}\right\} \\
&\overset{(b)}{\geq} \prod_{i=1}^{|\mathcal{X}|} \left(1 - 2\exp\left(-\frac{T\delta(x)^2}{2L^2}\right)\right) \\
&> 1 - 2\sum_{i=1}^{|\mathcal{X}|} \exp\left(-\frac{T\delta(x)^2}{2L^2}\right) \\
&\geq 1 - 2|\mathcal{X}| \cdot \exp\left(-\frac{T\delta_m^2}{2L^2}\right),
\end{aligned} \tag{8}$$

where (a) holds because the added noise signals are independent from each other and (b) follows from Hoeffding's inequality. Putting Equations (7) and (8) together, we have

$$\mathbb{P}\left\{\widehat{\mathcal{X}}_T^* = \mathcal{X}^*\right\} > 1 - 2|\mathcal{X}| \cdot \exp\left(-\frac{T\delta_m^2}{2L^2}\right). \tag{9}$$

If $1 - 2|\mathcal{X}| \cdot \exp\left(-\frac{T\delta_m^2}{2L^2}\right) \geq 1 - a$ or equivalently $T \geq \frac{2L^2}{\delta_m^2} \cdot \ln\left(\frac{2|\mathcal{X}|}{a}\right)$, we have

$$\mathbb{P}\left\{\widehat{\mathcal{X}}_T^* = \mathcal{X}^*\right\} > 1 - a, \tag{10}$$

from which the upper bound in Equation (4) follows. ∎

A. A Special Case for Unimodal Functions

A function f over an bounded set $\mathcal{X} \subseteq \mathbb{Z}$ is unimodal if it has only one global minimum $x^* \in \mathcal{X}$ such that $f(i) > f(j)$ for all $i < j \leq x^*$ while $f(i) < f(j)$ for all $x^* \leq i < j$. Assume that the unknown target function f is unimodal over $\mathcal{X} \subseteq \mathbb{Z}$, which implies it has a single global minimum. As mentioned earlier, the time-varying function $\widehat{f}_t$ may not even be unimodal for small values of t under disruptive noise, and therefore it could have multiple local extrema. However, the single global minimum of the function f becomes known after the hitting time with an associated confidence level. In this section, a new notion of hitting time is proposed for unimodal functions that captures the complexity of finding the global minimum of the function and does not take the local extrema of the estimated function $\widehat{f}_t$ into account.

We assume that the noise signals are continuous random variables, so the function $\widehat{f}_t$ has a single global minimum with probability one achieved at $\widehat{x}_t^* = \arg\min_{x \in \mathcal{X}} \ \widehat{f}_t(x)$. The hitting time $T_u(a)$ for the unimodal function f with its global minimum at $x^* = \arg\min_{x \in \mathcal{X}} \ f(x)$ for $0 < a \leq 1$ is defined as

$$T_u(a) = \min\left\{T : \mathbb{P}\big(\widehat{x}_T^* = x^*\big) \geq 1 - a\right\}. \tag{11}$$

The distance of the function value at point $x \in \mathcal{X}$ from the minimum function value is denoted by $\Delta(x)$, which is defined as

$$\Delta(x) = \begin{cases} f(x) - f(x^*), & \text{if } x \in \mathcal{X} \setminus \{x^*\}, \\ \min\{f(x^* - 1) - f(x^*), f(x^* + 1) - f(x^*)\}, & \\ & \text{if } x = x^*. \end{cases} \tag{12}$$

Theorem 2: Consider the time-varying function $\hat{f}_t$ defined in (3) with f being a unimodal function. The hitting time $T_u(a)$, defined in Equation (11), satisfies the inequality $T_u(a) \leq T$ where T is the smallest number such that

$$\exp\left(-\frac{T\delta_m^2}{2L^2}\right) + 2\sum_{i \in [\lfloor \frac{|\mathcal{X}|}{2} \rfloor]} \exp\left(-\frac{Ti^2\delta_m^2}{2L^2}\right) \leq a. \tag{13}$$

Such number T exists because the left-hand side approaches 0 when $T \to \infty$.

Proof: Note that $\Delta(x) > 0$ for all $x \in \mathcal{X}$ by construction. In order to find an upper bound on the hitting time $T_u(a)$, the hitting event used in Equation (11) satisfies the following condition:

$$\left\{\frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x) > -\frac{\Delta(x)}{2}, \forall x\in\mathcal{X}\setminus\{x^*\}\ \textbf{and}\right.$$
$$\left.\frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x^*) < \frac{\Delta(x^*)}{2}\right\} \subseteq \left\{\widehat{x}_T^* = x^*\right\}. \tag{14}$$

The probability of the event on the left-hand side of Equation (14) can be lower bounded as

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x) > -\frac{\Delta(x)}{2}, \forall x\in\mathcal{X}\setminus\{x^*\}\right.\\
&\qquad\left.\textbf{and } \frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x^*) < \frac{\Delta(x^*)}{2}\right\}\\
&\overset{(a)}{=} \mathbb{P}\left\{\frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x^*) < \frac{\Delta(x^*)}{2}\right\}\\
&\qquad\times\prod_{x\in\mathcal{X}\setminus\{x^*\}}\mathbb{P}\left\{\frac{1}{T}\cdot\sum_{t=1}^{T}N_t(x) > -\frac{\Delta(x)}{2}\right\}\\
&\overset{(b)}{\geq} \left(1-\exp\left(-\frac{T\Delta(x^*)^2}{2L^2}\right)\right)\\
&\qquad\times\prod_{x\in\mathcal{X}\setminus\{x^*\}}\left(1-\exp\left(-\frac{T\Delta(x)^2}{2L^2}\right)\right)\\
&> 1-\exp\left(-\frac{T\Delta(x^*)^2}{2L^2}\right) - \sum_{x\in\mathcal{X}\setminus\{x^*\}}\exp\left(-\frac{T\Delta(x)^2}{2L^2}\right)\\
&\overset{(c)}{\geq} 1-\exp\left(-\frac{T\delta_m^2}{2L^2}\right) - \sum_{x\in\mathcal{X}\setminus\{x^*\}}\exp\left(-\frac{T(x-x^*)^2\delta_m^2}{2L^2}\right)\\
&\overset{(d)}{\geq} 1-\exp\left(-\frac{T\delta_m^2}{2L^2}\right) - 2\sum_{i\in[\lfloor\frac{|\mathcal{X}|}{2}\rfloor]}\exp\left(-\frac{Ti^2\delta_m^2}{2L^2}\right)
\end{aligned} \tag{15}$$

where (a) holds true by the independence property of the added noise signals, (b) is due to Hoeffding's inequality, (c) is true because function f is unimodal, $\Delta(x^*)\geq\delta_m$, and $\Delta(x)\geq(x-x^*)\delta_m$, and (d) results from minimizing the equation with respect to the value of x^* that gives rise to $x^* = \left\lceil\frac{|\mathcal{X}|}{2}\right\rceil$ (taking the ceiling corresponding to the summation through $\left\lfloor\frac{|\mathcal{X}|}{2}\right\rfloor$). Putting Equations (14) and (15) together concludes the proof. ■

Remark 1: It can be verified that Theorem 2 provides a better bound than Theorem 1 because the properties of the unimodal functions are leveraged.

III. From Continuous Function to its Discretization

The hitting time analyzed in the previous two sections is defined over a discrete domain. However, with additional assumptions, the results can be carried over to continuous functions. The connection between continuous and discrete functions will be studied in this section via the notion of discretization, which is widely used in numerical analysis.

A function $f:\mathbb{R}^d\to\mathbb{R}$ is L-Lipschitz if $|f(x)-f(y)|\leq L\|x-y\|$ for all $x,y\in\mathbb{R}^d$. A finite subset $\mathcal{X}\subseteq\mathbb{R}^d$ is a gridding of a continuous function f if it can be decomposed into the Cartesian product

$$\mathcal{X} = \mathcal{X}_1\times\mathcal{X}_2\cdots\mathcal{X}_d, \tag{16}$$

where $\mathcal{X}_i = \{x_{i,1},\ldots,x_{i,l_i}\}\subseteq\mathbb{R}$ is a finite set for $i=1,2,\ldots,d$. The gridding is δ*-uniform* if $x_{i,k}=x_{i,1}+(k-1)\delta$ for all $k=1,2,\ldots,l_i$ and $i=1,2,\ldots d$. The gridding is *dyadic* if all grid coordinates $x_{i,k}$ are dyadic rational numbers that can be written in the form $a/2^b$ where a and b are integers. We use the notation $f|_{\mathcal{X}}$ to denote the restriction of f to a finite set $\mathcal{X}$, and identify $f|_{\mathcal{X}}$ with the n-dimensional vector $(f(x))_{x\in\mathcal{X}}$, where $n=|\mathcal{X}|$.

The notion of local minimum and global minimum of f and $f|_{\mathcal{X}}$ are defined in a way similar to Section II. Precisely, x^* is a global minimum of f over the domain $\mathcal{X}\subseteq\mathbb{R}^d$ if $f(x^*)\leq f(x)$ for all $x\in\mathcal{X}$. In the same fashion, x^* is a strict local minimum of f if there exists an $r>0$ and a neighborhood $B(x^*,r)=\{x\in\mathcal{X}:\|x-x^*\|\leq r\}$ such that $f(x^*)<f(x)$ for all $x\in B(x^*,r)\setminus\{x^*\}$. Given a δ-uniform gridding $\mathcal{X}$, we say that $y^*\in\mathcal{X}$ is a strict local minimum of $f|_{\mathcal{X}}$ if $f(y)>f(y^*)$ for all $y\in B(y^*,\delta)\cap\mathcal{X}\setminus\{y^*\}$. Equivalently, $f(y^*)<f(y^*\pm\delta e_i)$, where $e_1,\ldots,e_d$ are the standard basis of $\mathbb{R}^d$. The *depth* of y^* is defined as $\min_{y\in B(y^*,\delta)\cap\mathcal{X},y\neq y^*}f(y)-f(y^*)$.

Lemma 1: Suppose that $f:\mathbb{R}^d\to\mathbb{R}$ is L-Lipschitz continuous. The following statements hold:

- If x^* is a strict local minimum of f, then for every $r>0$, there exist a δ-uniform dyadic gridding $\mathcal{X}$ and a $y^*\in\mathcal{X}$ such that y^* is a local minimum of $f|_{\mathcal{X}}$ with $\|y^*-x^*\|<r$.
- Conversely, if y^* is a depth-p local minimum on a δ-uniform dyadic gridding of f and $p>L\delta\sqrt{\frac{d-1}{d}}$, then the function f has a local minimum x^* with $\|x^*-y^*\|<\delta$.

Proof: To prove the first statement, we construct a uniform dyadic gridding. Since x^* is a strict local minimum of f, there exists an $r>0$ such that $f(x^*)<f(x)$ for all $x\in B(x^*,r)$. We select an ϵ-uniform dyadic grid $\mathcal{X}(\epsilon)$ with $\epsilon<r/2$, and let $y^*(\epsilon)\in\mathcal{X}(\epsilon)$ be the global minimum of $f|_{X(\epsilon)\cap B(x^*,r)}$. If $B(y^*(\epsilon),\epsilon)\subseteq B(x^*,r)$, then $y^*(\epsilon)$ is a local minimum of $f|_{\mathcal{X}(\epsilon)}$. Otherwise, we refine the grid by halving ϵ and consider the global minima of $f|_{\mathcal{X}(\epsilon/2)}$, $f|_{\mathcal{X}(\epsilon/4)}$, etc.. Since x^* is a strict local minimum of f, when the grid is fine enough, all grid points close to the boundary of $B(x^*,r)$ will take values higher than the grid points close to x^*. As a result, there exists an integer $l>0$ such that the global minimum $y^*(\epsilon/2^l)$ of $f|_{\mathcal{X}(\epsilon/2^l)\cap B(x^*,r)}$ satisfies $B(y^*(\epsilon/2^l),\epsilon/2^l)\subseteq B(x^*,r)$. This implies that $y^*(\epsilon/2^l)$ is a local minimum of an $\epsilon/2^l$-uniform dyadic gridding of f.

To prove the second statement, consider the set $Y^*=B(y^*,\delta)\cap\mathcal{X}\setminus\{y^*\}=\{y^*\pm\delta e_i, i=1,2,\ldots,d\}$. From the definition of depth, $p\leq f(y)-f(y^*)$ for all $y\in Y^*$. Let x^* be a global minimum of $f|_{\mathrm{conv}(Y^*)}$, where $\mathrm{conv}(Y^*)$ is the convex hull of Y^*. If x^* is on the boundary of $\mathrm{conv}(Y^*)$, there exists a grid point $y \in Y^*$ such that $\|x^* - y\| \leq \delta\sqrt{\frac{d-1}{d}}$, and therefore $f(x^*) \geq f(y) - L\|x^* - y\| \geq f(y^*) + p - L\delta\sqrt{\frac{d-1}{d}} > f(y^*)$, which is a contradiction. Therefore, x^* is in the interior of $\operatorname{conv}(Y^*)$ and is a local minimum of f. ■

Lemma 1 states that any local minimum of f will appear in a fine discretization and, conversely, a deep local minimum of a discrete variant of f implies the existence of a continuous local minimum. As a result, Theorems 1 and 2 developed for the hitting time of the discrete functions can be used for the deep local minima of continuous Lipschitz functions by introducing a fine gridding.

IV. Linear Model of Time-Variation

The additive noise model studied in Sections II provides valuable information about the hitting time, but the time-variation of functions that arise in many real-world problems are of a non-linear nature. We argue in this section the generality of a linear model of time-variation, which is the basis of our study of hitting time under shape-dominant operators to follow.

Recall the standard fact in linear algebra that for any vectors $x, y \in \mathbb{R}^d$, there exists an affine transformation that satisfies $y = Ax + b$, and if $x \neq 0$, there exists a linear transformation that satisfies $y = Ax$. Similar results hold in the Hilbert space $L^2(\mathcal{X})$, where the inner product of f and $g \in L^2(\mathcal{X})$ is defined by $\langle f, g\rangle = \int_{\mathcal{X}} f(x)g(x)dx$. We use the same inner product notation when f and g are defined over a discrete domain. For any nonzero functions $f, g \in L^2$, there exists a bounded linear transformation $\mathcal{T} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ such that $\mathcal{T}f = g$. In fact, one such transformation is given by $\mathcal{T}h = \frac{\langle f,h\rangle}{\langle f,f\rangle} g$. Since the zero function is trivial to optimize, the restriction to linear transformation is a general framework that captures the varying nature of nonlinear functions in the examples of Section I.

We further note that for every scalar $\lambda > 0$, the functions f and λf share the same set of local minima. Rescaling by a positive number does not affect the complexity of the optimization problem. Hence, restricting the linear operators $\mathcal{T}$ to have norm 1 incurs no loss of generality.

In practice, the functions to be minimized are often not specified exactly, due to the rounding error of numerical computation or the inexact nature of the model. We model this limitation by random perturbation w sampled from some distribution. Given a sequence of linear operators $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{t-1}$ such that $\|\mathcal{A}_i\| = \sup_{f\neq 0} \frac{\|\mathcal{A}_i f\|}{\|f\|} = 1$ together with the perturbations $w_0, \ldots, w_{t-1}$, consider the following model of linear time variation:

$$f_{t+1} = \mathcal{T}_t f_t = \mathcal{A}_t f_t + w_t, \quad t = 0, 1, \ldots.$$

What properties the operators $\mathcal{T}_1, \mathcal{T}_2, \ldots$ should satisfy for f_t to reach a target f^* at time $t = \tau$? We will provide an answer with the notion of shape-dominant operator in the next section. To understand the importance of this problem, suppose that at time $t = 0$, we optimize f_0 around a poor local minimum x_0^*. If at $t = \tau$, the function f_τ becomes convex with a unique global minimum x_τ^*, then no matter how optimization is carried out for f_1 through $f_{\tau-1}$, minimizing f_τ will yield the same solution x_τ^*, which is globally optimal. The effect of minimizing f_τ cancels out the sub-optimality at time x_0. Moreover, under some technical conditions, the global solution at time τ can be used to find global solutions at future times using tracking methods [13], [20], [26]. In other words, the shape of f_τ affects the complexity of online optimization in the long run.

V. Shape Dominant Model

As explained in Section III, it suffices to study a discretized model where $f_t : \mathcal{X} \to \mathbb{R}$, for $t = 0, 1, \ldots$, are functions defined on a finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$. Equivalently, f_t is a vector in $\mathbb{R}^n$. Let $P_i(A, w)$ denote the joint distribution of A_i and w_i.

Definition 1: The joint distribution $P(A, w)$ is said to be $(\delta, \sigma, f^*, \phi^*)$ shape dominant if the following conditions hold with probability 1:

1) the unit vector f^* is the eigenvector of A associated with eigenvalue 1;
2) the unit vector ϕ^* is the eigenvector of $A^\top$ associated with eigenvalue 1;
3) $\langle f^*, \phi^*\rangle \neq 0$;
4) all other eigenvalues of A have norm less than $1 - \delta$;
5) conditioned on A, the noise w has zero mean and is sub-Gaussian with parameter σ^2 in the sense that for all $u \in \mathbb{R}^n$ with $\|u\| \leq 1$, we have $\mathbb{E}[\exp(su^\top w)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$.

To understand the conditions in Definition 1, consider the special case where A is a positive stochastic matrix whose column sums are all 1. The unit vector $\phi^* = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$ is the eigenvector of $A^\top$ associated with eigenvalue 1. By the Perron-Frobenius theorem, A also has an all-positive eigenvector f^* with eigenvalue 1, and all other eigenvalues of A have norm strictly less than 1. The vector f^* is the equilibrium distribution of a Markov chain whose transition matrix is A. Therefore, Conditions 1 and 3 are automatically satisfied. Moreover, Condition 2 amounts to requiring that, almost surely, the Markov chain defined by A has a fixed equilibrium f^*.

Theorem 3: Assume that $P_t(A, w)$ is $(\delta, \sigma_t, f^*, \phi^*)$ shape dominant and independent for $t = 1, 2, .., k$. Then,

$$f_k = \mathcal{T}_{k-1} \circ \cdots \circ \mathcal{T}_0 f_0 = \frac{\langle \phi^*, f_0 + \sum_{t=0}^{k-1} w_t\rangle}{\langle \phi^*, f^*\rangle} f^* + v + w,$$

where $\circ$ denotes the composition of linear operatoers and

$$\|v\| \leq (1-\delta)^k \left(\|f_0\| + \frac{\langle \phi^*, f_0\rangle}{\langle \phi^*, f^*\rangle}\right),$$

and w is sub-Gaussian with parameter $\sigma^2 = \left(1 + \frac{1}{\langle \phi^*, f^*\rangle^2}\right) \sum_{t=0}^{k-1} (1-\delta)^{2(k-t)} \sigma_t^2$.

Proof: For $i = 0, 1 \ldots, k-1$, consider the operator $\mathcal{T}_i f = A_i f + w_i$ that is $(\delta, \sigma_i, f^*, \phi^*)$ shape dominant. Construct the subspace

$$\mathcal{G} = \{g \in \mathbb{R}^n, \langle \phi^*, g\rangle = 0\}.$$

Since $\langle \phi^*, f^* \rangle \neq 0$, we have $f^* \notin \mathcal{G}$. Since ϕ^* is the eigenvector of $A_i^\top$, the following holds for all $g \in \mathcal{G}$

$$\langle \phi^*, A_i g \rangle = \langle A_i^\top \phi^*, g \rangle = \langle \phi^*, g \rangle = 0.$$

Therefore, $A_i g \in \mathcal{G}$, and $\mathcal{G}$ is an invariant subspace of A_i in $\mathbb{R}^n$. Let a basis of $\mathcal{G}$ be given by $\{g_1, \dots, g_{n-1}\}$. Then, $B = \{f^*, g_1, \dots, g_{n-1}\}$ is a basis of $\mathbb{R}^n$, under which the linear operator A_i takes the form

$$A_i = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & A_i' & \\ 0 & & & \end{bmatrix}, \tag{17}$$

where A_i' is a random matrix in $\mathbb{R}^{(n-1)\times(n-1)}$. With a slight abuse of notation, we regard A_i' as a linear transformation from $\mathcal{G}$ to $\mathcal{G}$. Note that $\|A'\|_2 \leq 1-\delta$ because all other eigenvalues of A_i have norm less than $1-\delta$.

Under the basis B, f_0 has the representation

$$f_0 = \frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle} f^* + g, \tag{18}$$

where $g \in \mathcal{G}$. As a result,

$$\begin{aligned} f_k &= \mathcal{T}_{k-1} \circ \dots \circ \mathcal{T}_0 f_0 \\ &= A_{k-1} \cdots A_0 f_0 + \sum_{i=0}^{k-1} A_{k-1} \cdots A_{i+1} w_i \\ &= \frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle} f^* + A_{k-1}' \dots A_1' g + \sum_{i=0}^{k-1} A_{k-1} \cdots A_{i+1} w_i. \end{aligned}$$

The norm estimate gives rise to

$$\begin{aligned} \left\| A_{k-1}' \dots A_1' g \right\| &\leq (1-\delta)^k \|g\| \\ &\leq (1-\delta)^k \left(\|f_0\| + \left| \frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle} \right| \right), \end{aligned}$$

where we used the triangle inequality. Similarly, one can write

$$w_i = \frac{\langle \phi^*, w_i \rangle}{\langle \phi^*, f^* \rangle} f^* + h_i,$$

where $h_i \in \mathcal{G}$. We have

$$A_{k-1} \cdots A_{i+1} w_i = \frac{\langle \phi^*, w_i \rangle}{\langle \phi^*, f^* \rangle} f^* + A_{k-1}' \cdots A_{i+1}' h_i.$$

For all $u \in \mathbb{R}^n$ with $\|u\| \leq 1$, it holds that

$$\begin{aligned} &\mathbb{E}[\exp(s\langle u, A_{k-1}' \cdots A_{i+1}' h_i \rangle)] \\ &= \mathbb{E}[\exp(s\langle A_{i+1}'^\top \cdots A_{k-1}'^\top u, h_i \rangle)] \\ &= \mathbb{E}[\exp(s\langle A_{i+1}'^\top \cdots A_{k-1}'^\top u, w_i - \frac{\langle \phi^*, w_i \rangle}{\langle \phi^*, f^* \rangle} f^* \rangle)] \\ &= \mathbb{E}\left[\exp(s\langle A_{i+1}'^\top \cdots A_{k-1}'^\top u, w_i \rangle) \right. \\ &\quad \left. \times \exp(s\langle -\frac{\langle A_{i+1}'^\top \cdots A_{k-1}'^\top u, f^* \rangle}{\langle \phi^*, f^* \rangle} \phi^*, w_i \rangle) \right] \\ &\leq \exp\left(\frac{\sigma_i^2 s^2 \left\| A_{i+1}'^\top \cdots A_{k-1}'^\top u \right\|^2}{2} \right) \\ &\quad \times \exp\left(\frac{\sigma_i^2 s^2}{2} \left(\frac{\langle A_{i+1}'^\top \cdots A_{k-1}'^\top u, f^* \rangle}{\langle \phi^*, f^* \rangle} \right)^2 \right) \\ &\leq \exp\left(\frac{\sigma_i^2 s^2 (1-\delta)^{2(k-i)} \left(1 + \frac{1}{\langle \phi^*, f^* \rangle^2}\right)}{2} \right). \end{aligned}$$

This implies that $A_{k-1}' \cdots A_{i+1}' h_i$ is sub-Gaussian with parameter $\sigma_i^2 (1-\delta)^{2(k-i)} \left(1 + \frac{1}{\langle \phi^*, f^* \rangle^2}\right)$, and thereby, $\sum_{i=0}^{k-1} A_{k-1}' \cdots A_{i+1}' h_i$ is sub-Gaussian with parameter $\sigma^2 = \left(1 + \frac{1}{\langle \phi^*, f^* \rangle^2}\right) \sum_{i=0}^{k-1} (1-\delta)^{2(k-i)} \sigma_i^2$. ■

Theorem 3 states that if the time-varying model is given by shape dominant operators, then f_k decomposes into the sum of the dominating shape f^*, a bias term v that gradually fades away, and a cumulating noise term that discounts noise in previous iterations.

We provide a bound for hitting time in the following theorem.

Theorem 4: Under the same assumptions made in Theorem 3, define

$$\tau_\epsilon = \inf\{k : \exists \lambda \in \mathbb{R} \text{ s.t. } \|f_k - \lambda f^*\| < \epsilon\}, \tag{19}$$

where $\epsilon > 0$. Suppose that $k > \frac{\log 2\left(\|f_0\| + \left|\frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle}\right|\right) - \log \epsilon}{\log \frac{1}{1-\delta}}$. Then,

$$\begin{aligned} &\mathbb{P}(\tau_\epsilon \geq k) \leq \\ &\quad C_n \exp\left(-\frac{\epsilon^2}{32\left(1 + \frac{1}{\langle \phi^*, f^* \rangle^2}\right) \sum_{i=0}^{k-1} (1-\delta)^{2(k-i)} \sigma_i^2} \right). \end{aligned}$$

where C_n is a universal constant depending only on n.

Proof: From the proof of Theorem 3 above, we note the following decomposition

$$f_k = \frac{\langle \phi^*, f_0 + \sum_{i=0}^{k-1} w_i \rangle}{\langle \phi^*, f^* \rangle} f^* + v^{(k)} + w^{(k)},$$

where $\left\| v^{(k)} \right\| < (1-\delta)^k (\|f_0\| + \left| \frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle} \right|)$ and

$$w^{(k)} = \sum_{i=0}^{k-1} A_{k-1}' \cdots A_{i+1}' h_i$$

is sub-Gaussian with parameter $\sigma^2 = \left(1 + \frac{1}{\langle \phi^*, f^* \rangle^2}\right) \sum_{i=0}^{k-1} (1-\delta)^{2(k-i)} \sigma_i^2$. From the definition of the hitting time in (19), we have

$$\mathbb{P}(\tau_\epsilon < k) \geq \mathbb{P}\left(\left\|v^{(k)}\right\| < \epsilon/2, \left\|w^{(k)}\right\| < \epsilon/2\right).$$

When $k > \frac{\log 2\left(\|f_0\| + \left|\frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle}\right|\right) - \log \epsilon}{\log \frac{1}{1-\delta}}$, the bound $\left\|v^{(k)}\right\| < \epsilon/2$ is satisfied. Since $w^{(k)}$ is sub-Gaussian with parameter σ^2, the tail-bound for $w^{(k)}$ yields

$$\begin{aligned}\mathbb{P}\left(\left\|w^{(k)}\right\| < \epsilon/2\right) &= 1 - \mathbb{P}\left(\|w_k\| > \epsilon/2\right) \\ &\geq 1 - C_n \exp\left(-\frac{\epsilon^2}{32\sigma^2}\right),\end{aligned}$$

where C_n is a universal constant depending only on n. ■

To understand the above bound, consider a fixed index k. When σ_i decreases, the bound becomes smaller. As a result, with a smaller random perturbation, it is more likely to reach the target function faster. As ϵ increases, the bound becomes smaller, which matches the intuition that a larger neighborhood is easier to reach than a smaller one.

VI. Conclusion

In this paper, we propose to study optimization of unknown functions that are revealed with noise over time. We propose two models of time revelation, define corresponding notions of hitting time, and prove probabilistic bounds for the hitting time. In the first model, noisy evaluations of the ground truth is revealed. In the second model, linear variations of functions given by shape-dominant models are revealed. We discuss the generality of the proposed models. This provides the first step toward understanding how the complexity of finding the global minima of time-varying functions is related to the properties of the evolution model of the problem.

References

[1] R. Sun, "Optimization for deep learning: Theory and algorithms." [Online]. Available: http://arxiv.org/abs/1912.08957

[2] F. Gu, H. Chang, W. Zhu, S. Sojoudi, and L. El Ghaoui, "Implicit graph neural networks," *To Appear in NeulPS 2020*, 2020.

[3] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," vol. 14, no. 1, pp. 3207–3260.

[4] J. Mulvaney-Kemp, S. Fattahi, and J. Lavaei, "Load variation enables escaping poor solutions of time-varying optimal power flow," 2020.

[5] S. Park, E. Glista, J. Lavaei, and S. Sojoudi, "Homotopy method for finding the global solution of post-contingency optimal power flow," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 3126–3133.

[6] C. Rao, J. Rawlings, and D. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," vol. 48, no. 2, pp. 246–258.

[7] E. Hazan, "Introduction to online convex optimization," vol. 2, no. 3-4, pp. 157–325.

[8] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, "A class of prediction-correction methods for time-varying convex optimization," vol. 64, no. 17, pp. 4576–4591.

[9] A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan, "Artificial Intelligence and Statistics," pp. 398–406.

[10] E. Hazan, K. Singh, and C. Zhang, "Efficient regret minimization in non-convex games," vol. 3, pp. 2278–2288.

[11] Y. Tang, E. Dall'Anese, A. Bernstein, and S. Low, "Running Primal-Dual Gradient Method for Time-Varying Nonconvex Problems." [Online]. Available: http://arxiv.org/abs/1812.00613

[12] J. Guddat, F. G. Vazquez, and H. T. Jongen, *Parametric Optimization: Singularities, Pathfollowing and Jumps*. Springer.

[13] S. Fattahi, C. Josz, R. Mohammadi, J. Lavaei, and S. Sojoudi, "Absence of spurious local trajectories in time-varying optimization." [Online]. Available: http://arxiv.org/abs/1905.09937

[14] R. Ge, J. D. Lee, and T. Ma, "Matrix Completion has No Spurious Local Minimum." [Online]. Available: http://arxiv.org/abs/1605.07272

[15] C. Josz, Y. Ouyang, R. Zhang, J. Lavaei, and S. Sojoudi, "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization," in *Advances in Neural Information Processing Systems 31*, pp. 2441–2449.

[16] L. Venturi, A. S. Bandeira, and J. Bruna, "Spurious valleys in one-hidden-layer neural network optimization landscapes," vol. 20, no. 133, pp. 1–34.

[17] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, pp. 797–842.

[18] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to Escape Saddle Points Efficiently," in *International Conference on Machine Learning*, pp. 1724–1732. [Online]. Available: http://proceedings.mlr.press/v70/jin17a.html

[19] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," vol. 31, no. 3, pp. 167–175.

[20] Y. Ding, J. Lavaei, and M. Arcak, "Escaping spurious local minimum trajectories in online time-varying nonconvex optimization." [Online]. Available: http://arxiv.org/abs/1912.00561

[21] F. Borrelli, A. Bemporad, and M. Morari, *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press.

[22] D. Masters and C. Luschi, "Revisiting Small Batch Training for Deep Neural Networks." [Online]. Available: http://arxiv.org/abs/1804.07612

[23] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis." [Online]. Available: http://arxiv.org/abs/1702.03849

[24] B. Kleinberg, Y. Li, and Y. Yuan, "An alternative view: When does SGD escape local minima?" in *International Conference on Machine Learning*, pp. 2698–2707.

[25] Y. Zhang, P. Liang, and M. Charikar, "A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics," in *Conference on Learning Theory*, pp. 1980–2022. [Online]. Available: http://proceedings.mlr.press/v65/zhang17b.html

[26] O. Massicot and J. Marecek, "On-line Non-Convex Constrained Optimization." [Online]. Available: http://arxiv.org/abs/1909.07492