ANALYSIS OF THE LANDSCAPE OF TIME-VARYING NON-CONVEX OPTIMIZATION PROBLEMS VIA LINEAR OPERATORS

Han Feng
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA, 94720
han_feng@berkeley.edu

Javad Lavaei
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA, 94720
lavaei@berkeley.edu

ABSTRACT
Time-varying optimization is an integral part of online learning, where we minimize a function whose shape changes over time. The complexity of solving a time-varying optimization problem depends on the changes in the geometric shape of the function. To understand what kind of variation guarantees that the function can be efficiently optimized at some point in the future, we first approximate the continuous function with a discrete function and then consider two models of finite rank operators that capture the variation of the function over time. In the first uniform model, we prove the existence of a large number of local minima in expectation and bound the likelihood of achieving a geometric shape with only a few local minima. The complexity of uniform perturbation motivates the development of a second model, where we define the notion of shape dominant operator. Under this model, the function is able to reach a desirable shape significantly faster. Under the further assumption of sub-Gaussian perturbation, we bound the hitting time for reaching the neighborhood of the target function.

1 Introduction
In many practical applications of optimization, such as those in the training of neural networks [1], in online advertising [2], and in the real-time state estimation of nonlinear systems [3], 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

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

where the random variable \(\xi\) models the uncertainty in the objective that comes from disturbance, inexactness of model, use of small batches, or injected noise. 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. The series of problems in (1) are normally solved in an online fashion, where the current solution \(x_t^*\) depends on previous solutions and the observations of the objective functions \([4]\).

When \(f_t(x)\) is convex, there is no difference between local and global minima. Efficient algorithms have been proposed to track the solution trajectories \([5]\). In particular, the field of online convex optimization has studied a wide range of algorithms with regret guarantees \([4, 6]\). When the objective \(f_t(x)\) is non-convex and changes over time, the
Notion of regret can be generalized to local regret [7]. Despite the use of numerical algorithms [8] in the non-convex setting, non-trivial definitions of local and global solution trajectories [9], the concept of no-spurious solutions and the singularity issue [10] all stand in the way of a clear picture of online optimization algorithms. The existing results on benign landscape of optimization [11] and the escape of saddle points [14] do not seem to have an obvious 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.

### 1.1 Empirical Risk Minimization

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

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

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 $S_t \subseteq \{1, 2, \ldots, n\}$ and then run one-step stochastic gradient descent (SGD) [16]:

$$x_{t+1} = x_t - \frac{\eta}{|S_t|} \sum_{i \in S_t} \nabla l(x_t, \xi_i),$$

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

where $\eta > 0$ is the step size. A large body of work addresses the issue of convergence and generalization properties of SGD [15] [17] [18] [19]. We observe that the iterates of SGD can be modeled as the solutions to a time-varying optimization problem.

### 1.2 Bandit Optimization

Consider the classical multi-armed bandit problem, where we are given a set $\mathcal{X}$ of arms to pull in each round. After a decision is made, a stochastic loss $l_t(x_t, \xi_t)$ is suffered. One aims to find a strategy that generates a series of arms $x_1, \ldots, x_t$ such that some notion of regret becomes sub-linear. One solution strategy is the online mirror descent [20], which can be written in the proximal form

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

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 $R$. This is another 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 [21] has provided theoretical analysis of its effect on eliminating spurious local minima.

### 1.3 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 [22]. One instance of MPC with horizon $T$ is given by

$$V_t(x_t) = \min_{\bar{u}_t, \ldots, \bar{u}_{t+T-1}} c_{t+T}(\bar{x}_{t+T}) + \sum_{k=t}^{t+T-1} c_k(\bar{x}_k, \bar{u}_k)$$

$$s.t. \quad \bar{x}_{k+1} = \bar{g}_k(\bar{x}_k, \bar{u}_k), \quad \text{for } t \leq k \leq t + T - 1$$

$$\bar{x}_t = x_t$$

$$\bar{x}_k \in X_k, \bar{u}_k \in U_k, \quad \text{for } t \leq k \leq t + T - 1,$$

where $X_k$ and $U_k$ are the sets of allowable states and actions; $\bar{g}_k(\cdot)$ and $c_k(\cdot)$ represent the dynamics and cost, respectively. After the solutions $\bar{u}_1, \ldots, \bar{u}_{t+T-1}$ are obtained, MPC applies the control action $\bar{u}_t = \bar{u}_t^*$, makes the transition to $x_{t+1}$ and re-solves the problem $V_{t+1}(x_{t+1})$ to obtain $u_{t+1}$. The series of actions $u_t, u_{t+1}, \ldots$ come from the solution of a time-varying optimization problem.
1.4 Reinforcement Learning

Our final example comes from Reinforcement Learning, where the goal is to finding an optimal policy for an uncertain Markov Decision Process [23]. Consider the tuple \((S, A, P, c)\), where \(S\) is the set of states, \(A\) is the set of actions, \(P\) is the transition probability kernel, and \(c(s, a)\) is the cost function. We aim to find a policy \(\pi_\theta(a, s) : A \times S \to [0, 1]\) that optimizes the total cost function of the form

\[
J(\theta) = \mathbb{E}[V_\theta(s)] = \mathbb{E} \left( \sum_{i=1}^{\infty} \gamma^i c(s_i, a_i) \right),
\]

where the expectation is taken over \(a_i \sim \pi_\theta(\cdot|s)\) and the transition \(s_{t+1} \sim P(\cdot|s_i, a_i)\). The value iteration algorithm, which is given by

\[
V^{(t+1)}(s) = \min_a \left\{ c(s, a) + \gamma \sum_{s'} P(s'|s, a) V^{(t)}(s') \right\},
\]

yields a time-varying optimization problem, where every \(V^{(t)}(s)\) and its corresponding optimal action \(a^*\) determine an optimal policy \(a^* = \mu^{(t)}(s)\) at iteration \(t\).

1.5 Contribution

In this work, we investigate the evolution of the shape of the time-varying function under different stochastic models. We first justify the generality of a linear operator model that captures the time-variation of \(f_t(x)\) in Section 2. The infinite-dimensional nature of the linear model motivates the discretization of the time-varying function in Section 3, where we discuss the relationship between the local minima of a continuous function and the local minima of its discrete counterpart. We then study two stochastic models for the time-variation of functions. In the first uniform perturbation model given in Section 4, we first bound the probability that the function becomes monotone and unimodal, and then bound the number of locally optimal solutions. The large number of local solutions in this model motivates the introduction of a second model in Section 5, where we define the 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.

2 Linear Model of Time-Variation

The above examples of time-varying functions arise from different contexts and appear to be of non-linear nature. We argue in this section that it suffices to consider a linear model of time-variation. 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 infinite-dimensional Hilbert space \(L^2\), where the inner product of \(f\) and \(g\) in \(L^2\) is defined by \(\langle f, g \rangle = \int f(x)g(x)dx\). For any nonzero functions \(f, g \in L^2\), there exists a bounded linear transformation \(T : L^2 \to L^2\) such that \(T f = g\). In fact, one such transformation is given by \(T h = \frac{(f, h)}{(f, f)} g\). Since the zero function is trivial to optimize, the restriction to linear transformation is a general framework that captures the varying nature of functions in the examples above.

We further note that for each scalar \(\lambda > 0\), the functions \(f\) and \(\lambda f\) have the same shape. Especially, they 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 \(T\) 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 functions \(w\) sampled from some distribution. Given a sequence of linear operators \(A_0, A_1, \ldots, A_{t-1}\) such that \(\|A_t\| = \sup_{f \neq 0} \frac{\|Tf\|}{\|f\|} = 1\) together with the perturbations \(w_0, \ldots, w_{t-1}\), consider the following model of linear time variation:

\[
f_{t+1} = T_t f_t = A_t f_t + w_t, \quad t = 0, 1, \ldots.
\]

The question we study in this work is: What properties the operators \(T_1, T_2, \ldots\) should satisfy for \(f_t\) to reach a desirable shape at time \(t = \tau\)?

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 \(f_\tau\) becomes convex with a unique global minimum \(x_\tau^*\), then no matter how optimization is
carried out for $f_1$ through $f_{r-1}$, minimizing $f_r$ will yield the same solution $x^*_r$, which is globally optimal. The effect of minimizing $f_r$ cancels out the sub-optimality at time $x_0$. Moreover, under some technical conditions, the global solution at time $\tau$ can be used to find global solutions at future times using tracking methods \cite{21,10,24}. In other words, the shape of $f_r$ affects the complexity of online optimization in the long run.

3 From Continuous Function to its Discretization

In what follows, we will consider a discretization of a continuous function to sidestep the technical issues associated with an infinite-dimensional linear spaces. After introducing the notations, Lemma\footnote{1} studies the sets of local minima of a continuous function and its discrete counterpart.

Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we say that $f$ is $L$-Lipschitz if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$. We say that 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

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

where $\mathcal{X}_i = \{x_{i,1}, \ldots, x_{i,d}\} \subseteq \mathbb{R}$ is a finite set for $i = 1, 2, \ldots, d$. The gridding is $\delta$-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 $n = |\mathcal{X}|$.

We say that $x^*$ is a global minimum of $f$ over the domain $\mathcal{X} \subseteq \mathbb{R}^d$ if $f(x^*) \leq f(x)$ for all $x \in \mathcal{X}$. We say that $x^*$ is a strict local minimum of $f$ if there exists a neighborhood $B(x^*, r) = \{x : \|x - x^*\| \leq r\}$ such that $f(x^*) < f(x)$ for all $x \in B(x^*, r) \setminus \{x^*\}$. Given a $\delta$-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, $f(y^*) < f(y^* + \epsilon e_i)$, 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 $\delta$-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 minimal on a $\delta$-uniform dyadic gridding of $f$ and $p > L\delta \sqrt{\frac{d+1}{4}}$, 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 $\epsilon > 0$ such that $f(x^*) < f(x)$ for all $x \in B(x^*, r)$. We select an $\epsilon$-uniform dyadic grid $\mathcal{X}(\epsilon)$ with $\epsilon < r/2$, and let $y^* \in \mathcal{X}(\epsilon)$ be the global minimum of $f|\mathcal{X}(\epsilon) \cap B(x^*, r)$. 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 $\epsilon$ and consider the global minimum 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 $Y^* = B(y^*, \delta) \setminus \{y^*\} = \{y^* \pm \delta e_i, i = 1, 2, \ldots, d\}$. From the definition of depth, $p \leq f(y) - f(y^*)$ for all $y \in Y^*$. Let $x^*$ be a global minimum of $f|\text{conv}(Y^*)$, where $\text{conv}(Y^*)$ is the convex-hull of $Y^*$. If $x^*$ is on the boundary of $\text{conv}(Y^*)$, there exists a grid point $y \in Y^*$ such that $\|x^* - y\| \leq \delta \sqrt{\frac{d+1}{4}}$, and therefore $f(x^*) \geq f(y) - L\|x^* - y\| \geq f(y^*) + p - L\delta \sqrt{\frac{d-1}{4}} > f(y^*)$, which is a contradiction. Therefore, $x^*$ is in the interior of $\text{conv}(Y^*)$ and is a local minimum of $f$. \hfill \Box

Lemma\footnote{1} roughly 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. In what follows, we will study two stochastic models of time-variation for a discretized functions. In the first uniform model, we prove the existence of a large number of local minima in expectation and bound the likelihood of the perturbed function to have only a few local minima. The complexity of the uniform model motivates the development of a second model, where we define the notion of shape dominant operator. Under this model, the function is able to reach a target function much faster. Under an additional assumption of sub-Gaussian perturbation, we bound the hitting time for reaching a neighborhood of the target function.
4 Uniform Perturbation Model

In this section, we analyze a uniform perturbation model. First, consider the case where the dimension $d = 1$ and the domain of function $\mathcal{X}$ has size $|\mathcal{X}| = n$. Without loss of generality, assume that $\mathcal{X} = [n]$, where $[n] = \{1, 2, \ldots, n\}$. The uniform perturbation model is given as follows

\begin{equation}
    f_{t+1}(x) = f_t(x) + \epsilon_t(x), \quad t = 0, 1, \ldots
\end{equation}

\begin{equation}
    f_0 \sim \mathcal{N}(0, I),
\end{equation}

where $x \in \mathcal{X} = [n]$, and $\mathcal{N}(0, 1)$ is the $n$-dimensional multi-variate normal distribution with zero mean and identity covariance. The perturbations $\epsilon_t(x)$’s are independent and identically distributed as

\begin{equation}
    \epsilon_1(x) = B1\{x = M\},
\end{equation}

where $M$ is uniform over $[n]$ and $B$ is uniform over the interval $[0, 1]$. The model \(3\) randomly selects a coordinate and perturbs it. We denote by $f_t \in \mathbb{R}^n$ the vector $(f_t(x))_{x \in \mathcal{X}}$.

The definition of desirable shapes of a function is introduced below.

**Definition 1.** The function $f : [n] \rightarrow \mathbb{R}$ is said to be

- strictly increasing if $f(x) < f(x+1)$ for all $x \in [n-1]$;
- strictly decreasing if $f(x) > f(x+1)$ for all $x \in [n-1]$;
- strictly monotone if $f(x)$ is either strictly increasing or strictly decreasing;
- strictly unimodal if there exists an $x_0 \in [n]$ such that $f(x) > f(x+1)$ for all $x < x_0$ and $f(x) < f(x+1)$ for all $x_0 \leq x < n-1$.

We drop the qualifier “strict” when the inequalities are not strict. Note that for $x, y \in [n]$ and $x \neq y$, $\mathbb{P}(f_t(x) = f_t(y)) = 0$ in model (3). Therefore, the function $f_t$ is equally likely to be monotone and strictly monotone. We further recall that for the gridding $\mathcal{X} = [n]$, a point $x \in [n]$ is said to be a strict local minimum of $f$ if the following holds:

\[ f(x) < \min\{f(x-1), f(x+1)\}, \quad \text{when } 1 < x < n \]

\[ f(x) < f(x+1), \quad \text{when } x = 1 \]

\[ f(x) < f(x-1), \quad \text{when } x = n. \]

Similarly, we drop the qualifier “strict” when the inequalities are not strict, and non-strict local minimum appears with zero probability in model (3). We use $L_f$ to denote the number of local minimum of $f$.

**Theorem 1.** Consider the uniform model \(3\) with $\mathcal{X} = [n]$ and suppose that $n \geq 6$. The following statements hold.

- $\mathbb{P}(f_t(x) \text{ is monotone}) = \frac{2}{n!}$
- $\mathbb{P}(f_t(x) \text{ is unimodal}) = \frac{2^{n-1}}{n!}$
- $\mathbb{E}L_{f_t} = \frac{1}{3}(n + 1)$
- $\text{Var}L_{f_t} = \frac{2}{45}(n + 1)$
- $\mathbb{P}(\left|L_{f_t} - \frac{1}{3}(n + 1)\right| > k) \leq \frac{2}{45} \frac{(n+1)}{k^2}.$

**Proof.** Let $S_n$ be the set of permutations of $[n]$. For any permutation $\pi \in S_n$, the random vectors $f_t = (f_t(1), \ldots, f_t(n))$ and $f_t^\pi = (f_t(\pi(1)), \ldots, f_t(\pi(n)))$ are identically distributed. Therefore,

\[ \mathbb{P}(f_t(1) < f_t(2) < \cdots < f_t(n)) = \mathbb{P}(f_t(\pi(1)) < f_t(\pi(2)) < \cdots < f_t(\pi(n))). \]

When $\pi$ is not identity, the events \{ $f_t(1) < f_t(2) < \cdots < f_t(n)$ \} and \{ $f_t(\pi(1)) < f_t(\pi(2)) < \cdots < f_t(\pi(n))$ \} are disjoint. Furthermore, because every entry of the vector $f_t$ follows a continuous distribution, the probability that two entries have the same value is zero. As a result,

\[ \mathbb{P}(f_t(1) < f_t(2) < \cdots < f_t(n)) = \frac{1}{n!} \sum_{\pi \in S_n} \mathbb{P}(f_t(\pi(1)) < f_t(\pi(2)) < \cdots < f_t(\pi(n))) = \frac{1}{n!}. \]
This implies that

\[ P(f_t \text{ is monotone}) = P(f_t(1) < f_t(2) < \cdots < f_t(n)) + P(f_t(1) > f_t(2) > \cdots > f_t(n)) = \frac{2}{n!}. \]

We now bound the probability of the function being unimodal. Let \( U \subseteq \mathbb{R}^n \) be the set of unimodal discrete functions, and let \( M \subseteq \mathbb{R}^n \) be the set of monotone functions. Consider the map \( \text{sort} : U \rightarrow M \) that is defined by sorting the vector \( x = (x_1, \ldots, x_n) \) to \( y = \text{sort}(x) \) so that \( \{y_1, \ldots, y_n\} = \{x_1, \ldots, x_n\} \) (equal as a multi-set) and \( y_1 \leq y_2 \ldots \leq y_n \). For any sorted vector \( y_1 < y_2 < \cdots < y_n \), there are \( 2^{n-1} \) ways to arrange them into \( x = (x_1, \ldots, x_n) \) in such a way that \( x \) becomes unimodal. This is because we may select any subset of \( \{y_2, \ldots, y_n\} \) and arrange them in decreasing order to form the decreasing part of \( x \), and arrange the remaining elements in increasing order to form the increasing part of \( x \). Therefore, \( \text{sort} \) defines a map where \( 2^{n-1} \) unimodal functions in \( \text{sort}^{-1}(y) \) are mapped to a monotonically increasing function. Furthermore, since \( \text{sort} \) is permutation, these \( 2^{n-1} \) unimodal functions in \( \text{sort}^{-1}(y) \) are equally likely to appear as the sorted vector \( y \). Therefore, \( \text{sort} \) is a \( 2^{n-1} \)-fold cover of \( M \) with \( U \), except for those vectors in \( U \) that have two equal coordinates, which occurs with zero probability. As a result,

\[ P(f_t \text{ is unimodal}) = 2^{n-1}P(\text{sort}(f_t) \text{ is strictly monotone}) = \frac{2^{n-1}}{n!}. \]

To calculate the expected number of local minima, we write \( L_{f_t} = \sum_{i=1}^{n} 1 \{i \text{ is a local minimum of } f_t\} \). By linearity of expectation and symmetry, we have

\[ \mathbb{E} L_{f_t} = \sum_{i=1}^{n} P(i \text{ is a local minimum of } f_t) \]

\[ = 2P(1 \text{ is a local minimum of } f_t) + (n-2)P(2 \text{ is a local minimum of } f_t) \]

\[ = 2P(f_t(1) < f_t(2)) + (n-2)P(f_t(2) < \min(f_t(1), f_t(3))) \]

\[ = 2 \cdot \frac{1}{2} + (n-2) \cdot \frac{1}{3} = \frac{1}{3}(n+1), \]
Similarly, we bound the second moment of \( L_{f_t} \) as follows:

\[
\mathbb{E}L_{f_t}^2 = \left( \sum_{i=1}^{n} 1(i \text{ is a local minimum of } f_i) \right)^2
\]

\[
= \sum_{i=1}^{n} \mathbb{P}(i \text{ is a local minimum of } f_i) + 2 \sum_{1 \leq i < j \leq n} \mathbb{P}(i, j \text{ are both local minima of } f_i)
\]

\( \equiv \) \( \mathbb{E}L_{f_t}^2 + 2 \sum_{1 \leq i < j \leq n} \mathbb{P}(i, j \text{ are both local minima of } f_i) \)

\( \equiv \) \[\frac{n+1}{3} + 2\mathbb{P}(1, n \text{ are both local minima of } f_i)
+ 4\mathbb{P}(1, 3 \text{ are both local minima of } f_i)
+ 4(n-4)\mathbb{P}(1, 4 \text{ are both local minima of } f_i)
+ 2(n-4)\mathbb{P}(2, 4 \text{ are both local minima of } f_i)
+ 2\left(\frac{(n-2)(n-3)}{2} - (n-3) - (n-4)\right)\mathbb{P}(2, 5 \text{ are both local minima of } f_i)\]

\[= \left[\frac{n+1}{3} + 2\mathbb{P}(f_1(1) < f_t(2), f_t(n) < f_t(n-1))
+ 4\mathbb{P}(f_t(1) < f_t(2), f_t(3) < \min(f_t(2), f_t(4)))
+ 4(n-4)\mathbb{P}(f_t(1) < f_t(2), f_t(4) < \min(f_t(3), f_t(5)))
+ 2(n-4)\mathbb{P}(f_t(2) < \min(f_t(1), f_t(3)), f_t(4) < \min(f_t(3), f_t(5)))
+ (n-4)(n-5)\mathbb{P}(f_t(2) < \min(f_t(1), f_t(3)), f_t(5) < \min(f_t(4), f_t(6)))\right]\]

\[\equiv \frac{n+1}{3} + 2\frac{1}{4} + 4\frac{5}{4!} + 4(n-4)\frac{1}{6}
+ 2(n-4)\frac{16}{5!} + (n-4)(n-5)\frac{1}{9}\]

\[= \frac{1}{9}n^2 + \frac{4}{15}n - \frac{7}{45},\]

where (a) uses the fact that neighboring \( i \) and \( j \) cannot both be strict local minima by definition. The equality (b) comes from symmetry. (c) follows from the fact that, among the \( 4! \) equally likely orderings of \( f_t(1), f_t(2), f_t(3), f_t(4) \), there are 5 instances that satisfy the relation \( f_t(1) < f_t(2), f_t(3) < \min(f_t(2), f_t(4)) \)

\[\text{Therefore,} \quad \text{Var}(L_{f_t}) = \mathbb{E}L_{f_t}^2 - (\mathbb{E}L_{f_t})^2 = \frac{2}{45}(n + 1)\]

The deviation bound of the last item follows from Chebyshev’s inequality.

Theorem 1 implies that under the uniform model (3), it is extremely unlikely for \( f_t(x) \) to have a desirable shape, and \( f_t(x) \) has a large number of local minima. We present similar results for a general dimension \( d \) and a uniform grid \( \mathcal{X} = [m]^d \) below.

**Theorem 2.** Consider the uniform model (3) with \( \mathcal{X} = [m]^d \). The following statements hold:

- \( \mathbb{E}L_{f_t} = \frac{1}{2d+1}(m^d + m^{d-1}) + o(m^{d-1}) \),

- There exists a constant \( c_d > 0 \) depending only on \( d \) such that

\[\text{Var}(L_{f_t}) = c_d m^d + o(m^d).\]

Hence, \( \Pr(|L_{f_t} - \mathbb{E}L_{f_t}| > k) \leq \frac{c_d m^{d-o(m^d)}}{k^2} \).

(the notation \( o(m^d) \) means \( o(m^d)/m^d \to 0 \) as \( m \to \infty \)).
Proof. For a general dimension $d$, we proceed with a similar argument as the one-dimensional case, but we only consider the highest order term of $L_{f_t}$. In other words, we only compute for those points mostly in the interior of the grid. Write $L_{f_t} = \sum_{i \in [m]^d} 1 \{i \text{ is a local minimum of } f_t\}$. Let $I = (1, 1, \ldots, 1) \in \mathbb{R}^d$ be the all-one vector and $e_1, \ldots, e_d$ be the standard basis. By linearity of expectation and symmetry, we calculate

\[
\mathbb{E} L_{f_t} = \sum_{i \in [m]^d} \mathbb{P}(i \text{ is a local minimum of } f_t)
= (m - 2)^d \mathbb{P}(2I \text{ is a local minimum of } f_t) + 2d(m - 2)^d - 1 \mathbb{P}(2I - e_1 \text{ is a local minimum of } f_t) + o(m^{d-1})
\]

\[
= (m^d - 2dm^{d-1}) \mathbb{P}(f_t(2I) \leq f_t(2I \pm e_i) \text{ for } 1 \leq i \leq d)
+ 2dm^{d-1} \mathbb{P}(f_t(2I - e_1) \leq f_t(2I - e_1 \pm e_i) \text{ for } 2 \leq i \leq d \text{ and } f_t(2I - e_1) \leq f_t(2I) + o(m^{d-1})
\]

\[
= (m^d - 2dm^{d-1}) \frac{1}{2d + 1} + 2dm^{d-1} \frac{1}{2d} + o(m^{d-1})
= (m^d + m^{d-1}) \frac{1}{2d + 1} + o(m^{d-1}).
\]

where $o(m^{d-1})$ denotes a term that goes to zero as $m \to \infty$. Moreover,

\[
\text{Var} L_{f_t} = \text{Var} \left( \sum_{i \in [m]^d} 1(i \text{ is a local minimum of } f_t) \right)
= \sum_{i,j \in [m]^d} \text{Cov} [1(i \text{ is a local minimum of } f_t), 1(j \text{ is a local minimum of } f_t)]
\]

\[
\overset{(a)}{=} \sum_{i,j \in [m]^d : B(i,1) \cap B(j,1) \neq \emptyset} \text{Cov} [1(i \text{ is a local minimum of } f_t), 1(j \text{ is a local minimum of } f_t)]
\]

\[
\overset{(b)}{=} c_d m^d + o(m^d)
\]

where (a) comes from the following argument. Note that whether $i$ is a local minimum of $f_t$ depends only on the relative values of $f_t(i)$ and $f_t(k)$ for $k \in B(i,1) \cap [m]^d$, where $B(i,1) \cap [m]^d$ is the neighborhood of $i$ in the grid. Since all orderings of $f_t$ are equally likely, the number of local minima remains the same when all coordinates of the vector $f_t$ are independently distributed. In that case, whenever $B(i,1) \cap B(j,1) = \emptyset$, the events $1(i \text{ is a local minimum of } f_t)$ and $1(j \text{ is a local minimum of } f_t)$ are independent, and hence they have a zero covariance. The equality (b) comes from the fact that there are $O(m^d)$ pairs of $i,j \in [m]^d$ whose neighborhoods intersect. The constant $c_d$ depends only on $d$ and can be given by the covariances in (a).

\[\square\]

5 Shape Dominant Model

As explained in Section 3, we focus on the discretized model

\[f_{t+1} = T_t f_t = A_t f_t + w_t, \tag{4}\]

where $f_t : X \to \mathbb{R}$, for $t = 0, 1, \ldots$, are functions defined on a finite set $X = \{x_1, \ldots, x_n\}$. Equivalently, $f_t$ is a vector in $\mathbb{R}^n$. We write $P_t(A, w)$ for the joint distribution of $A_t$ and $w_t$.

Definition 2. The distribution $P(A, w)$ is said to be $(\delta, \sigma, f^*, \phi^*)$ shape dominant if the following condition holds with probability 1:

1. the unit vector $f^*$ is the eigenvector of $A$ associated with eigenvalue 1;
2. the unit vector $\phi^*$ is the eigenvector of $A^\top$ associated with eigenvalue 1;
3. $(f^*, \phi^*) \neq 0$;
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 $\sigma^2$ in the sense that for all $u \in \mathbb{R}^n$ with $\|u\| \leq 1$, we have $\mathbb{E}[\exp(u^\top w)] \leq \exp \left( \frac{\sigma^2 u^2}{2} \right)$. 

8
To understand the conditions in Definition 2, consider the special case where $A$ is a positive stochastic matrix whose column sums are all 1. The unit vector $\phi^* = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ 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 $t = 1, 2, \ldots, k$. Then,

$$f_k = T_{k-1} \circ \cdots \circ T_0 f_0 = \frac{\langle \phi^*, f_0 + \sum_{i=0}^{k-1} w_i \rangle}{\langle \phi^*, f^* \rangle} f^* + v + w$$

where

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

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

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

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

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

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

Therefore, $A_i g \in G$, and $G$ is an invariant subspace of $A_i$ in $\mathbb{R}^n$. Since $\langle \phi^*, f^* \rangle \neq 0$, we have $f^* \notin G$. Let a basis of $G$ be given by $\{g_1, \ldots, g_{n-1}\}$. Then, $B = \{f^*, g_1, \ldots, g_{n-1}\}$ is a basis of $\mathbb{R}^n$, under which the linear operator $A_i$ takes the form

$$A_i = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & A_i' & 1
\end{bmatrix}$$

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 $G$ to $G$. Note that $\|A_i\|_2 \leq 1 - \delta$ because all other eigenvalues of $A_i$ have norm less than $1 - \delta$.

Under the basis $B$, $f_0$ has the representation

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

where $g \in G$. As a result,

$$f_k = T_{k-1} \circ \cdots \circ T_0 f_0 = A_{k-1} \cdots A_0 f_0 + \sum_{i=0}^{k-1} A_{k-1} \cdots A_{i+1} w_i$$

$$= \frac{\langle \phi^*, f_0 \rangle}{\langle \phi^*, f^* \rangle} f^* + A'_{k-1} \cdots A'_1 g + \sum_{i=0}^{k-1} A_{k-1} \cdots A_{i+1} w_i$$

The norm estimate gives rise to

$$\|A'_{k-1} \cdots A'_1 g\| \leq (1 - \delta)^k \|g\|$$

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

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

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

**Page 9**
We provide a hitting time theorem for the time-varying model under study.

For all \( u \in \mathbb{R}^n \) with \( \|u\| \leq 1 \), it holds that
\[
\begin{align*}
E[\exp(s(u, A_{k-1}' \cdots A_{i+1}'h_i))] \\
= E[\exp(s(A_{k-1}'^T \cdots A_{i+1}'^Tu, h_i))] \\
= E[\exp(s(A_{k-1}'^T \cdots A_{i+1}'u, w_i - \langle \phi^*, w_i \rangle \langle \phi^*, f^* \rangle))] \\
= E[\exp(s(A_{k-1}'^T \cdots A_{i+1}'u, w_i))] \\
\times \exp(s(-\frac{\langle A_{k-1}'^T \cdots A_{i+1}'u, f^* \rangle}{\langle \phi^*, f^* \rangle} \phi^*, w_i))]
\end{align*}
\]
\[
\leq \exp\left(\frac{\sigma_i^2 s^2 \|A_{k-1}'^T \cdots A_{i+1}'^Tu\|^2}{2}\right)
\times \exp\left(\frac{\sigma_i^2 s^2}{2} \left(\frac{\langle A_{k-1}'^T \cdots A_{i+1}'u, f^* \rangle}{\langle \phi^*, f^* \rangle}\right)^2\right)
\]
\[
\leq \exp\left(\frac{\sigma_i^2 s^2 (1 - \delta)^{2(k-i)} \left(1 + \frac{1}{(\phi^*, f^*)^2}\right)\sum_{i=0}^{k-1} (1 - \delta)^{2(k-i)} \sigma_i^2}{2}\right).
\]

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

Theorem 3 states that if the time-varying model is given by shape dominant operators, then \( f_k \) can be decomposed 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 hitting time theorem for the time-varying model under study.

**Theorem 4.** Consider the model \( \{u_i\} \). Under the same assumptions made in Theorem 3, define
\[
\tau_\epsilon = \inf \{ k : \exists \lambda \in \mathbb{R} \text{ s.t. } \|f_k - \lambda f^*\| < \epsilon \},
\]
(7)
where \( \epsilon > 0 \). Suppose that \( k > \frac{\log 2 (\|f_0\| + \|\phi^*, f^*\|^2) - \log \epsilon}{\log \frac{1}{\epsilon^2}} \).

Then,
\[
P(\tau_\epsilon \geq k) \leq C_n \exp\left(-\frac{\epsilon^2}{32 \left(1 + \frac{1}{(\phi^*, f^*)^2}\right) \sum_{i=0}^{k-1} (1 - \delta)^{2(k-i)} \sigma_i^2}\right).
\]

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 \( \|v(k)\| < (1 - \delta)^k (\|f_0\| + \|\phi^*, f^*\|) \) and
\[
w(k) = \sum_{i=0}^{k-1} A_{k-1}' \cdots A_{i+1}'h_i
\]
is sub-Gaussian with parameter $\sigma^2 = \left(1 + \frac{1}{(\sigma_1^*, f_1^*)^2}\right) \sum_{i=0}^{k-1} (1 - \delta)^{2(k-i)} \sigma_i^2$. From the definition of the hitting time in (7), we have
\[
\mathbb{P}(\tau_\epsilon < k) \geq \mathbb{P}\left( \|v^{(k)}\| < \epsilon/2, \|w^{(k)}\| < \epsilon/2 \right)
\]
when $k > \log 2 \left( \|f_0\| + \frac{(\sigma_1^*, f_0^*)}{\sigma_1^*} \right) - \log \epsilon$, then $\|v^{(k)}\| < \epsilon/2$ is satisfied. Since $w^{(k)}$ is sub-Gaussian with parameter $\sigma^2$, the tail-bound for $w^{(k)}$ yields
\[
\mathbb{P}\left( \|w^{(k)}\| < \epsilon/2 \right) = 1 - \mathbb{P}(\|w_k\| > \epsilon/2) 
\geq 1 - C_n \exp \left( - \frac{\epsilon^2}{32\sigma^2} \right).
\]
where $C_n$ is a universal constant depending only on $n$.

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

**Remark.** We note that the uniform model (1) is a special case of model (4) with $A_t = I$ and $w_t$ being independent bounded random variables. The linear operator $A_t = I$ satisfies all but the fourth condition of shape-dominant operator in Definition 2 that restricts the spectrum. This suggests that the spectral properties of the linear operator $A_t$ play a crucial role in determining the evolution and the eventual shape of the time-varying function.

**Remark.** Theorems 3 and Theorem 4 state that the shape of the discretized function $f_t|X$ would be close to the vector $f^*|X$. Taking the results back to the continuous domain, if there exists a continuous convex function $f^*$ whose discretization leads to $f^*|X$, the time-varying function $f_t$ in the original continuous time-varying optimization will become close to a convex function $f^*$. This does not imply that $f_t$ becomes convex, and it may have some sharp local minima. However, it is known that such local minima can be avoided using SGD [18].

6 Conclusion

In this paper, we modeled a time-variation of optimization problems via a linear operator that captures changes in the problem in the discrete domain. We studied two stochastic models of time-variation. The first model generates functions with a large number of local minima, whereas the second model reaches a target function significantly faster. The difference between the two models depends on the spectrum of the linear operator that models the time-variation. This provides the first step towards the understanding of random evolution of functions and their effect on optimization.

Acknowledgements

This work was supported by grants from NSF, AFOSR, ONR and ARO. The authors are grateful to Heyuan Liu, Haoyang Cao, and Salar Fattahi for many fruitful discussions.

References


