A Dynamical System Perspective for Escaping Sharp Local Minima in Equality Constrained Optimization Problems

Han Feng, Haixiang Zhang, and Javad Lavaei
Industrial Engineering and Operations Research, University of California, Berkeley

Abstract— **This paper provides a dynamical system perspective on the escape of sharp local minima in constrained optimization problems. The dynamical system view models a perturbed projected first-order optimization algorithm and translates the problem of escaping local minima in constrained optimization problems to that of escaping regions of attraction of the corresponding dynamical system. We develop the notion of biased perturbation and show that it gives a quantitative view of the notion of small regions of attraction that can be escaped. As a counterpart, we explain why the dynamics is stable in a wide region of attraction around a strongly stable equilibrium. Numerical examples are provided to illustrate the usefulness of the developed concepts.**

I. Introduction

The empirical success of optimization in the training of neural networks has motivated a large body of work on non-convex optimization algorithms, especially the stochastic gradient descent (SGD) algorithm that has been observed to not only converge faster than the gradient descent method but also converge to a better solution than the gradient descent method. Most of the literature has focused on unconstrained problems that are motivated from fitting statistical models. It is known that, almost surely with random initialization, a wide range of first-order methods never converge to saddle points [8], [9]. The effect of perturbation on the escape of saddle points has been quantified in various versions of perturbed gradient descent [4]. The effect of noise has also been studied under the model of Stochastic Gradient Langevin Dynamics [20], [13]. The recent work [17] argues that the Brownian-motion-based analysis may not be realistic for the SGD method and provides an alternative stochastic differential equation model driven by the Lévy motion with Symmetrical α-Stable ($\mathcal{S}\alpha\mathcal{S}$) distribution, which has infinite second-order momentum. Several existing works have been performed to analyze the role of the noise in the SGD method; one such explanation is provided in [6], where a change of variables introduces a new stochastic iteration based on a smoothed objective. Adaptive restart SGD method is proposed in [3] and shown to achieve both fast and robust convergence. Global exponential convergence of non-convex LQR is analyzed in [12]. Nevertheless, it lacks a satisfactory explanation of escaping *locally optimal solutions* and a quantitative view of sharp versus flat local minima. Moreover, the literature has largely ignored constrained optimization problems since the existing ideas are not readily generalizable to optimization over a feasible set.

This paper studies the escape of certain local minima for a general equality-constrained optimization problem, using tools in control theory. This approach has major advantages: (1) it translates local minima of the constrained optimization problem to equilibrium points of a continuous-time dynamical system, which can subsequently be analyzed using rich techniques in dynamical systems; (2) the continuous-time model allows deploying the notion of region of attraction to relate the sharpness of a local minimum to the size of the region of attraction. The continuous-time analysis of optimization algorithms has advanced the area of algorithm design [16], [18] and deepened our understanding of various phenomena in optimization. These include the analysis and the design of acceleration techniques via the Bregman Lagrangian [19], study of stochastic gradient via the Ornstein-Uhlenbeck process [11], and convergence analysis via integral quadratic constraints [10]. However, only a few works have studied the underlying problem in the constrained setting. The starting point of our formulation is [15], where multiple ordinary differential equation (ODE) models of gradient flow are proposed for the constrained case.

The paper is organized as follows. In Section II, we formalize a dynamical system model of a general equality-constrained optimization problem and discuss the correspondence between the local minima of the constrained optimization problem and the equilibrium points of the dynamical system. Section III introduces the noisy version of the model and develops sufficient conditions on the perturbation under which the constraints are approximately satisfied in the long run. Section IV introduces the key notion of biased perturbation and demonstrates the usefulness of this concept in the escape of local minima with a small region of attraction. For a wide region of attraction, Section V proposes the notion of strong equilibrium and proves that a small perturbation does not trigger the solution to escape the wide region of attraction around a strong equilibrium. This analysis enables the introduction of the notion of sharp minima, which can be escaped via small perturbations in the underlying algorithm. A numerical example is provided in Section VI. Concluding remarks are provided in Section VII.

Email: han_feng@berkeley.edu, haixiang_zhang@berkeley.edu, lavaei@berkeley.edu

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

II. Dynamical System Model of Constrained Optimization

The paper considers a general optimization problem with equality constraints:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.t. \quad & g_i(x) = 0, \quad i = 1, 2, \dots, m. \end{aligned} \tag{1}$$

We use $g(x) = [g_1(x), \dots, g_m(x)]^T$ to denote a vector-valued function and use

$$J_g(x) = [\nabla g_1(x), \dots, \nabla g_m(x)]^T$$

to denote the m-by-n Jacobian matrix of the equality constraints. We make a variant of the constraint qualifications assumption below.

Assumption 1. *The functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ are twice continuously differentiable for $i = 1, 2, \dots, m$. Furthermore, the m-by-n Jacobian matrix $J_g(x)$ has full row rank for all $x \in \mathbb{R}^n$.*

Remark 1. *The full rank assumption of the Jacobian matrix is part of a classical constraint qualification condition for the local minima of the constrained optimization problem to be KKT points [2, §4.3.8]. Moreover, Sard's theorem [14] ensures that for a sufficiently smooth map g, the set of values of g for which J_g is not full rank has measure 0.*

Under this assumption, every local minimum of the constrained optimization problem (1) satisfies the first-order stationary conditions:

$$\begin{aligned} \nabla f(x) + J_g(x)^T\lambda = 0, \\ g(x) = 0, \end{aligned}$$

for some vector $\lambda \in \mathbb{R}^m$. In this paper, we study the following dynamical system

$$\begin{aligned} \dot{x} = & - \left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \\ & - \alpha J_g(x)^T(J_g(x)J_g(x)^T)^{-1}g(x) \end{aligned} \tag{2}$$

where $\alpha \geq 0$ is a parameter. Under Assumption 1, the right-hand side of (2) is a continuously differentiable function of x. Therefore, it is locally Lipschitz. For any initial condition $x(0) = x_0$, the solution to (2) exists and is unique locally.

Proposition 1. *Consider the dynamical system* (2) *with $x(0) = x_0$. Assume that $x(t)$ exists for all $t \geq 0$. If x_0 is feasible for the constrained optimization problem* (1)*, then $x(t)$ is feasible for all $t \geq 0$. Furthermore, if $\alpha > 0$, then the equilibrium points of* (2) *are exactly the first-order stationary points of* (1)*. If x_0 is not feasible for the constrained optimization problem* (1)*, then the trajectory $x(t)$ will approach feasibility at an exponential rate.*

Proof. Taking the derivative of $g(\cdot)$ with respect to time along any trajectory $x(t)$ of the system (2) yields that

$$\begin{aligned} & \frac{d}{dt}g(x(t)) \\ & = -J_g(x(t))\left[I - J_g(x(t))^T(J_g(x(t))J_g(x(t))^T)^{-1}J_g(x(t))\right] \\ & \quad \times \nabla f(x(t)) - \alpha g(x(t)) \\ & = -\alpha g(x(t)), \end{aligned}$$

for all $t \geq 0$. Therefore, $g(x(t)) = e^{-\alpha t}g(x(0))$, which converges to zero as $t \to \infty$ when $\alpha > 0$ and is identical to zero if $g(x(0)) = 0$. For each equilibrium point x of the dynamical system (2), we have

$$\begin{aligned} & \left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \\ & \quad -\alpha J_g(x)^T(J_g(x)J_g(x)^T)^{-1}g(x) = 0 \end{aligned}$$

Since $\alpha \neq 0$, multiplying the above equation by $J_g(x)$ yields $g(x) = 0$. The first-order stationary condition is satisfied by setting

$$\lambda = -(J_g(x)J_g(x)^T)^{-1}J_g(x)\nabla f(x). \tag{3}$$

Conversely, if x is a first-order stationary point with the multiplier λ, then λ must be given by (3) due to Assumption 1. Hence, x is an equilibrium. □

Remark 2. *In the case $\alpha > 0$, it follows from Theorem 2.3 in [15] that, as long as* (2) *has finitely many equilibria, all bounded solutions to* (2) *exist for all $t \geq 0$. Therefore, the existence of $x(t)$ for all $t \geq 0$ is a mild assumption.*

Since the first-order stationary points for the constrained optimization problem (1) are the same as the equilibria of the dynamical system (2), the focus of the paper is on modifying this dynamics to eliminate some of its undesirable equilibria corresponding to sharp local minima.

III. Perturbation and Its Effect on Constraints

When solving an unconstrained optimization problem, the common practice is to inject noise or use momentum in the gradient algorithm to avoid undesirable local minima. As a counterpart of this technique, we add a perturbation w to the system dynamics. In its most general form, the perturbation can depend on the current state and time, which can be formally written as

$$\dot{x} = U_\alpha(x) + w(x, t), \tag{4}$$

where

$$\begin{aligned} U_\alpha(x) = & - \left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \\ & - \alpha J_g(x)^T(J_g(x)J_g(x)^T)^{-1}g(x) \end{aligned}$$

is the right-hand side of (2). In order for the perturbed dynamics to meaningfully solve the constrained optimization problem, one needs to ensure a near feasibility of the solution $x(t)$ as t goes to infinity. In what follows, we will provide a bound that gives sufficient conditions to guarantee that the solution to the perturbed dynamics approximately satisfies the constraints in the long run. We note that the solution to (4) is not always feasible for any $t \geq 0$ due to the perturbation term.

Proposition 2. *Consider the perturbed dynamics* (4) *with parameter* α. *Let* $y(t)$ *be the solution to* (4) *with an arbitrary initial condition* $y(0) = y_0$, *where* y_0 *is feasible for the optimization problem* (1). *Assume that the perturbation satisfies the bound*

$$\|J_g(y)w(y,t)\| \le \chi(t)\|g(y)\| + \xi(t) \tag{5}$$

for some $\chi(t) \ge 0$, $\xi(t) \ge 0$ *and all* $t \ge 0$. *Given a constant* $\epsilon > 0$, *if there is a number* $T(\epsilon) > 0$ *such that*

$$e^{-\alpha t}\int_0^t \xi(s)ds \cdot \exp\left(\int_0^t \chi(r)e^{-\alpha r}dr\right) \le \epsilon, \textit{ for all } t > T(\epsilon),$$

then $\|g(y(t))\| \le \epsilon$ *for all* $t > T(\epsilon)$.

Proof. One can write

$$\begin{aligned}\dot{g}(y(t)) &= -J_g(y)\left[I - J_g(y)^T(J_g(y)J_g(y)^T)^{-1}J_g(y)\right]\nabla f(y) \\ &\quad - \alpha g(y(t)) + J_g(y)w(y,t) \\ &= -\alpha g(y(t)) - J_g(y)w(y,t).\end{aligned}$$

Since $g(y(0)) = 0$, it holds that

$$\begin{aligned}\|g(y(t))\| &\le e^{-\alpha t}\int_0^t \|J_g(y(s))w(y(s),s)\|ds \\ &\le e^{-\alpha t}\int_0^t \chi(s)\|g(y(s))\|ds + e^{-\alpha t}\int_0^t \xi(s)ds.\end{aligned} \tag{6}$$

We proceed in the spirit of the proof of Grönwall's Inequality. Let $u(t) = \|g(y(t))\|$. Consider the function

$$v(s) := \exp\left(-\int_0^s \chi(r)e^{-\alpha r}dr\right)\int_0^s \chi(r)u(r)dr.$$

Taking the derivative of $v(s)$, we obtain

$$\begin{aligned}v'(s) &= \exp\left(-\int_0^s \chi(r)e^{-\alpha r}dr\right)\chi(s) \\ &\quad \times \left[u(s) - e^{-\alpha s}\int_0^s \chi(r)u(r)dr\right] \\ &\le \exp\left(-\int_0^s \chi(r)e^{-\alpha r}dr\right)\chi(s)e^{-\alpha s}\int_0^s \xi(r)dr,\end{aligned}$$

due to the bound (6) and the non-negativity of $\chi(s)$. By integrating both sides of the above inequality, we obtain

$$v(t) \le \int_0^t \exp\left(-\int_0^s \chi(r)e^{-\alpha r}dr\right)\chi(s)e^{-\alpha s}\int_0^s \xi(r)drds.$$

Therefore,

$$\begin{aligned}u(t) &\le e^{-\alpha t}\int_0^t \xi(s)ds + e^{-\alpha t}\exp\left(\int_0^t \chi(r)e^{-\alpha r}dr\right)v(t) \\ &\le e^{-\alpha t}\left[\int_0^t \xi(s)ds + \right. \\ &\quad \left.\int_0^t \exp\left(\int_s^t \chi(r)e^{-\alpha r}dr\right)\chi(s)e^{-\alpha s}\int_0^s \xi(r)drds\right] \\ &\le e^{-\alpha t}\int_0^t \xi(s)ds\left[1 - \int_0^t d(\exp\left(\int_s^t \chi(r)e^{-\alpha r}dr\right))\right] \\ &= e^{-\alpha t}\int_0^t \xi(s)ds\exp\left(\int_0^t \chi(r)e^{-\alpha r}dr\right),\end{aligned}$$

where we have used (6) in the first inequality and the non-negativity of $\xi(r)$ in the third inequality. This bound is less than ϵ for large values of t by assumption. □

Proposition 2 quantifies how the perturbation contributes to the deviation from the solutions of the optimization (1). Especially, when $\alpha > 0$ and the trajectory $y(t)$ is bounded, the condition of the proposition is satisfied for a bounded noise w by setting $\chi(t) = 0$ and $\xi(t) = C$ for a large enough constant C. In the following sections, we will study how to design the perturbation so that the trajectories of the dynamical system move away from undesirable local minima but stay close to desirable local minima.

IV. Escaping Sharp Local Minima

We refer to a local minimum of the optimization problem (1) as *sharp* if its associated region of attraction in the unperturbed dynamics (2) is relatively small. The threshold defining smallness will be quantified in Theorem 1. To simplify the presentation, we assume that the noise depends only on time and therefore study the following dynamics:

$$\dot{x} = U_\alpha(x) + w(t). \tag{7}$$

Without loss of generality, let $x = 0$ be the equilibrium under study. We will develop sufficient conditions for the solution of (7) to leave the region of attraction of the sharp local minimum. To this end, we introduce the key notion of biased perturbation. Intuitively, if $w(t)$ forces the dynamics to move along a certain direction for a certain amount of time, the dynamics will be perturbed away from 0.

Definition 1. *The function* $w(t)$ *is said to be* (δ, ϵ)*-biased over* $\mathcal{T}$ *if for all* $\tau \in \mathcal{T}$, *there exists a unit vector* $v(\tau) \in \mathbb{R}^n$ *such that* $\langle w(t), v(\tau)\rangle \ge \epsilon$ *for all* $t \in [\tau, \tau+\delta] \cap \mathcal{T}$. *Here we define the inner-product* $\langle w, v\rangle := \sum_{i=1}^n w_i v_i$.

Remark 3. *In many models of stochastic gradient descent, the sample estimator of the gradient is unbiased and the noise values at different iterations are independent of each other. However, there are many ways to introduce bias and dependence into the estimator and the theory to be developed below can be adapted to study avoiding undesirable local minima in this case.*

The following two lemmas demonstrate the generality of the definition.

Lemma 1. *If* $\|w(t)\| \ge \epsilon$ *for all* $t \in \mathcal{T}$ *and* $w(t)$ *is* L*-Lipschitz continuous, then the function* $w(t)$ *is* $(\delta, \epsilon - L\delta)$*-biased for* $\delta \in (0, \epsilon/L)$.

Proof. Select the unit vector $v(\tau) = \frac{w(\tau)}{\|w(\tau)\|}$, which is well-defined because $w(t)$ is assumed non-zero for all $t \in \mathcal{T}$. For $t \in [\tau, \tau+\delta] \cap \mathcal{T}$, we have

$$\langle w(t), v(\tau)\rangle \ge \langle w(\tau), v(\tau)\rangle - L|t - \tau| \ge \epsilon - L\delta.$$

□

Lemma 2. *Consider* $w(t) = B(t)$, *which is the* n*-dimensional Brownian motion with the initial condition*

$B(0) = 0$. *Given arbitrary numbers t_1 and t_2 such that $0 < t_1 < t_2$, the function $w(t)$ is (δ, ϵ)-biased over $\mathcal{T} = [t_1, t_2]$ with probability at least*

$$\left(1 - e^{-\epsilon^2/(2\delta)}\right) Ce^{-c\epsilon^2/t_1} \cdot (1 - 2e^{-t_1/(t_2-t_1)})^n,$$

where c and C are universal constants.

Proof. From the independent increment property of Brownian motion, the distribution of $w(t)$ is the same as $w(\tau) + B_{t-t_0}$, where B_t is an n-dimensional Brownian motion independent of $w(t)$. Select the unit vector $v(\tau) = \frac{w(\tau)}{\|w(\tau)\|}$, which is well-defined almost surely. For δ and $\epsilon > 0$, one can write

$$\begin{aligned}
&\mathbb{P}\left(\langle w(t), v(\tau)\rangle \geq \epsilon, \forall \tau \in \mathcal{T}, t \in [\tau, \tau+\delta] \cap \mathcal{T}\right) \\
&= \mathbb{P}\left(\langle w(\tau) + B_{t-\tau}, v(\tau)\rangle \geq \epsilon, \forall \tau \in \mathcal{T}, t \in [\tau, \tau+\delta] \cap \mathcal{T}\right) \\
&\overset{(a)}{\geq} \mathbb{P}\left(\langle w(\tau) + B_s, v(\tau)\rangle \geq \epsilon, \forall \tau \in \mathcal{T}, s \in [0, \delta]\right) \\
&= \mathbb{P}\left(\langle B_s, v(\tau)\rangle \geq \epsilon - \|w(\tau)\|, \forall \tau \in \mathcal{T}, s \in [0, \delta]\right) \\
&\overset{(b)}{=} \mathbb{P}\left(W_s \geq \epsilon - \|w(\tau)\|, \forall \tau \in \mathcal{T}, s \in [0, \delta]\right) \\
&\geq \mathbb{P}\left(W_s \geq \epsilon - r, \|w(\tau)\| \geq r, \forall \tau \in \mathcal{T}, s \in [0, \delta]\right) \\
&\overset{(c)}{=} \mathbb{P}\left(W_s \geq \epsilon - r, \forall s \in [0, \delta]\right)\mathbb{P}(\|w(\tau)\| \geq r, \forall \tau \in \mathcal{T}) \\
&\overset{(d)}{\geq} (1 - 2\mathbb{P}(W_\delta \leq \epsilon - r)) \\
&\qquad \mathbb{P}(\|w(t_1)\|^2 \geq r^2 + r'^2)\mathbb{P}(\|B_s\|^2 \leq r'^2, \forall s \in [0, t_2 - t_1]) \\
&\overset{(e)}{\geq} \left(1 - e^{-(\epsilon - r)^2/(2\delta)}\right) \cdot \\
&\qquad Ce^{-c(\frac{r^2 + r'^2}{t_1} - n)} \cdot \mathbb{P}(\sup_{s \in [0, t_2 - t_1]} |W_s| \leq r'/\sqrt{n})^n \\
&\geq \left(1 - e^{-(\epsilon - r)^2/(2\delta)}\right) Ce^{-c(\frac{r^2 + r'^2}{t_1} - n)} \cdot \\
&\qquad (1 - 2e^{-r'^2/(n(t_2 - t_1))})^n
\end{aligned}$$

In the above bounds, we enforce a stronger range of s in (a). We let $W_s = \langle B_s, v(\tau)\rangle$ in (b). W_s is a one-dimensional Brownian motion projected from an n-dimensional Brownian motion. The distribution of W_s does not depend on the projection vector $v(\tau)$ or $w(\tau)$. This independence factors the product in (c) and similarly in (d). (d) also uses the reflection principle of the Brownian motion. The bound (e) strengthens the bound on the norm of B_s to every coordinate. Since the squared norm $\|B_s\|_2^2$ obeys the χ_n^2 distribution, the constraints $r > \epsilon$ and $(r'^2 + r^2)/t_1 > n$ are required for lower bounding the tail probability [7]. Setting $r'^2 = nt_1$ and $r = 2\epsilon$ yields the desired probability bound. □

We next show that a biased perturbation escapes any small region of attraction, which is a well-defined notion for continuous-time dynamics. The region of attraction of the equilibrium (7) is defined as

$$\begin{aligned}
R = \{x_0 : x(t) \to 0 \text{ with } x(0) = x_0, \text{ where} \\
x(t) \text{ solves (7) with } w(t) = 0\}.
\end{aligned}$$

The notion of "smallness" of R is quantitatively described in the assumptions of Theorem 1 below. We assume that the region of attraction R is bounded in the following analysis.

Theorem 1. *Let $r = \sup_{x_0 \in R}\|x_0\|$ denote radius of the smallest ball containing the region of attraction. Define*

$$E = \sup_{\|x\| \leq r} \|U_\alpha(x)\|.$$

Assume that $w(t)$ is (δ, ϵ)-biased over $[0, \infty)$, where $\epsilon > E$ and $\delta(\epsilon - E) > 2r$. The perturbed solution $x(t)$ to (7) satisfies the property that whenever $x(\tau) \in R$ at any time τ, then there exists a time $t \in [\tau, \tau + \delta]$ such that $x(t) \notin R$.

Proof. We use the definition of (δ, ϵ)-biasness to find a unit vector $v(\tau)$ such that

$$\langle w(t), v(\tau)\rangle \geq \epsilon, \text{ for all } t \in [\tau, \tau + \delta].$$

When $x(\tau) \in R$, it holds that $\|x(\tau)\| \leq r$. We take the inner product of (7) with $v(\tau)$ to obtain

$$\begin{aligned}
\langle \dot{x}(t), v(\tau)\rangle &= \langle U(x), v(\tau)\rangle + \langle w(t), v(\tau)\rangle \\
&\geq \epsilon - \|U(x)\| \geq \epsilon - E, \text{ when } x \in R, t \in [t, \tau].
\end{aligned}$$

If $x(t) \in R$ for all $t \in [\tau, \tau + \delta]$, we arrive at the following contradiction:

$$\begin{aligned}
r &\geq \|x(\tau + \delta)\| \geq \langle x(\tau + \delta), v(\tau)\rangle \\
&\geq \langle x(\tau), v(\tau)\rangle + \int_\tau^{\tau+\delta} (e - E)dt \geq -r + (e - E)\delta.
\end{aligned}$$

□

V. Attraction to Wide Local Minima

In this section, we study wide local minima, which correspond to those equilibria of the unperturbed dynamics (2) whose regions of attraction are sufficiently large. As before, we focus on the dynamics (7) and with no loss of generality assume that the local minimum under study is $x^* = 0$. Let $\lambda_0 = -(J_g(0)J_g(0)^T)^{-1}J_g(0)\nabla f(0)$ and consider the augmented Lagrangian function

$$L_\beta(x) = f(x) - f(0) + \lambda_0^T g(x) + \frac{\beta}{2}\|g(x)\|^2.$$

We introduce the following notion of strong equilibrium.

Definition 2. *We say that $x^* = 0$ is a (γ_1, γ_2, r) strong equilibrium of the dynamics (2) if the following conditions hold for all x such that $\|x\| \leq r$:*

- $L_\beta(x) \geq \frac{\gamma_1}{2}\|x\|^2$
- *Either* $\|g(x)\|^2 \geq \gamma_2\|x\|^2$ *or*

$$\nabla f(x)^T\big[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\big]\nabla f(x) \geq \gamma_2\|x\|^2, \tag{8}$$

Remark 4. *It is known that if the equilibrium $x = 0$ satisfies the first- and the second-order sufficient conditions*

$$\begin{aligned}
&\nabla f(0) + J_g(x)^T\lambda_0 = 0, \quad g(0) = 0 \\
&y^T\nabla^2 L_0(0)y > 0 \textit{ for all } y \neq 0 \textit{ with } J_g(0)y = 0,
\end{aligned}$$

then for large $\beta > 0$, there exist $\epsilon > 0$ and $\gamma > 0$ such that $L_\beta(x) \geq L_\beta(0) + \frac{\gamma}{2}\|x\|^2$ for all x with $\|x\| \leq \epsilon$. This suggests that $L_\beta(x)$ is a plausible candidate for the Lyapunov function used in the proof of Theorem 2.

Remark 5. *The matrix $I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)$ is rank deficient; hence, one cannot expect solely that the inequality* (8) *holds for all x. In particular, it does not hold when $\nabla f(x)$ is in the range of $J_g^T(x)$ at a nonzero x. This is the reason why we include two possibilities in the second condition of strong equilibrium.*

Theorem 2. *Suppose that $x^* = 0$ is a (γ_1, γ_2, r)-strong equilibrium of the dynamics* (2)*. Then, there exist constants $\delta, k, \gamma, \theta, b, T$ such that, for all $\|x(t_0)\| < \theta r$, $\|w(t)\| \leq \delta$ and a small $\alpha > 0$, the solution $x(t)$ to* (7) *satisfies*

(1) $\|x(t)\| \leq k\exp[-\gamma(t - t_0)]\|x(t_0)\|$, *for $t_0 \leq t < t_0 + T$*
(2) $\|x(t)\| \leq b$, *for all $t \geq t_0 + T$*

The proof of the theorem makes use of the following lemma on the stability of perturbed dynamics.

Lemma 3 (Lemma 9.2 in [5])**.** *Let $V(t, x)$ be a Lyapunov function for* (2) *that satisfies*

$$c_1\|x\|^2 \leq V(t, x) \leq c_2\|x\|^2 \tag{9}$$

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}U_\alpha(x) \leq -c_3\|x\|^2 \tag{10}$$

$$\left\|\frac{\partial V}{\partial x}\right\| \leq c_4\|x\| \tag{11}$$

over $[0, \infty) \times D$, where $D = \{x \in \mathbb{R}^n : \|x\| \leq r\}$. Assume that the perturbation $w(t)$ satisfies $\|w(t)\| \leq \delta \leq \frac{c_3}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r$ for all $t \geq 0$, all $x \in D$ and some constant $\theta < 1$. Then, for all $\|x(t_0)\| < \sqrt{\frac{c_1}{c_2}}r$, the solution to (7) *satisfies*

(1) $\|x(t)\| \leq k\exp[-\gamma(t - t_0)]\|x(t_0)\|$, *for $t_0 \leq t < t_0 + T$*
(2) $\|x(t)\| \leq b$, *for all $t \geq t_0 + T$*

for some finite T and $k = \sqrt{\frac{c_2}{c_2}}$, $\gamma = \frac{(1-\theta)c_3}{2c_2}$, $b = \frac{c_4}{c_3}\sqrt{\frac{c_2}{c_1}}\frac{\delta}{\theta}$.

Proof of Theorem 2. We show that the conditions in Lemma 3 are satisfied for the Lyapunov function $V(t, x) = L_\beta(x)$ for a large enough $\beta > 0$. The existence of the constant c_1 in (9) follows from the assumption of strong equilibrium. The constant c_2 exists due to the first-order necessary condition, $L_\beta(0) = 0$ and $\nabla L_\beta(0) = 0$. Since $L_\beta(x)$ is twice continuously differentiable, the constant c_2 exists for every x in a neighborhood of the equilibirum $x^* = 0$ and can be selected large enough to apply to all points x with $\|x\| \leq r$. We compute

$$\nabla L_\beta(x) = \nabla f(x) + J_g(x)^T\lambda_0 + \beta J_g(x)^T g(x).$$

From the first-order necessary condition and twice differentiability, condition (11) can be satisfied for a large enough c_4. To verify condition (10), we compute the time derivative of L_β:

$$\begin{aligned}
&\dot{L}_\beta(x) = \nabla V(x) \cdot U_\alpha(x) \\
&= -\nabla f(x)^T\left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \\
&\quad - \alpha[\nabla f(x)^T J_g(x)^T(J_g(x)J_g(x)^T)^{-1} + \lambda_0]g(x) \\
&\quad - \alpha\beta\|g(x)\|^2 \\
&\leq -\nabla f(x)^T\left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \\
&\quad + \alpha L\|x\|^2 - \alpha\beta\|g(x)\|^2
\end{aligned}$$

for some constant L that depends on the Lipschitz constants of the functions $g(x)$ and $f(x)^T J_g(x)^T(J_g(x)J_g(x)^T)^{-1}$. The second condition of strong equilibrium leads us to consider two possibilities: when $\|g(x)\|^2 \geq \gamma_2\|x\|^2$, we have

$$\dot{L}_\beta(x) \leq -\alpha(\beta - L\gamma_2)\|x\|^2,$$

which is negative definite for a large β; when $\nabla f(x)^T\left[I - J_g(x)^T(J_g(x)J_g(x)^T)^{-1}J_g(x)\right]\nabla f(x) \geq \gamma_2\|x\|^2$, we have

$$\dot{L}_\beta(x) \leq -(\gamma_2 - \alpha L)\|x\|^2$$

which is negative definite for a small enough α. □

VI. Numerical Experiments

For the numerical experiment, we consider a variant of the Ackley Function [1], which is widely used for testing global optimization algorithms for non-convex problems:

$$\begin{aligned}
f(x) = ae^{-d} - a\exp\left(-\sqrt{\frac{1}{2}(x_1^2 + x_2^2) + d^2}\right) + \\
e - \exp\left(\frac{1}{2}(\cos(cx_1) + \cos(cx_2))\right).
\end{aligned} \tag{12}$$

Figure 1 plots the test function with selected parameter $a = 20$ and $d = 0.05$. As can be observed from the figure, there are numerous local minima, and the global minimum is achieved at $f(0, 0) = 0$. The global minimum lies in a large basin, which implies a large region of attraction for the constrained dynamics (2).

We incorporate the following equality constraint:

$$g(x) = x_1 - \frac{1}{2}x_2^2 = 0.$$

Note that we have focused on a two-dimensional optimization problem with a single constraint only for visualization purposes. The functions $f(x)$ and $g(x)$ satisfy Assumption 1, with

$$J_g(x) = [1, -x_2].$$

The function in (7) takes the form

$$U_\alpha(x) = -\begin{bmatrix} \frac{x_2^2}{1+x_2^2} & \frac{x_2}{1+x_2^2} \\ \frac{x_2}{1+x_2^2} & \frac{1}{1+x_2^2} \end{bmatrix}\nabla f(x) - \alpha\frac{x_1 - \frac{1}{2}x_2^2}{1 + x_2^2}\begin{bmatrix} 1 \\ -x_2 \end{bmatrix}$$

Figure 2 plots sample trajectories of the system without noise. The figure shows that, even without the noise, the dynamics of (2) is able to overcome the ridge of the function f in order to reach a feasible solution. Of the 298 uniformly

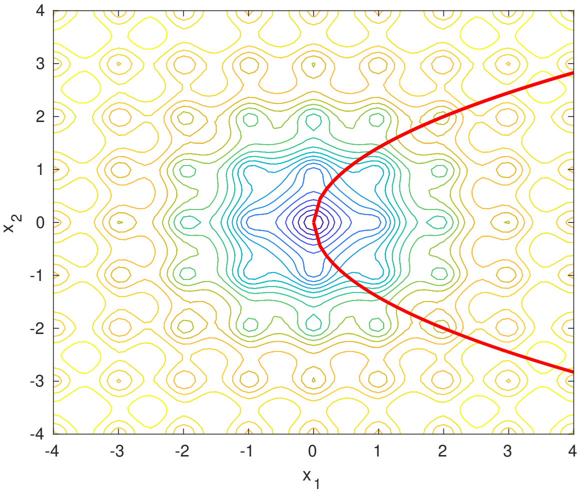

Fig. 1: Counter plot of the test function (12) overlaid with the locus of the feasible points.

random initializations in a square of side length 8, about 78% of the initializations achieve the "success", which is defined as reaching the unit ball[1] around the global minimum $x = (0, 0)$ by the time $t = 20$.

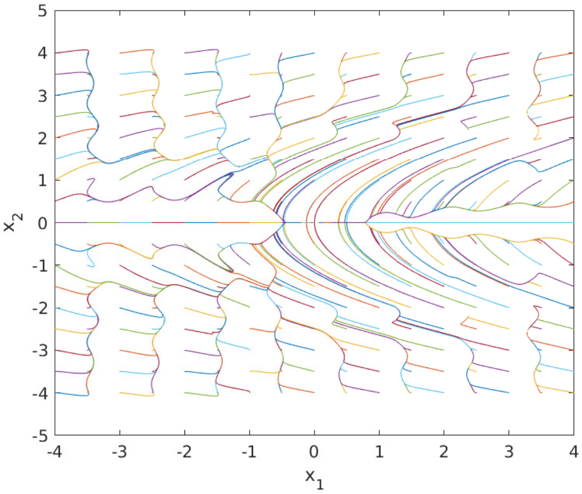

Fig. 2: Sample trajectories of the noiseless dynamical system (2). Some trajectories reach the global minimum while others do not.

Let the noise $w(t)$ be of the following form:

$$w(t) = A \begin{bmatrix} \cos(\omega t) \\ \sin(\omega t) \end{bmatrix},$$

where A and ω are parameters to be determined. The injected noise $\omega(t)$ is Lipschitz continuous and therefore biased by Proposition 1. Table I tallies the success rate of several choices of the frequency parameter ω. We note that when $A = 1$ and $\omega < 0.27$, the noise is varying so slowly that the success rate is worse than the case without perturbation, which is due to the fact that a persistent perturbation along one direction leads astray trajectories that may otherwise be able to escape to the global minimum. Table II tallies the success rate of several choices of the amplitude parameter A. Figure 3 and Figure 4 plot the sample trajectories corresponding to Table I and Table II, respectively. Since the trajectories are all bounded, by Proposition 2 and the continuity of ODE with respect to initial conditions, the constraint will eventually be satisfied with a close-to-feasible initialization, which is shown in Figure 3. In comparison, Figure 4 shows that when the initialization is far away from feasible and when the frequency ω is small, the paths are erratic due to unsuitable perturbations. In summary, the experiment shows that different choices of the parameters can significantly alter the behavior of the perturbed trajectory and their ability to explore many local minima. There is a sweet spot for the parameters (A, ω) that best improves the success rate.

TABLE I: Successfully rate by the frequency of noise

A	ω	success rate
1	0.1	71.28%
1	0.16681	76.12%
1	0.27826	93.77%
1	0.46416	94.46%
1	0.77426	86.51%
1	1.2915	80.97%
1	2.1544	84.78%
1	3.5938	88.24%
1	5.9948	83.39%
1	10	80.62%

TABLE II: Successfully rate by the amplitude of noise

A	ω	success rate
0.1	1	82.35%
0.14678	1	84.43%
0.21544	1	86.51%
0.31623	1	88.58%
0.46416	1	91%
0.68129	1	86.16%
1	1	80.28%
1.4678	1	87.89%
2.1544	1	88.58%
3.1623	1	87.2%

VII. Conclusion

This paper studies a dynamical system model of a general equality constrained optimization problem. This approach views sharp and wide local minima from the perspective of small and large regions of attraction of the dynamical system. We mathematically study the effect of injecting noise into the dynamics on small and large regions of attraction. In particular, we introduce the notion of biased perturbation and the notion of strong equilibrium, and show how these concepts can be used to quantify sharpness and wideness of a local minimum. This result enables us to understand how to escape undesirable local minima of a non-convex

[1] Note that from Figure 1, the unit ball is within the region of attraction of the equilibrium (0, 0).

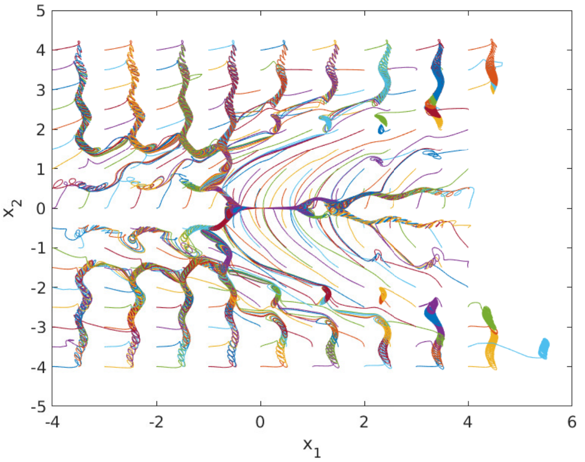

Fig. 3: Sample trajectories of the perturbed dynamics (7) corresponding to Table I. Thick ensembles of paths are formed when perturbation varies the gradient dynamics and allows the escape from local minima.

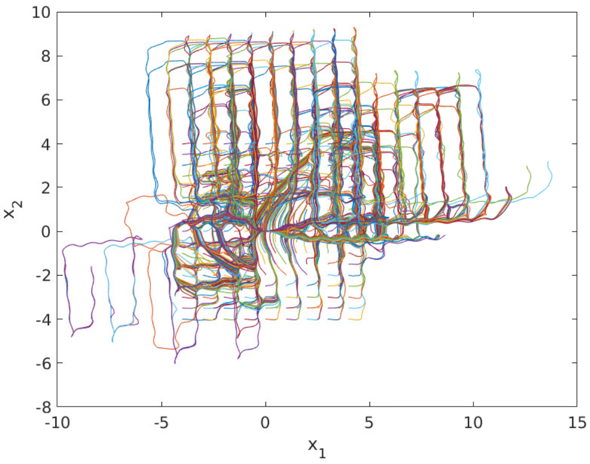

Fig. 4: Sample trajectories of the perturbed dynamics (7) corresponding to Table II. Unsuitable perturbation can cause wild wandering paths.

constrained optimization problem. The experiments illustrate the importance of adapting the noise to the optimization problem of interest. Future research includes the design of discrete-time numerical algorithms and a rigorous study of noise in the Stochastic Gradient Langevin model. It is also instructive to extend the study to accelerated methods, whose corresponding dynamics are typically time-varying.

VIII. Acknowledgment

The authors are grateful to Akhil Shetty for suggesting the test function library.

References

[1] Ernesto Padernal Adorio and Revised January. Mvf - multivariate test functions library in c for unconstrained global optimization. 2005.

[2] Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

[3] Ross Drummond and Stephen Duncan. Accelerated gradient methods with memory. *arXiv preprint arXiv:1805.09077*, 2018.

[4] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning*, pages 1724–1732, July 2017.

[5] Hassan K Khalil. Noninear systems. *Prentice-Hall, New Jersey*, 2(5):1–5, 1996.

[6] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2698–2707, 2018.

[7] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[8] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1):311–337, July 2019.

[9] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[10] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26(1):57–95, January 2016.

[11] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, January 2017.

[12] Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7474–7479. IEEE, 2019.

[13] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis. *arXiv:1702.03849 [cs, math, stat]*, June 2017.

[14] Arthur Sard. The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*, 48(12):883–890, 1942.

[15] Johannes Schropp and I. Singer. A dynamical systems approach to constrained minimization. *Numerical functional analysis and optimization*, 21(3-4):537–551, 2000.

[16] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[17] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837, 2019.

[18] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[19] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A Variational Perspective on Accelerated Methods in Optimization. *arXiv:1603.04245 [cs, math, stat]*, March 2016.

[20] Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Conference on Learning Theory*, pages 1980–2022, June 2017.