MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Finite-Time Convergence in Continuous-Time Optimization

Romero, Orlando; Benosman, Mouhacine

TR2020-100 July 14, 2020

Abstract

In this paper, we investigate a Lyapunov-like differential inequality that allows us to establish finite-time stability of a continuous-time state-space dynamical system represented via a multivariate ordinary differential equation or differential inclusion. Equipped with this condition, we synthesize first and second-order (in an optimization variable) dynamical systems that achieve finite-time convergence to the minima of a given sufficiently regular cost function. As a byproduct, we show that the q-rescaled gradient flow (q-RGF) proposed by Wibisono et al. (2016) is indeed finite-time convergent, provided the cost function is gradient dominated of order p E (1, q). This way, we effectively bridge a gap between the q-RGF and the finite-time convergent normalized gradient flow (NGF) (q = infinity) proposed by Cortes' (2006) in his seminal paper in the context of multiagent systems. We discuss strategies to discretize our proposed flows and conclude by conducting some numerical experiments to illustrate our results.

International Conference on Machine Learning (ICML)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Finite-Time Convergence in Continuous-Time Optimization

Orlando Romero [1] **Mouhacine Benosman** [2]

Abstract

In this paper, we investigate a Lyapunov-like differential inequality that allows us to establish finite-time stability of a continuous-time state-space dynamical system represented via a multivariate ordinary differential equation or differential inclusion. Equipped with this condition, we synthesize first and second-order (in an optimization variable) dynamical systems that achieve finite-time convergence to the minima of a given sufficiently regular cost function. As a byproduct, we show that the *q-rescaled gradient flow* (q-RGF) proposed by Wibisono et al. (2016) is indeed finite-time convergent, provided the cost function is gradient dominated of order $p \in (1, q)$. This way, we effectively bridge a gap between the q-RGF and the finite-time convergent *normalized gradient flow* (NGF) ($q = \infty$) proposed by Cortés (2006) in his seminal paper in the context of multi-agent systems. We discuss strategies to discretize our proposed flows and conclude by conducting some numerical experiments to illustrate our results.

1. Introduction

In recent years, there has been a surge of research papers aiming to leverage ideas from dynamical systems and control theory (both in continuous and discrete time) into optimization and machine learning. As a simple example to illustrate the connection between these fields, consider the *gradient flow* (GF)

$$\dot{x}(t) = -\nabla f(x(t)), \tag{1}$$

where $\dot{x}(t) \triangleq \frac{\mathrm{d}x(t)}{\mathrm{d}t}$, for a given convex and differentiable cost function (or functional) $f : \mathcal{X} \to \mathbb{R}$ defined over a smooth Banach space $\mathcal{X}$ (typically a Euclidean or otherwise finite-dimensional real vector space in nonlinear programming, but infinite-dimensional in optimal control, calculus of variations, and trajectory optimization). Indeed, this system has been long studied in the mathematical community due to its provable asymptotic stability (in the sense of Lyapunov), and thus its ability for solutions $x(t)$ to converge, as $t \to \infty$, to a minimum of f. This idea dates back to at least Hadamard (1908), as noted by Courant (1943), where solving the differential equation (1) as an optimization method is referred to as the "method of gradients." For a modern instances of papers dealing with the GF in abstract and infinite-dimensional (or otherwise non-Euclidean) spaces, see (Ambrosio et al., 2005), (Danieri & Savar, 2014), and (Feehan, 2016). On the other hand, the standard gradient descent (GD) algorithm

$$x_{k+1} = x_k - \eta \nabla f(x_k) \tag{2}$$

with fixed step size $\eta > 0$ (also known as *learning rate* in deep learning) is likely to be older even. Its origin, or at least its main inspiration, is often attributed to Cauchy (1847). Clearly, the GF and GD methods are connected since the GD (2) is nothing more than the forward-Euler discretization of the GF (1). Likewise, the backward-Euler discretization (also known as *implicit* discretization)

$$x_{k+1} = x_k - \eta \nabla f(x_{k+1}), \tag{3}$$

of the GF (1) can be readily rewritten as[1]

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}, \tag{4}$$

which is simply the usual proximal point algorithm (PPA).

In continuous-time optimization, an ordinary differential equation (ODE), partial differential equation (PDE), or differential inclusion is designed to be explicitly computable under assumed oracles of a cost function or some surrogate of it, in such a way to lead their solutions to converge to a minimizer or extremum of the cost function. The gradient flow (1) naturally becomes the archetype gradient-based system. To achieve this, tools from Lyapunov stability theory are often employed, mainly due to the rich body of work

[1]Department of Industrial and Systems Engineering, Rensselaer Polytechcnic Institute, Troy, NY, USA. [2]Mitsubishi Electric Research Laboratories, Cambridga, MA, USA.. Correspondence to: Orlando Romero <rodrio2@rpi.edu>, Mouhacine Benosman <benosman@merl.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]In this paper, $\|.\|$ denotes the ℓ_2-norm.

within the nonlinear systems and control theory community for this purpose. In particular, we often seek *asymptotically* Lyapunov stable gradient-based systems with an equilibrium (stationary point) at an isolated extremum of the given cost function, thus certifying local convergence. Naturally, *global* asymptotic stability leads to global convergence, though such an analysis will typically require the cost function to be strongly convex everywhere.

For early work in this direction, see the work of Botsaris (1978a;b), Zghier (1981), Snyman (1982; 1983), and Brown (1989). In particular, Brockett (1988) and, subsequently, Helmke & Moore (1994), studied relationships between linear programming, ODEs, and general matrix theory. Further, Schropp (1995) and Schropp & Singer (2000) explored several aspects linking nonlinear dynamical systems to gradient-based optimization, including nonlinear constraints. Cortés (2006) proposed two discontinuous normalized modifications of gradient flows to attain *finite-time convergence*. Later, Wang & Elia (2011) proposed a control-theoretic perspective on centralized and distributed convex optimization.

More recently, Su et al. (2014) derived a second-order ODE as the limit of Nesterov's accelerated gradient method, when the gradient step sizes vanish. This ODE is then used to study Nesterov's scheme from a new perspective, particularly in an larger effort to better understand acceleration without substantially increasing computational burden. Expanding upon the aforementioned idea, França et al. (2018) derived a first-order ODE that models the continuous-time limit of the sequence of iterates generated by the alternating direction method of multipliers (ADMM). Then, the authors employ Lyapunov theory to analyze the stability at critical points of the dynamical systems and to obtain associated convergence rates.

Later, França et al. (2019) analyzed general non-smooth and linearly constrained optimization problems by deriving equivalent (at the limit) non-smooth dynamical systems related to variants of the relaxed and accelerated ADMM. In particular, two new ADMM-like algorithms were proposed: one based on Nesterov's acceleration and the other inspired by Polyak's heavy ball method, and derive differential inclusions modeling these algorithms in the continuous-time limit. Using a non-smooth Lyapunov analysis, results on rate of convergence are obtained for these dynamical systems in the convex and strongly convex settings.

In the more traditional context of machine learning, there are multiple papers that have adopted the approach of explicitly borrowing or connecting ideas from control and dynamical systems. For unsupervised learning, Plumbley (1995) proposed Lyapunov stability theory as an approach to establish convergence of principal component algorithms. Pequito et al. (2011) and Aquilanti et al. (2019) proposed continuous-time generalized expectation-maximization (EM) algorithms, based on mean-field games, for clustering of finite mixture models. Romero et al. (2019) established convergence of the EM algorithm, and a class of generalized EM algorithms denoted δ-EM, via discrete-time Lyapunov stability theory. For supervised learning, Liu & Theodorou (2019) provided a review of deep learning from the perspective of control and dynamical systems, with a focus in optimal control. Zhu (2018) and Rahnama et al. (2019) explored connections between control theory and adversarial machine learning.

Statement of Contribution

In this work, we first provide a Lyapunov-based tool to both check and construct continuous-time dynamical systems that are finite-time stable represented via differential inclusions. We then use this condition to construct multiple dynamical systems with finite-time convergence to (strict) local minima that can be viewed as continuous-time optimization algorithms. In particular, for continuously differentiable and gradient dominated cost functions, we provide a first-order method that only assumes access to the cost function and its gradient. Finally, for twice continuously differentiable and strongly convex functions, we also provide a family of finite-time convergent second-order methods, whose convergence time can be prescribed near the desired minimum.

2. Finite-Time Convergence in Optimization via Finite-Time Stability

Consider some objective cost function $f : \mathbb{R}^n \to \mathbb{R}$ that we wish to minimize. In particular, let $x^\star \in \mathbb{R}^n$ be an arbitrary local minimum of f that is unknown to us. In continuous-time optimization, we typically proceed by designing a nonlinear state-space dynamical system

$$\dot{x} = F(t, x) \tag{5}$$

for which $F(t, x)$ can be computed without explicit knowledge[2] of $x^\star$ and for which (5) is certifiably *asymptotically* Lyapunov stable at $x^\star$. For instance, we often seek systems that use only up to second-order information on the cost function, thus we design F through an oracle $\mathcal{O}_f(x) = \{f(x), \nabla f(x), \nabla^2 f(x)\}$.

In this work, however, we seek dynamical systems for which (5) is certifiably *finite-time* Lyapunov stable at $x^\star$. As will be clear later, such systems need to be possibly discontinuous or non-Lipschitz, which can more naturally be expressed and analyzed in the lense of differential inclusions

[2] In other words, we typically design some $G(\cdot)$ that can be explicitly computed for any input, and we set $F(t, x) \triangleq G(t, \mathcal{O}_f(x))$, where $\mathcal{O}_f(\cdot)$ denotes some oracle function such that $\mathcal{O}_f(x)$ encompasses all available data regarding f near x.

instead of ODEs. To achieve this objective, our approach is largely based on exploiting the Lyapunov-like differential inequality

$$\dot{\mathcal{E}}(t) \leq -c\,\mathcal{E}(t)^{\alpha}, \quad \text{a.e. } t \geq 0, \tag{6}$$

with constants $c > 0$ and $\alpha < 1$, for absolutely continuous (AC) functions $\mathcal{E}$ such that $\mathcal{E}(0) > 0$. Indeed, under the aforementioned conditions, exact convergence $\mathcal{E}(t) \to 0$ will be reached in finite time $t \to t^\star \leq \frac{\mathcal{E}(0)^{1-\alpha}}{c(1-\alpha)} < \infty$.

We will restrict ourselves to the design of time-invariant systems $F(t,x) = F(x)$. We now summarize the problem statement:

Problem 1. *Given a sufficiently smooth cost function $f : \mathbb{R}^n \to \mathbb{R}$ with a sufficiently regular local minimizer $x^\star$, and an initial approximation $x_0 \in \mathbb{R}^n$ sufficiently near $x^\star$, solve the following tasks:*

1. *Design a sufficiently smooth[3] candidate Lyapunov function V which is defined and positive definite near and with respect to (w.r.t.) $x^\star$[4].*
2. *Design a function G that can be explicitly computed for any input and such that, for the (possibly discontinuous) system[5] (5) with $F \triangleq G \circ \mathcal{O}_f$, the differential inequality (6) holds for $\mathcal{E} \triangleq V \circ x(\cdot)$, with $x(\cdot)$ a Filippov solution[6] with $x(0) = x_0$.*

By following this strategy, we will therefore achieve (local and strong) *finite-time stability*, and thus finite-time convergence. Furthermore, if $V(x_0)$ can be upper bounded, then F can be readily tuned to achieve finite-time convergence under a *prescribed* range for the settling time, or even with *exact* prescribed settling time if $V(x_0)$ can be explicitly computed and (6) holds with equality.

One variant of Problem 1 that we do not explore in this paper is to remove the possible dependence of G on x_0 by replacing finite-time stability with *fixed-time* stability. In many (perhaps most) practical situations, we have full control on the initial approximation, so we find it reasonable to design optimization algorithms around it.

3. First-Order Convergent Flows

Given some continuously differentiable cost function $f : \mathbb{R}^n \to \mathbb{R}$ with a local minimizer and isolated stationary point $x^\star \in \mathbb{R}^n$, Cortés (2006) proposed the (discontinuous) normalized gradient flows

$$\dot{x} = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \tag{7}$$

and

$$\dot{x} = -\operatorname{sign}(\nabla f(x)), \tag{8}$$

where $\operatorname{sign}(\cdot)$ denotes the sign function (applied elementwise for real-valued vectors). He then established finite-time stability based on the candidate Lyapunov function $V(x) = f(x) - f^\star$, with $f^\star = f(x^\star)$, and two differential inequality assumptions: a first-order one akin to (6) for $\alpha = 0$, and another which essentially boils down to the corresponding energy function $\mathcal{E}(\cdot)$ being non-increasing and strongly convex. However, the first-order conditions proved insufficient to establish the finite-time convergence of his proposed flows, whereas the second-order condition is sufficient, but also requires twice continuously differentiability and that the Hessian of the cost function is positive definite near the local minimum of interest.

More precisely, Cortés (2006) showed that, if f is twice continuously differentiable and strongly convex in an open neighborhood $\mathcal{D} \subseteq \mathbb{R}^n$ of $x^\star$, then the solutions to his proposed flows (7) and (8) converge in finite time to $x^\star$, provided they start in some positively invariant compact subset $S \subset \mathcal{D}$. He further showed that the convergence times are upper bounded by

$$t^\star \leq \frac{\|\nabla f(x_0)\|}{\min\limits_{x \in S} \lambda_{\min}[\nabla^2 f(x)]} \tag{9}$$

for (7), and

$$t^\star \leq \frac{\|\nabla f(x_0)\|_1}{\min\limits_{x \in S} \lambda_{\min}[\nabla^2 f(x)]} \tag{10}$$

for (8).

We will now see how our approach can be used to generalize (7) and (8) while still maintaining finite-time convergence, but without requiring second differentiability or strong convexity, instead focusing on the notion of *gradient dominance*.

Borrowing terminology from Wilson et al. (2019), we say that a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-gradient dominated of order $p \in (1, \infty]$* (with $\mu > 0$) in some neighborhood $\mathcal{D}$ of a local minimizer $x^\star \in \mathbb{R}^n$ if

$$\frac{p-1}{p}\|\nabla f(x)\|^{\frac{p}{p-1}} \geq \mu^{\frac{1}{p-1}}(f(x) - f^\star) \tag{11}$$

for every $x \in \mathcal{D}$, where $f^\star = f(x^\star)$. When $\mu > 0$ is unknown or unimportant, but known to exist, we will omit it in the previous definition.

[3] At least locally Lipschitz continuous and *regular* (see supplementary material (SM)).

[4] In other words, there exists some open neighborhood $\mathcal{D}$ of $x^\star$ such that V is defined in $\mathcal{D}$ and satisfies $V(x) \geq 0$ with equality if and only if $x = x^\star$, for every $x \in \mathcal{D} \setminus \{x^\star\}$.

[5] Right-hand side defined at least a.e., Lebesgue measurable. Furthermore, we may require it to be locally essentially bounded to ensure existence of solutions.

[6] See SM.

For strongly convex functions, gradient dominance of order $p = 2$ can be established. In fact, gradient dominance is usually defined exclusively for order $p = 2$, often referred to as the Polyak-Łojasiewicz (PL) inequality, which was introduced by Polyak (1963) to relax the (strong) convexity assumption commonly used to show convergence of the GD algorithm (2). The PL inequality can also be used to relax convexity assumptions of similar gradient and proximal-gradient methods (Karimi et al., 2016). Our adopted generalized notion of gradient dominance is strongly tied to the Łojasiewicz gradient inequality from real analytic geometry, established by Łojasiewicz (1963; 1965)[7] independently and simultaneously from (Polyak, 1963), and generalizing the PL inequality. More precisely, this inequality is typically written as

$$\|\nabla f(x)\| \geq C \cdot |f(x) - f^\star|^\theta \tag{12}$$

for every $x \in \mathbb{R}^n$ in a small enough open neighborhood of the stationary point $x = x^\star$, for some $C > 0$ and $\theta \in \left(\frac{1}{2}, 1\right]$. This inequality is guaranteed for analytic[8] functions (Łojasiewicz, 1965). More precisely, when $x^\star$ is a local minimizer of f, the aforementioned relationship is explicitly given by

$$C = \left(\frac{p}{p-1}\right)^{\frac{p-1}{p}} \mu^{\frac{1}{p}}, \qquad \theta = \frac{p-1}{p}. \tag{13}$$

Therefore, analytic functions are always gradient dominated. However, while analytic functions are always smooth, smoothness is not required to attain gradient dominance. Continuous differentiability is still required, however.

Let us now consider the candidate Lyapunov function $V(x) \triangleq f(x) - f^\star$ from which we will construct first-order finite-time convergent flows. Notice that (6) with $\mathcal{E}(t) \triangleq V(x(t))$ then becomes

$$\nabla f(x(t)) \cdot \dot{x}(t) \leq -c(f(x(t)) - f^\star)^\alpha. \tag{14}$$

Naturally, an immediate candidate for $\dot{x}(t)$ is

$$\dot{x} = -c\frac{(f(x) - f^\star)^\alpha}{\|\nabla f(x)\|^2}\nabla f(x), \tag{15}$$

but, unfortunately, it requires knowledge of $f^\star$, which is usually not available. See SM for further discussion. More generally, to satisfy (14) without knowledge of $f^\star$, we could design

$$\dot{x} = -c\|\nabla f(x)\|^{\beta-2}\nabla f(x) \tag{16}$$

with β chosen such that $\|\nabla f(x)\|^\beta \geq (f(x) - f^\star)^\alpha$. Notice that the RHS of (16) is continuous (but non-Lipschitz, unless $\beta \geq 2$), provided that f is continuously differentiable near $x^\star$ and $\beta > 1$. From the gradient dominance, it follows that

$$\|\nabla f\|^{\frac{p}{p-1}} \geq \left(\frac{p}{p-1}\right)\mu^{\frac{1}{p-1}}(f - f^\star), \tag{17}$$

and thus

$$\|\nabla f(x)\|^{\alpha\left(\frac{p}{p-1}\right)} \geq \left(\frac{p}{p-1}\right)^\alpha \mu^{\frac{\alpha}{p-1}}(f(x) - f^\star)^\alpha \tag{18}$$

for every x near $x^\star$. Since $\nabla f(x) \to 0$ as $x \to x^\star$, then it clearly suffices to choose $\beta \leq \alpha\left(\frac{p}{p-1}\right)$. On the other hand, the particular choice of c and $\alpha < 1$ are unimportant to attain finite-time convergence, and thus we may choose any $\beta < \frac{p}{p-1}$. In other words, it suffices that $\beta = \frac{q}{q-1}$ with $q \in (p, \infty]$, which results in

$$\dot{x} = -c\frac{\nabla f(x)}{\|\nabla f(x)\|^{\frac{q-2}{q-1}}}, \tag{19}$$

known as the *q-rescaled gradient flow* (q-RGF), originally proposed by Wibisono et al. (2016). We can use a similar reasoning to generalize (8) into (q-SGF)

$$\dot{x} = -c\,\|\nabla f(x)\|_1^{\frac{1}{q-1}}\,\mathrm{sign}(\nabla f(x)), \tag{20}$$

which naturally coincides with (19) for $n = 1$. Due to inequalities between different ℓ_r-norms, we could also replace the norms in either of the proposed flows by $\|\cdot\|_r$, subject to a suitable range for $r \geq 1$. Other generalizations could include replacing the norms altogether by some function $x \mapsto \alpha(\|s\|)$ with α a class $\mathcal{K}$ function (Khalil, 2001).

Theorem 1. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and μ-gradient dominated of order $p \in (1, \infty)$ near a strict local minimizer $x^\star \in \mathbb{R}^n$. Let $c > 0$ and $q \in (p, \infty]$. Then, any maximal solution (in the sense of Filippov) to the q-RGF given by (19) or the q-SGF flow (20) will converge in finite time to $x^\star$, provided that $\|x(0) - x^\star\| > 0$ is sufficiently small. Furthermore, their convergence times are both upper bounded by*

$$t^\star \leq \frac{\|\nabla f(x_0)\|^{\frac{1}{\theta} - \frac{1}{\theta'}}}{cC^{\frac{1}{\theta}}\left(1 - \frac{\theta}{\theta'}\right)}, \tag{21}$$

where $x_0 = x(0)$ and $f^\star = f(x^\star)$, with C, θ given by (13) and $\theta' = \frac{q-1}{q}$. In particular, given any compact and positively invariant subset $S \subset \mathcal{D}$, both flows converge in finite with the aforementioned convergence time upper bound (which can be tightened by replacing $\overline{\mathcal{D}}$ with S) for any $x_0 \in S$. Furthermore, if $\mathcal{D} = \mathbb{R}^n$, then we have global finite-time convergence, i.e. *finite-time convergence to any maximal solution (in the sense of Filippov) $x(\cdot)$ with arbitrary $x_0 \in \mathbb{R}^n$.*

[7] For more modern treatments in English, see (Łojasiewicz & Zurro, 1999; Bolte et al., 2007)

[8] *Analytic* functions are functions that are locally given by convergent power series.

Proof. The main idea is to show that the differential inequality (6) is satisfied for the energy function $\mathcal{E}(t) \triangleq V(x(t))$, defined in terms of the candidate Lyapunov function $V(x) \triangleq f(x) - f^\star$. Refer to SM for full details. ■

Wibisono et al. (2016) showed, for convex cost functions, that solutions $x(\cdot)$ to the q-RGF (19) (with $c = 1$) satisfy the convergence rate $f(x(t)) - f^\star = \mathcal{O}\left(\frac{1}{t^{q-1}}\right)$. However, as concluded in Theorem 1, the q-RGF actually converges in finite time, provided that f is gradient dominated of order $p \in (1, q)$. Therefore, the q-RGF will be finite-time convergent for strongly convex functions, provided we choose $q > 2$. For analytic functions, the q-RGF is also finite-time convergent, provided that $q > 1$ is chosen large enough. More precisely, any $q > p = \frac{1}{1-\theta}$ achieves this. Unfortunately, however, bounding the exponent θ for non-strongly convex and non-polynomial functions appears to be mathematically intractable in general (D'Acunto & Kurdyka, 2005; Pham, 2012). Notice that, in principle, the convergence time can be prescribed, as $t^\star \leq T$ with used-selected $T > 0$ by appropriate choice of $c > 0$ and $q > p$ that make the RHS of (21) equal to T. However, the set of candidate hyperparameters c, q will naturally depend on μ, p. These observations are equally valid for the flow (8).

We believe a reciprocal result to Theorem 1 is likely to be true. More precisely, that, if f is not gradient dominated of order p for some $p \in (1, q)$, then the convergence of the q-RGF will only be asymptotic. To illustrate this notion, let us explore the objective function considered in Appendix F of (Wibisono et al., 2016), given by

$$f(x) = \frac{1}{p}\|x\|^p, \tag{22}$$

with $p \in (1, \infty)$, which is μ-gradient dominated of order p near $x^\star = 0$, with $\mu = (p-1)^{p-1}$. Furthermore, f is not gradient dominated of order p' for any $p' < p$. Therefore, in order to apply our theory and thus ensure finite-time convergence, we must choose $q > p$.

Notice that the p-RGF reduces to

$$\dot{x} = -c\, x, \tag{23}$$

and thus $x(t) = e^{-c\,t}x_0$. In other words, the solutions to the p-RGF converge asymptotically to the minimum $x^\star = 0$, and not in finite time.

On the other hand, the q-RGF for a general $q > 1$ becomes

$$\dot{x} = -c\|x\|^{-\varepsilon}x, \tag{24}$$

with $\varepsilon = \frac{q-p}{q-1}$. It appears that this ODE cannot be analytically solved in the multivariate case, so for simplicity we assume $x(t) \in \mathbb{R}$. The solutions thus become

$$x(t) = \operatorname{sign}(x_0)\max\{0, (|x_0|^\varepsilon - \varepsilon\, c\, t)^{\frac{1}{\varepsilon}}\}. \tag{25}$$

Clearly, if $q > p$, then $\varepsilon > 0$, and $x(t) \to t^\star$ as $t \to t^\star$, with $t^\star = \frac{1}{c\,\varepsilon}|x_0|^\varepsilon < \infty$. On the other hand, if $1 < q < p$, then $\varepsilon < 0$, and thus $x(t) \to x^\star$ only as $t \to \infty$. The case $q = p$, corresponds to $\varepsilon \to 0$, which leads to $x(t) = e^{-t}x_0$ as previously discussed.

In terms of the settling time upper bound (21) (assuming now $q > p$ and multivariate $x(t) \in \mathbb{R}^n$), in this case it turns out to hold with equality. Indeed, it simplifies to $t^\star \leq \frac{1}{c\,\varepsilon}\|x_0\|^\varepsilon$ with $\varepsilon = \frac{q-p}{q-1}$, which generalizes the exact settling time derived analytically in the scalar case. The settling time upper bound is tight in this case precisely because the inequality originating from the gradient dominance actually holds with equality, and thus the Lyapunov differential inequality (6) will hold with equality as well.

4. Second-Order Convergent Flows

Let us now investigate the candidate Lyapunov function $V(x) \triangleq \|\nabla f(x)\|^2$, with $x^\star$ now assumed a local minimizer and isolated stationary point. Setting $\mathcal{E}(t) \triangleq V(x(t))$, then, provided that f is twice continuously differentiable, (6) becomes

$$2\nabla f(x)^\top \nabla^2 f(x)\dot{x} \leq -c\|\nabla f(x)\|^{2\alpha}. \tag{26}$$

Clearly, there are many possible flows that can be constructed to satisfy the previous condition, so let us focus on a family constructed for strongly convex f. Given a symmetric and positive definite matrix $P \in \mathbb{R}^{n\times n}$ with SVD decomposition $P = V\Sigma V^\top$, $\Sigma = \operatorname{diag}(\lambda_1 \ldots, \lambda_n)$, $\lambda_1, \ldots, \lambda_n > 0$, we define $P^r \triangleq V\Sigma^r V^\top$, where $\Sigma^r \triangleq \operatorname{diag}(\lambda_1^r, \ldots, \lambda_n^r)$. Equipped with this definition, we propose the family

$$\dot{x} = -c\|\nabla f(x)\|^{2\alpha}\frac{[\nabla^2 f(x)]^r \nabla f(x)}{\nabla f(x)^\top [\nabla^2 f(x)]^{r+1}\nabla f(x)} \tag{27}$$

with $c > 0$, $\alpha < 1$, and $r \in \mathbb{R}$ tunable hyperparameters, as (27) leads directly to (26). In a similar fashion, we propose the family

$$\dot{x} = -\frac{c\|\nabla f(x)\|_1^{2\alpha-1}[\nabla^2 f(x)]^r \operatorname{sign}(\nabla f(x))}{\operatorname{sign}(\nabla f(x))^\top [\nabla^2 f(x)]^{r+1}\operatorname{sign}(\nabla f(x))} \tag{28}$$

with the same hyperparameters, constructed via the candidate Lyapunov function $V(x) \triangleq \|\nabla f(x)\|_1$.

We are now ready to state the finite-time convergence of these proposed flows.

Theorem 2. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and strongly convex in an open neighborhood $\mathcal{D} \subseteq \mathbb{R}^n$ of a stationary point $x^\star \in \mathbb{R}^n$. Let $c > 0$, $\alpha < 1$, and $r \in \mathbb{R}$. Then, any maximal Filippov solution to the discontinuous second-order generalized Newton-like flows (27) and (28) with sufficiently small $\|x_0 - x^\star\| > 0$*

(where $x_0 = x(0)$) will converge in finite time to $x^\star$. Furthermore, their convergence times are given exactly *by*

$$t^\star = \frac{\|\nabla f(x_0)\|^{2(1-\alpha)}}{2c(1-\alpha)}, \quad t^\star = \frac{\|\nabla f(x_0)\|_1^{2\alpha}}{2c\alpha}, \tag{29}$$

respectively. In particular, given any compact and positively invariant subset $S \subset \mathcal{D}$, the flows converge in finite for any $x_0 \in S$. Furthermore, if $\mathcal{D} = \mathbb{R}^n$, then we have global finite-time convergence.

Proof. Refer to SM for a detailed proof. ■

One point that we want to emphasise here is the fact that with these second-order flows, one can much more readily (compared to our first-order flows) prescribe the finite convergence time by appropriate choice of c, α. This is a clear advantage comparatively to the first-order methods, for which the obtained finite-time convergence upper bound (21) is less practical. In particular, we may prescribe $t^\star = T$ with arbitrary $T > 0$ by choosing, for instance, $\alpha = \frac{1}{2}$ and $c = \|\nabla f(x_0)\|/T$. In particular, for instance, we propose the *rescaled Newton flow* (RNF)

$$\dot{x} = -\frac{\|\nabla f(x_0)\|}{T} \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{\|\nabla f(x)\|} \tag{30}$$

by further choosing $r = -1$ in (27). Therefore, for (30), we obtain the prescribed finite-time convergence $x(t) \to x^\star$ as $t \to T$, where $T > 0$ fully tunable.

As a simple example, let us reconsider the function (22), this time only with $p = 2$. Indeed, the flow (30) reduces to $\dot{x} = -\frac{\|x_0\|}{T} \frac{x}{\|x\|}$. In particular, for $n = 1$, its solution is given by $x(t) = \max\left\{0, 1 - \frac{t}{T}\right\} x_0$, which clearly satisfies $x(t) \to x^\star = 0$ as $t \to T$.

5. Numerical Experiments

In this section, we illustrate the finite-time convergence properties of the q-RGF (19) and our designed second-order flow (27) on academic optimization test functions. Then, we investigate preliminary discretization strategies for the flows discussed in this paper.

5.1. First-Order Flow

Consider once again the cost function (22), in the scalar case $x \in \mathbb{R}$ and with $p = 3$. We will illustrate the performance of the RGF (19) with $c = 2$.

First, we fix $x_0 = 3/4$ and vary $q > 1$. The results are reported in Figure 1. As we can see, choosing any $q > p$ ensures finite-time convergence, but as $q \to p^+$, the convergence becomes purely asymptotic. Furthermore, we can

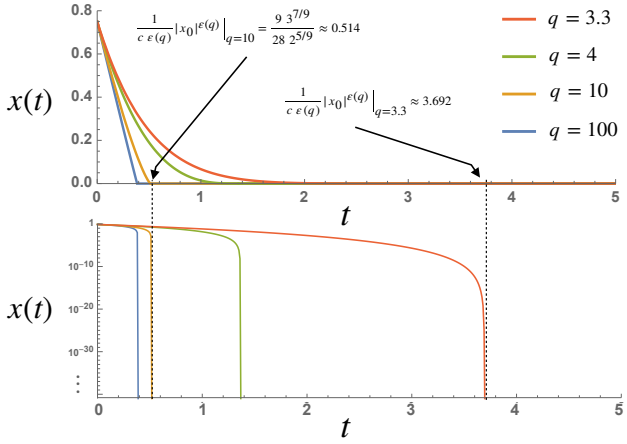

Figure 1. Solutions to the q-RGF (24) with $x_0 = 3/4$, $c = 2$, and various values of $q > 1$, on the cost function (22) (scalar case) with $p = 3$. Note: $\varepsilon(q) = (q-p)/(q-1)$.

see that the settling time upper bound (21) from Theorem 1, which simplifies to $t^\star \leq \frac{1}{c\,\varepsilon} \|x_0\|^\varepsilon$ with $\varepsilon = \frac{q-p}{q-1}$, is tight.

Next, we fix $q = 10$ and vary $x_0 \in \mathbb{R}$ near $x^\star = 0$, while maintaining every other parameter the same as before. The results are reported in Figure 2. As we can see, we have finite-time convergence near $x^\star = 0$ and the settling time upper bound (21) is tight. In reality, for this simple example, the finite-time convergence property holds globally, meaning that any $x_0 \in \mathbb{R}^n$ actually leads to $x(t) \to x^\star = 0$ as $t \to t^\star = \frac{1}{c\,\varepsilon} \|x_0\|^\varepsilon$.

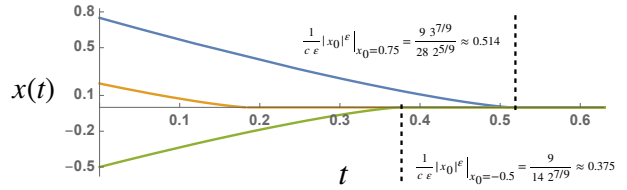

Figure 2. Solutions to the q-RGF (24) with $c = 2$, $q = 10$, and various values of $x_0 \in \mathbb{R}$ near $x^\star = 0$, on the cost function (22) (scalar case) with $p = 3$. Note: $\varepsilon = (q-p)/(q-1) = 7/9$.

5.2. Second-Order Flow

We now test the second-order flow (27) with $(c, \alpha, r) = (\|\nabla f(x_0)\|, 1/2, -1)$ on the optimization testbed function known as the Rosenbrock function, namely $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x_1, x_2) = (a - x_1)^2 + b(x_2 - x_1^2)^2, \tag{31}$$

with parameters $a, b \in \mathbb{R}$. This function admits exactly one stationary point $(x_1^\star, x_2^\star) = (a, a^2)$ for $b \geq 0$, which is a

strict global minimum for $b > 0$. If $b < 0$, then $(x_1^\star, x_2^\star)$ is a saddle point. Finally, if $b = 0$, then $\{(a, x_2) : x_2 \in \mathbb{R}\}$ are the stationary points of f, and they are all non-strict global minima.

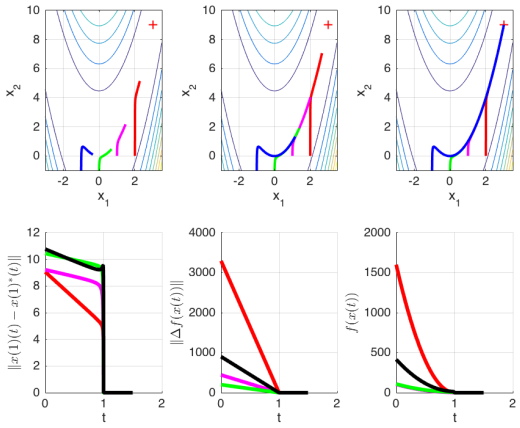

Figure 3. Trajectories of the proposed flow (27) with $(c, \alpha, r) = (\|\nabla f(x_0)\|, 1/2, -1)$ and four different initial conditions $x_0 \in \mathbb{R}^2$, for the Rosenbrock function with parameters $(a, b) = (3, 100)$, which has a unique minimum $x^\star = (a, a^2) = (3, 9)$.

As we can see in Figure 3, this flow converges correctly to the minimum $(a, a^2) = (3, 9)$ for all the tested initial conditions with an exact prescribed settling time $T = 1$. Note that, at any given point in the trajectory $x(\cdot)$, the functions $t \mapsto \|x(t) - x^\star\|$ and $t \mapsto |f(x(t)) - f(x^\star)|$ are not guaranteed to not increase, indeed only $t \mapsto \|\nabla f(x(t))\|$ can be guaranteed to do so, which explains the increase in (d) that could never have occurred in (e).

5.3. Preliminary Numerical Investigation of Potential Discretization Schemes

It is unclear if finite-time convergence in continuous time translates into some useful property when discretized, or if it should merely serve as a warning to better understand the limitations of a continuous-time representation and analysis of optimization algorithms that are ultimately intended to be implemented on a digital computer (and thus any continuous-time representation requires discretization to be implementable). Nevertheless, in this subsection we provide a preliminary investigation of potential discretization approaches that seek to combine the finite-time convergent flows studied in this paper together with well-established acceleration ideas for iterative (discrete-time) algorithms.

Recall that the Nesterov's accelerated GD is given by

$$y_k = x_k + \beta_k(x_k - x_{k-1}) \tag{32a}$$
$$x_{k+1} = y_k - \eta \nabla f(y_k) \tag{32b}$$

with $0 \leq \beta_k < 1$, often $\beta_k = \frac{k-1}{k+2}$ or $\beta_k = \beta \in [0, 1)$.

We argue that Nesterov's acceleration can be interpreted as actually being a modified Forward-Euler discretization of the GF (1), that thus achieves acceleration. More generally, given some *optimization flow* represented by the continuous-time system $\dot{x} = F(x)$, locally convergent to a local minimizer $x^\star \in \mathbb{R}^n$ of a cost function $f : \mathbb{R}^n \to \mathbb{R}$, then we can replicate Nesterov's acceleration of (1). More precisely, we obtain the algorithm

$$y_k = x_k + \beta_k(x_k - x_{k-1}) \tag{33a}$$
$$x_{k+1} = y_k + \eta F(y_k) \tag{33b}$$

where naturally choosing $F = -\nabla f$ results in Nesterov's accelerated GD.

We can now choose amongst the different flows discussed in this paper, in order to test this simple discretization idea. To achieve further acceleration, we can also make the step size $\eta > 0$ adaptive, *i.e.* we replace $\eta \leftarrow \eta_k$ in (33). The parameter $\mu > 0$ can also (and often is, for $F = -\nabla f$) be made adaptive. In particular, we will adopt an *accelerated* backtracking approach (and thus an inexact line search) borrowed from (Almeida et al., 1997). More precisely, we choose the (tunable) hyperparameters $0 < d < 1 < u$ and set $\eta_k = u\eta_{k-1}$ if

$$\begin{aligned} &f(y_k + u\eta_{k-1}F(y_k)) \\ &\qquad \leq \min\{f(x_k), f(y_k + \eta_{k-1}F(y_k))\}, \end{aligned} \tag{34}$$

and $\eta_k = d^{r_k}\eta_{k-1}$ otherwise, where r_k is defined by the smallest $r \in \{0, 1, \ldots\}$ such that

$$f(y_k + d^r \eta_{k-1} F(y_k)) \leq f(x_k) \tag{35}$$

In other words, we increase the step size by a factor $u > 0$ (*i.e.*, $\eta \leftarrow u \cdot \eta$) whenever helpful, but otherwise reduce it by a factor $d > 0$ (*i.e.*, $\eta \leftarrow d \cdot \eta$) repeatedly until the objective function actually decreases, or at least until it does not increase.

We now test this discretization idea with on the log-sum-exp function given by

$$f(x) = \rho \log\left[\sum_{i=1}^{n} \exp\left(\frac{a_i^\top x - b_i}{\rho}\right)\right] \tag{36}$$

with $n = 20$, $m = 50$, $\rho = 5$. Each entry of $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ is independently sampled from a $\mathcal{N}(0, 1)$ distribution.

The results are presented in Figures 4 and 5 and illustrate the potential of our proposed discretization approach, particularly when combined with the finite-time convergent flows studied in this paper. All of the hyperparameters were manually tuned to achieve near-optimal performance. The

figures represent the average result of 50 random initializations $x_0 \sim \mathcal{N}(0, 1)$ (component-wise, independently), with a fixed sample for $a_1, \ldots, a_m, b_1, \ldots, b_m$ as previously described.

We notice that, regarding the first-order flows, namely the q-RGF (19), we can slightly improve convergence of its discretization, compared to discretizations of the standard gradient flow (1), by tuning q, particularly so if we re-tune the other parameters in the discretization. However, we also noted in our experiments that backtracking does not appear to combine well with the first-order methods considered in this paper, including for the gradient flow case. However, once we move to the second-order flows, namely for the RNF (30), we see that (accelerated) backtracking synergizes remarkably well with the Nestorov-like forward-Euler discretization scheme proposed earlier. It is also interesting to note that a simple forward-Euler discretization with fixed step sizes of the RNF (30) originally appears slower than the standard Newton method, but indeed eventually surpasses it, for the example considered.

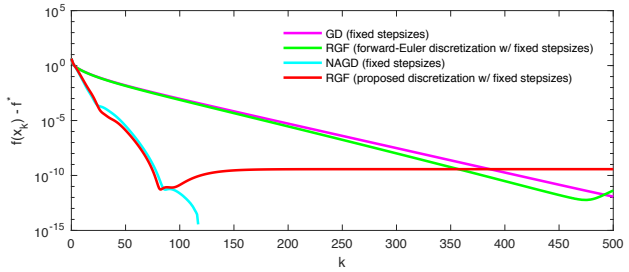

Figure 4. Numerical experiments for the first-order discrete algorithms

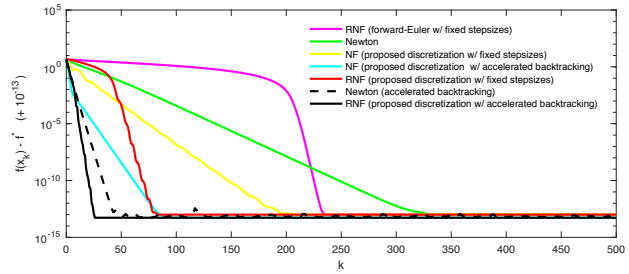

Figure 5. Numerical experiments for the second-order discrete algorithms

6. Conclusion

We have introduced a new family of discontinuous first-order and second-order flows for continuous-time optimization. The main characteristic of the proposed flows is their finite-time convergence guarantees, with, in some cases, an arbitrary *pre-defined* convergence time. To analyze these discontinuous flows, we first extended an exiting Lyapunov-based inequality condition for finite-time stability in the case of smooth dynamics to the case of non-smooth dynamics modeled by differential inclusions. We then derived and established finite-time stability (and thus convergence) for the proposed family of continuous-time optimization algorithms. We also proposed a preliminary discretization method of the proposed flows. Finally, we conducted numerical experiments on known optimization benchmarks.

Several questions remain open, which we will target in our future work. First, while we have used commonly available numerical solvers in part of our (small-scale) numerical experiments, and have proposed a first step towards a discretization method, more work will be done in this constructive discretization research direction. Furthermore, we also seek to adapt our methods to allow for linear and nonlinear constraints, and to develop distributed and decentralized variants. Lastly, many real-life problems that require a time-varying optimization framework, such as in motion planning or formation control in robotics, do not allow direct access to gradients, Hessian matrices, or time-derivatives of the gradient. Instead, these are typically estimated based on measurements (e.g. of the cost function) that often occur in discrete time and carry noisy perturbations. Therefore, future work will also be dedicated to the robustification of our proposed flows, including zeroth-order (gradient-free) schemes.

References

Almeida, L. B., Langlois, T., Amaral, J. D., and Redol, R. A. On-line step size adaptation. Technical report, INESC, 1997.

Ambrosio, L., Gigli, N., and Savare, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, January 2005.

Aquilanti, L., Cacace, S., Camilli, F., and De Maio, R. A mean field games approach to cluster analysis. July 2019.

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *Society for Industrial and Applied Mathematics*, 17:1205–1223, January 2007.

Botsaris, C. A class of methods for unconstrained minimization based on stable numerical integration techniques. *Journal of Mathematical Analysis and Applications*, 63 (3):729–749, 1978a.

Botsaris, C. Differential gradient methods. *Journal of Mathematical Analysis and Applications*, 63(1):177–198, 1978b.

Brockett, R. Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems. pp. 799–803, 1988.

Brown, A. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *Journal of Optimization Theory and Applications*, 62(2):211–224, August 1989.

Cauchy, A. Methode generale pour la resolution des systèmes dequations simultane. . *C. R. Acad. Sci. Paris*, 25:536–538, 1847.

Cortés, J. Finite-time convergent gradient flows with applications to network consensus. *Automatica*, 42(11): 1993–2000, November 2006.

Courant, R. Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49(1):1–23, January 1943.

D'Acunto, D. and Kurdyka, K. Explicit bounds for the Łojasiewicz exponent in the gradient inequality for polynomials. *Annales Polonici Mathematici*, 87:51–61, March 2005.

Danieri, S. and Savar, G. *Lecture notes on gradient flows and optimal transport*, pp. 100–144. London Mathematical Society Lecture Note Series. Cambridge University Press, 2014.

Feehan, P. M. N. Global existence and convergence of solutions to gradient systems and applications to yang-mills gradient flow. *arXiv preprint 1409.1525v4*, 2016.

França, G., Robinson, D., and Vidal, R. ADMM and accelerated ADMM as continuous dynamical systems. July 2018.

França, G., Robinson, D., and Vidal, R. A dynamical systems perspective on nonsmooth constrained optimization. *arXiv preprint 1808.04048*, 2019.

Hadamard, J. Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrée. *Mémoires présentés par divers savants à l'Académie des Sciences de l'Institut National de France, Series 2*, 33(4):1–128, 1908.

Helmke, U. and Moore, J. B. *Optimization and Dynamical Systems*. Springer-Verlag, 1994.

Karimi, H., Nutini, J., , and Schmidt, M. Linear convergence of gradient and proximal- gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Khalil, H. K. *Nonlinear Systems*. Prentice-Hall, Englewood Cliffs, New Jersey, 2001.

Liu, G.-H. and Theodorou, E. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint 1908.10920*, August 2019.

Łojasiewicz, S. A topological property of real analytic subsets (in French). *Les équations aux dérivées partielles*, pp. 87–89, 1963.

Łojasiewicz, S. *Ensembles semi-analytiques*. Centre de Physique Theorique de l'Ecole Polytechnique, 1965. URL https://perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf.

Łojasiewicz, S. and Zurro, M.-A. On the gradient inequality. *Bulletin of the Polish Academy of Sciences, Mathematics*, 47, January 1999.

Pequito, S. D., Aguiar, A. P., Sinopoli, B., and Gomes, D. A. Unsupervised learning of finite mixture models using mean field games. *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 321–328, 2011.

Pham, T. S. An explicit bound for the Łojasiewicz exponent of real polynomials. *Kodai Mathematical Journal*, 35(2): 311–319, 06 2012.

Plumbley, M. D. Lyapunov functions for convergence of principal component algorithms. *Neural Networks*, 8(1): 11–23, 1995.

Polyak, B. Gradient methods for the minimisation of functionals (in Russian). *USSR Computational Mathematics and Mathematical Physics*, 3:864–878, December 1963.

Rahnama, A., Nguyen, A. T., and Raff, E. Connecting lyapunov control theory to adversarial attacks. *arXiv preprint 1907.07732*, 2019.

Romero, O., Chaterjee, S., and Pequito, S. Convergence of the expectation-maximization algorithm through discrete-time lyapunov stability theory. *Proceedings of the American Control Conference (ACC)*, pp. 163–168, July 2019.

Schropp, J. Using dynamical systems methods to solve minimization problems. *Applied Numerical Mathematics*, 18(1):321–335, 1995.

Schropp, J. and Singer, I. A dynamical systems approach to constrained minimization. *Numerical Functional Analysis and Optimization*, 21:537–551, May 2000.

Snyman, J. A new and dynamic method for unconstrained minimization. *Applied Mathematical Modelling*, 6(6): 448–462, December 1982.

Snyman, J. An improved version of the original leap-frog dynamic method for unconstrained minimization: LFOP1(b). *Applied Mathematical Modelling*, 7(3):216–218, June 1983.

Su, W., Boyd, S., and Candes, E. J. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518. Curran Associates, Inc., 2014.

Wang, J. and Elia, N. A control perspective for centralized and distributed convex optimization. In *IEEE Conference on Decision and Control and European Control Conference*, pp. 3800–3805, December 2011.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

Wilson, A., Mackey, L., and Wibisono, A. Accelerating rescaled gradient descent: Fast optimization of smooth functions. In *Advances in Neural Information Processing Systems*. December 2019.

Zghier, A. K. *The use of differential equations in optimization*. PhD thesis, Loughborough University, 1981.

Zhu, X. An optimal control view of adversarial machine learning. *arXiv preprint 1811.04422*, 2018.