An Optimal Control Theory for Accelerated Optimization

I. M. Ross[1]

Naval Postgraduate School, Monterey, CA 93943

Abstract

Accelerated optimization algorithms can be generated using a double-integrator model for the search dynamics imbedded in an optimal control problem.

Keywords: accelerated Newton's method, control Lyapunov function, Lie derivative, Nesterov's accelerated gradient method, Polyak's heavy ball method, Riemannian metric

1. Introduction

In [1], we proposed an optimal control theory for solving a constrained optimization problem,

$$(N) \begin{cases} \underset{\boldsymbol{x}_f \in C \subseteq \mathbb{R}^{N_x}}{\text{Minimize}} \ E(\boldsymbol{x}_f) \end{cases} \tag{1}$$

where C is a constraint set in $\mathbb{R}^{N_x}$, $N_x \in \mathbb{N}^+$ and $E : \boldsymbol{x}_f \ni \mathbb{R}^{N_x} \to \mathbb{R}$ is an objective function. A key concept in this framework was to view an algorithmic map

$$\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}, \ldots$$

in terms of a discretization of a controllable, continuous-time trajectory, $t \mapsto \boldsymbol{x} \in \mathbb{R}^{N_x}$, whose dynamics is given by the single integrator model,

$$\dot{\boldsymbol{x}} = \boldsymbol{u} \tag{2}$$

[1] Distinguished Professor & Program Director, Control and Optimization

where $t \mapsto \boldsymbol{u} \in \mathbb{R}^{N_x}$ is a control trajectory that must be designed such that at some time $t = t_f$, $\boldsymbol{x}(t_f) = \boldsymbol{x}_f$ is a solution to the given optimization problem. Starting with this simple idea, it is possible to generate a wide variety of well-known algorithms such as Newton's method and the steepest descent method. Because continuous versions of "momentum" optimization methods involve second derivaties[2, 3], we explore the ramifications of replacing (2) by a double-integrator model,

$$\ddot{\boldsymbol{x}} = \boldsymbol{u} \tag{3}$$

In essence, we show that the application of the theory presented in [1] with (2) replaced by (3) generates accelerated optimization techniques.

Remark 1. Rewriting (3) in state-space form,

$$\dot{\boldsymbol{x}} = \boldsymbol{v}, \quad \dot{\boldsymbol{v}} = \boldsymbol{u} \tag{4}$$

it follows from (2) that a momentum method is essentially adding "inertia" to the "inertia-less" control of the single-integrator model.

Remark 2. A conjugate gradient (CG) method may also be viewed as an accelerated optimization technique in the context of (3). This observation follows by considering a generic CG method,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{v}_k \tag{5a}$$

$$\boldsymbol{v}_k = -\boldsymbol{g}_k + \beta_k^{CG} \boldsymbol{v}_{k-1} \tag{5b}$$

where $\alpha_k \geq 0$ is the step length, $\boldsymbol{v}_k$ is the search direction, $\boldsymbol{g}_k := \boldsymbol{g}(\boldsymbol{x}_k)$ is the gradient of the objective function function, and $\beta_k^{CG} \geq 0$ is the CG update parameter. Rewriting (5) as single equation,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \quad \beta_k := \left(\frac{\alpha_k \beta_k^{CG}}{\alpha_{k-1}}\right) \tag{6}$$

it follows that (6) may be viewed as a discretization of

$$\ddot{\boldsymbol{x}} = \boldsymbol{u}, \quad \boldsymbol{u} = -\gamma_a \boldsymbol{g}(\boldsymbol{x}) - \gamma_b \boldsymbol{v}, \quad \gamma_a \in \mathbb{R}^+, \gamma_b \in \mathbb{R}^+ \tag{7}$$

The non-control-theoretic, ordinary-differential-equation (ODE) form of (7),

$$\ddot{\boldsymbol{x}} + \gamma_a \boldsymbol{g}(\boldsymbol{x}) + \gamma_b \dot{\boldsymbol{x}} = \boldsymbol{0} \tag{8}$$

is Polyak's equation[2]. In the theory proposed in this paper, the function $(\boldsymbol{x}, \boldsymbol{v}) \mapsto -\gamma_a\, \boldsymbol{g}(\boldsymbol{x}) - \gamma_b\, \boldsymbol{v}$ turns out to be a specific "optimal" feedback controller $\boldsymbol{u}$ for the double integrator $\ddot{\boldsymbol{x}} = \boldsymbol{u}$.

Remark 3. Despite their mathematical equivalence, there is a sharp change in perspective between (5) and (7). Formula (5) suggests that the search variable is velocity. According to (7), the search variable is acceleration.

It will be apparent shortly that the objective of the proposed theory is not to take existing algorithms and interpret them as ODEs or control systems, rather, it is to use optimal control theory as a foundational concept for optimization and as a discovery tool for algorithms[1].

2. Background: Optimal Control Theory for Optimization

Consider some optimal control problem (M) whose cost functional is given by a "Mayer" cost function $E : \boldsymbol{x}_f \mapsto \mathbb{R}$, where, $\boldsymbol{x}_f = \boldsymbol{x}(t_f)$ is constrained to lie in a target set C. A transversality condition for Problem (M) is given by,

$$\boldsymbol{\lambda}_x(t_f) \in \nu_0\, \partial E(\boldsymbol{x}_f) + N_C(\boldsymbol{x}_f) \tag{9}$$

where, $\boldsymbol{\lambda}_x(t_f)$ is the final value of an adjoint arc $t \mapsto \boldsymbol{\lambda}_x$ associated with $t \mapsto \boldsymbol{x}$, $\nu_0 \geq 0$ is a cost multiplier and $N_C(\boldsymbol{x}_f)$ is the limiting normal cone[4] to the set C at $\boldsymbol{x}_f$. If Problem (M) is designed so that $\boldsymbol{\lambda}_x(t_f)$ vanishes, then the transversality condition (9) reduces to the necessary condition for Problem (N),

$$\boldsymbol{0} \in \nu_0\, \partial E(\boldsymbol{x}_f) + N_C(\boldsymbol{x}_f) \tag{10}$$

In [1], we showed the existence of Problem (M) by direct construction for the case when C is given by functional constraints,

$$C = \left\{ \boldsymbol{x} \in \mathbb{R}^{N_x} :\ \boldsymbol{e}^L \leq \boldsymbol{e}(\boldsymbol{x}) \leq \boldsymbol{e}^U \right\} \tag{11}$$

where, $\boldsymbol{e} : \boldsymbol{x} \mapsto \mathbb{R}^{N_e}$ is a given function, and $\boldsymbol{e}^L$ and $\boldsymbol{e}^U$ are the specified lower and upper bounds on the values of $\boldsymbol{e}$. In this paper, we briefly review and revise the results obtained in [1] in the context (4). Furthermore, for the purposes of brevity and clarity, we limit the discussions to the unconstrained "static" optimization problem given by,

$$(S)\left\{\underset{\boldsymbol{x}_f\in\mathbb{R}^{N_x}}{\text{Minimize}}\;\; E(\boldsymbol{x}_f)\right. \tag{12}$$

In following [1], we create a vector field by "sweeping" the function E backwards in time according to,

$$y(t) := E(\boldsymbol{x}(t)) \tag{13}$$

Differentiating (13) with respect to time we get,

$$\dot{y} = \left[\partial_{\boldsymbol{x}} E(\boldsymbol{x})\right]^T \dot{\boldsymbol{x}} = \left[\partial_{\boldsymbol{x}} E(\boldsymbol{x})\right]^T \boldsymbol{v}, \quad \dot{\boldsymbol{v}} := \boldsymbol{u} \tag{14}$$

Collecting all relevant equations, we define the following candidate optimal control problem (R) that purportedly solves the optimization problem (S):

$$(R)\left\{\begin{array}{lrl} \textsf{Minimize} & J[y(\cdot),\boldsymbol{x}(\cdot),\boldsymbol{v}(\cdot),\boldsymbol{u}(\cdot),t_f] := & y_f \\ \textsf{Subject to} & \dot{\boldsymbol{x}} = & \boldsymbol{v} \\ & \dot{\boldsymbol{v}} = & \boldsymbol{u} \\ & \dot{y} = & \left[\partial_{\boldsymbol{x}} E(\boldsymbol{x})\right]^T \boldsymbol{v} \\ & (\boldsymbol{x}(t_0), t_0) = & (\boldsymbol{x}^0, t^0) \\ & y(t_0) = & E(\boldsymbol{x}^0) \\ & \boldsymbol{v}(t_f) = & \boldsymbol{0} \end{array}\right. \tag{15}$$

where, $\boldsymbol{x}^0$ is an initial "guess" of the solution (to Problem (S)). The variables $t_f, \boldsymbol{x}(t_f)$ and $\boldsymbol{v}(t_0)$ are free.

Remark 4. The main difference between (15) and the optimal control problem for unconstrained optimization considered in [1] is the acceleration equation $\dot{\boldsymbol{v}} = \boldsymbol{u}$ and its associated endpoint condition $\boldsymbol{v}(t_f) = \boldsymbol{0}$.

Lemma 1. *Problem (R) has no abnormal extremals.*

Proof. The Pontryagin Hamiltonian[5] for this problem is given by,

$$H(\boldsymbol{\lambda}_x, \boldsymbol{\lambda}_v, \lambda_y, \boldsymbol{x}, \boldsymbol{v}, y, \boldsymbol{u}) := \boldsymbol{\lambda}_x^T \boldsymbol{v} + \boldsymbol{\lambda}_v^T \boldsymbol{u} + \lambda_y \left[\partial_{\boldsymbol{x}} E(\boldsymbol{x})\right]^T \boldsymbol{v} \tag{16}$$

where, $\boldsymbol{\lambda}_x, \boldsymbol{\lambda}_v$ and λ_y are costates that satisfy the adjoint equations,

$$\dot{\boldsymbol{\lambda}}_x = -\partial_{\boldsymbol{x}} H = -\lambda_y \, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x}) \, \boldsymbol{v} \tag{17a}$$

$$\dot{\boldsymbol{\lambda}}_v = -\partial_{\boldsymbol{v}} H = -\boldsymbol{\lambda}_x - \lambda_y \, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \tag{17b}$$

$$\dot{\lambda}_y = -\partial_y H = 0 \tag{17c}$$

The transversality conditions[5] for Problem (R) are given by,

$$\boldsymbol{\lambda}_x(t_f) = \boldsymbol{0} \tag{18a}$$

$$\boldsymbol{\lambda}_v(t_0) = \boldsymbol{0} \tag{18b}$$

$$\boldsymbol{\lambda}_y(t_f) = \nu_0 \geq 0 \tag{18c}$$

where, ν_0 is the cost multiplier. From (17c) and (18c) we have,

$$\lambda_y(t) = \nu_0 \tag{19}$$

If $\nu_0 = 0$, then $\lambda_y(t) \equiv 0$. This implies, from (17a) and (18a), that $\boldsymbol{\lambda}_x(t) \equiv \boldsymbol{0}$. Similarly, $\boldsymbol{\lambda}_v(t) \equiv \boldsymbol{0}$ from (17b) and (18b). The vanishing of all multipliers violates the nontriviality condition. Hence $\nu_0 > 0$. □

Theorem 1. *All extremals of Problem (R) are singular. Furthermore, the singular arc is of infinite order.*

Proof. The Hamiltonian is linear in the control variable and the control space is unbounded; hence, if $\boldsymbol{u}$ is optimal, it must be singular. Furthermore, from the Hamiltonian minimization condition we have the first-order condition,

$$\partial_{\boldsymbol{u}} H = \boldsymbol{\lambda}_v(t) = \boldsymbol{0} \qquad \forall t \in [t_0, t_f] \tag{20}$$

Differentiating (20) with respect to time, we get,

$$\frac{d}{dt}\partial_{\boldsymbol{u}} H = \dot{\boldsymbol{\lambda}}_v(t) = -\boldsymbol{\lambda}_x - \nu_0 \, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) = \boldsymbol{0} \tag{21}$$

Equation (21) does not generate an expression for the control function; hence, taking the second time derivative of $\partial_{\boldsymbol{u}} H$ we get,

$$\begin{aligned} \frac{d^2}{dt^2}\partial_{\boldsymbol{u}} H &= -\dot{\boldsymbol{\lambda}}_x - \nu_0\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \dot{\boldsymbol{x}} \\ &= -\dot{\boldsymbol{\lambda}}_x - \nu_0\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ &\equiv \mathbf{0} \end{aligned} \tag{22}$$

where, the last equality follows from (17a) and Lemma 1. Hence, we have,

$$\frac{d^k}{dt^k}\partial_{\boldsymbol{u}} H = \mathbf{0} \quad \text{for } k = 0, 1 \ldots$$

and no k yields an expression for $\boldsymbol{u}$. □

Theorem 2 (A Transversality Mapping Theorem)**.** *The necessary condition for Problem* (S) *is part of the transversality condition for Problem* (R)*.*

Proof. From (21), we have

$$\boldsymbol{\lambda}_x(t) = -\nu_0\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}(t)) \tag{23}$$

From (18a) and Lemma 1, it follows that $\partial_{\boldsymbol{x}_f} E(\boldsymbol{x}_f) = \mathbf{0}$. □

3. Minimum Principles for Accelerated Optimization

From the results of the previous section, it follows that the primal-dual control dynamical system generated by Problem (R) is given by,

$$\dot{\boldsymbol{x}} = \boldsymbol{v} \qquad \dot{\boldsymbol{\lambda}}_x = -\lambda_y\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{24a}$$

$$\dot{\boldsymbol{v}} = \boldsymbol{u} \qquad \dot{\boldsymbol{\lambda}}_v = -\boldsymbol{\lambda}_x - \lambda_y\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \tag{24b}$$

$$\dot{y} = \left[\partial_{\boldsymbol{x}} E(\boldsymbol{x})\right]^T \boldsymbol{v} \qquad \dot{\lambda}_y = 0 \tag{24c}$$

The boundary conditions for (24) are given by,

$$\boldsymbol{x}(t^0) = \boldsymbol{x}^0 \qquad \boldsymbol{v}(t_f) = \mathbf{0} \tag{25a}$$

$$y(t^0) = E(\boldsymbol{x}^0) \qquad \boldsymbol{\lambda}_x(t_f) = \mathbf{0} \tag{25b}$$

$$\boldsymbol{\lambda}_v(t^0) = \mathbf{0} \qquad \lambda_y(t_f) = \nu_0 > 0 \tag{25c}$$

Because the optimal control is singular of infinite order, any control trajectory that satisfies (24) and (25) is optimal. Along a singular arc, $\boldsymbol{\lambda}_v(t) \equiv \mathbf{0}$; hence, the auxiliary controllable dynamical system of interest[1] resulting from (24) is given by,

$$(A) \begin{cases} \dot{\boldsymbol{\lambda}}_x = -\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ \dot{\boldsymbol{v}} = \boldsymbol{u} \end{cases} \tag{26}$$

where, we have scaled the adjoint covector $\boldsymbol{\lambda}_x$ by $\nu_0 > 0$ (cf. Lemma 1). The target final-time condition for (A) is given by,

$$(T) \begin{cases} \boldsymbol{\lambda}_x(t_f) = \mathbf{0} \\ \boldsymbol{v}(t_f) = \mathbf{0} \end{cases} \tag{27}$$

That is, any singular control that satisfies (26) and (27) generates a candidate "optimal" continuous-time algorithm for Problem (S).

Remark 5. The auxiliary controllable dynamical system is equivalent to the time-derivative of the swept-back gradient function.

3.1. Application of a Minimum Principle Presented in [1]

Let $\boldsymbol{\beta}$ be a control vector field defined according to,

$$\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) := \begin{bmatrix} -\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ \boldsymbol{u} \end{bmatrix} \tag{28}$$

Let $V : (\boldsymbol{\lambda}_x, \boldsymbol{v}) \mapsto \mathbb{R}$ be a control Lyapunov function for the (A, T) pair. Let $\pounds_\beta V$ be the Lie derivative of V along the vector field $\boldsymbol{\beta}$. Then, a sufficient condition for producing a globally convergent algorithm[1] is to design a singular control function such that V is dissipative (when $\boldsymbol{x} \neq \boldsymbol{x}_f$),

$$\pounds_\beta V = \left[\partial V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]^T \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) < 0 \tag{29}$$

In [1], it is proposed that this objective can be achieved via the Minimum Principle,

$$(P) \begin{cases} \underset{\boldsymbol{u}}{\mathsf{Minimize}} & \pounds_\beta V := \left[\partial V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]^T \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) \\ \mathsf{Subject\ to} & \boldsymbol{u} \in \mathbb{U}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) \end{cases} \tag{30}$$

where, $\mathbb{U}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)$ is an appropriate compact set that may vary with respect to the tuple $(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)$. In an "unaccelerated" method, a solution to Problem (P) ensures the satisfaction of (29) when $\mathbb{U}$ is chosen to metricize the control space[1]. Because of the presence of a drift vector field in the auxiliary dynamical system (A), the Minimum Principle (P) cannot guarantee $\pounds_\beta V < 0$; this follows by simply inspecting the expression for $\pounds_\beta V$,

$$\pounds_\beta V = -\big[\partial_{\boldsymbol{\lambda}_x} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\big]^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} + \big[\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\big]^T \boldsymbol{u} \tag{31}$$

To ensure $\pounds_\beta V < 0$, we impose the following requirement on V,

$$\partial_{\boldsymbol{\lambda}_x} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\big]^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} > 0 \quad \text{if} \quad \partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}) = \boldsymbol{0} \text{ and } (\boldsymbol{\lambda}_x, \boldsymbol{v}) \neq \boldsymbol{0} \tag{32}$$

Furthermore, we set $\boldsymbol{u} = \boldsymbol{0}$ if $\partial_{\boldsymbol{v}} V = \boldsymbol{0}$. All of these results — in their general form — are well-known in nonlinear feedback control theory[6]; hence, they are, technically, not new. What is new is their application to static optimization.

3.2. An Alternative Minimum Principle

We can formulate an alternative Minimum Principle that essentially exchanges the cost and constraint functions in (30). Let $\rho : (\boldsymbol{\lambda}_x, \boldsymbol{v}, \boldsymbol{x}) \mapsto \mathbb{R}_+$ be an appropriate design function such that $-\rho$ specifies a rate of descent for $\dot{V}$. We propose to select a singular control $\boldsymbol{u}$ such that,

$$\pounds_\beta V = \big[\partial V(\boldsymbol{\lambda}_x, \boldsymbol{v})\big]^T \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) \leq -\rho(\boldsymbol{\lambda}_x, \boldsymbol{v}, \boldsymbol{x}) \tag{33}$$

That is, in contrast to (29), we seek a singular control that merely achieves a specified rate of descent given in terms of ρ. Let $D : (\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) \mapsto \mathbb{R}$ be an appropriate objective function. Then, a singular control $\boldsymbol{u}$ that solves the optimization problem,

$$(P^*) \begin{cases} \underset{\boldsymbol{u}}{\textsf{Minimize}} & D(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) \\ \textsf{Subject to} & \pounds_\beta V + \rho(\boldsymbol{\lambda}_x, \boldsymbol{v}, \boldsymbol{x}) \leq 0 \end{cases} \tag{34}$$

is a candidate (continuous-time) solution to the accelerated optimization problem.

Remark 6. Problem (P^*) has been widely used in control theory for generating feedback controls[6, 7]. Note also that condition on V specified by (32) is implicit in (34).

Remark 7. LaSalle's invariance principle[6] may be used to relax the positive definite condition on V and the negative definite condition on $\pounds_\beta V$.

4. Accelerated Optimization via Minimum Principles

Following [1], we consider

$$\mathbb{U}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) := \left\{\boldsymbol{u} : \ \boldsymbol{u}^T \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)\boldsymbol{u} \le \Delta(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)\right\} \tag{35}$$

where, $\Delta(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) \neq 0$ is a positive real number and $\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)$ is some appropriate positive definite matrix that metricizes the space $\mathbb{U}$. The quantities Δ and $\boldsymbol{W}$ may depend upon some or all of the variables $\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}$ and t. Applying the Minimum Principle given by (30), it is straightforward to show that if $\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}) \neq \boldsymbol{0}$, then $\boldsymbol{u}$ is given explicitly by,

$$\boldsymbol{u} = -\sigma[@t]\, \boldsymbol{W}^{-1}[@t]\, \partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}), \qquad \sigma[@t] > 0 \tag{36}$$

where,

$$\sigma^2[@t] := \frac{\Delta(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)}{\left[\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]^T \boldsymbol{W}^{-1}[@t]\left[\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]} \tag{37}$$

and $\boldsymbol{W}[@t] := \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)$. That is, the $[@t]$ notation is simply a convenient shorthand for the various implicit and explicit time dependencies[5].

Switching to an application of Minimum Principle (P^*) and using

$$D(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t) = \frac{1}{2}\left(\boldsymbol{u}^T \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)\boldsymbol{u}\right)$$

we get,

$$\boldsymbol{u} = -\sigma^*[@t]\, \boldsymbol{W}^{-1}[@t]\, \partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}), \qquad \sigma^*[@t] > 0 \tag{38}$$

where,

$$\sigma^*[@t] := \frac{\rho(\boldsymbol{\lambda}_x, \boldsymbol{v}, \boldsymbol{x}) - \left[\partial_{\boldsymbol{\lambda}_x} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v}}{\left[\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]^T \boldsymbol{W}^{-1}[@t]\left[\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\right]} \tag{39}$$

In other words, both minimum principles (P and P^*) generate the same functional form for $\boldsymbol{u}$ but with different interpretations for the control "multipliers" given by σ and σ^*.

Theorem 3. *Suppose we choose a quadratic positive definite Lyapunov function,*

$$V(\boldsymbol{\lambda}_x, \boldsymbol{v}) = (a/2)\boldsymbol{\lambda}_x^T\boldsymbol{\lambda}_x + (b/2)\boldsymbol{v}^T\boldsymbol{v} + c\boldsymbol{\lambda}_x^T\boldsymbol{v} \tag{40}$$

where,

$$a > 0, \quad b > 0, \quad c \neq 0, \quad ab - c^2 > 0 \tag{41}$$

are constants. Let $\boldsymbol{W}[@t] := \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x, \boldsymbol{v}, t)$ be a family of positive definite matrices that metricize the space $\mathbb{U}$. If $\partial_{\boldsymbol{x}}^2 E(\boldsymbol{x}) > 0$, then, the singular control resulting from either minimum principle (P or P^*) is given by,

$$\boldsymbol{u} = -\boldsymbol{W}^{-1}[@t]\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\boldsymbol{v}\big) \tag{42}$$

where, $\gamma_a \in \mathbb{R}^+$ and $\gamma_b \in \mathbb{R}^+$ are (variable) controller gains.

Proof. Applying (32) we get,

$$\partial_{\boldsymbol{v}} V = c\boldsymbol{\lambda}_x + b\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{\lambda}_x = -(b/c)\boldsymbol{v} \tag{43}$$

Hence,

$$\partial_{\boldsymbol{\lambda}_x} V(\boldsymbol{\lambda}_x, \boldsymbol{v})\big]^T \partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\,\boldsymbol{v} = [a\boldsymbol{\lambda}_x + c\boldsymbol{v}]^T \partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\,\boldsymbol{v} \tag{44}$$

$$= \left(\frac{-ab + c^2}{c}\right)\boldsymbol{v}^T \partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\,\boldsymbol{v} \tag{45}$$

Thus, for (32) to hold, it follows that $c < 0$ if $\partial_{\boldsymbol{x}}^2 E(\boldsymbol{x}) > 0$.

The control solution resulting from the Minimum Principle P or P^* can be written as,

$$\boldsymbol{u} = -\sigma_q[@t]\boldsymbol{W}^{-1}[@t]\big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \tag{46}$$

where σ_q is given by σ or σ^* depending upon the choice of P or P^* respectively. Substituting $\boldsymbol{\lambda}_x = -\partial_{\boldsymbol{x}} E(\boldsymbol{x})$ in (46) we get,

$$\boldsymbol{u} = -\boldsymbol{W}^{-1}[@t]\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\boldsymbol{v}\big) \tag{47}$$

where,

$$\gamma_a := -c\,\sigma_q[@t] \geq 0 \tag{48a}$$

$$\gamma_b := b\,\sigma_q[@t] \geq 0 \tag{48b}$$

are the (variable) controller gains. □

Corollary 1. *A family of continuous accelerated optimization methods, parameterized by $\boldsymbol{W}$, is given by the ODE,*

$$\ddot{\boldsymbol{x}} = -\boldsymbol{W}^{-1}[@t]\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\dot{\boldsymbol{x}}\big) \tag{49}$$

Equation (49) *generates:*

(a) Polyak's equation for the choice of a Euclidean metric (tensor) for $\boldsymbol{W}$ given by the identity matrix;

(b) a continuous accelerated Newton's method for a Riemannian $\boldsymbol{W}$ given by the Hessian, $\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})$; and,

(c) a continuous accelerated quasi-Newton method for the choice of $\boldsymbol{W}$ given by $\boldsymbol{B}(\boldsymbol{x}, \boldsymbol{\lambda}_x)$, a positive definite approximation to the Hessian.

The three special cases of (49) are given explicitly by:

(a) *Polyak's equation:*

$$\ddot{\boldsymbol{x}} = -\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\dot{\boldsymbol{x}}\big) \tag{50}$$

(b) *Continuous Accelerated Newton* $\big(\boldsymbol{W}(\boldsymbol{x}) = \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\big)$*:*

$$\ddot{\boldsymbol{x}} = -\left[\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\right]^{-1}\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\dot{\boldsymbol{x}}\big) \tag{51}$$

(b) *Continuous Accelerated Quasi-Newton* $\big(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{\lambda}_x) = \boldsymbol{B}(\boldsymbol{x}, \boldsymbol{\lambda}_x)\big)$:

$$\ddot{\boldsymbol{x}} = -\boldsymbol{B}^{-1}(\boldsymbol{x}, \boldsymbol{\lambda}_x)\big(\gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\,\dot{\boldsymbol{x}}\big) \tag{52}$$

Remark 8. From Remark 2, it follows that (50) may also be viewed as a derivation of the continuous version of a conjugate gradient method.

Remark 9. Rewriting (51) as,

$$\partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\,\ddot{\boldsymbol{x}} + \gamma_b\,\dot{\boldsymbol{x}} + \gamma_a\,\partial_{\boldsymbol{x}} E(\boldsymbol{x}) = \mathbf{0}$$

it follows that a mechanical-system analogy for the continuous accelerated Newton's method may be described in terms of a nonlinear mass-spring-damper system, where, the Hessian provides the variable inertia. Consequently, Polyak's equation may be viewed as the case corresponding to the use of a constant inertia. Note also that γ_a and γ_b are not necessarily constants; see (48).

Remark 10. In view of Corollary 1, (36) and (38), we define a family of generalized versions of the accelerated gradient, Newton and quasi-Newton methods according to:

(a) *Generalized Accelerated Gradient:*

$$\ddot{\boldsymbol{x}} = -\sigma_q[@t]\,\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}), \qquad \sigma_q[@t] > 0 \tag{53}$$

(b) *Generalized Accelerated Newton:*

$$\ddot{\boldsymbol{x}} = -\sigma_q[@t]\,\left[\partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\right]^{-1} \partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}), \qquad \sigma_q[@t] > 0 \tag{54}$$

(b) *Generalized Accelerated Quasi-Newton*:

$$\ddot{\boldsymbol{x}} = -\sigma_q[@t]\,\boldsymbol{B}^{-1}(\boldsymbol{x}, \boldsymbol{\lambda}_x)\,\partial_{\boldsymbol{v}} V(\boldsymbol{\lambda}_x, \boldsymbol{v}), \qquad \sigma_q[@t] > 0 \tag{55}$$

5. Accelerated Optimization via Direct Construction

As noted in Section 3, any singular control that satisfies (26) and (27) generates a candidate "optimal" continuous-time algorithm for Problem (S); hence, the Minimum Principles proposed in Section 3 are not necessary conditions. They are simply a convenient systematic procedure for generating continuous accelerated optimization methods. In fact, (29) is a sufficient condition; that is, any singular control $\boldsymbol{u}$ that renders $\pounds_\beta V < 0$ generates a globally convergent algorithm. In the case of the quadratic Lyapunov function given by (40), we have,

$$\pounds_\beta V = -[a\boldsymbol{\lambda}_x + c\boldsymbol{v}]^T \partial_{\boldsymbol{x}}^2 E(\boldsymbol{x})\,\boldsymbol{v} + [c\boldsymbol{\lambda}_x + b\boldsymbol{v}]^T \boldsymbol{u} \tag{56}$$

The optimal control resulting from either of the Minimum Principles does not directly incorporate the drift vector field for a generic metric tensor $\boldsymbol{W}$. Motivated by the intuition to design a control that directly incorporates the drift vector field, consider a feedback control strategy given by,

$$\boldsymbol{u} = K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v} + K_c\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{57}$$

where K_a, K_b and K_c are all (variable) real numbers that must be chosen so that $\pounds_\beta V$ is negative. Substituting (57) in (56) we get,

$$\begin{aligned} \pounds_\beta V \quad &= \big(-a + cK_c\big)\boldsymbol{\lambda}_x^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} + \big(-c + bK_c\big)\boldsymbol{v}^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ &\qquad + \big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big)^T \big(K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v}\big) \end{aligned} \tag{58}$$

Proposition 1. *Suppose $\partial^2_{\boldsymbol{x}} E(\boldsymbol{x}) > 0$. Let $c < 0$ in (40). If*

$$K_a > 0, \quad K_b < 0, \quad bK_a = cK_b, \quad \textit{and} \quad K_c = a/c < 0 \tag{59}$$

then, $\pounds_\beta V < 0$ for all $(\boldsymbol{\lambda}_x, \boldsymbol{v}) \neq \boldsymbol{0}$ and $\boldsymbol{u}$ given by (57).

Proof. Substituting $bK_a = cK_b$ in the third term of (58) generates,

$$\big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big)^T \big(K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v}\big) = \frac{K_b}{b}\big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big)^T \big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \leq 0 \tag{60}$$

where, the inequality in (60) follows from the assumption that $K_b < 0$.

With $K_c = a/c$, the first term of (58) vanishes. The second term simplifies to,

$$\big(-c + bK_c\big)\boldsymbol{v}^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} = \left(\frac{-c^2 + ab}{c}\right) \boldsymbol{v}^T \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{61}$$

Because $ab - c^2 > 0$ and $c < 0$, it follows that the second term of (58) is negative for a positive definite Hessian; hence, $\pounds_\beta V < 0$. □

Corollary 2. *Let,*

$$K_a := \gamma_a, \qquad \gamma_a > 0 \tag{62a}$$

$$K_b := -\gamma_b, \qquad \gamma_b > 0 \tag{62b}$$

$$K_c := -\gamma_c, \qquad \gamma_c > 0 \tag{62c}$$

then, the singular control law given by (57) *generates a family of (continuous) Nesterov-type accelerated gradient methods given by,*

$$\ddot{\boldsymbol{x}} + \gamma_a \, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b \, \dot{\boldsymbol{x}} + \gamma_c \, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x}) \, \dot{\boldsymbol{x}} = \boldsymbol{0} \tag{63}$$

Proof. Equation (63) follows from $\ddot{\boldsymbol{x}} = \boldsymbol{u}$, with $\boldsymbol{u}$ given by (57). The claim that the resulting ODE generates a family of Nesterov's accelerated gradient method[8] follows by considering a discretization of (63). To this end, consider first a discretization of the the last term on the left-hand-side of (63):

$$\gamma_c \, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x}) \, \dot{\boldsymbol{x}} = \gamma_c \, \frac{d}{dt} \Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x}) \Big) \longrightarrow \frac{\gamma_c}{h_k} \Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) - \partial_{\boldsymbol{x}} E(\boldsymbol{x}_{k-1}) \Big) \tag{64}$$

where, $h_k > 0$ is a discretization step. Next, consider the first three terms of (63). These are identical to Polyak's equation (Cf. (8)); hence it follows from (6) and (64) that (63) may be discretized as,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) + \beta_k (\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \gamma_k \Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) - \partial_{\boldsymbol{x}} E(\boldsymbol{x}_{k-1}) \Big) \tag{65}$$

Nesterov's method[8] is given by,

$$\boldsymbol{x}_k = \boldsymbol{y}_k - \alpha_k \, \partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) \tag{66a}$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_k + \beta_k (\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \tag{66b}$$

Substituting (66a) in (66b) generates

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \alpha_k \partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) + \beta_k (\boldsymbol{y}_k - \boldsymbol{y}_{k-1}) - \alpha_k \beta_k \Big(\partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) - \partial_{\boldsymbol{y}} E(\boldsymbol{y}_{k-1}) \Big) \tag{67}$$

which is the same as (65) with $\gamma_k = \alpha_k \beta_k$. □

Remark 11. Equation (63) was introduced and studied by Alvarez et al[9] as a "dynamical inertial Newton" system. Shi et al [10] generated this system as a "high-resolution" ODE that represents Nesterov's method[8] in continuous-time. Equation (63) differs from (8) by an additive "Hessian-driven damping" term[9] which has the effect of a "gradient correction"[10] a vis-à-vis Polyak's equation[2].

References

[1] I. M. Ross, An optimal control theory for nonlinear optimization, J. Comp. and Appl. Math., 354 (2019) 39–51.

[2] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Math. and Math. Phys., 4/5 (1964) 1–17 (Translated by H. F. Cleaves).

[3] W. Su, S. Boyd, E. J. Candes, A differential equation for modeling Nesterov's accelerated gradient method: theory and insights, J. machine learning research, 17 (2016) 1–43.

[4] R. B. Vinter, Optimal Control, Birkhäuser, Boston, MA, 2000

[5] I. M. Ross, A Primer on Pontryagin's Principle in Optimal Control, second ed., Collegiate Publishers, San Francisco, CA, 2015.

[6] E. D. Sontag, Mathematical Control Theory: Deterministic Finite Dimensional Systems, second ed., Springer, New York, NY, 1998.

[7] R. A. Freeman, P. V. Kokotović, Optimal nonlinear controllers for feedback linearizable systems, Proc. ACC, Seattle, WA, June 1995.

[8] Yu. E. Nesterov, A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, Soviet Math. Dokl., 27/2 (1983) 371–376 (Translated by A. Rosa).

[9] F. Alvarez, H. Attouch, J. Bolte, P. Redont, A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Applications to optimization and mechanics. J. Math. Pures Appl. 81 (2002) 747–779.

[10] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, Understanding the acceleration phenomenon via high-resolution differential equations, arXiv preprint (2018) arXiv:1810.08907.