# A Derivation of Nesterov's Accelerated Gradient Algorithm from Optimal Control Theory

I. M. Ross[1]

*Naval Postgraduate School, Monterey, CA 93943*

---

**Abstract**

Nesterov's accelerated gradient algorithm is derived from first principles. The first principles are founded on the recently-developed optimal control theory for optimization. The necessary conditions for optimal control generate a controllable dynamical system for accelerated optimization. Stabilizing this system via a control Lyapunov function generates an ordinary differential equation. An Euler discretization of the differential equation produces Nesterov's algorithm.

*Keywords:* accelerated optimization, singular optimal control theory, Lie derivative, control Lyapunov function

---

## 1. Introduction

In broad terms, Nesterov's accelerated gradient method for minimizing a function, $E : \mathbb{R}^{N_x} \to \mathbb{R}$, is given by[1],

$$\boldsymbol{x}_k = \boldsymbol{y}_k - \alpha_k \, \partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) \tag{1a}$$

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

where, $\boldsymbol{x}_k \in \mathbb{R}^{N_x}$, $\boldsymbol{y}_k \in \mathbb{R}^{N_x}$, $\alpha_k \in \mathbb{R}_+$, $\beta_k \in \mathbb{R}_+$ and $k \in \mathbb{N}$. There are many ways to "explain" this algorithm starting from a well-deserved attribution to Nesterov's penetrating insights in convex programming to discretizations of certain ordinary differential equations(ODEs)[2, 3, 4]. The use of ODEs to

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

explain algorithms has a long and rich history[5, 6, 7] going all the way back to Gavurin's pioneering work in the late 1950s[8]. The ODEs that explain algorithms are typically derived by considering the limiting cases of the algorithmic maps themselves. In other words, an ODE is usually generated after an algorithm is invented and not the other way around. Thus the question remains: can the ODEs be generated by some higher-level universal principle and without any a priori knowledge of algorithms? The recently-developed[9] optimal control theory for optimization answers this question in the affirmative. According this theory, controllable ODEs for optimization algorithms can be generated as an outcome of the necessary conditions for optimal control. Producing a practical algorithm is then reduced to using suitable semi-discretization methods. As shown in [9], these ideas generate a vast number of well-known but *unaccelerated* algorithms such as Newton's method and the gradient method. It was conjectured in [9] that the theory could also generate accelerated optimization methods by switching the dynamical model from a single integrator to a double integrator. In this note, we prove this conjecture by deriving Nesterov's accelerated gradient method using optimal control theory as a foundation for optimization.

## 2. Background: Optimal Control Theory for Optimization

We rely heavily on [9] in developing the basic framework while providing sufficient details for completeness. To this end, consider the unconstrained static optimization problem given by,

$$(S)\left\{\underset{\boldsymbol{x}_f\in\mathbb{R}^{N_x}}{\text{Minimize}}\; E(\boldsymbol{x}_f)\right. \tag{2}$$

To produce an optimal control problem that solves $(S)$, we create a vector field by "sweeping" the function $E$ backwards in time according to,

$$y(t) := E(\boldsymbol{x}(t)) \tag{3}$$

Differentiating (3) with respect to time we get,

$$\dot{y} = \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \cdot \dot{\boldsymbol{x}} \tag{4}$$

As shown in [9], if we set $\dot{\boldsymbol{x}} = \boldsymbol{u}$ to generate a controllable dynamical system, the resulting theory generates unaccelerated optimization methods. In pursuit of acceleration, we replace the single integrator model by a double integrator,

$$\dot{\boldsymbol{x}} = \boldsymbol{v}, \quad \dot{\boldsymbol{v}} = \boldsymbol{u} \tag{5}$$

This model generates a primal controllable dynamical system given by,

$$\dot{y} = \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \cdot \boldsymbol{v}, \quad \dot{\boldsymbol{x}} = \boldsymbol{v}, \quad \dot{\boldsymbol{v}} = \boldsymbol{u} \tag{6}$$

Taking generic initial conditions and a final rest "velocity," $\boldsymbol{v}(t_f) = \boldsymbol{0}$, we arrive at the following candidate optimal control problem $(R)$ that purportedly solves the optimization problem $(S)$:

$$(R) \left\{ \begin{array}{lrl} \text{Minimize} & J[y(\cdot), \boldsymbol{x}(\cdot), \boldsymbol{v}(\cdot), \boldsymbol{u}(\cdot), t_f] := & y(t_f) \\ \text{Subject to} & \dot{\boldsymbol{x}} = & \boldsymbol{v} \\ & \dot{\boldsymbol{v}} = & \boldsymbol{u} \\ & \dot{y} = & \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \cdot \boldsymbol{v} \\ & (\boldsymbol{x}(t_0), t_0) = & (\boldsymbol{x}^0, t^0) \\ & y(t_0) = & E(\boldsymbol{x}^0) \\ & \boldsymbol{v}(t_f) = & \boldsymbol{0} \end{array} \right. \tag{7}$$

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 all free.

*Remark* 1. The cost functional in Problem $(R)$ is given by the final value of the $y$ variable, which, by construction, is exactly equal to the objective function of Problem $(S)$. Consequently, a solution to Problem $(R)$ generates a solution to Problem $(S)$.

*Remark* 2. A solution to Problem $(R)$ generates an optimal $\boldsymbol{x}$-trajectory. A discretization of this continuous-time trajectory generates a practical algorithm for Problem $(S)$.

It follows from the preceding remarks that not only is Problem $(S)$ embedded in Problem $(R)$ but also that a solution to Problem $(R)$ automatically generates an algorithm for solving Problem $(S)$.

## 3. Necessary Conditions for Problem ($\boldsymbol{R}$)

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

*Proof.* The Pontryagin Hamiltonian[10] for this problem is given by,

$$H(\boldsymbol{\lambda}_x, \boldsymbol{\lambda}_v, \lambda_y, \boldsymbol{x}, \boldsymbol{v}, y, \boldsymbol{u}) := \boldsymbol{\lambda}_x \cdot \boldsymbol{v} + \boldsymbol{\lambda}_v \cdot \boldsymbol{u} + \lambda_y \, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \cdot \boldsymbol{v} \tag{8}$$

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

$$\dot{\boldsymbol{\lambda}}_x = -\partial_{\boldsymbol{x}} H = -\lambda_y \, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x}) \, \boldsymbol{v} \tag{9a}$$

$$\dot{\boldsymbol{\lambda}}_v = -\partial_{\boldsymbol{v}} H = -\boldsymbol{\lambda}_x - \lambda_y \, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \tag{9b}$$

$$\dot{\lambda}_y = -\partial_y H = 0 \tag{9c}$$

The transversality conditions[10] for Problem ($R$) are given by,

$$\boldsymbol{\lambda}_x(t_f) = \boldsymbol{0} \tag{10a}$$

$$\boldsymbol{\lambda}_v(t_0) = \boldsymbol{0} \tag{10b}$$

$$\boldsymbol{\lambda}_y(t_f) = \nu_0 \geq 0 \tag{10c}$$

where, $\nu_0$ is the cost multiplier. From (9c) and (10c) we have,

$$\lambda_y(t) = \nu_0 \tag{11}$$

If $\nu_0 = 0$, then $\lambda_y(t) \equiv 0$. This implies, from (9a) and (10a), that $\boldsymbol{\lambda}_x(t) \equiv \boldsymbol{0}$. Similarly, $\boldsymbol{\lambda}_v(t) \equiv \boldsymbol{0}$ from (9b) and (10b). 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 arcs are 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,

$$\partial_{\boldsymbol{u}} H = \boldsymbol{\lambda}_v(t) = \boldsymbol{0} \qquad \forall t \in [t_0, t_f] \tag{12}$$

Differentiating (12) with respect to time, we get,

$$\frac{d}{dt}\partial_{\boldsymbol{u}} H = \dot{\boldsymbol{\lambda}}_v(t) = -\boldsymbol{\lambda}_x - \nu_0\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) = \boldsymbol{0} \tag{13}$$

Equation (13) does not generate an expression for the control function; hence, taking the second time derivative of $\partial_{\boldsymbol{u}} H$ we get,

$$\begin{aligned}\frac{d^2}{dt^2}\partial_{\boldsymbol{u}} H &= -\dot{\boldsymbol{\lambda}}_x - \nu_0\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \dot{\boldsymbol{x}} \\ &= -\dot{\boldsymbol{\lambda}}_x - \nu_0\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ &\equiv \boldsymbol{0} \end{aligned} \tag{14}$$

where, the last equality follows from (9a) and Lemma 1. Hence, we have,

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

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

**Theorem 2** (A Transversality Mapping Theorem)**.** *The necessary condition for Problem* ($S$) *is embedded in the transversality condition for Problem* ($R$)*.*

*Proof.* From (13), we have

$$\boldsymbol{\lambda}_x(t) = -\nu_0\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}(t)) \tag{15}$$

From (10a) and Lemma 1, it follows that $\partial_{\boldsymbol{x}_f} E(\boldsymbol{x}_f) = \boldsymbol{0}$. □

Collecting all relevant equations, it follows that the primal-dual control dynamical system generated by Problem ($R$) is given by,

$$\dot{\boldsymbol{x}} = \boldsymbol{v} \qquad \dot{\boldsymbol{\lambda}}_x = -\lambda_y\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{16a}$$

$$\dot{\boldsymbol{v}} = \boldsymbol{u} \qquad \dot{\boldsymbol{\lambda}}_v = -\boldsymbol{\lambda}_x - \lambda_y\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \tag{16b}$$

$$\dot{y} = \partial_{\boldsymbol{x}} E(\boldsymbol{x}) \cdot \boldsymbol{v} \qquad \dot{\lambda}_y = 0 \tag{16c}$$

The boundary conditions for (16) are given by,

$$\boldsymbol{x}(t^0) = \boldsymbol{x}^0 \qquad \boldsymbol{v}(t_f) = \boldsymbol{0} \tag{17a}$$

$$y(t^0) = E(\boldsymbol{x}^0) \qquad \boldsymbol{\lambda}_x(t_f) = \boldsymbol{0} \tag{17b}$$

$$\boldsymbol{\lambda}_v(t^0) = \boldsymbol{0} \qquad \lambda_y(t_f) = \nu_0 > 0 \tag{17c}$$

Because the optimal control is singular of infinite order, any control trajectory that satisfies (16) and (17) is optimal. Along a singular arc, $\boldsymbol{\lambda}_v(t) \equiv \mathbf{0}$; hence, the auxiliary controllable dynamical system of interest[9] resulting from (16) is given by,

$$(A) \begin{cases} \dot{\boldsymbol{\lambda}}_x = -\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ \dot{\boldsymbol{v}} = \boldsymbol{u} \end{cases} \tag{18}$$

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

$$(T) \begin{cases} \boldsymbol{\lambda}_x(t_f) = \mathbf{0} \\ \boldsymbol{v}(t_f) = \mathbf{0} \end{cases} \tag{19}$$

That is, any singular control that satisfies (18) and (19) generates a candidate "optimal" continuous-time algorithm for Problem $(S)$.

## 4. A Feedback Controller for the $(A)$-$(T)$ System

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

$$\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) := \begin{bmatrix} -\partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ \boldsymbol{u} \end{bmatrix} \tag{20}$$

Let $V : (\boldsymbol{\lambda}_x, \boldsymbol{v}) \mapsto \mathbb{R}$ be a control Lyapunov function (CLF)[11] 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[11, 12] for globally guiding the pair $(\boldsymbol{\lambda}_x, \boldsymbol{v})$ to $(\mathbf{0}, \mathbf{0})$ is to render

$$\pounds_\beta V = \partial V(\boldsymbol{\lambda}_x, \boldsymbol{v}) \cdot \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{u}) < 0 \tag{21}$$

whenever $\boldsymbol{x} \neq \boldsymbol{x}_f$. Consequently, we seek to design a (singular) control function that satisfies (21).

For the remainder of this note, we choose the following positive definite CLF,

$$V(\boldsymbol{\lambda}_x, \boldsymbol{v}) = (a/2)\boldsymbol{\lambda}_x \cdot \boldsymbol{\lambda}_x + (b/2)\boldsymbol{v} \cdot \boldsymbol{v} + c\boldsymbol{\lambda}_x \cdot \boldsymbol{v} \tag{22}$$

where,

$$a > 0, \quad b > 0, \quad c < 0, \quad ab - c^2 > 0 \tag{23}$$

are constants. As a result, we have,

$$\pounds_\beta V = -[a\boldsymbol{\lambda}_x + c\boldsymbol{v}] \cdot \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} + [c\boldsymbol{\lambda}_x + b\boldsymbol{v}] \cdot \boldsymbol{u} \tag{24}$$

A generic linear feedback controller[11] is given by $\boldsymbol{u} = K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v}$, where $K_a$ and $K_b$ are real numbers. Motivated by the intuition to design a control that directly incorporates the drift vector field to render $\pounds_\beta V < 0$, consider a modification to the linear feedback control strategy given by,

$$\boldsymbol{u} = K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v} + K_c\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{25}$$

where $K_a, K_b$ and $K_c$ are all (variable) real numbers that must be chosen such that $\pounds_\beta V$ is negative.

**Proposition 1.** *Suppose $E$ is a convex function and $\boldsymbol{u}$ is given by* (25)*. If*

$$K_a > 0, \quad K_b < 0, \quad bK_a = cK_b, \quad \textit{and} \quad K_c = a/c \tag{26}$$

*then, $\pounds_\beta V < 0$ for all $(\boldsymbol{\lambda}_x, \boldsymbol{v}) \neq \boldsymbol{0}$.*

*Proof.* Substituting (25) in (24) we get,

$$\begin{aligned}\pounds_\beta V \quad &= \big(-a + cK_c\big)\boldsymbol{\lambda}_x \cdot \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} + \big(-c + bK_c\big)\boldsymbol{v} \cdot \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \\ &\qquad + \big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \cdot \big(K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v}\big)\end{aligned} \tag{27}$$

Substituting $bK_a = cK_b$ in the third term of (27) generates,

$$\big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \cdot \big(K_a\, \boldsymbol{\lambda}_x + K_b\, \boldsymbol{v}\big) = \frac{K_b}{b}\big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \cdot \big(c\boldsymbol{\lambda}_x + b\boldsymbol{v}\big) \leq 0 \tag{28}$$

where, the inequality in (28) follows from the assumption that $K_b < 0$.

With $K_c = a/c$, the first term of (27) vanishes. The second term simplifies to,

$$\big(-c + bK_c\big)\boldsymbol{v} \cdot \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} = \left(\frac{-c^2 + ab}{c}\right) \boldsymbol{v} \cdot \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \boldsymbol{v} \tag{29}$$

Because $ab - c^2 > 0$ and $c < 0$, it follows that the second term of (27) is negative for a positive definite Hessian; hence, $\pounds_\beta V < 0$. □

**Corollary 1.** *Let,*

$$K_a := \gamma_a, \qquad \gamma_a > 0 \tag{30a}$$

$$K_b := -\gamma_b, \qquad \gamma_b > 0 \tag{30b}$$

$$K_c := -\gamma_c, \qquad \gamma_c > 0 \tag{30c}$$

*then, the singular control law given by* (25) *generates the second order ODE,*

$$\ddot{\boldsymbol{x}} + \gamma_a\, \partial_{\boldsymbol{x}} E(\boldsymbol{x}) + \gamma_b\, \dot{\boldsymbol{x}} + \gamma_c\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \dot{\boldsymbol{x}} = \mathbf{0} \tag{31}$$

*Proof.* Equation (31) follows directly from (5), (25) and (30). □

## 5. Equation (31) Generates (1)

As shown by Shi et al[2], a discretization of (31) generates Nesterov's accelerated gradient method. To see this, consider first a discretization of the the last term on the left-hand-side of (31):

$$\gamma_c\, \partial^2_{\boldsymbol{x}} E(\boldsymbol{x})\, \dot{\boldsymbol{x}} = \gamma_c\, \frac{d}{dt}\Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x})\Big) \longrightarrow \frac{\gamma_c}{h_k}\Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) - \partial_{\boldsymbol{x}} E(\boldsymbol{x}_{k-1})\Big) \tag{32}$$

where, $h_k > 0$ is a discretization step. Next, consider the first three terms of (31). These are identical to Polyak's equation whose discretization generates the heavy ball method[13, 14],

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \tag{33}$$

Hence, (31) may be discretized as,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) - \gamma_k\Big(\partial_{\boldsymbol{x}} E(\boldsymbol{x}_k) - \partial_{\boldsymbol{x}} E(\boldsymbol{x}_{k-1})\Big) \tag{34}$$

Substituting (1a) in (1b), Nesterov's method may be rewritten as,

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \alpha_k \partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) + \beta_k(\boldsymbol{y}_k - \boldsymbol{y}_{k-1}) - \alpha_k\beta_k\Big(\partial_{\boldsymbol{y}} E(\boldsymbol{y}_k) - \partial_{\boldsymbol{y}} E(\boldsymbol{y}_{k-1})\Big) \tag{35}$$

Equation (35) is the same as (34) with $\gamma_k = \alpha_k\beta_k$.

*Remark* 3. Equation (31) was introduced and studied by Alvarez et al[3] as a "dynamical inertial Newton" system. Shi et al [2] generated this system as a "high-resolution" ODE that represents Nesterov's method[1]. See also [15] for further details on an optimal control theory for accelerated optimization.

## References


[1] 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).

[2] 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.

[3] 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.

[4] 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.

[5] P. T. Boggs, The solution of nonlinear system of equations by $A$-stable integration techniques, SIAM J. Numer. Anal. 8/4 (1971) 767–785.

[6] S. Smale, A convergent process of price adjustment and global Newton methods, J. mathematical economics, 3 (1976), 107–120.

[7] A. A. Brown, M. C. Bartholomew-Biggs, Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations, J. optimization theory and applications, 62/2 (1989) 211–224.

[8] M. K. Gavurin, Nonlinear functional equations and continuous analogues of iteration methods, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958) 18–31.

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

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

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

[12] F. Clarke, Lyapunov functions and feedback in nonlinear control. In: M.S. de Queiroz, M. Malisoff, P. Wolenski (eds) Optimal control, stabilization and nonsmooth analysis. Lecture Notes in Control and Information Science, vol 301. Springer, Berlin, Heidelberg (2004), 267–282.

[13] 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).

[14] B. Polyak, P. Shcherbakov, Lyapunov functions: an optimization theory perspective, IFAC PapersOnLine, 50-1 (2017) 7456–7461.

[15] I. M. Ross, An optimal control theory for accelerated optimization, arXiv preprint (2019) arXiv:1902.09004.