2020 59th IEEE Conference on Decision and Control (CDC)
Jeju Island, Republic of Korea, December 14-18, 2020

# High-Resolution Modeling of the Fastest First-Order Optimization Method for Strongly Convex Functions

Boya Sun, Jemin George and Solmaz Kia

*Abstract*— **Motivated by the fact that the gradient-based optimization algorithms can be studied from the perspective of limiting ordinary differential equations (ODEs), here we derive an ODE representation of the accelerated triple momentum (TM) algorithm. For unconstrained optimization problems with strongly convex cost, the TM algorithm has a proven faster convergence rate than the Nesterov's accelerated gradient (NAG) method but with the same computational complexity. We show that similar to the NAG method, in order to accurately capture the characteristics of the TM method, we need to use a high-resolution modeling to obtain the ODE representation of the TM algorithm. We propose a Lyapunov analysis to investigate the stability and convergence behavior of the proposed high-resolution ODE representation of the TM algorithm. We compare the rate of the ODE representation of the TM method with that of the NAG method to confirm its faster convergence. Our study also leads to a tighter bound on the worst rate of convergence for the ODE model of the NAG method. In this paper, we also discuss the use of the integral quadratic constraint (IQC) method to establish an estimate on the rate of convergence of the TM algorithm. A numerical example verifies our results.**

## I. Introduction

As it has been known in the classical optimization literature, improvement to the rate of convergence of optimization algorithms within a first-order framework can be obtained through methods such as quasi-Newton [1], [2], Polyak's heavy-ball [3], [4], and Nesterov's accelerated gradient (NAG) [5], [6] methods. Among these methods, because of its simple structure and its global convergence guarantees for convex objective functions, NAG has received much attention in the optimization and machine learning community. However, the quest for alternative fast converging first-order optimization algorithms is still an ongoing research topic. Recently, a new accelerated gradient-based method called the Triple Momentum (TM) method, which has the same computational complexity as the NAG method but with a proven faster convergence rate, was proposed in [7]. Our objective in this paper is to obtain a high-resolution continuous-time representation for the TM method and study its stability and convergence via control theoretic tools.

ODE representation and its analysis for optimization algorithms in the continuous-time domain have a long history going back to [8]; more discussions can be found in [9]–[13]. Continuous-time modeling comes with ease in theoretical analysis via powerful control theoretic tools such as Lyapunov analysis, perturbation theory, and the integral quadratic constraint (IQC) methods. Also, the continuous-time perspective provides intuition to design new algorithms, especially arriving at distributed algorithms in a systematic way from centralized solutions. Furthermore, the convergence analysis of several gradient-based Markov Chain Monte Carlo sampling schemes relies on the continuous-time approximation of such algorithms [14], [15]. Therefore, recently, ODE modeling has regained popularity to address the need to design new distributed gradient descent based optimization algorithms [16]–[20], as well as to analyze the new accelerated optimization algorithms [21]–[25]. In [21], a second-order ODE is presented as the limit of the NAG method. The connection between ODEs and discrete optimization algorithms is further strengthened in [22] by establishing an equivalence between the estimate sequence technique and Lyapunov function techniques. In [23], the authors propose a variational, continuous-time framework for understanding accelerated methods and show that there is a Lagrangian functional that generates a large class of accelerated methods in continuous time. NAG method and many of its generalizations can be viewed as a systematic way to go from the continuous-time curves generated by the Lagrangian functional to a family of discrete-time accelerated algorithms [23]. An ODE-based analysis of mirror descent given in [26] delivers new insights into the connections between acceleration and constrained optimization, averaging, and stochastic mirror descent. A deeper insight into the acceleration phenomenon via high-resolution ODE representation of various first-order methods is presented in [24]. These high-resolution ODEs are shown to permit a general Lyapunov function framework for convergence analysis in both continuous and discrete time [24]. Finally, in [25], the authors show that different types of proximal optimization algorithms based on fixed-point iteration can be derived from the gradient flow by using splitting methods for ODEs.

The connection between ODE representation of optimization algorithms and their discrete-time counterpart is often established by taking the step size of the discrete-time algorithm to zero and deriving a limiting ODE using first-order derivatives modeling. This approach works well for gradient descent and Newton algorithms (thus obtaining $\dot{x} = -\nabla f(x)$ and $\dot{x} = -\nabla^2 f(x)^{-1}\nabla f(x)$ from $x(k+1) = x(k) - s\nabla f(x)$ and $x(k+1) = x(k) - s\nabla^2 f(x(k))^{-1}\nabla f(x(k))$, respectively, where $s$ is the step size). However, recent literature has shown that first-order ODE modeling of accelerated algo-

B. Sun and S. Kia are with the Department of Mechanical and Aerospace Engineering, University of California Irvine, Irvine, CA 92697. J. George is with CCDC ARL, Adelphi, MD 20783, USA. The work of the authors from UCI was supported by NSF award ECCS-1653838. Corresponding author's email solmaz@uci.edu

978-1-7281-7447-1/20/$31.00 ©2020 IEEE

rithms such as the Polyak's heavy-ball and NAG methods fails to capture the true behavior of these algorithms [24]. In fact, it has been shown that the first-order ODE representation cannot differentiate between these two algorithms since it yields an identical limiting equation for both. Recent literature, therefore, has looked at higher-order resolution ODE representation of these algorithms [21], [24]. These high-resolution ODEs are more accurate since they better capture the characterizations of the discrete-time accelerated methods in their continuous-time counterpart representations.

In this paper, we derive a higher-resolution ODE representation for the accelerated TM method and show that the high-resolution ODE is able to accurately capture the characterizations of the TM method. For clarity, hereafter we refer to the TM method of [7] as the discrete-time TM. We first use an IQC framework as a convenient tool to assess the stability and obtain an estimate on the the rate of convergence of the continuous-time TM method. However, since the IQC provides only a sufficient condition for stability and convergence analysis, next, we use a Lyapunov analysis to study the stability and convergence behavior of the higher-resolution ODE representation of TM method. We further use the Lyapunov analysis to show that the TM method is robust to minor deviations in algorithm parameters. We also use our framework to estimate the rate of convergence of the TM algorithm and compare it to the NAG method, which confirms its faster convergence. Our work also leads to a tighter estimate on the rate of convergence of the ODE representation of the NAG method. Using a numerical example, we show the accuracy of our higher-order resolution ODE representation in capturing the accelerated behavior of the TM method and its faster convergence over the high-resolution ODE representation of the NAG method given in [24].

*Notations*: $\mathbb{R}$ and $\mathbb{R}_{>0}$ are the set of real and positive real numbers. $A^\top$ is the transpose of matrix $A$. We let $0_n$ denote the vector of $n$ zeros and $I_n$ denote the $n\times n$ identity matrix. When clear from the context, we do not specify the matrix dimensions. For a vector $x \in \mathbb{R}^n$, $\|x\| = \sqrt{x^\top x}$ is the standard Euclidean norm. The gradient of $f : \mathbb{R}^d \to \mathbb{R}$, is denoted by $\nabla f(x)$. The following relations hold for a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ that is $M$-strongly convex, $M \in \mathbb{R}_{>0}$, over $\mathbb{R}^d$ ,

$$f(\mathsf{y})-f(\mathsf{x})\leq\nabla f(\mathsf{x})^\top(\mathsf{y}-\mathsf{x})+\frac{1}{2M}\|\nabla f(\mathsf{y})-\nabla f(\mathsf{x})\|^2, \quad (1a)$$

$$M\|\mathsf{y}-\mathsf{x}\|^2 \leq (\mathsf{y}-\mathsf{x})^\top(\nabla f(\mathsf{y})-\nabla f(\mathsf{x})), \quad (1b)$$

$$M\|\mathsf{y}-\mathsf{x}\| \leq \|\nabla f(\mathsf{y})-\nabla f(\mathsf{x})\|, \quad (1c)$$

for any $\mathsf{x},\mathsf{y} \in \mathbb{R}^d$ [27]. When $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-Lipschitz continuous, $L \in \mathbb{R}_{>0}$, i.e., $\|\nabla f(\mathsf{y}) - \nabla f(\mathsf{x})\| \leq L\|\mathsf{y}-\mathsf{x}\|$, we have

$$f(\mathsf{y}) - f(\mathsf{x}) \leq \nabla f(\mathsf{x})^\top(\mathsf{y}-\mathsf{x})+\frac{L}{2}\|\mathsf{y}-\mathsf{x}\|^2, \quad (2a)$$

$$f(\mathsf{y})-f(\mathsf{x}) \geq \nabla f(\mathsf{x})^\top(\mathsf{y}-\mathsf{x})+\frac{1}{2L}\|\nabla f(\mathsf{y})-\nabla f(\mathsf{x})\|^2, \quad (2b)$$

for all $\mathsf{x},\mathsf{y} \in \mathbb{R}^d \times \mathbb{R}^d$ [27]. We represent the class of differentiable and $M$ strongly convex functions whose gradient is L-Lipschitz with $\mathcal{S}_{M,L}$.

## II. Problem definition

Consider

$$\mathsf{x}^\star = \operatorname{argmin}_{x\in\mathbb{R}^n} f(x), \quad (3)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $f \in \mathcal{S}_{L,M}$. We assume that $\mathsf{x}^\star$ exists and is reachable. The minimizer of this optimization problem is specified as follows.

**Lemma II.1** (Minimizer of (3) [1])**.** *Consider optimization problem* (3)*. A point* $\mathsf{x}^\star \in \mathbb{R}^n$ *is a unique solution of* (3) *if and only if* $\nabla f(\mathsf{x}^\star) = 0$*.*

In what follows, we let

$$\kappa = \frac{L}{M}, \quad \rho = 1 - \frac{1}{\sqrt{\kappa}}. \quad (4)$$

We refer to $\kappa$ as the condition number of the cost function $f$.

### A. Discrete-time TM Method

Here we consider the TM method, proposed in [7] as the fastest known globally convergent first-order method for solving strongly convex optimization problems. The TM method is an accelerated gradient-based optimization algorithm given as

$$\epsilon_{k+1} = (1+\beta)\epsilon_k - \beta\epsilon_{k-1} - \alpha\nabla f(y_k), \quad (5a)$$

$$y_k = (1+\gamma)\epsilon_k - \gamma\epsilon_{k-1}, \quad (5b)$$

$$x_k = (1+\delta)\epsilon_k - \delta\epsilon_{k-1}, \quad (5c)$$

where the algorithm parameters are given as (recall (4))

$$(\alpha,\beta,\gamma,\delta) = \left(\frac{1+\rho}{L}, \frac{\rho^2}{2-\rho}, \frac{\rho^2}{(1+\rho)(2-\rho)}, \frac{\rho^2}{1-\rho^2}\right), \quad (6)$$

and $\epsilon_0, \epsilon_{-1} \in \mathbb{R}^n$ are the initial conditions, $x \in \mathbb{R}^n$ is the output. The TM method has the same numerical complexity as the NAG method but converges faster. In [7], it is shown that starting from any initial conditions $\epsilon_0, \epsilon_{-1} \in \mathbb{R}^n$, the trajectories of $x_k, y_k, \epsilon_k$ converge to $\mathsf{x}^\star$ with the same rate but the convergence error is lowest for the output $x$. We observe the same trend in the high-resolution ODE representation of the TM method; see Section V for numerical examples.

Our objective in this paper is to derive a high-resolution ODE representation of the TM algorithm that accurately captures the performance characteristics of its discrete-time counterpart and establish its formal convergence guarantees using the Lyapunov stability analysis. To facilitate our discussions given next, we define a function $\mu(\alpha,\beta) : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}_{>0}$ (or simply $\mu$) as

$$\mu(\alpha,\beta) = \left(\frac{1-\beta}{\sqrt{\alpha}(1+\beta)}\right)^2. \quad (7)$$

Using the parameter relations given in (4) and (6) for the TM method, $\mu = \mu(\alpha,\beta)$ can also be written as

$$\mu = \frac{(9\,\kappa^2\sqrt{\kappa} - 6\,\kappa^2 + \kappa\sqrt{\kappa})L}{8\,\kappa^3\sqrt{\kappa} - 12\,\kappa^3 + 14\,\kappa^2\sqrt{\kappa} - 9\,\kappa^2 + 4\,\kappa\sqrt{\kappa} - \kappa}. \quad (8)$$

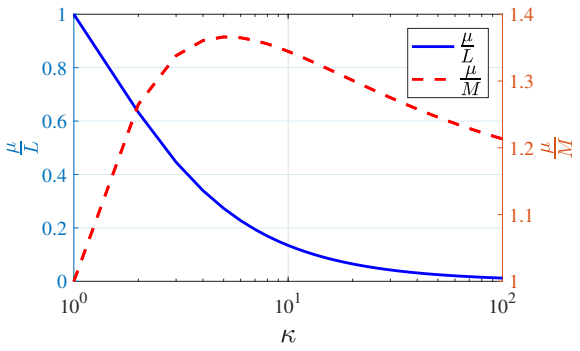



Fig. 1: $\mu/L$ (solid blue) and $\mu/M$ (dashed red) vs. $\kappa$.

As shown in Fig. 1, the maximum value of $\mu$ is $L$, which is attained at $\kappa=1$. When $\kappa\to\infty$, $\mu\to 0$. We can also show that

$$\mu(\alpha,\beta)\in(0,L]. \tag{9}$$

Replacing $L$ with $\kappa M$ in (7), we can also show that

$$\mu(\alpha,\beta)\in[M,1.3661M]. \tag{10}$$

**Remark II.1** (Role of parameter $\mu$)**.** Parameter $\mu$ plays a vital role in the analysis of the high-resolution ODE representation of the TM method. Also note that after substituting appropriate $\alpha$ and $\beta$ values into (7), we obtain $\mu_{\text{NAG}}$=$M$ for NAG method while $\mu\geq M$ for the TM method. Therefore the parameter $\mu$ also plays an important role when comparing the convergence rate between the high-resolution NAG and TM methods. □

## III. Continuous-Time Representation of the TM Method

Let $t_k=k\sqrt{\alpha}$ and $y_k=Y(t_k)$ for some sufficiently smooth curve $Y(t)$. Now the Taylor series expansion at both $y_{k+1}$ and $y_{k-1}$ with the step size $\sqrt{\alpha}$ are

$$\begin{aligned} y_{k+1}=Y(t_{k+1})&=Y(t_k)+\dot{Y}(t_k)\sqrt{\alpha}\\ &\quad+\frac{1}{2}\ddot{Y}(t_k)(\sqrt{\alpha})^2+\mathcal{O}((\sqrt{\alpha})^3), \end{aligned} \tag{11}$$

$$\begin{aligned} y_{k-1}=Y(t_{k-1})&=Y(t_k)-\dot{Y}(t_k)\sqrt{\alpha}\\ &\quad+\frac{1}{2}\ddot{Y}(t_k)(\sqrt{\alpha})^2+\mathcal{O}((\sqrt{\alpha})^3). \end{aligned} \tag{12}$$

Combining (11) and (12) yields

$$Y(t_{k+1})+Y(t_{k-1})-2Y(t_k)=\alpha\ddot{Y}(t_k)+\mathcal{O}(\alpha^2). \tag{13}$$

Next, we note that we can rewrite (5) as

$$\epsilon_{k+1}=\epsilon_k+\beta\left(\epsilon_k-\epsilon_{k-1}\right)-\alpha\nabla f(y_k), \tag{14a}$$

$$y_k=\epsilon_k+\gamma\left(\epsilon_k-\epsilon_{k-1}\right), \tag{14b}$$

$$x_k=\epsilon_k+\delta\left(\epsilon_k-\epsilon_{k-1}\right). \tag{14c}$$

From (14a), we obtain

$$\beta\left(\epsilon_{k-1}-\epsilon_k\right)+\left(\epsilon_{k+1}-\epsilon_k\right)+\alpha\nabla f(y_k)=0. \tag{15}$$

Now adding and subtracting $\beta\left(\epsilon_{k+1}-\epsilon_k\right)$ and dividing both sides of (15) with $\beta\alpha$ yields $\frac{\left(\epsilon_{k+1}+\epsilon_{k-1}-2\epsilon_k\right)}{\alpha}+\frac{(1-\beta)}{\beta\alpha}\left(\epsilon_{k+1}-\epsilon_k\right)+\frac{1}{\beta}\nabla f(y_k)=0$. Substituting $\epsilon_k=\varepsilon(t_k)$ and $y_k=Y(t_k)$ and (13) yields

$$\begin{aligned} \ddot{\varepsilon}(t_k)+\mathcal{O}(\alpha)+\frac{1-\beta}{\beta\sqrt{\alpha}}\left(\dot{\varepsilon}(t_k)+\frac{1}{2}\ddot{\varepsilon}(t_k)\sqrt{\alpha}+\mathcal{O}(\alpha)\right)\\ +\frac{1}{\beta}\nabla f(Y(t_k))=0, \end{aligned} \tag{16}$$

where we used $\varepsilon(t_{k+1})-\varepsilon(t_k)=\dot{\varepsilon}(t_k)\sqrt{\alpha}+\frac{1}{2}\ddot{\varepsilon}(t_k)(\sqrt{\alpha})^2+\mathcal{O}((\sqrt{\alpha})^3)$. If we consider the limit of (16), when $\alpha\to 0$, we then obtain the low-resolution representation for the TM algorithm as

$$\ddot{\varepsilon}(t_k)+2\sqrt{\mu}\,\dot{\varepsilon}(t_k)+\nabla f(Y(t_k))=0, \tag{17}$$

where we used $\beta=\frac{1-\sqrt{\mu\alpha}}{1+\sqrt{\mu\alpha}}$. The low-resolution representation in (17) is exactly the same as the low-resolution ODE obtained for the NAG and heavy ball methods in [24]. Therefore the low-resolution ODE fails to distinguish the TM method from the NAG and heavy ball methods. Next, we derive a high-resolution ODE that captures the characteristics of the TM method, i.e., shows a faster convergence in comparison to the NAG and heavy ball methods.

### A. High-resolution ODE of TM Method

We obtain a high-resolution ODE for the TM method by ignoring $\mathcal{O}(\alpha)$ terms but keeping $\sqrt{\alpha}$ in (16), which results in

$$\frac{1+\beta}{2\beta}\ddot{\varepsilon}(t_k)+\frac{1-\beta}{\beta\sqrt{\alpha}}\dot{\varepsilon}(t_k)+\frac{1}{\beta}\nabla f(Y(t_k))=0. \tag{18}$$

Now multiplying both sides of (18) by $\frac{2\beta}{1+\beta}$ and substituting (7) yields

$$\ddot{\varepsilon}(t_k)+2\sqrt{\mu}\dot{\varepsilon}(t_k)+\left(1+\sqrt{\mu\alpha}\right)\nabla f(Y(t_k))=0,$$

where we used $\frac{2}{1+\beta}=1+\sqrt{\mu\alpha}$. Next, we note that from (12) we have

$$\varepsilon(t_k)-\varepsilon(t_{k-1})=\dot{\varepsilon}(t_k)\sqrt{\alpha}+\mathcal{O}(\alpha). \tag{19}$$

Ignoring the $\mathcal{O}(\alpha)$ term and substituting (19) into (14b) yields $Y=\varepsilon+\sqrt{\alpha}\gamma\dot{\varepsilon}$. Let $x_k=X(t_k)$. Similarly, from (14c) we have $X=\varepsilon+\sqrt{\alpha}\delta\dot{\varepsilon}$. Thus, we obtain

$$\ddot{\varepsilon}+2\sqrt{\mu}\,\dot{\varepsilon}+(1+\sqrt{\mu\alpha})\nabla f(Y)=0, \tag{20a}$$

$$Y=\varepsilon+\sqrt{\alpha}\gamma\,\dot{\varepsilon}, \tag{20b}$$

$$X=\varepsilon+\sqrt{\alpha}\delta\,\dot{\varepsilon}. \tag{20c}$$

as a high-resolution ODE that maintains the main characteristics of the TM method with the appropriate initial conditions $\varepsilon_0$ and $Y_0$. Note that differentiating (20a) yields

$$\dddot{\varepsilon}+2\sqrt{\mu}\ddot{\varepsilon}+(1+\sqrt{\mu\alpha})\nabla^2 f(Y)\dot{Y}=0. \tag{21}$$

Substituting (20b) and its first and second derivative $\dot{Y}=\dot{\varepsilon}+\sqrt{\alpha}\gamma\ddot{\varepsilon}$ and $\ddot{Y}=\ddot{\varepsilon}+\sqrt{\alpha}\gamma\dddot{\varepsilon}$ into (21) yields the high-resolution representation of the TM method in terms of output $Y$ as

$$\begin{aligned} \ddot{Y}+2\sqrt{\mu}\dot{Y}+\gamma(1+\sqrt{\mu\alpha})\sqrt{\alpha}\nabla^2 f(Y)\dot{Y}\\ +(1+\sqrt{\mu\alpha})\nabla f(Y)=0. \end{aligned} \tag{22}$$

In what follows, we use (22) to analyze the stability and convergence of the ODE representation of the TM method

in (20) and compare its rate of convergence to that of the high-resolution ODE representation of the NAG method given in [24] as

$$\ddot{Y}+2\sqrt{M}\dot{Y}+\sqrt{s}\nabla^2 f(Y)\dot{Y}+(1+\sqrt{Ms})\nabla f(Y)=0, \quad (23)$$

where $s=\frac{1}{L}$. One can think of $s$ as the equivalent of $\alpha$ in the TM method (5), i.e., it is the step-size multiplying the gradient term. In comparing the TM method to the NAG method, it is interesting to recall (10). It is important to note that the main difference between the NAG method given in (23) and the TM methods in (22) is in the coefficient multiplying the gradient correction term $\nabla^2 f(Y)\dot{Y}$. Even though it is not discussed in [24], it is worth mentioning that by introducing an appropriate intermediate variable similar to (20b), one can write the NAG method in an equivalent form that does not require $\nabla^2 f(Y)$. In the ODE representation of the TM and NAG algorithms we also refer to the parameters $\alpha$ and $s$ as stepsize.

## IV. Convergence Analysis

In this section, we analyze the stability of (20) and obtain an estimate on its convergence rate. Before we begin the analysis, note the equilibrium point of (20a), where $\ddot{\varepsilon}=\dot{\varepsilon}=0$. It follows from (20b) and (20c) that $\varepsilon_{\text{eq}}=Y_{\text{eq}}=X_{\text{eq}}$. As a result, at the equilibrium point, from (20a) we obtain $\nabla f(\varepsilon_{\text{eq}})=0$, and thereby $\nabla f(Y_{\text{eq}})=\nabla f(X_{\text{eq}})=0$. Consequently, by virtue Lemma II.1 we obtain

$$X_{\text{eq}}=Y_{\text{eq}}=\varepsilon_{\text{eq}}=\mathsf{x}^\star. \quad (24)$$

### A. Analysis via IQC

We note that algorithm (20) can be cast as an LTI system

$$\dot{\xi}(t)=A\xi(t)+Bq(t), \quad (25a)$$

$$Y(t)=C\xi(t)+Dq(t), \quad (25b)$$

with state $\xi(t)=[\dot{\varepsilon}(t)\quad \varepsilon(t)]^\top\in\mathbb{R}^{2n}$, input $q(t)=\nabla f(Y(t))\in\mathbb{R}^n$, and output $Y(t)\in\mathbb{R}^n$, where $A=\begin{bmatrix}-2\sqrt{\mu} & 0\\ 1 & 0\end{bmatrix}\otimes I_n$, $B=\begin{bmatrix}-(1+\sqrt{\mu\alpha})\\ 0\end{bmatrix}\otimes I_n$, $C=\begin{bmatrix}\sqrt{\alpha}\gamma & 1\end{bmatrix}\otimes I_n$, $D=0_{n\times n}$. Given (25) representation, the IQC analysis provides a convenient framework to assess the stability of continuous-time TM (20) and obtain an estimate on its convergence rate. When $f\in\mathcal{S}_{M,L}$, [28] shows that the nonlinear map $q(t)=\nabla f(Y)$ satisfies the so-called point-wise IQC condition cast as

$$\begin{bmatrix}Y-Y_{\text{eq}}\\ \nabla f(Y)-\nabla f(Y_{\text{eq}})\end{bmatrix}^\top Q_f\begin{bmatrix}Y-Y_{\text{eq}}\\ \nabla f(Y)-\nabla f(Y_{\text{eq}})\end{bmatrix}\succeq 0, \quad (26)$$

where $Q_f=\left(\begin{smallmatrix}-2ML & L+M\\ L+M & -2\end{smallmatrix}\right)\otimes I_n$. Recall from (24) that $Y_{\text{eq}}=\mathsf{x}^\star$.

**Lemma IV.1** (An estimate on the rate of convergence of $Y$ in (20) using an IQC based solution)**.** Given the LTI representation of the continuous-time TM in (25) satisfies the point-wise IQC condition (26), the exponential convergence rate of $\|Y(t)-\mathsf{x}^\star\|$ to zero is $p_{\text{IQC}}$ if

$$\begin{bmatrix}A^\top P+PA+p_{\text{IQC}}P & PB\\ B^\top P & 0\end{bmatrix}+\sigma\begin{bmatrix}C^\top & 0\\ D^\top & I\end{bmatrix}Q_f\begin{bmatrix}C & D\\ 0 & I\end{bmatrix}\preceq 0 \quad (27)$$

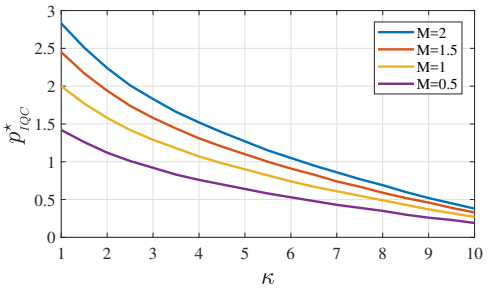



Fig. 2: Convergence rate of the TM method given by the IQC method of Lemma IV.1.

is feasible for some $\sigma\in\mathbb{R}_{\geq 0}$, $p_{\text{IQC}}\in\mathbb{R}_{>0}$, $P\succ 0$, $P\in\mathbb{R}^{n\times n}$. A tighter estimate $p^\star_{\text{IQC}}$ on the rate of convergence can be obtained by maximizing $p_{\text{IQC}}$ subject to (27). □

Given (26), the proof of Lemma IV.1 follows form standard IQC stability results [29]. Figure 2 shows the convergence rate $p^\star_{\text{IQC}}$ the IQC method of Lemma IV.1 for various values of $M$ and $\kappa$. As we can see, the IQC approach shows that similar to the gradient descent method, the rate of convergence of the TM method also decreases as the condition number $\kappa$ increases. We should mention here though that the IQC approach offers a sufficient condition for stability and convergence analysis, which is not guaranteed to yield a solution for every value of $M$ and $L$.

### B. Lyapunov Analysis

In this section, we establish the exponential stability of the continuous-time TM algorithm (20) and give an estimate on its rate of convergence using a framework regardless of the value of $M\in\mathbb{R}_{>0}$ and $L\in\mathbb{R}_{>0}$.

**Theorem IV.1** (Stability and convergence analysis the ODE TM)**.** The following two statements hold for the optimization problem (3) and the algorithm (20):

(i) For $\alpha,\beta,\gamma,\delta\in\mathbb{R}_{>0}$, $\beta\neq 1$, starting from any initial condition $\varepsilon(0),\dot{\varepsilon}(0)\in\mathbb{R}^n$ the trajectories of $t\mapsto\varepsilon$, $t\mapsto X$ and $t\mapsto Y$ converge exponentially fast to $\mathsf{x}^\star$, the minimizer of (3). Moreover, $f(Y)-f(\mathsf{x}^\star)$ vanishes exponentially with a rate no worse than $p^\star$ where

$$p^\star=\max_{\phi\in\mathbb{R}_{>0}} p(\phi), \quad (28)$$

$$p(\phi)=\min\left\{\frac{\sqrt{\mu}}{2},\frac{3L}{4\kappa(1+\phi)\sqrt{\mu}},\frac{1}{\gamma\sqrt{\alpha}(1+\frac{1}{\phi})},\frac{4\sqrt{\mu}}{3+\frac{2}{\phi}}\right\}.$$

(ii) If $\alpha,\beta,\gamma,\delta>0$ is set to the parameters of the TM method in (6) and the algorithm is initialized at $\varepsilon(0)=Y_0-\frac{\alpha\gamma^2(1+\sqrt{\mu\alpha})\nabla f(Y_0)}{(1-2\gamma\sqrt{\mu\alpha})}$ and $\dot{\varepsilon}(0)=\frac{\sqrt{\alpha}\gamma(1+\sqrt{\mu\alpha})\nabla f(Y_0)}{(1-2\gamma\sqrt{\mu\alpha})}$, where $Y_0=Y(0)\in\mathbb{R}^n$, then the trajectory $t\mapsto Y$ of (20) satisfies

$$f(Y(t))-f(\mathsf{x}^\star)\leq\frac{1.5\|Y_0-\mathsf{x}^\star\|^2}{\alpha}\,\mathrm{e}^{-p^\star_{\text{TM}}t},\quad t\in\mathbb{R}_{\geq 0},$$

where $p^\star_{\text{TM}}$ is $p^\star$ evaluated at $\mu$ given by (8), and $\alpha$ and $\gamma$ of the TM method. □

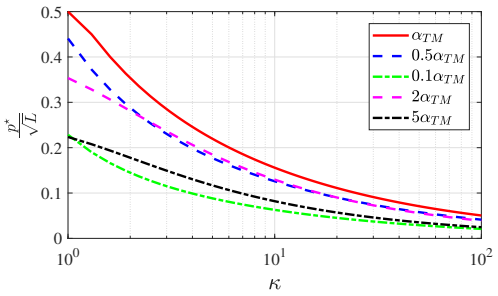



Fig. 3: $p^\star/\sqrt{L}$ vs. $\kappa$, where $p^\star$ is given in (28). Here, we use $\beta$ and $\gamma$ of the TM method but implement different values for $\alpha$. $\alpha_{TM}$ corresponds to the $\alpha$ of the TM method.

The proof of Theorem IV.1 relies on studying convergence of (22) using radially unbounded Lyapunov function candidate $V(t) = (1+\sqrt{\mu\alpha})(f(Y)-f(\mathrm{x}^\star)) + \frac{1}{4}\|\dot{Y}\|^2 + \frac{1}{4}\|\dot{Y} + 2\sqrt{\mu}(Y-\mathrm{x}^\star) + \gamma(1+\sqrt{\mu\alpha})\sqrt{\alpha}\nabla f(Y)\|^2$. Due to space limitation the details are provided in [30]. Theorem IV.1 shows that (20) has robustness to deviations from the TM parameters. But, an interesting observation about our rate of convergence analysis is that our simulation study of the rate $p$ in Theorem IV.1 indicates that the best rate is obtained when we use $\alpha, \beta, \gamma$ of the TM method given in (6), see Fig. 3 for some example scenarios.

Next, we note that the rate of convergence established for the ODE representation of the NAG method (23) in [24] is $\frac{\sqrt{M}}{4}$. Our analysis show that a tighter bound of

$$p^\star_{\text{NAG}} = \max_{\phi\in\mathbb{R}_{>0}} p_{\text{NAG}}(\phi), \tag{29}$$

$$p_{\text{NAG}}(\phi) = \min\left\{\frac{\sqrt{L}}{2\sqrt{\kappa}}, \frac{3\sqrt{L}}{4\sqrt{\kappa}(1+\phi)}, \frac{\sqrt{L}}{\sqrt{\kappa}(1+\frac{1}{\phi})}, \frac{4\sqrt{L}}{\sqrt{\kappa}(3+\frac{2}{\phi})}\right\}.$$

can be obtained for the convergence rate $p^\star_{\text{NAG}}$ of the ODE NAG method; for details see [30]. Given $\frac{L}{\kappa} = M$, we can write $p_{\text{NAG}}(\phi) = \min\left\{\frac{1}{2}, \frac{3}{4(1+\phi)}, \frac{1}{(1+\frac{1}{\phi})}, \frac{4}{(3+\frac{2}{\phi})}\right\}\sqrt{M}$. Figure 4 shows how each of the four elements varies with $\phi$ and the optimal $\phi$ for which the minimum among the four elements is at its maximum. As can be seen and also shown analytically $p^\star_{\text{NAG}} = \frac{3}{7}\sqrt{M}$ is attained at $\phi^\star = \frac{3}{4} = 0.75$. Thus, $p^\star_{\text{NAG}}$ is a tighter bound than $\frac{\sqrt{M}}{4}$ that is established in [24] as the rate of convergence for the ODE NAG method. On the other hand, Fig. 5 compares $\frac{p_{\text{TM}}}{\sqrt{L}}$ with $\frac{p_{\text{NAG}}}{\sqrt{L}}$ at different values of $\kappa$. As we can see the TM method attains a better convergence rate than the NAG method. In comparing the rate of convergences of the TM and NAG methods, it is worth to remember (8) and (9). It is also interesting to note that similar to the gradient descent method, the rate of convergence of the TM and the NAG methods decreases as $\kappa$ increases. Finally note that $\phi^\star$ corresponding to $p^\star_{\text{TM}}$ can be obtained as $\phi^\star = \frac{9L-16\mu\kappa+\sqrt{256(\mu\kappa)^2+96\mu kL+81L^2}}{32\mu k}$.

## V. Simulation Results

Let the cost function in (3) be given by $f(x) = \frac{x^2}{2\log(2+x^2)} - x$. For this cost, we have $M = 0.038$ and $L = 1.443$. Thus,

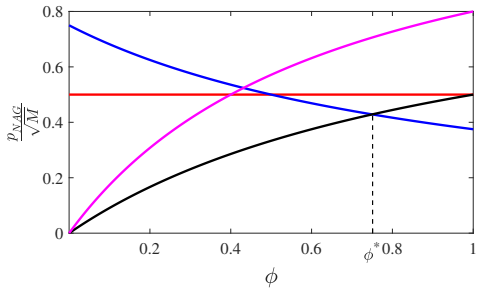



Fig. 4: Variation of the elements of $\left\{\frac{1}{2}, \frac{3}{4(1+\phi)}, \frac{1}{(1+\frac{1}{\phi})}, \frac{4}{(3+\frac{2}{\phi})}\right\}$ with $\phi$. $\frac{p^\star_{\text{NAG}}}{\sqrt{M}} = \frac{3}{7} = 0.4286$ is attained at $\phi^\star = 0.75$.

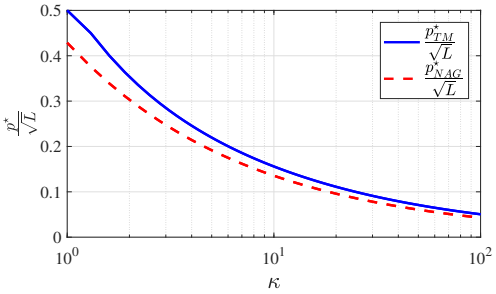
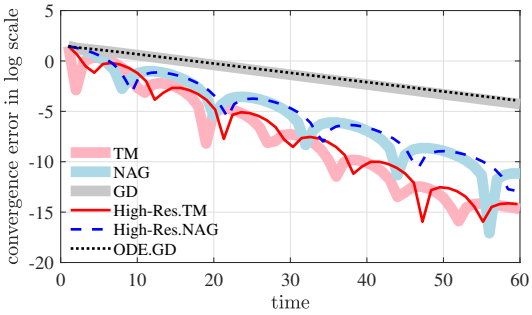



Fig. 5: Rate of convergence of the ODE representations of the TM and the NAG methods at different values of $\kappa$.

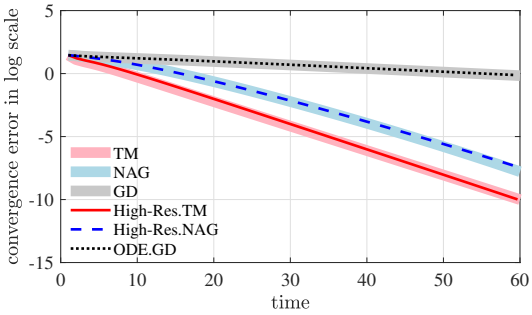
$\kappa = 37.713$. Figure 6(a) shows the convergence error for the TM, NAG, gradient descent with stepsize $1/L$ (GD), high-resolution ODE representations of TM (20) and NAG (23) methods, and continuous-time gradient descent (ODE GD) algorithms. Figure 6(b) shows the same plot when a smaller stepsize is used for all the algorithms. As we can see in these plots, the high-resolution ODE representation of the TM algorithm closely captures the characteristics of the discrete-time TM. Moreover, we can see from the plots that for both cases, the TM algorithm converges faster than the gradient descent and the NAG methods. We can also see that using a smaller stepsize removes the oscillatory behavior that we see in the trajectories of the TM and NAG methods however as expected and predicted by our analysis using a smaller stepsize results in a slower convergence.

## VI. Conclusion

In this paper, we presented the ODE representation of the the TM method, which is considered as the fastest first-order optimization method for strongly convex functions. We showed that to obtain an accurate continuous-time representation of the TM method, we need to use a higher-resolution limiting ODE representation. We also presented a Lyapunov analysis to prove the exponential convergence of the developed continuous-time model of the TM method. We compared the rate of this ODE model of the TM with that of the Nesterov method and showed that the Lyapunov analysis also confirms that the TM method has a faster convergence than the Nesterov method. We also discussed how an IQC




(a) When the parameters of the algorithms are set to their respective exact values




(b) When the parameters of the algorithms are set to their respective exact values except for stepsizes which are scaled down by a factor of 0.3

Fig. 6: Convergence error for the TM, NAG and gradient descent (GD) algorithms.

approach can be used to obtain an estimate on the convergence rate of the ODE representation of the TM method. We validated our theoretical results through several numerical simulations. Since control theoretic tools in continuous-domain generally provide a convenient framework for design and analysis of algorithms, our future work includes first devising a distributed version of the continuous-time TM method that can be used for distributed optimizations. Then, our objective is to discretize this algorithm to obtain an iterative solution that can be implemented over networks with wireless communication.

## REFERENCES


[1] D. Bertsimas, *Nonlinear Programming: 3rd Edition*. Athena Scientific, 2019.

[2] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. Springer US, 2016.

[3] B. Polyak, "Some methods of speeding up the convergence of iteration methods," USSR *Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[4] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, "Global convergence of the heavy-ball method for convex optimization,," *in 2015 European Control Conference*, vol. 17, pp. 310–315, July 2015.

[5] Y. Nesterov, "Accelerating the cubic regularization of newtons method on convex problems," *Mathematical Programming*, vol. 112, no. 1, pp. 159–181, 2008.

[6] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization, Springer US, 2013.

[7] B. Van Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions," *IEEE Control Systems Letters*, vol. 2, pp. 49–54, Jan 2018.

[8] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and nonlinear programming*. Stanford University Press, 1958.

[9] U. Helmke and J. Moore, "Optimization and dynamical systems," *Proceedings of the IEEE*, vol. 84, p. 907, Jun 1996.

[10] J. Schropp and I. Singer, "A dynamical systems approach to constrained minimization," *Numerical Functional Analysis and Optimization*, vol. 21, no. 3-4, pp. 537–551, 2000.

[11] H. T. Jongen and O. Stein, "Constrained global optimization: Adaptive gradient flows," in *Frontiers in Global Optimization* (C. A. Floudas and P. Pardalos, eds.), (Boston, MA), pp. 223–236, Springer US, 2004.

[12] V. Shikhman and O. Stein, "Constrained optimization: Projected gradient flows," *Journal of Optimization Theory and Applications*, vol. 140, pp. 117–130, Jan 2009.

[13] U. Helmke, R. Brockett, and J. Moore, *Optimization and Dynamical Systems*. Communications and Control Engineering, Springer London, 2012.

[14] X. Cheng and P. L. Bartlett, "Convergence of langevin mcmc in kl-divergence," *Proceedings of Machine Learning Research*, no. 83, pp. 186–211, 2018.

[15] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, "Sampling can be faster than optimization," *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 20881–20885, 2019.

[16] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *IEEE Conf. on Decision and Control*, (FL, USA), 2011.

[17] J. Lu and C. Tang, "Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case," *IEEE Transactions on Automatic Control*, vol. 57, no. 9, pp. 2348–2354, 2012.

[18] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2014.

[19] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-Raphson consensus for distributed convex optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 4, pp. 994 – 1009, 2015.

[20] S. S. Kia, "Distributed optimal in-network resource allocation algorithm design via a control theoretic approach," *Systems & Control Letters*, vol. 107, pp. 49–57, 2017.

[21] W. Su, S. Boyd, and E. J. Candès, "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[22] A. C. Wilson, B. Recht, and M. I. Jordan, "A Lyapunov Analysis of Momentum Methods in Optimization," *arXiv e-prints,* `arXiv:1611.02635`, Nov 2016.

[23] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *PNAS*, vol. 113, no. 47, pp. E7351–E7358, 2016.

[24] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, "Understanding the Acceleration Phenomenon via High-Resolution Differential Equations," *arXiv e-prints,* `arXiv:1810.08907`, Oct 2018.

[25] G. França, D. P. Robinson, and R. Vidal, "Gradient Flows and Accelerated Proximal Splitting Methods," *arXiv e-prints,* `arXiv:1908.00865`, Aug 2019.

[26] W. Krichene, A. Bayen, and P. L. Bartlett, "Adaptive averaging in accelerated descent dynamics," in *Advances in Neural Information Processing Systems 29*, pp. 2991–2999, Curran Associates, Inc., 2016.

[27] X. Zhou, "On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient," *arXiv e-prints,* `arXiv:1803.06573`, Mar 2018.

[28] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[29] Z. E. Nelson and E. Mallada, "An integral quadratic constraint framework for real-time steady-state optimization of linear time-invariant systems," in *American Control Conference*, (Milwaukee, Wisconsin), pp. 597–603, Jun 2018.

[30] B. Sun, J. George, and S. Kia, "High-Resolution Modeling of the Fastest First-Order Optimization Method for Strongly Convex Functions," *arXiv e-prints,* `arXiv:2008.11199`, Aug. 2020.