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

On the transient growth of Nesterov's accelerated method for strongly convex optimization problems

Samantha Samuelson, Hesameddin Mohammadi, and Mihailo R. Jovanović

Abstract— **Compared to standard descent-based algorithms, accelerated first-order methods for strongly convex smooth optimization problems may exhibit large transient responses. For quadratic problems, this phenomenon arises from the presence of non-normal dynamics and the modes that yield an algebraic growth in early iterations. In this paper, we employ the framework of integral quadratic constraints to examine the transient response of Nesterov's accelerated method. We prove that a bound on the largest value of the Euclidean distance between the optimization variable and the global minimizer is proportional to the square root of the condition number. For problems with large condition numbers we demonstrate tightness of this bound up to constant factors, thereby establishing the merits of our approach.**

Index Terms— **Convex optimization, Gradient descent, Integral quadratic constraints, Nesterov's accelerated method, Nonnormal dynamics, Transient growth.**

I. Introduction

First-order optimization algorithms are commonly used in a large variety of applications including statistics, signal and image processing, optimal control, and machine learning [1]–[6]. Accelerated first-order algorithms achieve a faster rate of convergence when compared to gradient descent, while preserving low per-iteration complexity. A large body of literature exists which investigates convergence results of accelerated algorithms for various stepsizes and acceleration parameters, including [7]–[10]. Many recent works also study the robustness of these algorithms under uncertainty [11]–[17], determining that acceleration in first-order algorithms increases sensitivity to uncertainty in gradient evaluation. In addition to sensitivity to gradient uncertainty, accelerated first-order algorithms can exhibit aberrant transient growth in early iterations. Unlike gradient descent, which is a contraction mapping for strongly convex problems with suitable stepsize [18], accelerated algorithms are not monotonically decreasing. Accordingly, it is important to study not only the convergence behavior in asymptotic regimes, but also the transient behavior in non-asymptotic regimes. This is particularly important for applications such as ADMM and distributed optimization methods, which may use only a few iterations of an accelerated algorithm, making analysis of their behavior in early iterations crucial.

In this paper, we provide upper bounds on the possible transient growth of Nesterov's accelerated method for the class of general strongly convex problems. Our goal is to establish an analytic bound for the possible magnitude of the optimization error. We build on our previous work on the strongly convex quadratic problems [19] where we developed tight analytical bounds on the magnitude and iteration number of the peak of the transient response and showed that these bounds scale with the square root of the condition number. Similar results with extensions to the Wasserstein distance have been recently reported in [20]. Here, we extend our previous study to the case of general strongly convex problems. We use linear matrix inequalities to create a bound on the Euclidean distance between the optimization variable and the global minimizer, which holds for all iterations. We determine that, as in the case of quadratic problems, the bound scales with the square root of the condition number.

Previous work on non-asymptotic bounds on Nesterov's accelerated method includes [21], which presents bounds on the objective error in terms of condition number. In contrast to our work, this reference introduces an assumption on the initial conditions. In addition, while reference [22] presents numeric bounds on the value of the estimated optimizer, we provide analytical bounds on the non-asymptotic value of the estimated optimizer.

The rest of the paper is structured as follows. In Section II, we provide background on accelerated first-order algorithms. In section III, we present our main result. In Section IV, we discuss the theoretical lower bounds by focusing on the worst-case transient growth for quadratic problems. In Section V, we use IQCs to derive a bound on the transient response in terms of the condition number for general problem. We offer thoughts and future direction in Section VI.

II. Motivation and background

Consider the unconstrained optimization problem

$$\underset{x}{\text{minimize}} \;\; f(x) \tag{1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function with an L-Lipschitz continuous gradient ∇f. This problem can be solved using first-order methods including gradient descent

$$x^{t+1} \;=\; x^t \;-\; \alpha \nabla f(x^t) \tag{2a}$$

and Nesterov's accelerated method

$$\begin{aligned} x^{t+2} \;=\; x^{t+1} \;+\; \beta(x^{t+1} - x^t) \;-\; \\ \alpha \nabla f\big(x^{t+1} + \beta(x^{t+1} - x^t)\big) \end{aligned} \tag{2b}$$

where t is the iteration index, $\alpha > 0$ is the stepsize, and $\beta \in (0,1)$ is the extrapolation parameter.

Let us denote the set of functions f that are m-strongly convex and L-smooth by $\mathcal{F}_m^L$; $f \in \mathcal{F}_m^L$ means that $f(x) -$

Financial support from the National Science Foundation under awards ECCS-1708906 and ECCS-1809833 is gratefully acknowledged.

The authors are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. E-mails: ({sasamuel, hesamedm, mihailo}@usc.edu).

$\frac{m}{2}\|x\|^2$ is convex and that the gradient ∇f is L-Lipschitz continuous. In particular, for a twice continuously differentiable function f with the Hessian matrix $\nabla^2 f$, we have

$$f \in \mathcal{F}_m^L \;\Leftrightarrow\; mI \preceq \nabla^2 f(x) \preceq LI, \quad \forall\, x \in \mathbb{R}^n.$$

For $f \in \mathcal{F}_m^L$, the parameters α and β can be selected such that gradient descent and Nesterov's accelerated method converge to the global minimum $x^\star$ of (1) at a linear rate,

$$\|x^t - x^\star\| \;\leq\; c\,\rho^t\,\|x^0 - x^\star\| \tag{3}$$

for all t and some positive scalar $c > 0$, where $\|\cdot\|$ is the Euclidean norm. Table I provides the conventional values of these parameters and the corresponding guaranteed convergence rates [23]. Gradient descent achieves the convergence rate $\rho_{\rm gd} = \sqrt{1-2/(\kappa+1)}$, where $\kappa := L/m$ is the condition number associated with $\mathcal{F}_m^L$. Thus, for reaching the accuracy level $\|x^t - x^*\| \leq \epsilon$, gradient descent requires $O(\kappa \log(1/\epsilon))$ iterations. This dependence on the condition number can be significantly improved using Nesterov's accelerated method which achieves the rate

$$\rho_{\rm na} \;=\; \sqrt{1 - 1/\sqrt{\kappa}} \;\leq\; 1 - 1/(2\sqrt{\kappa})$$

thereby, requiring $O(\sqrt{\kappa}\log(1/\epsilon))$ iterations. This convergence rate is *orderwise optimal* in the sense that for any first-order algorithm, there are problem instances $f \in \mathcal{F}_m^L$ for which $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ iterations are necessary [23, Theorem 2.1.13].

In spite of a significant improvement in the rate of convergence, acceleration may deteriorate performance on finite time intervals and lead to large transient responses. In particular, the constant c in (3) may become significantly larger than 1 for Nesterov's accelerated algorithm whereas $c = 1$ for gradient descent because of its contractive property for strongly convex problems.

In this paper, we study the transient growth of Nesterov's accelerated method by quantifying the largest ratio of the error at all iterations to the initial error for all $f \in \mathcal{F}_m^L$,

$$J \;:=\; \sup_{\{f\in\mathcal{F}_m^L;\; t\in\mathbb{N};\; z^0,z^1\in\mathbb{R}^n\}} \frac{\|z^t\|}{\sqrt{\|z^0\|^2 + \|z^1\|^2}} \tag{4}$$

where $z^t := x^t - x^\star$ determines the error to the optimal solution $x^\star$ at the iteration t. We establish analytical upper and lower bounds on J and show that both of them grow linearly with $\sqrt{\kappa}$. Our bounds are *almost tight* in the sense that they differ only by a factor of 4.3 for $\kappa \gg 1$. In his seminal work [23], Nesterov showed the upper bound $\sqrt{\kappa+1}$ on J, under the assumption that the initial condition is confined to the subspace $x^0 = x^1$. In our analysis, we remove this assumption and establish that similar trends hold for general initial conditions.

Notation: We write $g = \Omega(h)$ (or, equivalently, $h = O(g)$) to denote the existence of positive constants c_i such that, for any $y > c_2$, the functions g and h that map $\mathbb{R}$ to $\mathbb{R}$ satisfy $g(y) \geq c_1 h(y)$. We write $g = \Theta(h)$, or more informally $g \approx h$, if both $g = \Omega(h)$ and $g = O(h)$.

Method	Parameters	Linear rate
Gradient	$\alpha = \frac{1}{L}$	$\rho = \sqrt{1 - \frac{2}{\kappa+1}}$
Nesterov	$\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$\rho = \sqrt{1 - \frac{1}{\sqrt{\kappa}}}$

TABLE I: Conventional values of parameters and the corresponding rates for $f \in \mathcal{F}_m^L$, $\|x^t - x^\star\| \leq c\,\rho^t\,\|x^0 - x^\star\|$, where $\kappa := L/m$ and $c > 0$ is a constant [23, Theorems 2.1.15, 2.2.1].

III. Main result

Herein, we present the main result of the paper.

Theorem 1: For Nesterov's accelerated method with parameters provided in Table I, the largest ratio of the error at all iterations to the initial error, defined in (4), satisfies

$$\frac{\sqrt{2}\,(\kappa - 2\sqrt{\kappa} - 1)}{\mathrm{e}\,\sqrt{\kappa}} \;\leq\; J \;\leq\; \sqrt{5\kappa}$$

where $\kappa = L/m$ is the condition number associated with the set of L-smooth m-strongly convex objective functions $\mathcal{F}_m^L$.

While we only consider parameters provide by Table I, our framework and proof technique for Theorem 1 is general enough to handle other values of parameters. We observe that both upper and lower bounds grow linearly with $\sqrt{\kappa}$. This scaling demonstrates a potential drawback of using Nesterov's accelerated method in scenarios with limited time budget. To investigate the tightness of these bounds, we examine the ratio of the upper and lower bounds which indicates how far the upper bound may be from the true maximum transient possible for all $f \in \mathcal{F}_m^L$. As $\kappa \to \infty$, this quantity converges to $\mathrm{e}\sqrt{5}/\sqrt{2} \approx 4.3$.

The rest of the paper is devoted to proving Theorem 1. Our lower bounds are obtained by restricting the set $\mathcal{F}_m^L$ to the class of strongly convex quadratic problems, which we describe next. In Section V, we utilize linear matrix inequalities to establish upper bounds for general strongly convex problems with Lipschitz continuous objective functions.

IV. A lower bound: Quadratic problems

For quadratic objective functions,

$$f(x) \;=\; \frac{1}{2}\,(x - x^\star)^T Q (x - x^\star)$$

where $Q = Q^T \succ 0$ is a positive definite matrix, the algorithms in (2) are linear time-invariant (LTI) systems and the dynamics of the error $z^t := x^t - x^\star$ can be described by state-space models of the form

$$\psi^{t+1} \;=\; A\,\psi^t \tag{5a}$$

$$z^t \;=\; C\,\psi^t. \tag{5b}$$

Here, ψ^t is the state, and A, C are constant matrices of appropriate dimensions. For example, we can select $\psi^t = z^t$

for gradient descent and $\psi^t := \begin{bmatrix} (z^t)^T & (z^{t+1})^T \end{bmatrix}^T$ for Nesterov's accelerated method. These choices of the state variable ψ^t, respectively, correspond to $A = I - \alpha Q$, $C = I$ for gradient descent, and

$$A \;=\; \begin{bmatrix} 0 & I \\ -\beta(I-\alpha Q) & (1+\beta)(I-\alpha Q) \end{bmatrix},\; C \;=\; \begin{bmatrix} I & 0 \end{bmatrix}$$

for Nesterov's accelerated method.

The linearity of the underlying dynamics allows us to fully characterize the transient growth in terms of the eigenvalues/vectors of the matrix A. In particular, the response of LTI system (5) with an initial state ψ^0 is determined by

$$z^t \;=\; CA^t\psi^0 \;=\; \Phi(t)\psi^0.$$

Here, A^t is the state-transition matrix of system (5), i.e., the tth power of the matrix A, and $\Phi(t) := CA^t$ is the mapping from the initial condition ψ^0 to the performance output $z^t = x^t - x^\star$. The rate of convergence of system (5) is determined by the spectral radius of the matrix A, i.e.,

$$\rho(A) \;:=\; \max_i \; |\mu_i(A)|$$

where μ_i are the eigenvalues of the matrix A.

The fastest convergence rate is achieved by optimizing the spectral radius over α and β. The optimal values of these parameters and the corresponding convergence rates $\rho < 1$ are provided in Table II. This commonly used metric for evaluating performance of optimization algorithms only determines the asymptotic behavior and it does not provide insights into transient responses. To study the performance of Nesterov's accelerated method on finite-time intervals, we can quantify the transient growth, i.e., the worst-case ratio of the energy of z^t to the energy of the initial condition ψ^0. For quadratic objective functions, this quantity is determined by the largest singular value of the matrix $\Phi(t)$

$$\sup_{\psi^0 \neq 0} \; \frac{\|z^t\|}{\|\psi^0\|} \;=\; \sup_{\|\psi^0\|=1} \; \|\Phi(t)\psi^0\| \;=\; \sigma_{\max}(\Phi(t)).$$

While for general matrices A and C the analytical calculation of $\sigma_{\max}(\Phi(t))$ is challenging, for the first-order algorithms in (2), the underlying structure allows us to use a unitary transformation to express the dynamics via n decoupled dynamical systems that are parameterized by α, β, and the eigenvalues of the matrix Q. This decomposition can then be used to obtain an analytical expression for both A^t and $\Phi(t)$. This approach was recently utilized in [19] to prove the following result.

Theorem 2: For Nesterov's accelerated algorithm with the parameters provided in Tables I and II, and the rate of convergence $\rho \in [1/\mathrm{e}, 1)$, we have

$$-\frac{\sqrt{2}\rho}{\mathrm{e}\log\rho} \;\leq\; \max_t \; \sigma_{\max}(\Phi(t)) \;\leq\; -\frac{\sqrt{2}}{\mathrm{e}\,\rho\log\rho}.$$

Combining this result with the explicit value of ρ provided in Table II, we obtain upper and lower bounds in terms of the condition number

$$\frac{\sqrt{3\kappa+1}-2}{\sqrt{2}\,\mathrm{e}} \;\leq\; \max_t \; \sigma_{\max}(\Phi(t)) \;\leq\; \frac{\sqrt{3\kappa+1}}{\sqrt{2}\,\mathrm{e}}.$$

We can also obtain similar upper and lower bounds on $\max_t \sigma_{\max}(\Phi(t))$ for the values of α and β in Table I. In this case, the rate of convergence for quadratic problems is given by $\rho = 1 - 1/\sqrt{\kappa}$; see [24] for a proof. This yields

$$\frac{\sqrt{2}(\kappa - 2\sqrt{\kappa} - 1)}{\mathrm{e}\,\sqrt{\kappa}} \;\leq\; \max_t \; \sigma_{\max}(\Phi(t)) \;\leq\; \frac{\sqrt{2}\kappa}{\mathrm{e}\,(\sqrt{\kappa}-1)}$$

and, in both cases, the upper and lower bounds scale as $\sqrt{\kappa}$ for $\kappa \gg 1$.

Method	Parameter choice		Rate
Gradient	$\alpha = \frac{2}{L+m}$		$\frac{\kappa-1}{\kappa+1}$
Nesterov	$\alpha = \frac{4}{3L+m}$	$\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$\frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}}$

TABLE II: Optimal parameters and the linear convergence rate bounds for m-strongly convex quadratic objective functions with L-Lipschitz gradients and $\kappa := L/m$.

Since the quadratic objective functions with $Q \succ 0$ are strongly convex and smooth, the lower bound established in this section holds for J defined in (4) that corresponds to the set $\mathcal{F}_m^L$. This completes the proof of our lower bound presented in Theorem 1. However, the above derived upper bound does not carry over to the general case. We next describe a method based on linear matrix inequalities that we utilize to determine the upper bound on J in (4).

V. An upper bound: Linear matrix inequalities

For the class $\mathcal{F}_m^L$ of m-strongly convex objective functions with L-Lipschitz continuous gradients, algorithms in (2) are invariant under translation, i.e., if we let $\tilde{x} := x - \bar{x}$ and $g(\tilde{x}) := f(\tilde{x} + \bar{x})$, then (2b), for example, satisfies

$$\tilde{x}^{t+2} \;=\; \tilde{x}^{t+1} \;+\; \beta(\tilde{x}^{t+1} - \tilde{x}^t) \;-\; \alpha\nabla g\big(\tilde{x}^{t+1} \;+\; \beta(\tilde{x}^{t+1} - \tilde{x}^t)\big).$$

Thus, in what follows, without loss of generality, we assume that $x^\star = 0$ is the unique minimizer of (1) and $f(x^\star) = 0$.

We present a general framework based on Linear Matrix Inequalities (LMIs) that allows us to obtain non-asymptotic bounds on the error. This framework combines certain Integral Quadratic Constraints (IQCs) [25] and Lyapunov functions of the form

$$V(\psi) \;=\; \psi^T X \psi \;+\; f(C\psi) \qquad (6)$$

which consists of a standard quadratic function of the state ψ, where X is a positive semidefinite matrix and the objective function evaluated at $C\psi$. The IQC framework provides a convex control-theoretic approach to analyzing optimization algorithms [24] and it was employed to study convergence and robustness of the first-order algorithms [12]–[15], [22], [26], [27]. This type of generalized Lyapunov functions was

introduced in [22], [28] and used to study convergence of optimization algorithms for non-strongly convex problems. For Nesterov's accelerated algorithm (2b), we next demonstrate that this approach yields *orderwise-tight* analytical upper bounds on the norm of the error in the iterates.

For $f \in \mathcal{F}_m^L$, the nonlinear mapping $\Delta\colon \mathbb{R}^n \to \mathbb{R}^n$

$$\Delta(y) \;:=\; \nabla f(y) \;-\; m\,y$$

satisfies the quadratic inequality [24, Lemma 6]

$$\begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix} \;\geq\; 0 \qquad (7)$$

for all $y,\, y_0 \in \mathbb{R}^n$, where the matrix Π is given by

$$\Pi \;:=\; \left[\begin{array}{cc} 0 & (L\,-\,m)I \\ (L\,-\,m)I & -2I \end{array}\right]. \qquad (8)$$

Nesterov's accelerated algorithm (2) admits a time-invariant state-space form

$$\begin{aligned} \psi^{t+1} \;&=\; A\,\psi^t \;+\; B_u u^t \\ \left[\begin{array}{c} z^t \\ y^t \end{array}\right] \;&=\; \left[\begin{array}{c} C_z \\ C_y \end{array}\right] \psi^t \\ u^t \;&=\; \Delta(y^t) \end{aligned} \qquad (9a)$$

that contains a feedback interconnection of linear and nonlinear components. Figure 1 illustrates the block diagram of system (9a), where ψ^t is the state, z^t is the performance output, and u^t is the output of the nonlinear term $\Delta(y^t)$. In particular, if we let

$$\psi^t \;:=\; \left[\begin{array}{c} x^t \\ x^{t+1} \end{array}\right],\; z^t \;:=\; x^t,\; y^t \;:=\; -\beta x^t + (1+\beta)x^{t+1}$$

and define the corresponding matrices as

$$\begin{aligned} A \;&=\; \left[\begin{array}{cc} 0 & I \\ -\beta(1-\alpha\,m)I & (1+\beta)(1-\alpha\,m)I \end{array}\right], \\ B_u \;&=\; \left[\begin{array}{c} 0 \\ -\alpha\,I \end{array}\right], \\ C_z \;&=\; \left[\begin{array}{cc} I & 0 \end{array}\right], \quad C_y \;=\; \left[\begin{array}{cc} -\beta\,I & (1+\beta)I \end{array}\right] \end{aligned} \qquad (9b)$$

then (9a) represents Nesterov's method (2b).

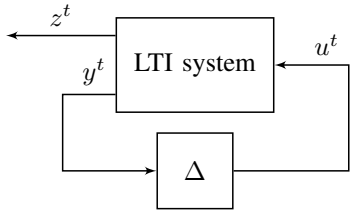

Fig. 1: Block diagram of system (9a).

We next demonstrate how property (7) of the nonlinear mapping Δ in conjunction with a suitable Lyapunov function of the form (6) allow us to obtain upper bounds on $\|z^t\|$.

Lemma 1: Let the matrix $M(m, L; \alpha, \beta)$ be defined as

$$M \;:=\; N_1^T \left[\begin{array}{cc} L\,I & I \\ I & 0 \end{array}\right] N_1 \;+\; N_2^T \left[\begin{array}{cc} -m\,I & I \\ I & 0 \end{array}\right] N_2$$

where

$$\begin{aligned} N_1 \;&:=\; \left[\begin{array}{ccc} \alpha\,m\,\beta\,I & -\alpha\,m(1+\beta)\,I & -\alpha\,I \\ -m\,\beta\,I & m(1+\beta)\,I & I \end{array}\right], \\ N_2 \;&:=\; \left[\begin{array}{ccc} -\beta\,I & \beta\,I & 0 \\ -m\,\beta\,I & m(1+\beta)\,I & I \end{array}\right]. \end{aligned}$$

Consider state-space model (9) for algorithm (2b) and let Π be given by (8). Then, for any positive semidefinite matrix X and nonnegative scalars λ_1 and λ_2 that satisfy

$$\begin{aligned} &\left[\begin{array}{cc} A^T X\,A - X & A^T X\,B_u \\ B_u^T\,X\,A & B_u^T\,X\,B_u \end{array}\right] \;+ \\ &\qquad \lambda_1 \left[\begin{array}{cc} C_y^T & 0 \\ 0 & I \end{array}\right] \Pi \left[\begin{array}{cc} C_y & 0 \\ 0 & I \end{array}\right] \;+\; \lambda_2\,M \;\preceq\; 0 \end{aligned} \qquad (10)$$

for all $t \geq 1$ the transient growth is upper bounded by

$$\|x^t\|^2 \;\leq\; \frac{\lambda_{\max}(X)\|x^0\|^2 + (\lambda_{\max}(X) + L\lambda_2)\,\|x^1\|^2}{\lambda_{\min}(X) + m\lambda_2}. \qquad (11)$$

Proof: See Appendix A. ∎

Lemma 1 exploits the Lyapunv function candidate $V(\psi) := \psi^T X \psi + \lambda_2 f([\,0\;\; I\,]\psi)$ to show that the state of the algorithm ψ^t is confined in the sublevel set

$$\{\psi \;\in\; \mathbb{R}^{2n} \mid V(\psi) \;\leq\; V(\psi^0)\}.$$

associated with $V(\psi^0)$. To provide insight into this result, let us consider the special case with $\lambda_2 = 0$. In this case, the resulting upper bound on $\|x^t\|^2$ is determined by the condition number of the matrix X, which is a standard result for quadratic Lyapunov functions [19]. The bound in Lemma 1 also leads to the following corollary.

Corollary 1: In the setting of Lemma 1, for any $t \geq 1$,

$$\|x^t\|^2 \;\leq\; (\mathrm{cond}(X) \;+\; \kappa)\,\left(\|x^0\|^2 \;+\; \|x^1\|^2\right).$$

We can solve LMI (10) in Lemma 1 both numerically and analytically to establish an upper bound on J defined in (4). In particular, expression (11) allows us to write

$$\begin{aligned} J^2 \;&\leq \\ &\sup_{\|x^0\|^2 + \|x^1\|^2 \,\leq\, 1} \frac{\lambda_{\max}(X)\|x^0\|^2 + (\lambda_{\max}(X) + L\lambda_2)\,\|x^1\|^2}{\lambda_{\min}(X) + m\lambda_2} \\ &\qquad\qquad = \;\frac{\lambda_{\max}(X) \;+\; L\lambda_2}{\lambda_{\min}(X) \;+\; m\lambda_2}. \end{aligned}$$

Thus, we can obtain an upper bound on J by solving

$$\begin{aligned} \underset{X,\,\lambda_1,\,\lambda_2}{\text{minimize}} \quad & \frac{\lambda_{\max}(X) \;+\; L\lambda_2}{\lambda_{\min}(X) \;+\; m\lambda_2} \\ \text{subject to} \quad & \text{LMI (10)},\; X \;\succeq\; 0,\; \lambda_1 \;\geq\; 0,\; \lambda_2 \;\geq\; 0. \end{aligned} \qquad (12)$$

Even though the objective function in (12) is nonconvex, we can still find a feasible point which allows us to establish

$\|x^t\| = \Omega(\sqrt{\kappa})$, which is of the same order as theoretical lower bound in Section IV.

Theorem 3: Nesterov's accelerated method with parameters provided in Table I satisfies

$$\|x^t\|^2 \;\leq\; 4\kappa\|x^0\|^2 \;+\; 5\kappa\|x^1\|^2. \tag{13}$$

Proof: See Appendix B. ∎

Our upper bound in Theorem 1 is a direct consequence of Theorem 3. In particular, the largest coefficient on the right hand side of Eq. (13) provides an upper bound on the quantity of interest J, which completes the proof of Theorem 1.

VI. Concluding remarks

We have examined the transient response of Nesterov's accelerated method for strongly convex optimization problems. The framework of integral quadratic constraints was utilized to establish that an upper bound on the largest value of the Euclidean distance between the optimization variable and the global minimizer is proportional to the square root of the condition number. Unlike gradient descent which is always contractive, our analysis reveals that there are always quadratic problem instances for which the accelerated method generates a large transient response that meets our theoretical upper bound up to a constant factor. Future directions include extending our analysis to nonsmooth accelerated methods and devising algorithms that balance acceleration with quality of transient responses.

Appendix

A. Proof of Lemma 1

In order to prove Lemma 1, we present a technical lemma which along the lines of results of [22] provides us with an upper bound on the difference between the objective value at two consecutive iterations.

Lemma 2: Let $f \in \mathcal{F}_m^L$ and $\kappa := L/m$. Then, Nesterov's accelerated method, with the notation introduced in Section V, satisfies

$$f(x^{t+2}) \;-\; f(x^{t+1}) \;\leq\; \frac{1}{2}\begin{bmatrix} \psi^t \\ u^t \end{bmatrix}^T M \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}$$

where N_1 and N_2 are defined in Lemma 1.

Proof: For any $f \in \mathcal{F}_m^L$, by the L-Lipschitz continuity of the gradient ∇f, we have

$$f(x^{t+2}) \;-\; f(y^t) \;\leq\; \frac{1}{2}\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix} \tag{14a}$$

and by the m-strong convexity of f, we have

$$f(y^t) \;-\; f(x^{t+1}) \;\leq\; \frac{1}{2}\begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix}^T \begin{bmatrix} -m\,I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix}. \tag{14b}$$

Moreover, the state and output equations in (9) lead to

$$\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix} \;=\; N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}, \tag{15a}$$

$$\begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix} \;=\; N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}. \tag{15b}$$

Adding inequalities (14a) and (14b) in conjunction with (15) completes the proof. ∎

We are now ready to prove Lemma 1. It is straightforward to verify that

$$\begin{bmatrix} y^t \\ u^t \end{bmatrix} \;=\; \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \eta^t$$

where $\eta^t := [\,(\psi^t)^T\;(u^t)^T\,]^T$. Combining this equation and inequality (7) yields the quadratic inequality

$$(\eta^t)^T \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \eta^t \;\geq\; 0. \tag{16}$$

We can now pre and post multiply LMI (10) by $(\eta^t)^T$ and η^t, respectively, to obtain

$$\begin{aligned} 0 \;\geq\; & (\eta^t)^T \begin{bmatrix} A^T X\,A - X & A^T X\,B_u \\ B_u^T X\,A & B_u^T X\,B_u \end{bmatrix} \eta^t + \lambda_2 (\eta^t)^T M \eta^t \\ & +\; \lambda_1\,(\eta^t)^T \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \eta^t \\ \geq\; & (\eta^t)^T \begin{bmatrix} A^T X\,A - X & A^T X\,B_u \\ B_u^T X\,A & B_u^T X\,B_u \end{bmatrix} \eta^t + \lambda_2 (\eta^t)^T M \eta^t \end{aligned}$$

where the second inequality follows from (16) . Rearranging terms yields

$$0 \;\leq\; \hat{V}(\psi^t) \;-\; \hat{V}(\psi^{t+1}) \;-\; \lambda_2 (\eta^t)^T M\,\eta^t \tag{17}$$

where $\hat{V}(\psi) := \psi^T X \psi$ is a positive semidefinite function. Also, Lemma 2 implies

$$-(\eta^t)^T M\,\eta^t \;\leq\; 2\left(f(x^{t+1}) \;-\; f(x^{t+2})\right). \tag{18}$$

Combining inequalities (17) and (18) yields

$$\hat{V}(\psi^{t+1}) \;+\; 2\lambda_2 f(x^{t+2}) \;\leq\; \hat{V}(\psi^t) \;+\; 2\lambda_2 f(x^{t+1}).$$

Thus, using induction, we obtain the uniform upper bound

$$\hat{V}(\psi^t) \;+\; 2\lambda_2 f(x^{t+1}) \;\leq\; \hat{V}(\psi^0) \;+\; 2\lambda_2 f(x^1). \tag{19}$$

We can bound the function $\hat{V}$ using the smallest and largest eigenvalues of X,

$$\lambda_{\min}(X)\|\psi\|^2 \;\leq\; \hat{V}(\psi) \;\leq\; \lambda_{\max}\|\psi\|^2. \tag{20a}$$

We can also bound the objective function with parameters of Lipschitz continuity L and strong convexity m,

$$m\|x\|^2 \;\leq\; 2f(x) \;\leq\; L\|x\|^2. \tag{20b}$$

Finally, combining (19) and (20) yields

$$\begin{aligned} \lambda_{\min}(X)\|\psi^t\|^2 \;+\; m\,\lambda_2\|x^{t+1}\|^2 & \\ \leq\; \lambda_{\max}(X)\|\psi^0\|^2 \;+\; L\,\lambda_2\|x^1\|^2. & \end{aligned}$$

Rearranging terms and noting that $\|x^{t+1}\| \leq \|\psi^t\|$ completes the proof.

B. Proof of Theorem 3

The proof works by finding a feasible solution for λ_1, λ_2 and X in terms of problem condition number κ. Let us define

$$\begin{aligned} X &:= \begin{bmatrix} x_1 I & x_0 I \\ x_0 I & x_2 I \end{bmatrix} \\ x_0 &:= -L\lambda_2,\ x_1 := L\lambda_2 p_1(\sqrt{\kappa}) \\ x_2 &:= L^2\lambda_1 + L\lambda_2,\ \lambda_1 := \lambda_2 p_2(\sqrt{\kappa})/m \end{aligned} \tag{21}$$

where the rational functions

$$p_1(r) := \frac{2r^3 - 4r^2 + 3r - 1}{2r^3}$$

$$p_2(r) := \frac{4r^2 - 3r + 1}{2r^5 - 4r^4 + 4r^2 - 2r}$$

are nonnegative for $r \geq 1$. This yields

$$\det(X) = \frac{(L\lambda_2)^2(12\kappa^{\frac{3}{2}} - 17\kappa + 9\kappa^{\frac{1}{2}} - 2)}{4\kappa^{\frac{3}{2}}(\kappa^{\frac{1}{2}} - 1)^2(\kappa^{\frac{1}{2}} + 1)}$$

It is straightforward to verify that both x_1 and $\det(X)$ are nonnegative for $\kappa \geq 1$. Thus, by Schur complement, we obtain $X \succeq 0$ and the left-hand-side of LMI becomes

$$-\lambda_1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & m^2(2\kappa - 1)I & 0 \\ 0 & 0 & I \end{bmatrix} \preceq 0.$$

Thus, the choice of $(\lambda_1, \lambda_2, X)$ in (21) satisfies the conditions of Lemma 1. Hence, we obtain

$$\begin{aligned} \|x^t\|^2 &\leq \frac{\lambda_{\max}(X)\|x^0\|^2 + (\lambda_{\max}(X) + L\lambda_2)\,\|x^1\|^2}{\lambda_{\min}(X) + m\lambda_2} \\ &\leq w\|x^0\|^2 + (w + \kappa)\|x^1\|^2 \end{aligned} \tag{22}$$

where $w := (|x_0| + |x_1| + |x_2|)/(m\lambda_2)$. Here, the first inequality follows from Lemma 1 and the second inequality is obtained using $\lambda_{\min}(X) \geq 0$, $\lambda_{\max}(X) \leq |x_0| + |x_1| + |x_2|$. Using (21), we can upper bound w as

$$\begin{aligned} w &= \frac{|x_0| + |x_1| + |x_2|}{m\lambda_2} \\ &= \kappa\left(2 + p_1(\sqrt{\kappa}) + \kappa\, p_2(\sqrt{\kappa})\right) \leq 4\kappa \end{aligned} \tag{23}$$

The above inequality uses the fact that $p_1(\sqrt{\kappa}) \leq 1$ and $\kappa p_2(\sqrt{\kappa}) \leq 1$ for all $\kappa \geq 4$. Combining (22) and (23) competes the proof.

REFERENCES

[1] L. Bottou and Y. Le Cun, "On-line learning for very large data sets," *Appl. Stoch. Models Bus. Ind.*, vol. 21, no. 2, pp. 137–151, 2005.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[3] Y. Nesterov, "Gradient methods for minimizing composite objective functions," *Math. Program.*, vol. 140, no. 1, pp. 125–161, 2013.

[4] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, 2016.

[5] L. Bottou, F. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[6] S. Hassan-Moghaddam and M. R. Jovanović, "Topology design for stochastically-forced consensus networks," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1075–1086, September 2018.

[7] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, 2013, pp. 1139–1147.

[8] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. & Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[9] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," in *Dokl. Akad. Nauk SSSR*, vol. 27, 1983, pp. 543–547.

[10] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[11] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program.*, vol. 146, no. 1-2, pp. 37–75, 2014.

[12] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *Proc. ICML*, vol. 70, 2017, pp. 1549–1557.

[13] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems," in *Proceedings of the 57th IEEE Conference on Decision and Control*, 2018, pp. 5753–5758.

[14] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Performance of noisy Nesterov's accelerated method for strongly convex optimization problems," in *Proceedings of the 2019 American Control Conference*, Philadelphia, PA, 2019, pp. 3426–3431.

[15] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," *IEEE Trans. Automat. Control*, 2020, doi: 10.1109/TAC.2020.3008297; also arXiv:1905.11011.

[16] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimization algorithms: An integral quadratic constraint approach," 2019, arXiv:1905.00279.

[17] J. I. Poveda and N. Li, "Robust hybrid zero-order optimization algorithms with acceleration via averaging in continuous time," 2019, arXiv:1909.00265.

[18] D. P. Bertsekas, *Convex optimization algorithms*. Athena Scientific, 2015.

[19] S. Samuelson, H. Mohammadi, and M. R. Jovanović, "Transient growth of accelerated first-order methods," in *Proceedings of the 2020 American Control Conference*, Denver, CO, 2020, pp. 2858–2863.

[20] B. Can, M. Gurbuzbalaban, and L. Zhu, "Accelerated linear convergence of stochastic momentum methods in Wasserstein distances," in *Proceedings of Machine Learning Research*, vol. 97, 2019, pp. 891–901.

[21] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004, vol. 87.

[22] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2654–2689, 2018.

[23] Y. Nesterov, *Lectures on convex optimization*. Springer Optimization and Its Applications, 2018, vol. 137.

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

[25] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, 1997.

[26] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *Proceedings of the 2018 American Control Conference*, 2018, pp. 1376–1381.

[27] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Automat. Control*, vol. 64, no. 7, pp. 2861–2868, 2019.

[28] B. T. Polyak and P. Shcherbakov, "Lyapunov functions: An optimization theory perspective," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7456–7461, 2017.