Transient growth of accelerated optimization algorithms

Hesameddin Mohammadi, Samantha Samuelson, and Mihailo R. Jovanović

Abstract—Optimization algorithms are increasingly being used in applications with limited time budgets. In many real-time and embedded scenarios, only a few iterations can be performed and traditional convergence metrics cannot be used to evaluate performance in these non-asymptotic regimes. In this paper, we examine the transient behavior of accelerated first-order optimization algorithms. For convex quadratic problems, we employ tools from linear systems theory to show that transient growth arises from the presence of non-normal dynamics. We identify the existence of modes that yield an algebraic growth in early iterations and quantify the transient excursion from the optimal solution caused by these modes. For strongly convex smooth optimization problems, we utilize the theory of integral quadratic constraints (IQCs) to establish an upper bound on the magnitude of the transient response of Nesterov's accelerated algorithm. We show that both the Euclidean distance between the optimization variable and the global minimizer and the rise time to the transient peak are proportional to the square root of the condition number of the problem. Finally, for problems with large condition numbers, we demonstrate tightness of the bounds that we derive up to constant factors.

Index Terms—Convex optimization, first-order optimization algorithms, heavy-ball method, integral quadratic constraints, Nesterov's accelerated method, non-asymptotic behavior, non-normal matrices, transient growth.

I. INTRODUCTION

First-order optimization algorithms are widely used in a variety of fields including statistics, signal and image processing, control, and machine learning [1]–[8]. Acceleration is often utilized as a means to achieve a faster rate of convergence relative to gradient descent while maintaining low per-iteration complexity. There is a vast literature focusing on the convergence properties of accelerated algorithms for different stepsize rules and acceleration parameters, including [9]–[12]. There is also a growing body of work which investigates robustness of accelerated algorithms to various types of uncertainty [13]–[19]. These studies demonstrate that acceleration increases sensitivity to uncertainty in gradient evaluation.

In addition to deterioration of robustness in the face of uncertainty, asymptotically stable accelerated algorithms may also exhibit undesirable transient behavior [20]. This is in contrast to gradient descent which is a contraction for strongly convex problems with suitable stepsize [21]. In real-time optimization and in applications with limited time budgets, the transient growth can limit the appeal of accelerated methods. In addition, first-order algorithms are often used as a building block in multi-stage optimization including ADMM [22] and distributed optimization methods [23]. In these settings, at each stage we can perform only a few iterations of first-order

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

updates on primal or dual variables and transient growth can have a detrimental impact on the performance of the entire algorithm. This motivates an in-depth study of the behavior of accelerated first-order methods in non-asymptotic regimes.

It is widely recognized that large transients may arise from the presence of resonant modal interactions and non-normality of linear dynamical generators [24]. Even in the absence of unstable modes, these can induce large transient responses, significantly amplify exogenous disturbances, and trigger departure from nominal operating conditions. For example, in fluid dynamics, such mechanisms can initiate departure from stable laminar flows and trigger transition to turbulence [25], [26].

In this paper, we consider the optimization problem

$$\operatorname*{minimize}_{x} \; f(x) \qquad (1)$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex and smooth function, and we focus on a class of accelerated first-order algorithms

$$x^{t+2} = x^{t+1} + \beta(x^{t+1} - x^t) - \alpha\nabla f(x^{t+1} + \gamma(x^{t+1} - x^t)) \quad (2)$$

where t is the iteration index, α is the stepsize, and β is the momentum parameter. In particular, we are interested in Nesterov's accelerated and Polyak's heavy-ball methods that correspond to $\gamma = \beta$ and $\gamma = 0$, respectively. While these algorithms have faster convergence rates compared to the standard gradient descent ($\gamma = \beta = 0$), they may suffer from large transient responses; see Fig. 1 for an illustration. To quantify the transient behavior, we examine the ratio of the largest error in the optimization variable to the initial error.

For convex quadratic problems, (2) can be cast as a linear time-invariant (LTI) system for which modal analysis of the state-transition matrix can be performed. For both accelerated algorithms, we identify non-normal modes that create large transient growth, derive analytical expressions for the state-transition matrices, and establish bounds on the transient response in terms of the convergence rate and the iteration number. We show that both the peak value of the transient response and the rise time to this value increase with the square
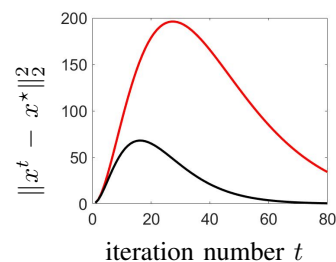

Fig. 1. Error in the optimization variable for Polyak's heavy-ball (black) and Nesterov's (red) algorithms with the parameters that optimize the convergence rate for a strongly convex quadratic problem with the condition number 10^3 and a unit norm initial condition with $x^0 \neq x^\star$.

root of the condition number of the problem. Moreover, for general strongly convex problems, we combine a Lyapunov-based approach with the theory of IQCs to establish an upper bound on the transient response of Nesterov's accelerated algorithm. As for quadratic problems, we demonstrate that this bound scales with the square root of the condition number.

This work builds on our recent conference papers [27], [28]. In contrast to these preliminary results, we provide a comprehensive analysis of transient growth of accelerated algorithms for convex quadratic problems and address the important issue of eliminating transient growth of Nesterov's accelerated algorithm with the proper choice of initial conditions. Adaptive restarting, which was introduced in [20] to address the oscillatory behavior of Nesterov's accelerated method, provides heuristics for improving transient responses. In [29], the transient growth of second-order systems was studied and a framework for establishing upper bounds was introduced, with a focus on real eigenvalues. The result was applied to the heavy-ball method but was not applicable to quadratic problems in which the dynamical generator may have complex eigenvalues. We account for complex eigenvalues and conduct a thorough analysis for Nesterov's accelerated algorithm as well. Furthermore, for convex quadratic problems, we provide tight upper and lower bounds on transient responses in terms of the condition number and identify the initial condition that induces the largest transient response. Similar results with extensions to the Wasserstein distance have been recently reported in [30]. Previous work on non-asymptotic bounds for Nesterov's accelerated algorithm includes [31], where bounds on the objective error in terms of the condition number were provided. However, in contrast to our work, this result introduces a restriction on the initial conditions. Finally, while [32] presents computational bounds we develop analytical bounds on the non-asymptotic value of the estimated optimizer.

II. CONVEX QUADRATIC PROBLEMS

In this section, we examine transient responses of accelerated algorithms for convex quadratic objective functions,

$$f(x) = \frac{1}{2} x^T Q x \tag{3a}$$

where $Q = Q^T \succeq 0$ is a positive semi-definite matrix. In what follows, we first bring (2) into a standard LTI state-space form and then utilize appropriate coordinate transformation to decompose the dynamics into decoupled subsystems. Using this decomposition, we provide analytical expressions for the state-transition matrix and establish sharp bounds on the transient growth and the location of the transient peak for accelerated algorithms. We also examine the influence of initial conditions on transient responses and relegate the proofs to Appendix A.

A. LTI formulation

The matrix Q admits an eigenvalue decomposition, $Q = V \Lambda V^T$, where Λ is the diagonal matrix of eigenvalues with

$$\begin{aligned} L &:= \lambda_1 \geq \cdots \geq \lambda_r =: m > 0 \\ \lambda_i &= 0 \text{ for } i = r+1, \ldots, n \end{aligned} \tag{3b}$$

Method	Optimal parameters		Linear rate ρ
Nesterov	$\alpha = \frac{4}{3L+m}$	$\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$1 - \frac{2}{\sqrt{3\kappa+1}}$
Polyak	$\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$	$\beta = \frac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)^2}$	$1 - \frac{2}{\sqrt{\kappa}+1}$

TABLE I
PARAMETERS THAT PROVIDE OPTIMAL CONVERGENCE RATES FOR A
CONVEX QUADRATIC OBJECTIVE FUNCTION (3) WITH $\kappa := L/m$.

and V is the unitary matrix of the corresponding eigenvectors. We define the condition number $\kappa := L/m$ as the ratio of the largest and smallest non-zero eigenvalues of the matrix Q. For f in (3a), we have $\nabla f(x) = Qx$, and the change of variables $\hat{x}^t := V^T x^t$ brings dynamics (2) to

$$\hat{x}^{t+2} = (I - \alpha\Lambda)\hat{x}^{t+1} + (\beta I - \gamma\alpha\Lambda)(\hat{x}^{t+1} - \hat{x}^t). \tag{4}$$

This system can be represented via n decoupled second-order subsystems of the form,

$$\hat{\psi}_i^{t+1} = A_i \hat{\psi}_i^t, \quad \hat{x}_i^t = C_i \hat{\psi}_i^t \tag{5a}$$

where $\hat{x}_i^t$ is the ith element of the vector $\hat{x}^t \in \mathbb{R}^n$, $\hat{\psi}_i^t := \begin{bmatrix} \hat{x}_i^t & \hat{x}_i^{t+1} \end{bmatrix}^T$, $C_i := \begin{bmatrix} 1 & 0 \end{bmatrix}$, and

$$A_i = \begin{bmatrix} 0 & 1 \\ -(\beta - \gamma\alpha\lambda_i) & 1 - \alpha\lambda_i + (\beta - \gamma\alpha\lambda_i) \end{bmatrix}. \tag{5b}$$

B. Linear convergence of accelerated algorithms

The minimizers of (3a) are determined by the null space of the matrix Q, $x^\star \in \mathcal{N}(Q)$. The constant parameters α and β can be selected to provide stability of subsystems in (5) for all $\lambda_i \in [m, L]$, and guarantee convergence of $\hat{x}_i^t$ to $\hat{x}_i^\star := 0$ with a linear rate determined by the spectral radius $\rho(A_i) < 1$. On the other hand, for $i = r+1, \ldots, n$ the eigenvalues of A_i are β and 1. In this case, the solution to (5) is given by

$$\hat{x}_i^t = \frac{1 - \beta^t}{1 - \beta} (\hat{x}_i^1 - \hat{x}_0^0) + \hat{x}_i^0 \tag{6a}$$

and the steady-state limit of $\hat{x}_i^t$,

$$\hat{x}_i^\star := \frac{1}{1 - \beta} (\hat{x}_i^1 - \hat{x}_i^0) + \hat{x}_i^0 \tag{6b}$$

is achieved with a linear rate $\beta < 1$. Thus, the iterates of (2) converge to the optimal solution $x^\star = V\hat{x}^\star \in \mathcal{N}(Q)$ with a linear rate $\rho < 1$ and Table I provides the parameters α and β that optimize the convergence rate [33, Proposition 1].

C. Transient growth of accelerated algorithms

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. This is in contrast to gradient descent which is a contraction [21]. At any t, we are interested in the worst-case ratio of the two norm of the error of the optimization variable $z^t := x^t - x^\star$ to the two norm of the initial condition $\psi^0 - \psi^\star = \begin{bmatrix} (z^0)^T & (z^1)^T \end{bmatrix}^T$,

$$J^2(t) := \sup_{\psi^0 \neq \psi^\star} \frac{\|x^t - x^\star\|_2^2}{\|\psi^0 - \psi^\star\|_2^2}. \tag{7}$$

Proposition 1: For accelerated algorithms applied to convex quadratic problems, $J(t)$ in (7) is determined by

$$J^2(t) = \max\left\{\max_{i \leq r} \|C_i A_i^t\|_2^2, \ \beta^{2t}/(1 + \beta^2)\right\}. \quad (8)$$

Proof: Since V is unitary and dynamics (5) that govern the evolution of each $\hat{x}_i^t$ are decoupled, $J(t)$ is determined by

$$J^2(t) = \max_i \sup_{\hat{\psi}_i^0 \neq \hat{\psi}_i^\star} \frac{(\hat{x}_i^t - \hat{x}_i^\star)^2}{\|\hat{\psi}_i^0 - \hat{\psi}_i^\star\|_2^2} \quad (9)$$

where $\hat{\psi}_i^\star := \begin{bmatrix} \hat{x}_i^\star & \hat{x}_i^\star \end{bmatrix}^T$. Furthermore, the mapping from $\hat{\psi}_i^0 - \hat{\psi}_i^\star$ to $\hat{x}_i^t - \hat{x}_i^\star$ is given by $\Phi_i(t) := C_i A_i^t$ where the state-transition matrix A_i^t is determined by the tth power of A_i,

$$\hat{x}_i^t - \hat{x}_i^\star = C_i A_i^t(\hat{\psi}_i^0 - \hat{\psi}_i^\star) =: \Phi_i(t)(\hat{\psi}_i^0 - \hat{\psi}_i^\star). \quad (10)$$

For $\lambda_i \neq 0$, $\hat{\psi}_i^0 - \hat{\psi}_i^\star = \hat{\psi}_i^0$ is an arbitrary vector in $\mathbb{R}^2$. Thus,

$$\sup_{\hat{\psi}_i^0 \neq \hat{\psi}_i^\star} \frac{(\hat{x}_i^t - \hat{x}_i^\star)^2}{\|\hat{\psi}_i^0 - \hat{\psi}_i^\star\|_2^2} = \|C_i A_i^t\|_2^2, \quad i = 1, \ldots, r. \quad (11)$$

This expression, however, does not hold when $\lambda_i = 0$ in (5) because $\psi_i^0 - \psi_i^\star$ is restricted to a line in $\mathbb{R}^2$. Namely, from (6),

$$\hat{x}_i^t - \hat{x}_i^\star = \frac{-\beta^t}{1 - \beta}(\hat{x}_i^1 - \hat{x}_0^0)$$
$$\psi_i^0 - \psi_i^\star = \begin{bmatrix} \hat{x}_i^0 - \hat{x}_i^\star \\ \hat{x}_i^1 - \hat{x}_i^\star \end{bmatrix} = \frac{-(\hat{x}_i^1 - \hat{x}_i^0)}{1 - \beta}\begin{bmatrix} 1 \\ \beta \end{bmatrix} \quad (12)$$

which, for any initial condition with $\hat{x}_i^0 \neq \hat{x}_i^1$, leads to

$$\frac{(\hat{x}_i^t - \hat{x}_i^\star)^2}{\|\psi_i^0 - \psi_i^\star\|_2^2} = \frac{\beta^{2t}}{1 + \beta^2}, \quad i = r + 1, \ldots, n. \quad (13)$$

Finally, substitution of (11) and (13) to (9) yields (8). ∎

D. Analytical expressions for transient response

We next derive analytical expressions for the state-transition matrix A_i^t and the response matrix $\Phi_i(t) = C_i A_i^t$ in (5).

Lemma 1: Let μ_1 and μ_2 be the eigenvalues of the matrix

$$M = \begin{bmatrix} 0 & 1 \\ a & b \end{bmatrix}$$

and let t be a positive integer. For $\mu_1 \neq \mu_2$,

$$M^t = \frac{1}{\mu_2 - \mu_1}\begin{bmatrix} \mu_1\mu_2(\mu_1^{t-1} - \mu_2^{t-1}) & \mu_2^t - \mu_1^t \\ \mu_1\mu_2(\mu_1^t - \mu_2^t) & \mu_2^{t+1} - \mu_1^{t+1} \end{bmatrix}.$$

Moreover, for $\mu := \mu_1 = \mu_2$, the matrix M^t is determined by

$$M^t = \begin{bmatrix} (1 - t)\mu^t & t\mu^{t-1} \\ -t\mu^{t+1} & (t + 1)\mu^t \end{bmatrix}. \quad (14)$$

Lemma 1 with $M = A_i$ determines explicit expressions for A_i^t. These expressions allow us to establish a bound on the norm of the response for each decoupled subsystem (5). In Lemma 2, we provide a tight upper bound on $\|C_i A_i^t\|_2^2$ for each t in terms of the spectral radius of the matrix A_i.

Lemma 2: The matrix M in Lemma 1 satisfies

$$\|\begin{bmatrix} 1 & 0 \end{bmatrix} M^t\|_2^2 \leq (t - 1)^2\rho^{2t} + t^2\rho^{2t-2} \quad (15)$$

where ρ is the spectral radius of M. Moreover, (15) becomes equality if M has repeated eigenvalues.

Remark 1: For Nesterov's accelerated algorithm with the parameters that optimize the convergence rate (cf. Table I), the matrix $\hat{A}_r$, which corresponds to the smallest non-zero eigenvalue of Q, $\lambda_r = m$, has an eigenvalue $1 - 2/\sqrt{3\kappa + 1}$ with algebraic multiplicity two and incomplete sets of eigenvectors. Similarly, for both $\lambda_1 = L$ and $\lambda_r = m$, $\hat{A}_1$ and $\hat{A}_r$ for the heavy-ball method with the parameters provided in Table I have repeated eigenvalues which are, respectively, given by $(1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$ and $-(1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$.

We next use Lemma 2 with $M = A_i$ to establish an analytical expression for $J(t)$.

Theorem 1: For accelerated algorithms applied to convex quadratic problems, $J(t)$ in (7) satisfies

$$J^2(t) \leq \max\left\{(t - 1)^2\rho^{2t} + t^2\rho^{2(t-1)}, \ \beta^{2t}/(1 + \beta^2)\right\}$$

where $\rho := \max_{i \leq r} \rho(A_i)$. Moreover, for the parameters provided in Table I

$$J^2(t) = (t - 1)^2\rho^{2t} + t^2\rho^{2(t-1)}. \quad (16)$$

Theorem 1 highlights the source of disparity between the long and short term behavior of the response. While the geometric decay of ρ^t drives x^t to $x^\star$ as $t \to \infty$, early stages are dominated by the algebraic term which induces a transient growth. We next provide tight bounds on the time $t_{\max}$ at which the largest transient response takes place and the corresponding peak value $J(t_{\max})$. Even though we derive the explicit expressions for these two quantities, our tight upper and lower bounds are more informative and easier to interpret.

Theorem 2: For accelerated algorithms with the parameters provided in Table I, let $\rho \in [1/e, 1)$. Then the rise time $t_{\max} := \text{argmax}_t J(t)$ and the peak value $J(t_{\max})$ satisfy

$$-1/\log(\rho) \leq t_{\max} \leq 1 - 1/\log(\rho)$$

$$-\frac{\sqrt{2}\rho}{e\log(\rho)} \leq J(t_{\max}) \leq -\frac{\sqrt{2}}{e\,\rho\log(\rho)}.$$

For accelerated algorithms with the parameters provided in Table I, Theorem 2 can be used to determine the rise time to the peak in terms of condition number κ. We next establish that both $t_{\max}$ and $J(t_{\max})$ scale as $\sqrt{\kappa}$.

Proposition 2: For accelerated algorithms with the parameters provided in Table I, the rise time $t_{\max} := \text{argmax}_t J(t)$ and the peak value $J(t_{\max})$ satisfy

(i) Polyak's heavy-ball method with $\kappa \geq 4.69$

$$(\sqrt{\kappa} - 1)/2 \leq t_{\max} \leq (\sqrt{\kappa} + 3)/2$$

$$\frac{(\sqrt{\kappa} - 1)^2}{\sqrt{2}\,e(\sqrt{\kappa} + 1)} \leq J(t_{\max}) \leq \frac{(\sqrt{\kappa} + 1)^2}{\sqrt{2}\,e(\sqrt{\kappa} - 1)}$$

(ii) Nesterov's accelerated method with $\kappa \geq 3.01$

$$(\sqrt{3\kappa + 1} - 2)/2 \leq t_{\max} \leq (\sqrt{3\kappa + 1} + 2)/2$$

$$\frac{(\sqrt{3\kappa + 1} - 2)^2}{\sqrt{2}\,e\sqrt{3\kappa + 1}} \leq J(t_{\max}) \leq \frac{3\kappa + 1}{\sqrt{2}\,e(\sqrt{3\kappa + 1} - 2)}.$$

In Proposition 2, the lower-bounds on κ are only required to ensure that the convergence rate ρ satisfies $\rho \geq 1/e$, which
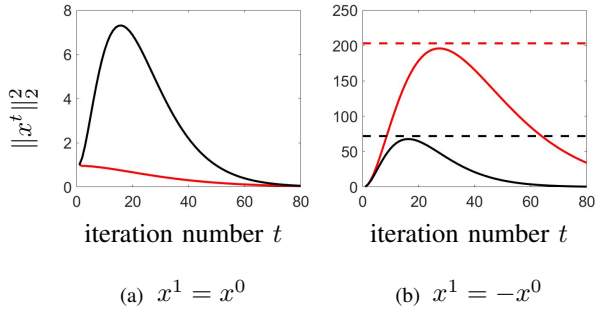
(a) $x^1 = x^0$ (b) $x^1 = -x^0$

Fig. 2. Dependence of the error in the optimization variable on the iteration number for the heavy-ball (black) and Nesterov's methods (red), as well as the peak magnitudes (dashed lines) obtained in Proposition 2 for two different initial conditions with $\|x^1\|_2 = \|x^0\|_2 = 1$.

allows us to apply Theorem 2. We also note that the upper and lower bounds on $t_{\max}$ and $J(t_{\max})$ are *tight* in the sense that their ratio converges to 1 as $\kappa \to \infty$.

E. The role of initial conditions

The accelerated algorithms need to be initialized with x^0 and $x^1 \in \mathbb{R}^n$. This provides a degree of freedom that can be used to potentially improve their transient performance. To provide insight, let us consider the quadratic problem with $Q = \mathrm{diag}\,(\kappa, 1)$. Figure 2 shows the error in the optimization variable for Polyak's and Nesterov's algorithms as well as the peak magnitudes obtained in Proposition 2 for two different types of initial conditions with $x^1 = x^0$ and $x^1 = -x^0$, respectively. For $x^1 = -x^0$, both algorithms recover their worst-case transient responses. However, for $x^1 = x^0$, Nesterov's method shows no transient growth.

Our analysis shows that large transient responses arise from the existence of non-normal modes in the matrices A_i. However, such modes do not move the entries of the state transition matrix A_i^t in arbitrary directions. For example, using Lemma 1, it is easy to verify that A_r in (5b), associated with the smallest non-zero eigenvalue $\lambda_r = m$ of Q in Nesterov's algorithm with the parameters provided by Table I has the repeated eigenvalue $\mu = 1 - 2/\sqrt{3\kappa + 1}$ and A_r^t is determined by (14) with $M = A_r$. Even though each entry of A_r^t experiences a transient growth, its row sum is determined by

$$A_r^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 + 2t/(\sqrt{3\kappa + 1} - 2) \\ 1 + 2t/\sqrt{3\kappa + 1} \end{bmatrix} (1 - 2/\sqrt{3\kappa + 1})^t$$

and entries of this vector are monotonically decaying functions of t. Furthermore, for $i < r$, it can be shown that the entries of $A_i^t \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ remain smaller than 1 for all i and t. In Theorem 3, we provide a bound on the transient response of Nesterov's method for *balanced* initial conditions with $x^1 = x^0$.

Theorem 3: For convex quadratic optimization problems, the iterates of Nesterov's accelerated method with a balanced initial condition $x^1 = x^0$ and parameters provided in Table I satisfy $\|x^t - x^\star\|_2 \leq \|x^0 - x^\star\|_2$.

Proof: See Appendix B. ∎

It is worth mentioning that the transient growth of the heavy-ball method cannot be eliminated with the use of balanced

initial conditions. To see this, we note that the matrices A_r^t and A_1^t for the heavy-ball method with parameters provided in Table I also take the form in (14) with $\mu = (1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$ and $\mu = -(1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$, respectively. In contrast to $A_r^t \begin{bmatrix} 1 & 1 \end{bmatrix}^T$, which decays monotonically,

$$A_1^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 + 2t\sqrt{\kappa}/(1 - \sqrt{\kappa}) \\ 1 + 2t\sqrt{\kappa}/(1 + \sqrt{\kappa}) \end{bmatrix} \frac{(1 - \sqrt{\kappa})^t}{(1 + \sqrt{\kappa})^t}$$

experiences transient growth. It was recently shown that an averaged version of the heavy-ball method experiences smaller peak deviation than the heavy-ball method [34]. We also note that adaptive restarting provides effective heuristics for reducing oscillatory behavior of accelerated algorithms [20].

Remark 2: For accelerated algorithms with the parameters provided in Table I, the initial condition that leads to the largest transient growth at any time τ is determined by

$$\hat{\psi}_r^0 = c \begin{bmatrix} (1 - \tau)\rho^\tau & \tau\rho^{\tau-1} \end{bmatrix}^T, \quad \hat{\psi}_i^0 = 0 \text{ for } i \neq r$$

where $c \neq 0$ and $\hat{\psi}_r^0$ is the principal right singular vector of $C_r A_r^\tau$. Thus, the largest peak $J(t_{\max})$ occurs for $\{\hat{\psi}_i^0 = 0, i \neq r\}$ and $\hat{\psi}_r^0 = c \begin{bmatrix} (1 - t_{\max})\rho^{t_{\max}} & t_{\max}\rho^{t_{\max}-1} \end{bmatrix}^T$, where tight bounds on $t_{\max}$ are established in Proposition 2.

Remark 3: For $\lambda_i = 0$ in (5), $|\hat{x}_i^t - \hat{x}_i^\star|$ decays monotonically with a linear rate β and only non-zero eigenvalues of Q contribute to the transient growth. Furthermore, for the parameters provided in Table I, our analysis shows that $J^2(t) = \max_{i \leq r} \|C_i A_i^t\|_2^2$. In what follows, we provide bounds on the largest deviation from the optimal solution for Nesterov's algorithm for general strongly convex problems.

III. GENERAL STRONGLY CONVEX PROBLEMS

In this section, we combine a Lyapunov-based approach with the theory of IQCs to provide bounds on the transient growth of Nesterov's accelerated algorithm for the class $\mathcal{F}_m^L$ of m-strongly convex and L-smooth functions. When f is not quadratic, first-order algorithms are no longer LTI systems and eigenvalue decomposition cannot be utilized to simplify analysis. Instead, to handle nonlinearity and obtain upper bounds on J in (7), we augment standard quadratic Lyapunov functions with the objective error.

For $f \in \mathcal{F}_m^L$, algorithm (2) is invariant under translation. Thus, without loss of generality, we assume that $x^\star = 0$ is the unique minimizer of (1) with $f(0) = 0$. In what follows, we present a framework based on Linear Matrix Inequalities (LMIs) that allows us to obtain time-independent bounds on the error in the optimization variable. This framework combines certain IQCs [35] with Lyapunov functions of the form

$$V(\psi) = \psi^T X \psi + \theta f(C\psi) \tag{17}$$

which consist of the objective function evaluated at $C\psi$ and a quadratic function of ψ, where X is a positive definite matrix.

The theory of IQCs provides a convex control-theoretic approach to analyzing optimization algorithms [33] and it was recently employed to study convergence and robustness of the first-order methods [14], [17], [32], [36]–[38]. The type of Lyapunov functions in (17) was introduced in [32], [39] to

study convergence for convex problems. For Nesterov's accelerated algorithm, we demonstrate that this approach provides *orderwise-tight* analytical upper bounds on $J(t)$.

Nesterov's accelerated algorithm can be viewed as a feedback interconnection of linear and nonlinear components

$$
\begin{aligned}
\psi^{t+1} &= A\,\psi^t + B\,u^t \\
y^t &= C_y\,\psi^t, \quad u^t = \Delta(y^t)
\end{aligned}
\tag{18a}
$$

where the LTI part of the system is determined by

$$
A = \begin{bmatrix} 0 & I \\ -\beta I & (1+\beta)I \end{bmatrix}
\tag{18b}
$$

$$
B = \begin{bmatrix} 0 \\ -\alpha I \end{bmatrix}, \quad C_y = \begin{bmatrix} -\beta I & (1+\beta)I \end{bmatrix}
$$

and the nonlinear mapping $\Delta \colon \mathbb{R}^n \to \mathbb{R}^n$ is $\Delta(y) := \nabla f(y)$. Moreover, the state vector ψ^t and the input y^t to Δ are determined by

$$
\psi^t := \begin{bmatrix} x^t \\ x^{t+1} \end{bmatrix}, \quad y^t := (1+\beta)x^{t+1} - \beta x^t.
\tag{18c}
$$

For smooth and strongly convex functions $f \in \mathcal{F}_m^L$, Δ satisfies the quadratic inequality [33, Lemma 6]

$$
\begin{bmatrix} y - y_0 \\ \Delta(y) - \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y - y_0 \\ \Delta(y) - \Delta(y_0) \end{bmatrix} \geq 0
\tag{19a}
$$

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

$$
\Pi := \begin{bmatrix} -2mLI & (L+m)I \\ (L+m)I & -2I \end{bmatrix}.
\tag{19b}
$$

Using $u^t := \Delta(y^t)$ and $y^t := C_y\psi^t$ and evaluating (19a) at $y = y^t$ and $y_0 = 0$ leads to,

$$
\begin{bmatrix} \psi^t \\ u^t \end{bmatrix}^T M_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \geq 0
\tag{19c}
$$

where

$$
\begin{aligned}
M_1 &:= \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} -2mLC_y^T C_y & (L+m)C_y^T \\ (L+m)C_y & -2I \end{bmatrix}.
\end{aligned}
\tag{19d}
$$

In Lemma 3, we provide an upper bound on the difference between the objective function at two consecutive iterations of Nesterov's algorithm. In combination with (19), this result allows us to utilize Lyapunov function of the form (17) to establish an upper bound on transient growth. We note that variations of this lemma have been presented in [32, Lemma 5.2] and in [17, Lemma 3].

Lemma 3: Along the solution of Nesterov's accelerated algorithm (18), the function $f \in \mathcal{F}_m^L$ with $\kappa := L/m$ satisfies

$$
f(x^{t+2}) - f(x^{t+1}) \leq \frac{1}{2} \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}^T M_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}
\tag{20a}
$$

where the matrix M_2 is given by

$$
\begin{aligned}
M_2 &:= \begin{bmatrix} -mC_2^T C_2 & C_2^T \\ C_2 & -\alpha(2-\alpha L)I \end{bmatrix} \\
C_2 &:= \begin{bmatrix} -\beta I & \beta I \end{bmatrix}.
\end{aligned}
\tag{20b}
$$

Using Lemma 3, we next demonstrate how a Lyapunov function of the form (17) with $\theta := 2\theta_2$ and $C := \begin{bmatrix} 0 & I \end{bmatrix}$ in conjunction with property (19) of the nonlinear mapping Δ can be utilized to obtain an upper bound on $\|x^t\|_2^2$.

Lemma 4: Let M_1 be given by (19d) and let M_2 be defined in Lemma 3. Then, for any positive semi-definite matrix X and nonnegative scalars θ_1 and θ_2 that satisfy

$$
W := \begin{bmatrix} A^T X A - X & A^T X B \\ B^T X A & B^T X B \end{bmatrix} + \theta_1 M_1 + \theta_2 M_2 \preceq 0
\tag{21}
$$

the transient growth of Nesterov's accelerated algorithm (18) for all $t \geq 1$ is upper bounded by

$$
\|x^t\|_2^2 \leq \frac{\lambda_{\max}(X)\|x^0\|_2^2 + (\lambda_{\max}(X) + L\theta_2)\|x^1\|_2^2}{\lambda_{\min}(X) + m\theta_2}.
\tag{22}
$$

In Lemma 4, the Lyapunov function candidate $V(\psi) := \psi^T X \psi + 2\theta_2 f(\begin{bmatrix} 0 & I \end{bmatrix}\psi)$ is used to show that the state vector ψ^t is confined within the sublevel set $\{\psi \in \mathbb{R}^{2n} \,|\, V(\psi) \leq V(\psi^0)\}$ associated with $V(\psi^0)$. We next establish an *order-wise* tight upper bound on $\|x^t\|_2$ that scales linearly with $\sqrt{\kappa}$ by finding a feasible point to LMI (21) in Lemma 4.

Theorem 4: For $f \in \mathcal{F}_m^L$ with the condition number $\kappa := L/m$, the iterates of Nesterov's accelerated algorithm (18) for any stabilizing parameters $\alpha \leq 1/L$ and $\beta < 1$ satisfy

$$
\|x^t\|_2^2 \leq \kappa \left(\frac{1+\beta^2}{\alpha\beta L}\|x^0\|_2^2 + (1 + \frac{1+\beta^2}{\alpha\beta L})\|x^1\|_2^2 \right).
\tag{23a}
$$

Furthermore, for the conventional values of parameters

$$
\alpha = 1/L, \quad \beta = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)
\tag{23b}
$$

the largest transient error, defined in (7), satisfies

$$
\frac{\sqrt{2}\,(\sqrt{\kappa} - 1)^2}{e\sqrt{\kappa}} \leq \sup_{\{t \in \mathbb{N},\, f \in \mathcal{F}_m^L\}} J(t) \leq \sqrt{3\kappa + \frac{4\kappa}{\kappa - 1}}.
\tag{23c}
$$

For balanced initial conditions, i.e., $x^1 = x^0$, Nesterov established the upper bound $\sqrt{\kappa + 1}$ on J in [12]. Theorem 4 shows that similar trends hold without restriction on initial conditions. Linear scaling of the upper and lower bounds with $\sqrt{\kappa}$ illustrates a potential drawback of using Nesterov's accelerated algorithm in applications with limited time budgets. As $\kappa \to \infty$, the ratio of these bounds converges to $e\sqrt{3/2} \approx 3.33$, thereby demonstrating that the largest transient response for all $f \in \mathcal{F}_m^L$ is within the factor of 3.33 relative to the bounds established in Theorem 4.

IV. CONCLUDING REMARKS

We have examined the impact of acceleration on transient responses of first-order optimization algorithms. Without imposing restrictions on initial conditions, we establish bounds on the largest value of the Euclidean distance between the optimization variable and the global minimizer. For convex quadratic problems, we utilize the tools from linear systems theory to fully capture transient responses and for general strongly convex problems, we employ the theory of integral quadratic constraints to establish an upper bound on transient growth. This upper bound is proportional to the square root

of the condition number and we identify quadratic problem instances for which accelerated algorithms generate transient responses which are within a constant factor of this upper bound. Future directions include extending our analysis to nonsmooth optimization problems and devising algorithms that balance acceleration with quality of transient responses.

APPENDIX

A. Proofs of Section II

We first present a technical lemma that we use in our proofs.

Lemma 5: For any $\rho \in [1/e, 1)$, $a(t) := t\rho^t$ satisfies

$$\underset{t \geq 1}{\operatorname{argmax}} \; a(t) \;=\; -1/\log(\rho), \quad \max_{t \geq 1} a(t) \;=\; -1/(e \log(\rho)).$$

Proof: Follows from the fact that $da/dt = \rho^t (1 + t \log(\rho))$ vanishes at $t = -1/\log(\rho)$. ∎

1) Proof of Lemma 1: For $\mu_1 \neq \mu_2$, the eigenvalue decomposition of M is determined by

$$M = \frac{1}{\mu_2 - \mu_1} \begin{bmatrix} 1 & 1 \\ \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \begin{bmatrix} \mu_2 & -1 \\ -\mu_1 & 1 \end{bmatrix}.$$

Computing the tth power of the diagonal matrix and multiplying throughout completes the proof for $\mu_1 \neq \mu_2$. For $\mu_1 = \mu_2 =: \mu$, M admits the Jordan canonical form

$$M = \begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix} \begin{bmatrix} \mu & 1 \\ 0 & \mu \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\mu & 1 \end{bmatrix}$$

and the proof follows from

$$\begin{bmatrix} \mu & 1 \\ 0 & \mu \end{bmatrix}^t = \begin{bmatrix} \mu^t & t\mu^{t-1} \\ 0 & \mu^t \end{bmatrix}.$$

2) Proof of Lemma 2: From Lemma 1, it follows

$$\begin{bmatrix} 1 & 0 \end{bmatrix} M^t = \begin{bmatrix} -\sum_{i=0}^{t-2} \mu_1^{i+1} \mu_2^{t-1-i} & \sum_{i=0}^{t-1} \mu_1^i \mu_2^{t-1-i} \end{bmatrix},$$

where μ_1 and μ_2 are the eigenvalues of M. Moreover,

$$\left| \sum_{i=0}^{t-2} \mu_1^{i+1} \mu_2^{t-1-i} \right| \leq \sum_{i=0}^{t-2} |\mu_1^{i+1} \mu_2^{t-1-i}| \leq \sum_{i=0}^{t-2} \rho^t \leq (t-1)\rho^t$$

$$\left| \sum_{i=0}^{t-1} \mu_1^i \mu_2^{t-1-i} \right| \leq \sum_{i=0}^{t-1} |\mu_1^i \mu_2^{t-1-i}| \leq \sum_{i=0}^{t-1} \rho^{t-1} \leq t\rho^{t-1}$$

by triangle inequality. Finally, for $\mu_1 = \mu_2 \in \mathbb{R}$, we have $\rho = |\mu_1| = |\mu_2|$ and these inequalities become equalities.

3) Proof of Theorem 1: Let μ_{1i} and μ_{2i} be the eigenvalues and let $\rho_i = \max\{|\mu_{1i}|, |\mu_{2i}|\}$ be the spectral radius of A_i. We can use Lemma 2 with $M := A_i$ to obtain

$$\begin{aligned} \max_{i \leq r} \|C_i A_i^t\|_2^2 &\leq \max_{i \leq r} \left((t-1)^2 \rho_i^{2t} + t^2 \rho_i^{2t-2} \right) \\ &\leq (t-1)^2 \rho^{2t} + t^2 \rho^{2t-2} \end{aligned} \tag{24}$$

where $\rho := \max_{i \leq r} \rho_i$. For the parameters provided in Table I, the matrices A_1 and A_r, that correspond to the largest and smallest non-zero eigenvalues of Q, i.e., $\lambda_1 = L$ and $\lambda_r = m$, respectively, have the largest spectral radius [17, Eq. (64)],

$$\rho = \rho_1 = \rho_r \geq \rho_i, \; i = 2, \ldots, r-1 \tag{25}$$

and A_r has repeated eigenvalues. Thus, we can write

$$\begin{aligned} \max_{i \leq r} \|C_i A_i^t\|_2^2 &\geq \| \begin{bmatrix} 1 & 0 \end{bmatrix} A_r^t \|_2^2 = \\ (t-1)^2 \rho_r^{2t} + t^2 \rho_r^{2t-2} &= (t-1)^2 \rho^{2t} + t^2 \rho^{2t-2} \end{aligned} \tag{26}$$

where the first equality follows from Lemma 2 applied to $M := A_r$ and the second equality follows from (25). Finally, combining (24) and (26) with $\beta < \rho$ and Proposition 1 completes the proof.

4) Proof of Theorem 2: Let $a(t) := t\rho^t$. Theorem 1 implies $J^2(t) = \rho^2 a^2(t-1) + \rho^{-2} a^2(t)$ and, for $t \geq 1$, $J(t)$ has only one critical point, which is a maximizer. Moreover, since $dJ^2(t)/dt$ is positive at $t = -1/\log(\rho)$ and negative at $t = 1 - 1/\log(\rho)$, we conclude that the maximizer lies between $-1/\log(\rho)$ and $1 - 1/\log(\rho)$. Regarding $\max_t J(t)$, we note that $\sqrt{2}\rho a(t-1) \leq J(t) \leq \sqrt{2} a(t)/\rho$ and the proof follows from $\max_{t \geq 1} a(t) = -1/(e \log(\rho))$ (cf. Lemma 5).

5) Proof of Proposition 2: Since for all $a \leq 1$, we have [40]

$$a \leq -\log(1-a) \leq a/(1-a)$$

$\rho_{\mathrm{hb}} = 1 - 2/(\sqrt{\kappa}+1)$ and $\rho_{\mathrm{na}} = 1 - 2/(\sqrt{3\kappa+1})$ satisfy

$$2/(\sqrt{\kappa}+1) \leq -\log(\rho_{\mathrm{hb}}) \leq 2/(\sqrt{\kappa}-1)$$
$$2/\sqrt{3\kappa+1} \leq -\log(\rho_{\mathrm{na}}) \leq 2/(\sqrt{3\kappa+1}-2).$$

The conditions on κ ensure that ρ_{hb} and ρ_{na} are not smaller than $1/e$ and we combine the above bounds with Theorem 2 to complete the proof.

B. Proof of Theorem 3

The condition $x_0 = x_1$ is equivalent to $\hat{x}_i^0 = \hat{x}_i^1$ in (5). Thus, for $\lambda_i = 0$, equation (12) yields $\hat{x}_i^t = \hat{x}_i^0 = \hat{x}_i^\star$. For $\lambda_i \neq 0$, we have $\hat{\psi}_i^0 - \hat{\psi}_i^\star = \begin{bmatrix} \hat{x}_i^0 & \hat{x}_i^0 \end{bmatrix}^T$ and, hence,

$$\frac{\|x^t - x^\star\|_2}{\|x^0 - x^\star\|_2} \leq \max_{i \leq r} \frac{|\hat{x}_i^t - \hat{x}_i^\star|}{|\hat{x}_0^t - \hat{x}_i^\star|} = \max_{i \leq r} \left| C_i A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right| \tag{27a}$$

where the equality follows from (10). To bound the right-hand side, we use Lemma 1 with $M = A_i$ to obtain

$$C_i A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix} A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \omega_t(\mu_{1i}, \mu_{2i}) \tag{27b}$$

where μ_{1i} and μ_{2i} are the eigenvalues of A_i and

$$\omega_t(z_1, z_2) := \sum_{i=0}^{t-1} z_1^i z_2^{t-1-i} - \sum_{i=1}^{t-1} z_1^i z_2^{t-i} \tag{28}$$

for any $t \in \mathbb{N}$ and $z_1, z_2 \in \mathbb{C}$.

For Nesterov's accelerated method, the characteristic polynomial $\det(zI - A_i) = z^2 - (1+\beta)h_i z + \beta h_i$ yields $\mu_{1i}, \mu_{2i} = ((1+\beta)h_i \pm \sqrt{(1+\beta)^2 h_i^2 - 4\beta h_i})/2$, where λ_i is the ith the eigenvalue of Q and $h_i := 1 - \alpha\lambda_i$. For the parameters provided in Table I, it is easy to show that:

- For $\lambda_i \in [m, 1/\alpha]$, we have $h_i \in [0, 4\beta/(1+\beta)^2]$ and μ_{1i} and μ_{2i} are complex conjugates of each other and lie on a circle of radius $\beta/(1+\beta)$ centered at $z = \beta/(1+\beta)$.
- For $\lambda_i \in (1/\alpha, L]$, μ_{1i} and μ_{2i} are real with opposite signs and can be sorted to satisfy $|\mu_{2i}| < |\mu_{1i}|$ with $-1 \leq \mu_{1i} \leq 0 \leq \mu_{2i} \leq 1/3$.

The next lemma provides a unit bound on $|\omega_t(\mu_{1i}, \mu_{2i})|$ for both of the above cases.

Lemma 6: For any $z = l\cos(\theta)e^{i\theta} \in \mathbb{C}$ with $|\theta| \le \pi/2$ and $0 \le l \le 1$, and for any real scalars (z_1, z_2) such that $-1 \le z_1 \le 0 \le z_2 \le 1/3$, and $z_2 < -z_1$, the function ω_t in (28) satisfies $|\omega_t(z, \bar{z})| \le 1$ and $|\omega_t(z_1, z_2)| \le 1$ for all $t \in \mathbb{N}$, where $\bar{z}$ is the complex conjugate of z.

Proof: Since $\omega_1(z_1, z_2) = 1$, we assume $t \ge 2$. We first address $\theta = 0$, i.e., $z = l \in \mathbb{R}$ and $\omega_t(z, \bar{z}) = tl^{t-1} - (t-1)l^t$. We note that $\mathrm{d}\omega_t/\mathrm{d}l = t(t-1)(l^{t-2} - l^{t-1}) = 0$ only if $l \in \{0, 1\}$. This in combination with $l \in [0, 1]$ yield $|\omega_t(l, l)| \le \max\{|\omega_t(1, 1)|, |\omega_t(0, 0)|\} \le 1$.

To address $\theta \ne 0$, we note that $b(t) := \sin(t\theta)/t$ satisfies

$$|b(t)| \le |\sin(\theta)| \qquad (29)$$

which follows from

$$|\sin(t\theta)| = |\sin((t-1)\theta)\cos(\theta) + \cos((t-1)\theta)\sin(\theta)|$$
$$\le |\sin((t-1)\theta)| + |\sin(\theta)|.$$

For $z = l\cos(\theta)e^{i\theta}$, we have

$$\omega_t(z, \bar{z}) = (z^t - \bar{z}^t - z\bar{z}(z^{t-1} - \bar{z}^{t-1}))/(z - \bar{z}) =$$
$$(l\cos(\theta))^{t-1}(\sin(t\theta) - l\cos(\theta)\sin((t-1)\theta))/\sin(\theta).$$

Thus, $\mathrm{d}\omega_t/\mathrm{d}l = 0$ only if $l = 0$, 1, or $l^\star := b(t)/(b(t-1)\cos(\theta))$. Moreover, it is easy to show that $\omega_t(z, \bar{z})$ is equal to 0 for $l = 0$; to $(\cos(\theta))^{t-1}\cos((t-1)\theta)$ for $l = 1$; and to $(l^\star \cos(\theta))^{t-1}b(t)/\sin(\theta)$ for $l = l^\star$. Combining this with (29) completes the proof for complex z.

To address the case of $z_1, z_2 \in \mathbb{R}$, we note that $\omega_t(z_1, z_2) = (z_1^t(1 - z_2) - z_2^t(1 - z_1))/(z_1 - z_2)$. Thus, differentiating with respect to z_1 yields

$$\frac{\mathrm{d}\omega_t}{\mathrm{d}z_1} = (1 - z_2)\frac{(t-1)z_1^{t-1} - z_2\sum_{i=0}^{t-2}z_1^{t-2-i}z_2^i}{z_1 - z_2}.$$

Moreover, from $|z_2| < |z_1|$, it follows that

$$(t-1)|z_1^{t-1}| > |z_2|\sum_{i=0}^{t-2}|z_1^{t-2-i}z_2^i| > |z_2\sum_{i=0}^{t-2}z_1^{t-2-i}z_2^i|.$$

Therefore, $\mathrm{d}\omega_t/\mathrm{d}z_1 \ne 0$ over our range of interest for z_1, z_2. Thus, $\omega_t(z_1, z_2)$ may take its extremum only at the boundary $z_1 \in \{0, -1\}$, i.e. $|\omega_t(z_1, z_2)| \le \max\{|\omega_t(0, z_2)|, |\omega_t(1, z_2)|\}$. Finally, it is easy to show that $|\omega_t(0, z_2)| = |z_2^{t-1}| < 1$, and $|\omega_t(-1, z_2)| = |(-1)^t(z_2 - 1) + 2z_2^t|/(1 + z_2) \le 1$. ∎

We complete the proof of Theorem 3 by noting that the eigenvalues of A_i for Nesterov's algorithm with parameters provided in Table I satisfy the conditions in Lemma 6.

C. Proofs of Section III

1) Proof of Lemma 3: For any $f \in \mathcal{F}_m^L$, the L-Lipschitz continuity of the gradient ∇f,

$$f(x^{t+2}) - f(y^t) \le (\nabla f(y^t))^T(x^{t+2} - y^t) + \frac{L}{2}\|x^{t+2} - y^t\|_2^2 \qquad (30a)$$

and the m-strong convexity of f,

$$f(y^t) - f(x^{t+1}) \le (\nabla f(y^t))^T(y^t - x^{t+1}) - \frac{m}{2}\|y^t - x^{t+1}\|_2^2 \qquad (30b)$$

can be used to show that (20) holds along the solution of Nesterov's accelerated algorithm (18). In particular, for (18)

we have $u^t := \nabla f(y^t)$ and

$$x^{t+2} - y^t = -\alpha u^t$$
$$y^t - x^{t+1} = \beta(x^{t+1} - x^t) = \begin{bmatrix} -\beta I & \beta I \end{bmatrix}\psi^t. \qquad (31)$$

Substituting (31) into (30a) and (30b) and adding the resulting inequalities completes the proof.

2) Proof of Lemma 4: Pre- and post-multiplication of LMI (21) by $(\eta^t)^T$ and $\eta^t := [(\psi^t)^T \ (u^t)^T]^T$ yields

$$0 \ge (\eta^t)^T \begin{bmatrix} A^T X A - X & A^T X B \\ B^T X A & B^T X B \end{bmatrix}\eta^t + \theta_1(\eta^t)^T M_1 \eta^t$$
$$+ \theta_2(\eta^t)^T M_2 \eta^t \ge$$
$$(\eta^t)^T \begin{bmatrix} A^T X A - X & A^T X B \\ B^T X A & B^T X B \end{bmatrix}\eta^t + \theta_2(\eta^t)^T M_2 \eta^t$$

where the second inequality follows from (19c). This yields

$$0 \le \hat{V}(\psi^t) - \hat{V}(\psi^{t+1}) - \theta_2(\eta^t)^T M_2 \eta^t \qquad (32)$$

where $\hat{V}(\psi) := \psi^T X \psi$. Also, since Lemma 3 implies

$$-(\eta^t)^T M_2 \eta^t \le 2\left(f(x^{t+1}) - f(x^{t+2})\right) \qquad (33)$$

combining (32) and (33) yields

$$\hat{V}(\psi^{t+1}) + 2\theta_2 f(x^{t+2}) \le \hat{V}(\psi^t) + 2\theta_2 f(x^{t+1}).$$

Thus, using induction, we obtain the uniform upper bound

$$\hat{V}(\psi^t) + 2\theta_2 f(x^{t+1}) \le \hat{V}(\psi^0) + 2\theta_2 f(x^1). \qquad (34)$$

This allows us to bound $\hat{V}$ by writing

$$\lambda_{\min}(X)\|\psi\|_2^2 \le \hat{V}(\psi) \le \lambda_{\max}(X)\|\psi\|_2^2. \qquad (35a)$$

We can also upper and lower bound $f \in \mathcal{F}_m^L$ as

$$m\|x\|_2^2 \le 2f(x) \le L\|x\|_2^2. \qquad (35b)$$

Finally, combining (34) and (35) yields

$$\lambda_{\min}(X)\|\psi^t\|_2^2 + m\theta_2\|x^{t+1}\|_2^2 \le$$
$$\lambda_{\max}(X)\|\psi^0\|_2^2 + L\theta_2\|x^1\|_2^2.$$

We complete the proof by noting that $\|x^{t+1}\|_2 \le \|\psi^t\|_2$.

3) Proof of Theorem 4: To prove (23a), we need to find a feasible solution for θ_1, θ_2 and X in terms of the condition number κ. Let us define

$$X := \begin{bmatrix} x_1 I & x_0 I \\ x_0 I & x_2 I \end{bmatrix} = x_2 \begin{bmatrix} \beta^2 I & -\beta I \\ -\beta I & I \end{bmatrix}$$
$$\theta_2 := \theta_1(L + m)\beta/(1 - \beta) \qquad (36)$$
$$x_2 := ((L + m)\theta_1 + \theta_2)/\alpha = \theta_2/(\alpha\beta).$$

If (36) holds, it is easy to verify that $X \succeq 0$ with $\lambda_{\min}(X) = 0$, $\lambda_{\max}(X) = (1 + \beta^2)x_2 = \theta_2(1 + \beta^2)/(\alpha\beta)$, and $A^T X A - X = 0$. Moreover, the matrix W on the left-hand-side of (21) is block-diagonal, $W := \mathrm{diag}(W_1, W_2)$, and negative semi-definite for all $\alpha \le 1/L$, where

$$W_1 = -m(2\theta_1 L C_y^T C_y + \theta_2 C_2^T C_2) \preceq 0$$
$$W_2 = -((2 - \alpha(L + m))\theta_1 + \alpha(1 - \alpha L)\theta_2)I \preceq 0.$$

Thus, the choice of (θ_1, θ_2, X) in (36) satisfies the con-

ditions of Lemma 4. Using the expressions for the largest and smallest eigenvalues of the matrix X in equation (22) in Lemma 4, leads to the upper bound for $\|x^t\|_2^2$ in (23a). Furthermore, from (23a) we have

$$\|x^t\|_2^2 \leq \kappa \left(1 + (1 + \beta^2)/(\alpha\beta L)\right) \|\psi^0\|_2^2$$

and the upper bound in (23c) follows from the fact that, for α and β in (23b), $1 + (1 + \beta^2)/(\alpha\beta L) = 3 + 4/(\kappa - 1)$.

To obtain the lower bound in (23c), we employ our framework for quadratic objective functions in Section II. In particular, for the parameters α and β in (23b), the largest spectral radius $\rho(A_i)$ corresponds to A_n, which is associated with the smallest eigenvalue $\lambda_n = m$ of Q. Since A_n has repeated real eigenvalues $\rho = 1 - 1/\sqrt{\kappa}$, using similar arguments as in Theorem 1 for quadratic problems we obtain,

$$
\begin{aligned}
J(t_{\max}) &= \sqrt{(t_{\max} - 1)^2 \rho^{2t_{\max}} + t_{\max}^2 \rho^{2(t_{\max}-1)}} \\
&\geq \sqrt{2}\,(t_{\max} - 1)\,\rho^{t_{\max}} \geq \sqrt{2}(\sqrt{\kappa} - 1)^2/(e\sqrt{\kappa})
\end{aligned}
$$

which completes 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] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, September 2013.

[7] 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.

[8] A. Zare, H. Mohammadi, N. K. Dhingra, T. T. Georgiou, and M. R. Jovanović, "Proximal algorithms for large-scale statistical modeling and sensor/actuator selection," *IEEE Trans. Automat. Control*, vol. 65, no. 8, pp. 3441–3456, August 2020.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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*, Miami, FL, 2018, pp. 5753–5758.

[16] 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.

[17] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," *IEEE Trans. Automat. Control*, vol. 66, no. 6, pp. 2480–2495, June 2021.

[18] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimisation algorithms: an integral quadratic constraint approach," *Int. J. Control*, pp. 1–24, 2020.

[19] J. I. Poveda and N. Li, "Robust hybrid zero-order optimization algorithms with acceleration via averaging in time," *Automatica*, p. 109361, 2021.

[20] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Found. Comput. Math.*, vol. 15, pp. 715–732, 2015.

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

[22] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao, "An accelerated linearized alternating direction method of multipliers," *SIAM J. Imaging Sci.*, vol. 8, no. 1, pp. 644–681, 2015.

[23] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, 2014.

[24] L. N. Trefethen and M. Embree, *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton: Princeton University Press, 2005.

[25] M. R. Jovanović and B. Bamieh, "Componentwise energy amplification in channel flows," *J. Fluid Mech.*, vol. 534, pp. 145–183, July 2005.

[26] M. R. Jovanović, "From bypass transition to flow control and data-driven turbulence modeling: An input-output viewpoint," *Annu. Rev. Fluid Mech.*, vol. 53, no. 1, pp. 311–345, January 2021.

[27] 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.

[28] S. Samuelson, H. Mohammadi, and M. R. Jovanović, "On the transient growth of Nesterov's accelerated method for strongly convex optimization problems," in *Proceedings of the 59th IEEE Conference on Decision and Control*, Jeju Island, Republic of Korea, 2020, pp. 5911–5916.

[29] B. T. Polyak and G. V. Smirnov, "Transient response in matrix discrete-time linear systems," *Autom. Remote Control*, vol. 80, no. 9, pp. 1645–1652, 2019.

[30] B. Can, M. Gurbuzbalaban, and L. Zhu, "Accelerated linear convergence of stochastic momentum methods in Wasserstein distances," in *International Conference on Machine Learning*. PMLR, 2019, pp. 891–901.

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

[32] 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.

[33] 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.

[34] M. Danilova and G. Malinovsky, "Averaged heavy-ball method," 2021, arXiv:2111.05430.

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

[36] 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.

[37] 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, July 2019.

[38] S. Hassan-Moghaddam and M. R. Jovanović, "Proximal gradient flow and Douglas-Rachford splitting dynamics: global exponential stability via integral quadratic constraints," *Automatica*, vol. 123, p. 109311, January 2021.

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

[40] F. Topsok, "Some bounds for the logarithmic function," *Inequal. Theory Appl.*, vol. 4, p. 137, 2006.