2023 American Control Conference (ACC)
San Diego, CA, USA. May 31 - June 2, 2023

Performance of noisy higher-order accelerated gradient flow dynamics for strongly convex quadratic optimization problems

Samantha Samuelson, Hesameddin Mohammadi, and Mihailo R. Jovanović

***Abstract*— We study performance of momentum-based accelerated first-order optimization algorithms in the presence of additive white stochastic disturbances. For strongly convex quadratic problems with a condition number κ, we determine the best possible convergence rate of continuous-time gradient flow dynamics of order n. We also demonstrate that additional momentum terms do not affect the tradeoffs between convergence rate and variance amplification that exist for gradient flow dynamics with $n = 2$.**

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

I. Introduction

Accelerated first-order optimization algorithms [1]–[4] are the workhorse for large-scale problems [5]–[7]. Momentum-based algorithms enjoy favorable asymptotic behavior [8]–[12] and, for different noise models, tradeoffs between acceleration and robustness have also been studied [13]–[20]. These references suggest that accelerated methods are more sensitive to noise than gradient descent.

The connection between ordinary differential equations and iterative optimization algorithms is also well established [21]–[29]. Recently, a second-order continuous-time dynamical system with constant coefficients for which a certain implicit-explicit Euler discretization yields Nesterov's accelerated algorithm was introduced in [30]. For strongly convex problems, these accelerated gradient flow dynamics were shown to be exponentially stable with rate $1/\sqrt{\kappa}$, where κ is the condition number of the problem. A more recent work [31] examined the tradeoffs between convergence rate and robustness to additive white noise of accelerated gradient flow dynamics and established a lower bound on the product between steady-state variance of the error in the optimization variable and the settling time that scales with κ^2. For this class of accelerated dynamical systems, there appears to be a fundamental limitation between convergence rate and variance amplification imposed by the condition number. In addition, similar phenomena was shown to persist in the discrete-time setting for the class of noisy two-step momentum algorithms [31], [32]. Corroborating results for d1iscrete-time algorithms were also presented in [33] by examining a parameterized family of two-step momentum algorithms that enable systematic tradeoffs between these quantities.

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).

In this paper, we extend the results in [30], [31] by considering a nth-order accelerated gradient flow dynamics that generalizes the system presented in [30]. For strongly convex quadratic problems, we analyze convergence properties of this system and sensitivity to additive white noise. In particular, we establish the optimal convergence rate $\rho = \kappa^{-1/n}$ and identify constant algorithmic parameters that achieve the optimal rate. In addition, we derive analytical expressions for the noise amplification in terms of the entries of the Routh-Hurwitz table [34]. This characterization allows us to show that the product of variance amplification J (in the error of the optimization variable) and settling time $1/\rho$ is lower bounded by $\kappa^2/(2n)$.

Previous work [35] obtained similar results regarding the parameters which achieve the convergence rate $\rho = \kappa^{-1/n}$. We provide additional analysis of system behavior, including extending results to the noise amplification properties of this class of accelerated gradient flow dynamics, particularly the trade-off between settling time and noise amplification.

The rest of the paper is structured as follows. In Section II, we provide preliminaries and background material. In Section III, we present our results regarding convergence rate and steady-state variance amplification. We first determine the optimal rate of exponential convergence ρ in terms of the condition number κ and system order n, and show that there exists a set of parameters which achieves this rate. Next, we determine an analytical expression for the steady-state variance amplification in terms of Routh-Hurtwitz coefficients and identify a lower bound on the product between variance amplification and settling time which scales with κ^2. The Proof is given in the appendix.

II. Motivation and background

We consider a class of dynamical systems,

$$x^{(n)}(t) + \sum_{k=0}^{n-1} \beta_k x^{(k)}(t) \,+\, \alpha g\left(\sum_{k=0}^{n-1} \gamma_k x^{(k)}(t)\right) \,=\, w(t) \tag{1}$$

where $x^{(k)}(t)$ is the kth derivative of x with respect to time t, g is a nonlinear function, α, β_k, and γ_k are constant parameters, and w is a white-noise input with

$$\mathbb{E}[w(t)] \,=\, 0, \quad \mathbb{E}[w(t_1)w(t_2)] \,=\, \sigma^2 I \delta(t_1 - t_2) \tag{2}$$

and $\delta(\cdot)$ is the Kronecker delta. Our motivation for studying system (1) comes from optimization. In the absence of noise, we can use system (1) with $g(x) := \nabla f(x)$ to solve

unconstrained optimization problems

$$\underset{x}{\text{minimize}} \;\; f(x) \tag{3}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is an m-strongly convex function with an L-Lipschitz continuous gradient ∇f. Throughout the paper, we make the following assumptions:

Assumption 1: The parameters in system (1) satisfy

$$\beta_0 \;=\; 0,\;\; \gamma_0 \;=\; 1. \tag{4}$$

Assumption 1 ensures that the equilibrium points $x^\star$ of system (1) satisfy the first-order optimality conditions for (3),

$$g(x^\star) \;=\; \nabla f(x^\star) \;=\; 0. \tag{5}$$

As varying the parameter α is a matter of time-scaling, we set $\alpha = 1/L$ without loss of generality, and note that, for $n = 1$, system (1) simplifies to the gradient flow dynamics,

$$x^{(1)}(t) \;+\; (1/L)\,\nabla f(x(t)) \;=\; w(t)$$

and, for $n = 2$, noisy accelerated gradient flow dynamics is obtained [30],

$$x^{(2)}(t) + \beta_1 x^{(1)}(t) + (1/L)\,\nabla f(x(t) + \gamma_1 x^{(1)}(t)) \;=\; w(t).$$

A. Quadratic optimization problems

For strongly convex quadratic optimization problems,

$$f(x) \;=\; \frac{1}{2}\,x^T Q x \;-\; q^T x \tag{6}$$

with $Q \in \mathbb{R}^{d\times d}$, the parameters of strong convexity and Lipschitz continuity, m and L, are respectively determined by the smallest and the largest eigenvalues of the Hessian matrix Q,

$$mI \;\preceq\; Q \;\preceq\; LI$$

and the condition number is given by $\kappa := L/m$. In this case, differential equation (1) with $g = \nabla f$ becomes linear,

$$x^{(n)}(t) + \sum_{k\,=\,0}^{n-1} (\beta_k I \;+\; \gamma_k \alpha Q) x^{(k)}(t) \;=\; w(t) \tag{7}$$

and the optimization algorithm admits an LTI state-space representation,

$$\begin{aligned} \dot{\psi} \;&=\; A\psi \;+\; Bw \\ z \;&=\; C\psi \end{aligned} \tag{8a}$$

where $z := x - x^\star$ is the error in the optimization variable, ψ is the state vector defined by

$$\begin{aligned} \psi \;&=\; \begin{bmatrix} \psi_1^T & \psi_2^T \end{bmatrix}^T \\ \psi_1 \;&:=\; z,\; \psi_2 \;:=\; \begin{bmatrix} (x^{(1)})^T & \cdots & (x^{(n-1)})^T \end{bmatrix}^T \end{aligned} \tag{8b}$$

and A, B, C and constant matrices that are partitioned conformably with the state vector ψ,

$$\begin{aligned} A \;&=\; \begin{bmatrix} 0 & I \\ A_{21} & A_{22} \end{bmatrix},\; B \;=\; \begin{bmatrix} 0 \\ I \end{bmatrix},\; C \;=\; \begin{bmatrix} I & 0 \end{bmatrix} \\ A_{21} \;&=\; -(\beta_0 I + \gamma_0 \alpha Q) \\ A_{22} \;&=\; \begin{bmatrix} -(\beta_1 I + \gamma_1 \alpha Q) & \cdots & -(\beta_{n-1} I + \gamma_{n-1} \alpha Q) \end{bmatrix}. \end{aligned} \tag{8c}$$

The eigenvalue decomposition of the Hessian matrix, $Q = V\Lambda V^T$, can be utilized to bring matrices in (8) into their block diagonal forms, where V is an orthogonal matrix of the eigenvectors of Q and Λ is a diagonal matrix of its eigenvalues. In particular, the change of variables,

$$\hat{x} \;:=\; V^T x,\;\; \hat{w} \;:=\; V^T w \tag{9}$$

allows us to transform system (8) into a parameterized family of d decoupled subsystems indexed by $i = 1, \ldots, d$,

$$\begin{aligned} \dot{\hat{\psi}}_i \;&=\; \hat{A}(\lambda_i)\hat{\psi}_i \;+\; \hat{B}\hat{w}_i \\ z_i \;&=\; \hat{C}\psi_i \end{aligned} \tag{10a}$$

where λ_i is the ith eigenvalue of the matrix $Q \in \mathbb{R}^{d\times d}$, $\hat{w}_i$ is the ith component of the vector $\hat{w}$,

$$\begin{aligned} \hat{A}(\lambda) \;&=\; \begin{bmatrix} 0 & I \\ -a_0(\lambda) & [-a_1(\lambda) \;\cdots\; -a_{n-1}(\lambda)] \end{bmatrix} \\ \hat{B} \;&=\; \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T,\;\; \hat{C} \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{10b}$$

Here,

$$a_k(\lambda) \;:=\; \beta_k \;+\; \gamma_k \alpha \lambda,\;\; k \;=\; \{0, \ldots, n-1\} \tag{10c}$$

and the characteristic polynomial of $\hat{A}(\lambda)$ is given by,

$$F(s) \;:=\; \sum_{k\,=\,0}^{n} a_k(\lambda) s^k \;=\; \prod_{k\,=\,1}^{n} (s \;-\; \mu_k) \tag{10d}$$

where we let $a_n(\lambda) := 1$ and μ_k are the eigenvalues of $\hat{A}(\lambda)$.

B. Exponential stability

System (8) is exponentially stable in the absence of noise if all eigenvalues of the matrix A have negative real parts, i.e., if A is Hurwitz,

$$\|\psi(t)\| \;=\; \|\mathrm{e}^{At}\psi(0)\| \;\leq\; c\,\mathrm{e}^{-\rho t}\|\psi(0)\| \tag{11}$$

and the convergence rate ρ is determined by

$$\rho \;=\; |\max\, \Re\,(\mathrm{eig}(A))| \tag{12}$$

where $\Re\,(\mathrm{eig}(\cdot))$ denotes the real part of the eigenvalues of a given matrix. From the modal decomposition (10), it can be seen that the convergence rate of system (8) is determined by the slowest mode of matrices $\hat{A}(\lambda)$,

$$\rho \;=\; \min_{\lambda\in[m,L]} \hat{\rho}(\lambda).$$

where $\hat{\rho}(\lambda) := |\max \Re(\mathrm{eig}(\hat{A}(\lambda)))|$. Our goal is to examine the impact of the algorithmic parameters

$$\theta \;:=\; [\,\beta_1 \;\cdots\; \beta_{n-1} \;\; \gamma_1 \;\cdots\; \gamma_{n-1}\,]^T \tag{13}$$

on the eigenvalues $\mu_k(\lambda)$ of $\hat{A}(\lambda)$ and, thus, stability of the system. Under Assumption 1 on (β_0, γ_0) and with $\alpha = 1/L$, the vector of parameters θ define the coefficients $a_k(\lambda)$ in (10c) of the characteristic polynomial (10d).

The Routh-Hurwitz (RH) criterion provides necessary and sufficient conditions for coefficients $a_k(\lambda)$ in (10c) to ensure stability. Furthermore, by introducing the shifted characteristic polynomial

$$F_\rho(s) \;:=\; \sum_{k=0}^{n} a_k(\lambda)(s \,-\, \rho)^k \;=\; \prod_{k=1}^{n} (s \,-\, \nu_k) \qquad (14)$$

we can utilize the RH criterion to determine conditions for ρ-exponential stability. In particular, since the roots of $F(s)$ and $F_\rho(s)$ are related by $\nu_k = \mu_k + \rho$, $F(s)$ is ρ-exponentially stable (i.e., all its roots have real parts smaller than $-\rho$) if and only if $F_\rho(s)$ is stable. Thus, it suffices to examine the RH conditions on the coefficients of

$$\begin{aligned} F_\rho(s) \;&:=\; \sum_{k=0}^{n} \tilde{a}_k^\rho(\lambda) s^k \\ \tilde{a}_k^\rho(\lambda) \;&:=\; \sum_{i=0}^{n-k} a_{n-i}(\lambda) \binom{n-i}{k} (-\rho)^{(n-k-i)} \end{aligned} \qquad (15)$$

where $\tilde{a}_n^\rho(\lambda) := 1$.

In Section III, we determine the largest rate of convergence for a given condition number κ and identify the vector of parameters θ that achieves this rate.

C. Variance amplification

In addition to the convergence rate, we are also interested in quantifying the steady-state variance of the error in the optimization variable (variance amplification),

$$J \;:=\; \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{E}\left(\|x(\tau) \,-\, x^\star\|^2\right) \mathrm{d}\tau. \qquad (16)$$

For the LTI system (8), the state covariance matrix at the steady-state is determined by

$$X \;=\; \lim_{t\to\infty} \mathbb{E}\left[\psi(t)(\psi(t))^T\right] \qquad (17)$$

where X solves the algebraic Lyapunov equation

$$AX \;+\; XA^T \;=\; -\sigma^2 BB^T. \qquad (18)$$

The eigenvalue decomposition of the Hessian matrix Q can be utilized to express the variance amplification as

$$J \;=\; \mathrm{trace}\left(CXC^T\right) \;=\; \sum_{i=1}^{d} \hat{J}(\lambda_i) \qquad (19)$$

where $\hat{J}(\lambda_i)$ denotes the contribution of the ith eigenvalue λ_i of Q to the variance amplification,

$$\hat{J}(\lambda_i) \;:=\; \mathrm{trace}\left(\hat{C}\hat{X}(\lambda_i)\hat{C}^T\right) \;=\; \hat{X}_{11}(\lambda_i). \qquad (20)$$

In Section III, we derive the expression for the 11-element $\hat{X}_{11}(\lambda_i)$ of the matrix $\hat{X}(\lambda_i)$ that solves,

$$\hat{A}(\lambda_i)\hat{X}(\lambda_i) \;+\; \hat{X}(\lambda_i)\hat{A}^T(\lambda_i) \;=\; -\sigma^2 \hat{B}\hat{B}^T \qquad (21)$$

in terms of the coefficients of the Routh-Hurwitz table and utilize this relation to determine lower bounds on J in terms of the convergence rate ρ and condition number κ.

III. Main results

In this section, we present our main results regarding the convergence rate and steady-state variance amplification for the class of nth-order gradient flow dynamics described by (7). For strongly convex quadratic problems with condition number κ, we establish the optimal rate of exponential convergence $\rho = \kappa^{-1/n}$ and identify algorithmic parameters that achieve this optimal rate. We also provide analytical expressions for the modal contributions $\hat{J}(\lambda)$ to the variance amplification in terms of the entries of the Routh-Hurwitz table. We use these expressions to establish that a lower bound on the product between the variance amplification and the settling time scales as κ^2. Our results extend the observations made in [31] that only considered the case $n = 2$ to general n and they recover the same tradeoff between the settling time and variance amplification. Proofs of all results are relegated to the Appendix.

Theorem 1 establishes the optimal rate of convergence and determines parameters that achieve the optimal rate.

Theorem 1: For strongly convex quadratic objective function f with condition number κ, under Assumption 1 regarding the optimality conditions and the normalization condition $\alpha = 1/L$, the optimal rate of exponential convergence of system (7) is given by

$$\rho \;=\; \kappa^{-1/n}$$

and is achieved by parameters

$$\gamma_k \;=\; \rho^{-k}\binom{n-1}{k}, \quad \beta_k \;=\; \binom{n}{k}\rho^{n-k} - \gamma_k \kappa^{-1} \qquad (22)$$

for $k = 0, \ldots, n-1$.

As we demonstrate in the proof, the constraint on the optimal rate established by Theorem 1 is imposed by

$$\alpha\lambda \;=\; \prod_{k=1}^{n} (-\mu_k)$$

where μ_k are the eigenvalues of $\hat{A}(\lambda)$ in (10) for all $\lambda \in [m, L]$. In addition, the optimal parameters are determined by setting $n-1$ eigenvalues of $\hat{A}(\lambda)$ at $-\rho$ for all $\lambda \in [m, L]$. This result generalizes previous works that considered the cases $n = 1$ [30] and $n = 2$ [30], [31]. Note that while we employ eigenvalue decomposition for analysis, the parameters in (22) rely only on the condition number, and eigenvalue decomposition is unnecessary for algorithm implementation.

The next theorem presents an analytical expression for the modal contribution to variance amplification $\hat{J}(\lambda)$ for all λ in terms of the entries of the RH table; see [34].

Theorem 2: For strongly convex quadratic objective function f, under Assumption 1 regarding the optimality conditions, the modal contribution $\hat{J}(\lambda)$ to the steady-state variance amplification of system (7) with stabilizing parameters

θ and $\alpha = 1/L$ can be expressed as

$$\hat{J}(\lambda) \;=\; \frac{\sigma^2}{2a_0(\lambda)r(\lambda)}$$

where $r(\lambda)$ is the first (and only) entry in the nth row of the Routh-Hurwitz table associated with the characteristic polynomial $F(s)$ in (10d) of the matrix $\hat{A}(\lambda)$ in (10).

As the entries of the RH table are themselves defined in terms of the coefficients of the characteristic polynomial (10d), the analytical expression we derive above provides a way to examine how the steady-state noise amplification is determined by our choice of parameters θ. We will use Theroem 2 to lead directly to our next result.

Theorem 3: For strongly convex quadratic objective function f, under Assumption 1 regarding the optimality conditions, with stabilizing parameters θ and the normalization condition $\alpha = 1/L$, the product of the modal contribution to the steady-state variance amplification and the settling time $T_s = 1/\rho$ is lower bounded by

$$\hat{J}(\lambda)/\rho \;>\; \sigma^2/(2\,n\,(\alpha\lambda)^2).$$

For the case $\lambda = m$, this simplifies to

$$\hat{J}(m)/\rho \;>\; \sigma^2\kappa^2/(2\,n).$$

We observe that this lower bound is decreasing with λ and scales as κ^2 for $\lambda = m$. This is consistent with [31] where the authors examined the case of $n = 2$.

Corollary 1: Under the settings of Theorem 3, we have the lower bound

$$\frac{J}{\rho} \;>\; \sum_{i\,=\,1}^{d} \frac{\sigma^2}{2\,n\,(\alpha\lambda_i)^2} \;\geq\; \frac{\sigma^2\kappa^2}{2\,n} + \frac{\sigma^2(d-1)}{2\,n}$$

The proof is immediate by combining Theorem 3 with equation (19). Based on this result, we conclude that for systems of type (7), there is a fundamental tradeoff between settling time and variance amplification for bounded order $n \ll \kappa$.

IV. Concluding remarks

Our results demonstrate that regardless of the number of momentum terms in accelerated gradient flow dynamics, the product between variance amplification (of the error in the optimization variable) and settling time scales as κ^2. Our ongoing work focuses on examining how additional momentum terms affect the behavior of familiar discrete time accelerated algorithms.

Appendix

A. Proof of Theorem 1

The full proof is omitted due to page limitations; we include an outline here.

Proof: We first prove that the optimal rate cannot exceed $\rho \leq \kappa^{-\frac{1}{n}}$. We begin by examining the best possible rate $\rho(\lambda)$ associated with the matrix $\hat{A}(\lambda)$ for a fixed λ. By matching the constant terms in (10d) of the product and summation expressions of $F(s)$, we can write

$$a_0(\lambda) = \alpha\lambda = \prod_{k=1}^{n}(-\mu_k) \tag{23}$$

Notice that the coefficient a_0 is fixed with respect to λ, while the freedom to choose β_k and γ_k without constraints allows for any desired placement of $a_k(\lambda)$ for $k \neq 0$. In essence, we wish to place the real part of the largest eigenvalue as far from the imaginary axis as possible, given that the product of all eigenvalues is fixed. The solution to this problem is given by

$$\mu_k \;=\; -(\alpha\lambda)^{1/n}, \quad k = 1, \ldots, n \tag{24}$$

which yields $\rho(\lambda) \leq (\alpha\lambda)^{1/n}$. The result is apparent when μ_k are real, and the extension to imaginary roots is straightforward upon noting that all roots must come in complex conjugate pairs, whose product is real and greater than the product of the real parts.

Then, the rate of the system (8) is upper bounded by

$$\rho \;=\; \min_{\lambda\in[m,L]} \rho(\lambda) \;\leq\; \min_{\lambda\in[m,L]} (\alpha\lambda)^{\frac{1}{n}} \;=\; (m/L)^{\frac{1}{n}} \;=\; \kappa^{-\frac{1}{n}}$$

since $\alpha = 1/L$. It follows that for a fixed λ, parameters θ can be chosen to achieve $\rho \leq \kappa^{-1/n}$.

In total, we must select $2n-2$ parameters β_k, γ_k in order to design the linear functions $a_k(\lambda) = \beta_k + \gamma_k\alpha\lambda$ for $k = 0, \ldots, n-1$. This amounts to placing the line segment

$$a(\lambda) \;:=\; [a_0(\lambda), \cdots, a_{n-1}(\lambda)]^T, \;\; \lambda \in [m, L]$$

in $\mathbb{R}^n$. We begin the process of selecting these parameters by noticing that the end point $a(m)$ of this line segment is fixed, as a unique set of $a_k(m)$ allows $\rho(m) = \kappa^{-1/n}$.

This follows directly from the optimal solution (24) for $\lambda = m$. The relationship between $a(m)$ and eigenvalues $\mu_k(\hat{A}(m))$ shown in (10d) yields

$$a_k(m) = \binom{n}{k}\rho^{n-k} \tag{25}$$

where $\rho = \kappa^{-1/n}$. Now that we have determined values of a_k which give the desired rate of convergence at $\lambda = m$, we examine the conditions under which this margin of stability is maintained as λ increases.

As stated in Section II, we will determine conditions for ρ-exponential stability by imposing the RH stability criterion on the coefficients of the shifted characteristic polynomial shown in (15). We decompose the solution into two parts: given that rate of convergence ρ is achieved at $\lambda = m$, we will determine the slope parameters γ_k which ensure the stability constraints are not violated as $\lambda : m \to L$. Once we have obtained the values of γ, we use equation (10c) that directly determines the values of β, given that $a_k(m)$ are designed according to (25).

■

B. Proof of Theorem 2

Proof: We provide a proof for the case where n is even. The case of odd n can be proven in a similar way and is

omitted for brevity. Recall $\hat{X}(\lambda)$ solves

$$\hat{A}(\lambda)\hat{X}(\lambda) + \hat{X}(\lambda)\hat{A}(\lambda)^T = -\sigma^2 \hat{B}\hat{B}^T.$$

Using the block structure of $B = [0,1]^T$, where $0 \in \mathbb{R}^{n-1}$, we can define the block matrix

$$\hat{Z}(\lambda) \;:=\; \hat{A}(\lambda)\hat{X}(\lambda) + \hat{X}(\lambda)\hat{A}(\lambda)^T$$

whose entries are given by

$$z_{i,j} \;=\; \begin{cases} x_{i+1,j} + x_{j+1,i}, & i \neq n \neq j \\ x_{i+1,n} \;-\; \sum_{k=1}^{n}(a_{k-1}x_{i,k}), & i \neq n = j \\ -2\sum_{k=1}^{n} a_{k-1}x_{k,n}, & i = n = j. \end{cases}$$

Here, we have dropped the (λ) indicators for ease of reading, and without loss of generality, set $\sigma = 1$.

We begin by noting all diagonal terms of $\hat{Z}$ except the nth must be zero, yielding

$$x_{i,i+1} \;=\; 0, \quad i \;=\; 1, \ldots, n-1. \tag{26a}$$

Due to our definitions of $z_{i,j}$ for the $1-1$ block of $\hat{Z}$, (26a) forces additional zero constraints. In particular, for $j = i+2$,

$$0 \;=\; z_{i,j} \;=\; x_{i+1,j} + x_{j+1,i} \;=\; x_{i+1,i+2} + x_{i+3,i}.$$

Together with (26a), this yields $x_{i,i+3} = 0$, for $i = 1, \ldots, n-1$. Repeating this procedure for $j = i+2m$ yields

$$x_{i,i+(2m-1)} \;=\; 0, \quad \text{for } m, i \geq 1. \tag{26b}$$

The $z_{i,n}$ can now be written as linear functions of $x_{i,j}$

$$z_{i,n} \;=\; \begin{cases} x_{i+1,n} - \sum_{k=0}^{n/2-1} a_{2k}x_{i,2k+1} & i \text{ odd} \\ -\sum_{k=0}^{n/2-1}(a_{2k+1}x_{i,2(k+1)}) & i \text{ even} \\ -2\sum_{k=0}^{n/2-1} a_{2k+1}x_{2(k+1),n} & i \;=\; n. \end{cases}$$

We observe that, for $i < n$, the coefficients of the $z_{i,n}$ replicate the first and second rows

$$\begin{aligned} a_e \;&:=\; [a_0, a_2, \cdots, a_n] \\ a_o \;&:=\; [a_1, a_3, \cdots, a_{n-1}] \end{aligned} \tag{27}$$

of the RH table associated with the characteristic polynomial $F(s)$ in (10d), containing even and odd coefficients a_k.

After the considerations in (26), the remaining terms in the $1-1$ block of $\hat{Z}$ must also equal zero, requiring

$$x_{i,1+2m} \;=\; -x_{i+1,1+2m-1}, \quad \text{for } m, i \geq 1. \tag{28}$$

This allows us to reduce the number of unknown $x_{i,j}$ to n. To see this, note that as $\hat{X}$ is symmetric, we begin with $\frac{n(n+1)}{2}$ variables $x_{i,j}$. The eliminations in (26b) set $\frac{n^2}{4}$ variables to zero and the eliminations in (28) fix $\frac{n(n-2)}{4}$ variables in terms of others, leaving n free variables. By iterating across the matrix in a column-wise order, we can denote the unknown variables by $\bar{x} = [\hat{x}_1 \;\; \hat{x}_2 \;\; \ldots \;\; \hat{x}_n]^T$ such that $\hat{x}_1 := x_{1,1}$, $\hat{x}_n := x_{n,n}$, and

$$z_{i,n} \;=\; \begin{cases} \sum_{k=0}^{n/2} a_{2k}\,\hat{x}_{(i+1)/2+k} & i \quad \text{odd} \\ \sum_{k=0}^{n/2-1} a_{2k+1}\,\hat{x}_{(i/2)+1+k} & i \quad \text{even}. \end{cases}$$

Combining the above expressions with $2z_{n,n} = -1$ and $z_{1,n}, \ldots, z_{n-1,n} = 0$ brings us to

$$\bar{a}_i^T \bar{x} = y_i \tag{29a}$$

where we let $y_1 := -1/2$, $y_i := 0$ for $i = 1, \ldots, n$, and

$$\bar{a}_i^T \;=\; \begin{cases} \begin{bmatrix} 0_{(n-i+1)/2} & a_o & 0_{(i-1)/2} \end{bmatrix} & i \;\text{ odd} \\ \begin{bmatrix} 0_{(n-i)/2} & a_e & 0_{(i-2)/2} \end{bmatrix} & i \;\text{ even}. \end{cases} \tag{29b}$$

Here, the zero vector $0_k \in \mathbb{R}^k$, and a_o and a_e are given by (27).

We next use the same technique of polynomial quotients and remainders that is used to derive the RH coefficients to recursively generate linear equations of the form (29a) that have fewer non-zero coefficients and ultimately obtain the value of

$$\hat{J}(\lambda) = \text{trace}(\hat{C}\hat{X}(\lambda)\hat{C}^T) = \hat{x}_1. \tag{30}$$

Based on the coefficients given in the third row of the RH table as seen in [34], we are motivated to define

$$\bar{b}_i \;:=\; \bar{a}_{2i} \;-\; \frac{a_n}{a_{n-1}}\bar{a}_{2i-1} \quad i = 1, \ldots, n/2.$$

Note that the vectors $\bar{b}_i$ are of the same structure as those in (29b) except they have $n/2$ non-zero entries that constitute the 3^{rd} row of the RH table. In addition, combining (29b) and the definition of $\bar{b}_i$ yields

$$\bar{b}_i^T \bar{x} \;=\; \begin{cases} 0 & i = 1, \ldots, n/2 - 1 \\ (-1)^n a_0/(2a_1) & i = n/2 \end{cases}$$

We continue this procedure to recover all rows of the RH array. To generalize to any system of size n, let

$$q^k \;:=\; \frac{\bar{f}_1^{k-2}(n-k+2)}{\bar{f}_1^{k-1}(n-k+2)}, \quad \bar{f}_i^k \;:=\; \bar{f}_{i+1}^{k-2} - q^k \bar{f}_i^{k-1} \tag{31a}$$

initialized with

$$\bar{f}_i^1 \;=\; \bar{a}_{2(i-1)}, \quad \bar{f}_i^2 \;=\; \bar{a}_{2i-1} \tag{31b}$$

where in $\bar{f}_i^k(\cdot)$, the superscript k denotes the recursion index, the subscript i ranges from 1 to $\lfloor (n-k+2)/2 \rfloor$, and the argument denotes the entry number. It is easy to verify that the vectors $\bar{f}_i^k$ are of the same structure as those in (29b) except they have $\lfloor (n-k+2)/2 \rfloor$ non-zero entries that constitute the kth row of the RH table. The subscript i indicates the position of the non-zero entries in the vector.

It is now straightforward to show that

$$(\bar{f}_i^k)^T \bar{x} \;=\; \begin{cases} 0 & i \neq 1 \\ (-1)^k \frac{a_n}{2\bar{f}_1^{k-1}(1)} & i = 1 \end{cases} \tag{32}$$

At iteration $k = n+1$, we are left with a single vector $\bar{f}_1^{n+1}$ with a single non-zero element $\bar{f}_1^{n+1}(1)$ in the first position. This allows us to solve for $\hat{x}_1$ and obtain

$$(\bar{f}_1^{n+1})^T \bar{x} \;=\; \bar{f}_1^{n+1}(1)\,\hat{x}_1 \;=\; \frac{a_n}{2\bar{f}_1^n(2)} \tag{33}$$

which yields $\hat{x}_1 = a_n/(2\bar{f}_1^n(2)\bar{f}_1^{n+1}(1))$. By construction,

$\bar{f}_1^n(2)$ and $\bar{f}_1^{n+1}(1)$ are the last two terms in the first column, in addition to $\bar{f}_1^n(2) =: r$ being the first (and only) entry in the nth row of the RH table. It is now easy to verify that the last term $\bar{f}_1^{n+1}(1)$ is given by a_0. Combining this with $a_n = 1$ and (30) completes the proof.

■

C. Proof of Theorem 3

Proof: By Theorem 2, $\hat{J}(\lambda) = \sigma^2/(2r(\lambda)a_0(\lambda))$. According to the definition given in (31a), we can write

$$r(\lambda) = \bar{f}_1^n(2) = \bar{f}_2^{n-2}(2) - q^n \bar{f}_1^{n-1}(2) < \bar{f}_2^{n-2}(2)$$

leading to the general chain of inequalities

$$r(\lambda) \;=\; \bar{f}_1^n(2) \;\leq\; \bar{f}_2^{n-2}(2) \;\leq\; \ldots \leq \bar{f}_{n/2}^2(2) \;=\; a_1(\lambda)$$

which follows from positivity of q^k and positivity of the term $f_i^k(2) = a_0(\lambda)$ at $i = \lfloor (n-k+2)/2 \rfloor$. Thus, we obtain that

$$\hat{J}(\lambda) \;\geq\; \sigma^2/(2a_1(\lambda)a_0(\lambda)). \tag{34a}$$

According to (10d), it is easy to verify that

$$a_0(\lambda) \;=\; \prod_{k=1}^{n}(-\mu_k), \quad a_1(\lambda) \;=\; a_0(\lambda)\sum_{k=1}^{n} -1/\mu_k. \tag{34b}$$

In addition, from $\Re(-\mu_k) \geq \rho$, it follows that

$$\sum_{k=1}^{n} -1/\mu_k \;\leq\; n/\rho. \tag{34c}$$

Combining $\alpha\lambda = a_0(\lambda)$ with (34a), (34b), and (34c) completes the proof.

■

REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] B. V. Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions," *IEEE Control Syst. Lett.*, vol. 2, no. 1, pp. 49–54, 2018.

[11] A. Badithela and P. Seiler, "Analysis of the heavy-ball algorithm using integral quadratic constraints," in *Proceedings of the 2019 American Control Conference*. IEEE, 2019, pp. 4081–4085.

[12] 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.

[13] Y. Bengio, "Gradient-Based Optimization of Hyperparameters," *Neural Computation*, vol. 12, no. 8, pp. 1889–1900, 08 2000. [Online]. Available: https://doi.org/10.1162/089976600300015187

[14] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *Proc. ICML*, 2015, pp. 2113–2122.

[15] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh, "On optimal generalizability in parametric learning," in *NIPS*, 2017.

[16] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Ann. Oper. Res.*, vol. 46, no. 1, pp. 157–178, 1993.

[17] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, pp. 400–407, 1951.

[18] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[19] O. Devolder, "Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization," Ph.D. dissertation, Louvain-la-Neuve, 2013.

[20] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," *J. Optimiz. Theory App.*, vol. 171, no. 1, pp. 121–145, 2016.

[21] L. H. K. J. Arrow and H. Uzawa, "Studies in linear and non-linear programming," 1958.

[22] A. A. Brown and M. C. Bartholomew-Biggs, "Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations," *J. Optim. Theory Appl.*, vol. 62, no. 2, pp. 211–224, 1989.

[23] W. Su, S. Boyd, and E. Candes, "A differential equation for modeling nesterov's accelerated gradient method: theory and insights," *Proc. Neural Information Processing (NIPS)*, vol. 27, 2014.

[24] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.

[25] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *2011 50th IEEE conference on decision and control and European control conference*. IEEE, 2011, pp. 3800–3805.

[26] A. K. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal-dual dynamics," *Syst. Control. Lett.*, vol. 87, pp. 10–15, 2015.

[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, July 2019.

[28] 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 (7 pages), January 2021.

[29] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "A second order primal-dual method for nonsmooth convex composite optimization," *IEEE Trans. Automat. Control*, vol. 67, no. 8, pp. 4061–4076, August 2022.

[30] M. Muehlebach and M. Jordan, "A dynamical systems perspective on Nesterov acceleration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4656–4662.

[31] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms," 2022, arXiv:2209.11920v1.

[32] 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.

[33] B. V. Scoy and L. Lessard, "The speed-robustness trade-off for first-order methods with additive gradient noise," 2021, arXiv:2109.05059.

[34] K. Ogata, *Discrete-time control systems*. New Jersey: Prentice-Hall, 1994.

[35] M. Muehlebach and M. I. Jordan, "Continuous-time lower bounds for gradient-based algorithms," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.