The Fastest Known First-Order Method for Minimizing Twice Continuously Differentiable Smooth Strongly Convex Functions

Bryan Van Scoy, *Member, IEEE*, and Laurent Lessard, *Senior Member, IEEE*

***Abstract*— We consider iterative gradient-based optimization algorithms applied to functions that are smooth and strongly convex. The fastest globally convergent algorithm for this class of functions is the Triple Momentum (TM) method. We show that if the objective function is also twice continuously differentiable, a new, faster algorithm emerges, which we call C^2-Momentum (C2M). We prove that C2M is globally convergent and that its worst-case convergence rate is strictly faster than that of TM, with no additional computational cost. We validate our theoretical findings with numerical examples, demonstrating that C2M outperforms TM when the objective function is twice continuously differentiable.**

***Index Terms*— Optimization algorithms, robust control**

I. INTRODUCTION

WE CONSIDER the well-studied optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is continuously differentiable. A popular approach to solving (1), particularly when the dimension d is large, is to use iterative gradient-based methods, such as Gradient Descent (GD) and its accelerated variants.

A central question in the study of iterative methods is that of *worst-case convergence rate* over a class of functions $\mathcal{F}$. In this letter, we consider the *root-convergence factor* (also known as geometric convergence rate), denoted $\rho\in(0,1)$, a notion we make precise in Section II. Associated with the root-convergence factor are two important concepts:

a) Lower bounds: ρ is a *lower bound* for $\mathcal{F}$ if for any algorithm, there exists $f\in\mathcal{F}$ and an algorithm initialization such that the algorithm converges no faster than ρ.

b) Upper bounds: ρ is an *upper bound* for $\mathcal{F}$ if there exists an algorithm such that for all $f\in\mathcal{F}$ and algorithm initializations, the algorithm converges at least as fast as ρ.

If $\mathcal{F}$ has matching lower and upper bounds, this ρ and the corresponding algorithm that achieves it are said to be *minimax optimal* for $\mathcal{F}$.

This material is based upon work supported by the National Science Foundation under Grant No. 2347121.

B. Van Scoy is with the Department of Electrical and Computer Engineering, Miami University, Oxford, OH 45056, USA. (e-mail: bvanscoy@miamioh.edu)

L. Lessard is with the Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115, USA. (e-mail: l.lessard@northeastern.edu)

Generally, adding more structure to a function class, such as convexity or Lipschitz properties, makes the minimax rate faster because iterative algorithms can exploit the additional structure to converge more rapidly. We now provide a brief survey of different function classes and their minimax rates. The relationship between these classes is illustrated in the Venn diagram of Fig. 1.

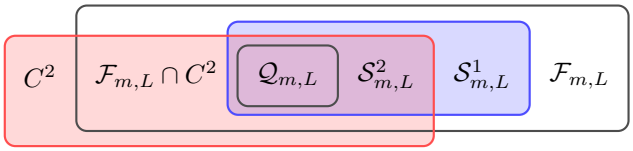

Fig. 1. Venn diagram of different function classes. Blue region: strongly convex functions. Red region: twice continuously differentiable functions. This letter focuses on the shaded intersection of these sets, $\boldsymbol{\mathcal{S}^2_{m,L}}$.

The class $\mathcal{F}_{m,L}$ consists of continuously differentiable functions with sector-bounded gradients. Specifically, there exists $x_\star\in\mathbb{R}^d$ (the optimal point) and constants $0<m\leq L$ such that $\big(L(x-x_\star)-\nabla f(x)\big)^{\mathsf{T}}\big(\nabla f(x)-m(x-x_\star)\big)\geq 0$ for all $x\in\mathbb{R}^d$. Functions in this class may be nonconvex but nevertheless have a unique local (and global) minimizer. The minimax rate for $\mathcal{F}_{m,L}$ is $\rho=\frac{\kappa-1}{\kappa+1}$ where $\kappa:=\frac{L}{m}$, and is achieved by GD with stepsize $\alpha=\frac{2}{L+m}$.

The class $\mathcal{S}^1_{m,L}$ consists of functions that have Lipschitz gradient with Lipschitz constant L and are strongly convex with parameter m. The superscript "1" indicates that $f\in C^1$, which follows from Lipschitz gradients. One can show that $\mathcal{S}^1_{m,L}\subset\mathcal{F}_{m,L}$. The minimax rate for $\mathcal{S}^1_{m,L}$ is $\rho=1-\frac{1}{\sqrt{\kappa}}$, and was recently proved in [1] using an exact characterization of $\mathcal{S}^1_{m,L}$ via interpolation conditions and the Performance Estimation paradigm [2]. The same lower bound was obtained in a parallel line of work by viewing algorithms as discrete-time Lur'e systems and applying integral quadratic constraints (IQCs) or dissipativity theory [3]–[5]. Specifically, the set $\mathcal{S}^1_{m,L}$ was over-approximated using Zames–Falb IQCs [6], [7], leading to an upper bound that turned out to be exact. The minimax rate for $\mathcal{S}^1_{m,L}$ is achieved by the Triple Momentum (TM) Method [8] and the Information Theoretic Exact Method (ITEM) [9].

The class $\mathcal{Q}_{m,L}\subset\mathcal{S}^1_{m,L}$ consists of quadratic functions of the form $f(x)=x^{\mathsf{T}}Qx+p^{\mathsf{T}}x+r$, with $mI_d\preceq Q\preceq LI_d$, and we have $\mathcal{Q}_{m,L}\subset\mathcal{S}^1_{m,L}$. The minimax rate for $\mathcal{Q}_{m,L}$ is

$\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. The lower bound was proved by Nemirovsky [10] and Nesterov [11, §2.1.4]. There are several minimax optimal methods for $\mathcal{Q}_{m,L}$, the simplest of which is Polyak's Heavy Ball (HB) method [12, §3.2.1]. Polyak used Lyapunov's indirect method to show that HB converges *locally* for any $f \in \mathcal{S}^1_{m,L}$ provided that f is twice continuously differentiable ($f \in C^2$) in a neighborhood of the optimal point. In other words, HB converges on $\mathcal{S}^1_{m,L}$ when initialized sufficiently close to the optimal point and enjoys the same fast rate as for $\mathcal{Q}_{m,L}$! If incorrectly initialized, HB need not converge at all on $\mathcal{S}^1_{m,L}$ [3], [8].

The aforementioned minimax optimal algorithms are described in Section II-A and summarized in Table I.

Polyak's observation raises an interesting possibility, which forms the starting point for the present work. If we consider the function class $\mathcal{S}^2_{m,L} \coloneqq \mathcal{S}^1_{m,L} \cap C^2$, then by Lyapunov's indirect method, any globally convergent method will converge at its *local rate*, which may be faster than the minimax rate of $\mathcal{S}^1_{m,L}$. This function class satisfies $\mathcal{Q}_{m,L} \subset \mathcal{S}^2_{m,L} \subset \mathcal{S}^1_{m,L}$ and may be characterized succinctly as functions whose Hessians satisfy $mI_d \preceq \nabla^2 f(x) \preceq LI_d$. Functions of interest in this category include regularized logistic loss, exponential family negative log-likelihoods with bounded natural parameters, and Moreau envelope smoothing of any $f \in \mathcal{S}^1_{m,L}$.

Our main result is a new algorithm, C^2-Momentum (C2M). We show that C2M achieves an upper bound of $\max\left\{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \rho_{\text{C2M}}\right\}$ on $\mathcal{S}^2_{m,L}$, where $\rho_{\text{C2M}} < 1 - \sqrt{\frac{2}{\kappa}}$. This corresponds to an iteration complexity that is faster than the minimax rate of $\mathcal{S}^1_{m,L}$ by a factor of $\sqrt{2}$.

Notable related works are the recent papers [13], [14], which use the same idea of optimizing the local convergence rate while enforcing global convergence. Specifically, these works develop re-tunings of HB and TM that converge globally on $\mathcal{F}_{m,L}$ but have optimized local rates because they also assume $f \in C^2$ locally near the optimal point.

The rest of this letter is organized as follows. In Section II we describe C2M, in Section III we prove convergence results, in Section IV we present some numerical results, and in Section V we discuss implications and future directions.

II. MAIN RESULT

In this section, we describe our proposed algorithm, state its main convergence result, and use root locus arguments to provide intuition behind the algorithm parameters.

A. Algorithm Form

We consider iterative first-order algorithms parameterized by $\alpha, \beta, \eta \in \mathbb{R}$ of the form

$$y_k = x_k + \eta\,(x_k - x_{k-1}) \tag{2a}$$
$$u_k = \nabla f(y_k) \tag{2b}$$
$$x_{k+1} = x_k + \beta\,(x_k - x_{k-1}) - \alpha\,u_k \tag{2c}$$

for $k \geq 0$ with initial conditions $x_0, x_{-1} \in \mathbb{R}^d$. We can interpret such an algorithm as a linear time-invariant (LTI) system G in feedback with the gradient ∇f, where the transfer function[1] from the gradient u_k to the point y_k at which the gradient is evaluated is

$$G(z) = g(z)I_d \quad \text{where} \quad g(z) = -\alpha\frac{(1+\eta)z - \eta}{(z-1)(z-\beta)}. \tag{3}$$

A minimal state-space realization of the reduced system g is

$$\left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right] = \left[\begin{array}{cc|c} 1+\beta & -\beta & -\alpha \\ 1 & 0 & 0 \\ \hline 1+\eta & -\eta & 0 \end{array}\right]. \tag{4}$$

Despite its simplicity, the form (2) can represent *all* algorithms referenced in Section I when α, β, η are suitably chosen (GD, HB, TM, ITEM, GHB, GAG). Table I shows parameters for the minimax methods discussed in Section I.

TABLE I
MINIMAX-OPTIMAL METHODS FOR SEVERAL FUNCTION CLASSES.

Function class	Minimax method	α	β	η	Minimax rate ρ
$\mathcal{F}_{m,L}$	GD	$\frac{1-\rho}{m}$	0	0	$\frac{\kappa-1}{\kappa+1}$
$\mathcal{S}^1_{m,L}$	TM [8]	$\frac{1+\rho}{L}$	$\frac{\rho^2}{2-\rho}$	$\frac{\rho^2}{(1+\rho)(2-\rho)}$	$1 - \frac{1}{\sqrt{\kappa}}$
$\mathcal{Q}_{m,L}$	HB [12]	$\frac{(1-\rho)^2}{m}$	ρ^2	0	$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

B. C^2-Momentum

Definition 1 (C2M): Given parameters $m, L, \rho \in \mathbb{R}$ with $0 < m \leq L$, $\kappa \coloneqq \frac{L}{m}$, and $\rho \in (0,1)$, the C^2-Momentum (C2M) algorithm is of the form (2) with parameters

$$\alpha = \tfrac{(1-\rho)^2}{m}, \qquad \beta = \tfrac{\rho}{\kappa-1}\big(1 - \tfrac{\kappa\,(1-3\rho)}{1+\rho}\big), \qquad \eta = \tfrac{\rho}{\kappa-1}\big(\tfrac{1+\rho}{(1-\rho)^2} - \tfrac{\kappa}{1+\rho}\big). \tag{5}$$

The C2M parameters depend on ρ, which we choose based on the condition number κ of the objective function:

$$\begin{cases} \rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} & \text{if } \kappa < 9 + 4\sqrt{5} \\ \rho \in \left(\rho_{\text{C2M}}, 1 - \sqrt{\frac{2}{\kappa}}\right] & \text{if } \kappa \geq 9 + 4\sqrt{5} \end{cases} \tag{6}$$

where ρ_{C2M} is the smallest positive root of the polynomial

$$\begin{aligned} p(\kappa,\rho) \coloneqq\ & 8\kappa(\kappa+1)\rho^7 - (23\kappa^2 + 18\kappa + 7)\rho^6 \\ & + 2(5\kappa^2 - 14\kappa - 7)\rho^5 + (31\kappa^2 + 50\kappa + 15)\rho^4 \\ & - 4(11\kappa^2 - 4\kappa - 11)\rho^3 + (23\kappa^2 - 30\kappa + 23)\rho^2 \\ & - 2(\kappa-1)(3\kappa+1)\rho + (\kappa-1)^2. \end{aligned} \tag{7}$$

When $\kappa < 9 + 4\sqrt{5}$, the parameters of C2M reduce to those of HB [12] in Table I. For $\kappa \geq 9 + 4\sqrt{5}$, we in general want to pick ρ as small as possible, but we will see that proving global asymptotic stability requires a strict inequality, so in practice we can choose $\rho = \rho_{\text{C2M}} + \varepsilon$ for some small $\varepsilon > 0$. The C2M stepsizes are defined in terms of the root ρ_{C2M} of the polynomial $p(\kappa,\rho)$ in (7). The following result (i) shows that this quantity is well defined in that the polynomial does

[1] As a slight abuse of notation, we use the same symbol to refer to both an LTI system and its transfer function.

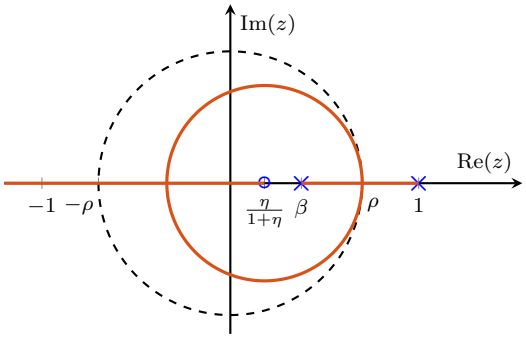

Fig. 2. Root locus of C2M. The locus has a double root at $z = \rho$ at gain m and a single root at $z = -\rho$ at gain L.

have a positive root, and (ii) provides bounds on this root that will be used in the analysis. The proof is in Appendix A.

Lemma 1: Suppose $\kappa \geq 9 + 4\sqrt{5}$. The polynomial $p(\kappa, \rho)$ defined in (7) has exactly one real root ρ_{C2M} in the open interval $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, 1 - \sqrt{\frac{2}{\kappa}}\right)$. Moreover, ρ_{C2M} is the smallest positive root and $p(\kappa, \rho) < 0$ for all $\rho \in \left(\rho_{\text{C2M}}, 1 - \sqrt{\frac{2}{\kappa}}\right]$.

C. Main Result

To describe our main result, we first define the root-convergence factor of an algorithm, which is a way to characterize its rate of convergence; see [15, §9.2].

Definition 2: Let $\{x_k\}$ be a sequence that converges to a point $x_\star$. Then, the root-convergence factor of $\{x_k\}$ is

$$\rho = \limsup_{k \to \infty} \|x_k - x_\star\|^{1/k}.$$

Moreover, the worst-case root-convergence factor of an algorithm (2) over a function class $\mathcal{F}$ is the supremum of the root-convergence factors over all sequences produced by the algorithm when applied to a function $f \in \mathcal{F}$.

We now state our main convergence result for C2M over the class $\mathcal{S}^2_{m,L}$. A full proof is included in Section III.

Theorem 1 (Upper bound for C2M): Consider the C2M method defined in (5) with parameter ρ chosen according to (6). An upper bound for the worst-case root-convergence factor of C2M over the function class $\mathcal{S}^2_{m,L}$ is ρ.

D. Root Locus Interpretation

Before rigorously analyzing the convergence of C2M, we first provide intuition behind the C2M parameters (5) using a root locus argument.

Consider the general algorithm (2) applied to a function $f \in \mathcal{Q}_{m,L} \subset \mathcal{S}^2_{m,L}$ with Hessian Q. By diagonalizing the Hessian, the iterates separate into d decoupled systems, each in (positive) feedback with an eigenvalue q_i of Q. Since the objective function is L-smooth, m-strongly convex, and twice continuously differentiable, its Hessian has eigenvalues in the interval $[m, L]$. Therefore, we can study worst-case local convergence by analyzing the eigenvalues of $A + qBC$ for $q \in [m, L]$. These closed-loop eigenvalues are solutions of the root locus $0 = 1 - q\,g(z)$ for $q \in [m, L]$. The parameters of C2M are the solutions to the following conditions:

1) The root locus passes through $z = -\rho$ when $q = L$.
2) The root locus has a double root at $z = \rho$ when $q = m$.

The visual reasoning for these two conditions is illustrated in Fig. 2, which shows the root locus of $1 - q\,g(z)$ as q varies. As $q \to 0$, the roots are the poles of $g(z)$, which are β and 1. These roots meet at $z = \rho$, circle around the zero at $z = \frac{\eta}{1+\eta}$, then break in on the negative real axis, with one root converging to the zero and the other going to $-\infty$ along the real axis. By enforcing the above two conditions, the root locus remains entirely inside the ρ-circle for all $q \in [m, L]$. In terms of the transfer function, these conditions are that

$$L\,g(-\rho) = 1, \qquad m\,g(\rho) = 1, \qquad \frac{\mathrm{d}g(z)}{\mathrm{d}z}\bigg|_{z=\rho} = 0, \tag{8}$$

where the last two equations are for the double root. Straightforward calculations show that the parameters (5) for C2M are the unique solution to the equations (8).

III. CONVERGENCE ANALYSIS

We now prove the main convergence result for C2M from Theorem 1. Our proof consists of two steps. First, we show that the algorithm is globally asymptotically stable, meaning that the iterates converge to the minimizer of f for all initial conditions. Once we have global convergence, we then show that the worst-case root-convergence factor is ρ by analyzing the linearization of the algorithm about its equilibrium.

A. Global Stability via Frequency-Domain Analysis

It is convenient to shift the dynamics of the algorithm (2) about its optimal point $x_\star$, which satisfies $\nabla f(x_\star) = 0$. To this effect, define $\tilde{x}_k \coloneqq x_k - x_\star$, $\tilde{y}_k \coloneqq y_k - x_\star$, $\tilde{u}_k = u_k$, and $\tilde{f}(y) \coloneqq f(y + x_\star)$. Then, we can rewrite (2) as:

$$\tilde{y}_k = \tilde{x}_k + \eta\,(\tilde{x}_k - \tilde{x}_{k-1}) \tag{9a}$$
$$\tilde{u}_k = \nabla \tilde{f}(\tilde{y}_k) \tag{9b}$$
$$\tilde{x}_{k+1} = \tilde{x}_k + \beta\,(\tilde{x}_k - \tilde{x}_{k-1}) - \alpha\,\tilde{u}_k \tag{9c}$$

Convergence of the algorithm G applied to f is therefore equivalent to convergence of G applied to $\tilde{f}$. In other words, we may assume without loss of generality that $x_\star = 0$.

To verify global asymptotic stability, we use integral quadratic constraints (IQCs) [16]. In discrete time, these are defined as follows (see [17]), where ℓ_2^n denotes the space of square-summable sequences on $\mathbb{R}^n$.

Definition 3: Signals $y \in \ell_2^{n_y}$ and $u \in \ell_2^{n_u}$ with associated z-transforms $\hat{y}(z)$ and $\hat{u}(z)$ satisfy the IQC defined by a measurable, bounded, and Hermitian matrix-valued function $\Pi : \mathbb{T} \to \mathbb{C}^{(n_y+n_u)\times(n_y+n_u)}$ if

$$\int_{\mathbb{T}} \begin{bmatrix} \hat{y}(z) \\ \hat{u}(z) \end{bmatrix}^* \Pi(z) \begin{bmatrix} \hat{y}(z) \\ \hat{u}(z) \end{bmatrix} \mathrm{d}z \geq 0, \tag{10}$$

where $\mathbb{T} \coloneqq \{z \in \mathbb{C} \mid |z| = 1\}$ is the unit circle in the complex plane. A bounded operator $\Delta : \ell_2^{n_y} \to \ell_2^{n_u}$ satisfies the IQC defined by Π if (10) holds for all $y \in \ell_2^{n_y}$ with $u = \Delta(y)$.

It is well known (see for example [17], [18]) that the gradient of a smooth strongly convex function can be described using IQCs.

Proposition 1: The operator $\Delta : \ell_2^d \to \ell_2^d$ defined by $(\Delta(y))_k := \nabla\tilde{f}(y_k)$ for all $k \geq 0$ and $y \in \ell_2^d$, where $\tilde{f} \in \mathcal{S}_{m,L}^1$ and $\nabla\tilde{f}(0) = 0$ satisfies the O'Shea–Zames–Falb IQC $\Pi_{m,L} \otimes I_d$, where

$$\Pi_{m,L} := \begin{bmatrix} -mL(2-h-h^*) & L(1-h^*)+m(1-h) \\ L(1-h)+m(1-h^*) & -(2-h-h^*) \end{bmatrix}$$

and $h(z)$ is any transfer function with impulse response $\{h_k\}$ satisfying $\|h\|_1 = \sum_{k=-\infty}^{\infty} |h_k| \leq 1$ and $h_k \geq 0$ for all k.

While Δ satisfies the IQC $\Pi_{m,L} \otimes I_d$, to analyze the interconnection of the algorithm G with Δ using the main IQC theorem (see Proposition 2), we will first need to perform a loop transformation (see, e.g., [19, §6.6]) so that the zero operator is contained in the class of transformed uncertainties. Doing so, the feedback interconnection of G and Δ is equivalent to the feedback interconnection of $\tilde{G}$ and $\tilde{\Delta}$, where

$$\tilde{G}(z) = \tilde{g}(z) \otimes I_d \quad \text{where} \quad \tilde{g}(z) = \frac{\frac{L-m}{2} g(z)}{1 - \frac{L+m}{2} g(z)}$$

and the transformed operator is given by

$$\tilde{\Delta}(x) = \tfrac{2}{L-m}\big(\Delta(x) - \tfrac{L+m}{2} x\big).$$

Using properties of shifting and scaling the gradient of smooth strongly convex functions [2, §2.4], $\tilde{\Delta}$ satisfies the IQC $\Pi_{-1,1} \otimes I_d$ if and only if Δ satisfies the IQC $\Pi_{m,L} \otimes I_d$. We are now ready to apply the following main IQC result.

Proposition 2 (Discrete-time IQC result [17, Thm. 2]): Fix $\rho \in (0,1)$. Suppose that $\tilde{G}$ is stable, $\tilde{\Delta}$ is a bounded causal operator, and

(i) the interconnection of $\tilde{G}$ and $\tilde{\Delta}$ is well-posed,
(ii) for every $\tau \in [0,1]$, $\tau\tilde{\Delta}$ satisfies the IQC Π, and
(iii) the following frequency-domain inequality holds:

$$\begin{bmatrix} \tilde{G}(z) \\ I \end{bmatrix}^* \Pi(z) \begin{bmatrix} \tilde{G}(z) \\ I \end{bmatrix} < 0 \quad \text{for all } z \in \mathbb{T}.$$

Then the feedback interconnection of $\tilde{G}$ and $\tilde{\Delta}$ is stable.

Applying Proposition 2 therefore yields the following.

Proposition 3: Algorithm (2) is globally asymptotically stable for all $f \in \mathcal{S}_{m,L}^1$ if $\big(1 - \frac{L+m}{2} g(z)\big)^{-1}$ is stable and the following frequency-domain inequality (FDI) holds:

$$\begin{bmatrix} g(z) \\ 1 \end{bmatrix}^* \Pi_{m,L}(z) \begin{bmatrix} g(z) \\ 1 \end{bmatrix} < 0 \quad \text{for all } z \in \mathbb{T}, \qquad (11)$$

where $h(z)$ satisfies $\|h\|_1 \leq 1$ and $h_k \geq 0$ for all $k \in \mathbb{Z}$.

Proof: The stability condition is equivalent to stability of $\tilde{G}$. It is straightforward to verify that the interconnection of $\tilde{G}$ and $\tilde{\Delta}$ is well-posed and that $\tau\tilde{\Delta}$ satisfies the IQC $\Pi = \Pi_{-1,1} \otimes I_d$ for all $\tau \in [0,1]$. Therefore, the first two conditions in Proposition 2 hold for the transformed system $\tilde{G}$ and the IQC Π. It remains to show that the FDI in (iii) is equivalent to that in (11). To that end, we first write the numerator and denominator of $\tilde{g}$ as

$$\begin{bmatrix} \frac{L-m}{2} g \\ 1 - \frac{L+m}{2} g \end{bmatrix} = \begin{bmatrix} \frac{L-m}{2} & 0 \\ -\frac{L+m}{2} & 1 \end{bmatrix} \begin{bmatrix} g \\ 1 \end{bmatrix} = M \begin{bmatrix} g \\ 1 \end{bmatrix}.$$

Using this relationship along with $M^{\mathsf{T}} \Pi_{-1,1} M = \Pi_{m,L}$, the FDI in (iii) of Proposition 2 is

$$\begin{bmatrix} \tilde{g} \\ 1 \end{bmatrix}^* \Pi_{-1,1} \begin{bmatrix} \tilde{g} \\ 1 \end{bmatrix} = \frac{1}{|1 - \frac{L+m}{2} g|^2} \begin{bmatrix} g \\ 1 \end{bmatrix}^* M^{\mathsf{T}} \Pi_{-1,1} M \begin{bmatrix} g \\ 1 \end{bmatrix}.$$

Therefore, the FDI in (iii) is equivalent to that in (11). From Proposition 2, the interconnection of $\tilde{G}$ and $\tilde{\Delta}$ is stable, which via loop shifting implies the interconnection of G and Δ is stable. Finally, (input-output) stability means that all signals have bounded norms. Therefore, $\|\tilde{x}\| < \infty \implies \lim_{k\to\infty} \tilde{x}_k = 0 \implies \lim_{k\to\infty} x_k = x_\star$. ∎

We use Proposition 3 with $h(z) = z^{-1}$ to show that C2M is convergent by directly verifying the FDI (11). First, the condition that $\big(1 - \frac{L+m}{2} g(z)\big)^{-1}$ is bounded follows from the root locus argument in Section II-D; see the following Section III-B for a more rigorous argument. Letting $z = x + i\sqrt{1-x^2}$ for $x \in [-1,1]$ and substituting the C2M parameters, it is straightforward to verify that the FDI is satisfied when $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ and $\kappa < 9 + 4\sqrt{5}$. In the other case with $\kappa > 9 + 4\sqrt{5}$, the FDI reduces to the inequality

$$\begin{aligned} 0 > &-4\rho\left(\kappa(1-\rho)^2 - (1+\rho)\right) x^2 \\ &- 2(1-\rho)\left(\kappa(1-\rho)^2(1+2\rho) - (1+\rho)^2\right) x \\ &- (1+\rho)\left(\kappa(1-\rho)^2(1-4\rho+\rho^2) + 6\rho - 2\rho^3\right). \end{aligned} \qquad (12)$$

The right-hand side of (12) is a quadratic in x. To show that this inequality holds, we will use the following.

Lemma 2: Suppose $\rho \in [0,1]$ and $\kappa > 1$. Then,

$$\rho > \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \quad \iff \quad \kappa(1-\rho)^2 < (1+\rho)^2,$$
$$\rho < 1 - \sqrt{\frac{2}{\kappa}} \quad \iff \quad \kappa(1-\rho)^2 > 2.$$

Since $\rho < 1 - \sqrt{\frac{2}{\kappa}}$ by assumption, it follows from Lemma 2 that $\kappa(1-\rho)^2 > 2 > 1 + \rho$. Therefore, the leading coefficient of the quadratic is negative. Maximizing the right-hand side of (12), this inequality holds if

$$\frac{p(\kappa,\rho)}{4\rho(\kappa(1-\rho)^2 - (1+\rho))} < 0,$$

where $p(\kappa,\rho)$ is the polynomial in (7). The denominator is positive from the prior argument. Moreover, $p(\kappa,\rho)$ is negative for any $\rho \in \big(\rho_{\text{C2M}}, 1 - \sqrt{\frac{2}{\kappa}}\big]$ by Lemma 1, so the FDI is satisfied. Therefore, C2M is globally asymptotically stable for any ρ satisfying (6).

B. Local Convergence

While the root locus interpretation provides intuition behind the local convergence of C2M, we now use Lyapunov's indirect method along with the Jury criterion to systematically prove local convergence; see [20] for similar analyses in other settings. We begin by characterizing the worst-case root convergence factor in terms of the system matrices.

Lemma 3: The worst-case root-convergence factor of the algorithm (2) over the function class $\mathcal{S}_{m,L}^2$ is the maximum spectral radius of $A + qBC$ over $q \in [m, L]$.

Proof: From the linear convergence theorem [15, Thm. 10.1.4], the root-convergence factor of the algorithm (2) is the spectral radius of its linearization evaluated at the equilibrium. In particular, let $\{x_k\}$ denote the sequence produced by applying algorithm (2) to a function $f \in \mathcal{S}^2_{m,L}$ for some initial conditions $x_0, x_{-1} \in \mathbb{R}^d$. Let Q denote the Hessian of f evaluated at the minimizer of f. Then the root-convergence factor of the sequence $\{x_k\}$ is the spectral radius of the linearization $A \otimes I_d + (BC) \otimes Q$, where $\otimes$ denotes the Kronecker product. Since Q is real and symmetric, it is diagonalizable. Applying this diagonalization to the linearized system yields $A \otimes I_d + (BC) \otimes \operatorname{diag}(q_1, \ldots, q_d)$, where $q_1, \ldots, q_d$ are the eigenvalues of Q. Therefore, the worst-case root-convergence factor over the function class $\mathcal{S}^2_{m,L}$ is the maximum spectral radius of $A + qBC$ over $q \in [m, L]$. ∎

Based on Lemma 3, we can characterize the worst-case root-convergence factor using the eigenvalues of $A+qBC$. We next analyze these eigenvalues using the Jury criterion. Recall that a polynomial $z^2 + a_1\, z + a_0$ with real coefficients has roots in the closed unit disk if and only if [21, §4.5][2]

$$1 + a_1 + a_0 \geq 0, \quad 1 - a_1 + a_0 \geq 0, \quad |a_0| \leq 1. \tag{13}$$

The characteristic polynomial of the closed-loop system matrix $A + qBC$ is the quadratic $\chi(z) = z^2 + (q\alpha(1+\eta) - (1+\beta))z + (\beta - q\alpha\eta)$. Applying the Jury criterion (13) to the scaled polynomial $\chi(\rho z)$, the closed-loop eigenvalues are in the closed ρ-disk if and only if

$$\begin{aligned} (1-\rho)(\beta-\rho) + \alpha\,(\eta\rho - \eta + \rho)q \geq 0, \\ (1+\rho)(\beta+\rho) - \alpha\,(\eta\rho + \eta + \rho)q \geq 0, \\ \rho^2 + \beta - \alpha\eta q \geq 0, \qquad \rho^2 - \beta + \alpha\eta q \geq 0, \end{aligned}$$

for all $q \in [m, L]$. Since each inequality is linear in q, it suffices to enforce the inequality at the endpoints $q \in \{m, L\}$. Substituting the C2M parameters, this system of inequalities reduces to

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \leq \rho \leq \frac{\kappa - 1}{\kappa + 1}.$$

The lower bound on ρ is the minimax rate for $\mathcal{Q}_{m,L}$. Since all parameters ρ in (6) satisfy these conditions, we have that all eigenvalues of $A+qBC$ are in the ρ-disk. Therefore, from Lemma 3, the parameter ρ is the worst-case root-convergence factor of C2M, which completes the proof of Theorem 1.

C. Iteration Complexity

It is common in optimization to characterize algorithm convergence using *iteration complexity* [11, §1.1.2]. Iteration complexity is an expression for how the worst-case number of iterations N required to reach a specified error ε scales as a function of problem parameters such as κ, expressed asymptotically as $\varepsilon \to 0$ and $\kappa \to \infty$. If the convergence rate is ρ as defined in Definition 2, then $\|x_k - x_\star\| \leq c(k)\rho^k$, where $c(k)$ grows sub-exponentially in k. We seek the smallest N such that $c(N)\rho^N \leq \varepsilon$. Rearranging, we obtain $\log c(N) + N \log \rho \leq \log \varepsilon$. Since $c(N)$ is sub-exponential, it is dominated by the linear term in N as $\varepsilon \to 0$ (and therefore $N \to \infty$), so we neglect it. We are left with $N \geq \frac{-1}{\log \rho} \log \frac{1}{\varepsilon}$. Next, we expand $\frac{-1}{\log \rho}$ as a function of $\kappa \to \infty$, keeping only the most significant term. For example, if $\rho = 1 - \frac{c}{\sqrt{\kappa}}$,

$$\frac{-1}{\log \rho} = \frac{-1}{\log\left(1 - \frac{c}{\sqrt{\kappa}}\right)} = \frac{\sqrt{\kappa}}{c} - \frac{1}{2} - \frac{c}{12\sqrt{\kappa}} - \cdots \approx \frac{\sqrt{\kappa}}{c}.$$

Based on the minimax rate of TM in Table I ($c = 1$), we conclude that $N_{\text{TM}} \gtrsim \sqrt{\kappa} \log \frac{1}{\varepsilon}$. Similarly, for HB, we have $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}+1} \approx 1 - \frac{2}{\sqrt{\kappa}}$ ($c = 2$), so $N_{\text{HB}} \gtrsim \frac{\sqrt{\kappa}}{2} \log \frac{1}{\varepsilon}$.

For C2M, we do not have a nice expression for ρ_{C2M}, but we can nevertheless find an asymptotic analytic expansion for it about $\kappa \to \infty$, which leads to the bounds

$$1 - \sqrt{\frac{2}{\kappa}} - \frac{1 + 2\sqrt{2}}{4\kappa} < \rho_{\text{C2M}} < 1 - \sqrt{\frac{2}{\kappa}}.$$

Therefore, $c = \sqrt{2}$ and $N_{\text{C2M}} \gtrsim \frac{\sqrt{\kappa}}{\sqrt{2}} \log \frac{1}{\varepsilon}$. In other words, C2M is faster than TM by a factor of $\sqrt{2}$.

In contrast, GD has iteration complexity $N_{\text{GD}} \gtrsim \frac{\kappa}{2} \log \frac{1}{\varepsilon}$. In the optimization literature, methods with the $\sqrt{\kappa}$ factor instead of merely κ are called *accelerated methods*. We can visualize iteration complexity by plotting $\frac{-1}{\log \rho}$ versus κ on a log-log scale (we omit the $\log \frac{1}{\varepsilon}$ factor); see Fig. 3. We also included a plot for GD (see Table I).

We see in Fig. 3 that non-accelerated methods (GD, GAG) have an asymptotic slope of 1 whereas accelerated methods (C2M, TM) have an asymptotic slope of $\frac{1}{2}$.

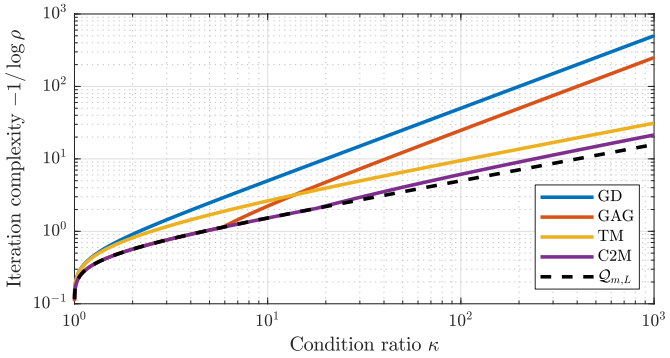

Fig. 3. Iteration complexity of several iterative methods applied to $\boldsymbol{\mathcal{S}^2_{m,L}}$. The proposed C2M method outperforms TM [8], which is minimax optimal on $\boldsymbol{\mathcal{S}^1_{m,L}}$, by exploiting a faster local convergence rate. Similarly, GAG [14] outperforms GD, which is minimax optimal on $\boldsymbol{\mathcal{F}_{m,L}}$.

IV. NUMERICAL VALIDATION

We simulate our proposed algorithm C2M along with several other first-order methods on a function chosen to showcase worst-case behavior. We used the function [8, §IV]

$$f(x) = (L - m) \sum_{i=1}^{p} g(a_i^{\mathsf{T}} x - b_i) + \frac{m}{2} \|x\|^2,$$

where $g(w)$ is $\frac{1}{2}w^2 e^{-r/w}$ if $w > 0$ and zero if $w \leq 0$. When $r > 0$ and $0 < m \leq L$ and $\left\| \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} \right\| = 1$, such

[2]The reference states the results for the roots to be contained in the *open* unit disk, which is described by the corresponding *strict* inequalities. Since the roots of a polynomial depend continuously on its coefficients, the corresponding result for the closed unit disk holds with non-strict inequalities.

functions satisfy $f \in \mathcal{S}^2_{m,L}$. We chose the parameters $L = 1$, $m = 10^{-3}$, $r = 10^{-3}$, $p = 2$, $a_1 = (1, 0)$, $a_2 = (0, 0.002)$, and $b_1 = b_2 = 100$. All methods were initialized at $x_0 = 0$.

In Fig. 4, we plot error as a function of iteration. The function f elicits worst-case behavior from GD, HB, and TM. In other words, GD and TM converge at their respective minimax rates for $\mathcal{F}_{m,L}$ and $\mathcal{S}^1_{m,L}$. Since $f \notin \mathcal{Q}_{m,L}$, HB is only locally convergent. In our simulation, we see that HB does not converge; however, if we were to initialize HB sufficiently close to $x_\star$, then it would converge at least as fast as the minimax rate for $\mathcal{Q}_{m,L}$. Our proposed C2M exploits additional smoothness in the objective to converge globally at a rate that is always faster than the minimax $\mathcal{S}^1_{m,L}$ rate. Likewise, GAG, which is globally convergent on $\mathcal{F}_{m,L}$, is slightly faster than GD, which is minimax-optimal on $\mathcal{F}_{m,L}$.

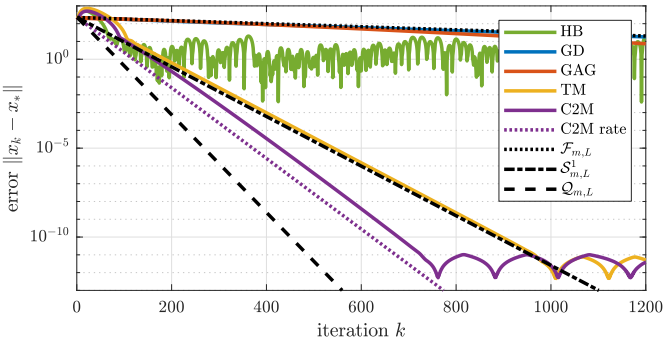

Fig. 4. Simulation results for a function $f \in \mathcal{S}^2_{m,L}$ (see Section IV). Solid lines are simulation results for the specified method; black lines are minimax rates for different function classes (see Table I); the dotted purple line is our theoretical upper bound (worst-case) rate for C2M.

V. DISCUSSION

The proposed C2M algorithm is the first method, to the best of the authors' knowledge, that is designed specifically for the function class $\mathcal{S}^2_{m,L}$. The minimax rate for this function class, however, is not known, in contrast to the function classes $\mathcal{S}^1_{m,L}$ and $\mathcal{Q}_{m,L}$. Finding this minimax rate or even lower bounds are interesting open problems.

The parameters of C2M are related to two other algorithms from the literature. As we have already seen, C2M reduces to HB when $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Moreover, the general C2M parameters are identical (after appropriate transformations) to those of GAG [14, Cor. 1.1]. This makes sense, since the work [14] also considers the family of algorithms (2) and is optimizing for local convergence. The two cases differ, however, in the choice of ρ, since GAG is optimized over the function class $\mathcal{F}_{m,L}$ defined in Section I rather than $\mathcal{S}^2_{m,L}$.

APPENDIX

A. Proof of Lemma 1

We apply Sturm's theorem [22, Thm. 2.62] to $p(\kappa, \rho)$ as a polynomial in ρ. Define the Sturm sequence

$$p_0 = p, \quad p_1 = \frac{\mathrm{d}p}{\mathrm{d}\rho}, \quad p_{i+1} = -\mathrm{rem}(p_{i-1}, p_i) \text{ for } i \geq 1,$$

where rem() denotes the remainder after polynomial division (considered as polynomials in ρ), and the sequence terminates when p_i is constant, which occurs for $i \leq 7$ since p is degree 7 in ρ. Evaluating the Sturm sequence at $\rho = 0$ and $\rho = 1$ yields 5 sign changes and 3 sign changes, respectively. Therefore, there are two real roots in the interval $(0, 1)$. Moreover, p is positive when $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, negative when $\rho = 1 - \sqrt{\frac{2}{\kappa}}$, and positive when $\rho = 1$. By the intermediate value theorem, we conclude that there is exactly one real root in each interval $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, 1 - \sqrt{\frac{2}{\kappa}}\right)$ and $\left(1 - \sqrt{\frac{2}{\kappa}}, 1\right)$, and the value of p is negative for all $\rho \in \left(\rho_{\mathrm{C2M}}, 1 - \sqrt{\frac{2}{\kappa}}\right]$. ■

REFERENCES

[1] Y. Drori and A. Taylor, "On the oracle complexity of smooth strongly convex minimization," *J. Complexity*, vol. 68, p. 101590, 2022.

[2] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first-order methods," *Math. Program.*, vol. 161, pp. 307–345, 2017.

[3] 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.

[4] L. Lessard, "The analysis of optimization algorithms: A dissipativity approach," *IEEE Control Syst. Mag.*, vol. 42, no. 3, pp. 58–72, Jun. 2022.

[5] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimisation algorithms: An integral quadratic constraint approach," *Int. J. Control*, vol. 94, no. 11, pp. 2956–2979, 2021.

[6] C. Scherer and C. Ebenbauer, "Convex synthesis of accelerated gradient algorithms," *SIAM J. Control Optim.*, vol. 59, no. 6, pp. 4615–4645, 2021.

[7] L. Lessard and P. Seiler, "Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate," in *Proc. Amer. Control Conf.*, Jul. 2020, pp. 119–125.

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

[9] A. Taylor and Y. Drori, "An optimal gradient method for smooth strongly convex minimization," *Math. Program.*, vol. 199, no. 1, pp. 557–594, 2023.

[10] A. S. Nemirovsky, "Information-based complexity of linear operator equations," *J. Complexity*, vol. 8, no. 2, pp. 153–175, 1992.

[11] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[12] 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.

[13] V. Ugrinovskii, I. Petersen, and I. Shames, "Global convergence and asymptotic optimality of the heavy ball method for a class of nonconvex optimization problems," *IEEE Control Syst. Lett.*, vol. 6, pp. 2449–2454, 2022.

[14] A. X. Wu, I. R. Petersen, V. Ugrinovskii, and I. Shames, "A generalized accelerated gradient optimization method," in *Proc. Amer. Control Conf.*, 2024, pp. 1904–1908.

[15] J. Ortega and W. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Academic Press, 1970.

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

[17] M. Fetzer and C. W. Scherer, "Absolute stability analysis of discrete time feedback interconnections," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8447–8453, 2017.

[18] W. Heath and A. Wills, "Zames–Falb multipliers for quadratic programming," in *Proc. IEEE Conf. Decis. Control*, 2005, pp. 963–968.

[19] H. K. Khalil, *Nonlinear systems*. Upper Saddle River, N.J.: Prentice Hall, 2002.

[20] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms," *IEEE Trans. Autom. Control*, vol. 70, no. 2, pp. 889–904, 2025.

[21] M. Fadali and A. Visioli, *Digital Control Engineering: Analysis and Design*. Academic Press, 2009.

[22] S. Basu, R. Pollack, and M. Coste-Roy, *Algorithms in Real Algebraic Geometry*, ser. Algorithms Comput. Math. Springer, 2007.