The Fastest Known Globally Convergent First-Order Method for Minimizing Strongly Convex Functions

Bryan Van Scoy, *Member, IEEE*, Randy A. Freeman, *Senior Member, IEEE*, and Kevin M. Lynch, *Fellow, IEEE*

***Abstract*—We design and analyze a novel gradient-based algorithm for unconstrained convex optimization. When the objective function is m-strongly convex and its gradient is L-Lipschitz continuous, the iterates and function values converge linearly to the optimum at rates ρ and ρ^2, respectively, where $\rho = 1 - \sqrt{m/L}$. These are the fastest known guaranteed linear convergence rates for globally convergent first-order methods, and for high desired accuracies the corresponding iteration complexity is within a factor of two of the theoretical lower bound. We use a simple graphical design procedure based on integral quadratic constraints to derive closed-form expressions for the algorithm parameters. The new algorithm, which we call the triple momentum method, can be seen as an extension of methods such as gradient descent, Nesterov's accelerated gradient descent, and the heavy-ball method.**

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

I. Introduction

CONSIDER the optimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, strongly convex with parameter m, and has a Lipschitz continuous gradient with Lipschitz constant L. Since f is strongly convex, it has a unique global minimizer $x_\star \in \mathbb{R}^n$. We consider first-order (gradient-based) algorithms to solve (1).

Perhaps the simplest algorithm which solves (1) is gradient descent with constant step size, which has the form

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \quad x_0 \in \mathbb{R}^n.$$

Using $\alpha = 2/(L+m)$, the iterates converge globally and linearly to the optimizer with rate $(L-m)/(L+m)$.[1]

Due to the slow convergence of gradient descent, many methods have been proposed to obtain faster convergence. In general, faster convergence rates can be achieved by introducing *momentum*. Examples of methods which incorporate momentum include the heavy-ball method [1],

$$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f(x_k),$$

and Nesterov's accelerated gradient descent [2],

$$\begin{aligned} x_{k+1} &= y_k - \alpha \nabla f(y_k) \\ y_k &= (1+\beta)x_k - \beta x_{k-1}. \end{aligned}$$

It remains an open question how to choose the parameters α and β to achieve global convergence while optimizing the convergence rate. For the heavy-ball method, one can choose parameters to achieve a *local* convergence rate of $(\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$, but the resulting method does not converge globally [3]. For other parameter choices the method converges globally to the optimizer with a linear rate, although a tight bound on the rate has not been found [4].

In his book [2], Nesterov gives several choices of both constant and time-varying parameters which guarantee that the function values generated by his algorithm converge with linear rate $1 - \sqrt{m/L}$ if f is strongly convex and sublinearly as $\mathcal{O}(1/k^2)$ if f is weakly convex.[2] The derived bound on the corresponding iteration complexity (i.e., the number of iterations required to minimize the objective function to within a given tolerance) is proportional to a theoretical lower bound, so his method is often called optimal [2, Th. 2.2.2]. It has recently been shown, however, that other algorithms have smaller bounds on the iteration complexity when the objective function is weakly convex [5], [6].

[1]Throughout this letter, the phrase "linear convergence with rate ρ" means R-linear convergence after some finite iteration, i.e., having errors bounded by $c\rho^k$ for some constant $c > 0$ and for all $k \geq k_0 \geq 0$ where k_0 is finite.

[2]Throughout this letter, "weakly convex" means convex but not necessarily strongly convex.

Manuscript received March 6, 2017; revised May 22, 2017; accepted June 16, 2017. Date of publication July 3, 2017; date of current version July 21, 2017. This work was supported by the Office of Naval Research under Project N00014-16-1-2106. Recommended by Senior Editor G. Yin. *(Corresponding author: Bryan Van Scoy.)*

B. Van Scoy is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: bryanvanscoy2012@u.northwestern.edu).

R. A. Freeman is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA, and also with the Northwestern Institute on Complex Systems, Evanston, IL 60208 USA (e-mail: freeman@eecs.northwestern.edu).

K. M. Lynch is with the Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA, and also with the Northwestern Institute on Complex Systems, Evanston, IL 60208 USA (e-mail: kmlynch@northwestern.edu).

Digital Object Identifier 10.1109/LCSYS.2017.2722406

TABLE I
PARAMETERS OF OPTIMIZATION ALGORITHMS IN THE FORM OF EQ. (2) (UP TO A CHANGE OF VARIABLES)

Method	Parameters $(\alpha, \beta, \gamma, \delta)$
Gradient descent	$(\alpha, 0, 0, 0)$
Heavy-ball method [1], [4]	$(\alpha, \beta, 0, 0)$
Nesterov's accelerated gradient descent [2]	$(\alpha, \beta, \beta, 0)$
Algorithm in [3, Eq. 6.1]	$(\alpha, \beta, \gamma, 0)$
Triple momentum method (Defn. 2)	$(\alpha, \beta, \gamma, \delta)$

To gain intuition into the acceleration process, other accelerated methods have recently been designed based on both geometric descent [7] and optimal quadratic averaging [8]. Both methods achieve the same rate as Nesterov's method when the objective function is strongly convex.

In this letter, we develop a novel algorithm to solve (1) for strongly convex objective functions with known parameters m and L. Our algorithm, called the triple momentum (TM) method, uses three momentum terms to achieve global linear convergence to the optimizer with the fastest known rate bound, improving on Nesterov's bound by a factor of two. We give the constant algorithm parameters in closed-form. Inspired by [3], we use integral quadratic constraints from robust control to design our algorithm, although we provide convergence proofs which do not rely on control theory.

We describe our algorithm in Section II and prove the error bound in Section III. We verify our algorithm with simulations in Section IV, and conclude in Section V. Furthermore, we motivate the design of the TM method using integral quadratic constraints in the Appendix.

Notation: $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ denotes the 2-norm. Define ℓ_{2e}^n as the set of all one-sided sequences $x : \mathbb{N} \to \mathbb{R}^n$. The unit circle in the complex plane is denoted $\mathbb{T}$. The angle of a complex number is denoted $\angle(re^{j\theta}) = \theta$.

II. MAIN RESULT

Many gradient-based algorithms for solving problem (1) can be written using the recursion

$$\begin{aligned} \xi_{k+1} &= (1+\beta)\xi_k - \beta\xi_{k-1} - \alpha\nabla f(y_k) \\ y_k &= (1+\gamma)\xi_k - \gamma\xi_{k-1} \\ x_k &= (1+\delta)\xi_k - \delta\xi_{k-1} \end{aligned} \tag{2}$$

where $\xi \in \ell_{2e}^n$ is the internal state, the gradient is applied to $y \in \ell_{2e}^n$, the output is $x \in \ell_{2e}^n$, and $\xi_0, \xi_{-1} \in \mathbb{R}^n$ are the initial conditions. In this letter we assume the parameters α, β, γ, and δ are constant (i.e., they do not change with k). Table I shows how some known methods are of the form (2) with particular constraints on these parameters. For comparison, we plot the convergence rates of the iterates in Fig. 1.

Definition 1 (Function Class): Define $\mathcal{S}_{m,L}$ to be the set of functions $f : \mathbb{R}^n \to \mathbb{R}$ that are continuously differentiable, strongly convex with parameter m, and have Lipschitz gradients with Lipschitz constant L. Furthermore, $\kappa = L/m$ is called the *condition number* of $f \in \mathcal{S}_{m,L}$.

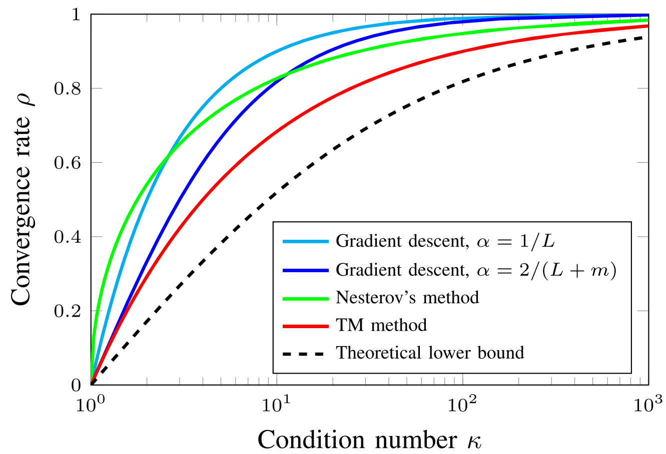


Fig. 1. Convergence rates of gradient-based optimization algorithms for $f \in \mathcal{S}_{m,L}$. Shown are gradient descent with $\alpha = 1/L$ (cyan) and $\alpha = 2/(L+m)$ (blue), Nesterov's method with $\alpha = 1/L$ and $\beta = (\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$ (green), and the TM method (red). Nesterov's lower bound (dashed black) is also shown. The heavy-ball method with $\alpha = 4/(\sqrt{L} + \sqrt{m})^2$ and $\beta = (\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$ converges *locally* with rate equal to the lower bound, but does not converge *globally*.

Definition 2 (TM Method): Let $\rho = 1 - 1/\sqrt{\kappa}$. We call the algorithm in (2) with constant parameters

$$(\alpha, \beta, \gamma, \delta) = \left(\frac{1+\rho}{L}, \frac{\rho^2}{2-\rho}, \frac{\rho^2}{(1+\rho)(2-\rho)}, \frac{\rho^2}{1-\rho^2}\right) \tag{3}$$

the *triple momentum method* (or TM method).

We now state our main theorem which gives error bounds for the TM method. The proof is in Section III.

Theorem 1 (Triple Momentum Method): Let $f \in \mathcal{S}_{m,L}$ with $0 < m \le L$ and let $x_\star \in \mathbb{R}^n$ be the unique minimizer of f. For any initial condition $\xi_0, \xi_{-1} \in \mathbb{R}^n$, the TM method produces iterates which satisfy

$$\|x_k - x_\star\| \le c\,\rho^k \tag{4}$$

$$f(x_k) - f(x_\star) \le c^2\,\frac{L}{2}\,\rho^{2k} \tag{5}$$

for all $k \ge 1$ where $\rho = 1 - 1/\sqrt{\kappa}$ and

$$c = \rho^{-1}\left(\|x_1 - x_\star\|^2 - \frac{1}{mL}p_m(y_0)^T p_L(y_0)\right)^{1/2} \tag{6}$$

with $p_r(y) = \nabla f(y) - r\,(y - x_\star)$.

Corollary 1: In Thm. 1, ξ also converges to $x_\star$ with rate ρ.

Proof: The transfer function from ξ to x is $(1+\delta) - \delta\,z^{-1}$. The inverse of this transfer function, i.e., the transfer function from x to ξ, has a pole at $\delta/(1+\delta) = \rho^2$ which is no greater than ρ for any $\kappa \ge 1$. Since the decay rate of this pole is faster than that of x, then ξ converges with the same rate as x. The transfer function from x to ξ has unit dc gain and x converges to $x_\star$ with rate ρ, so ξ also converges to $x_\star$ with rate ρ. ■

We now use the bound for the error of the iterates to establish the corresponding iteration complexity. Suppose we have a bound of the form $\|x_k - x_\star\| \le c\,\rho^k$. Then the number of iterations k_ϵ required to guarantee that $\|x_k - x_\star\| \le \epsilon$ for all $k \ge k_\epsilon$ is

$$k_\epsilon = -\frac{\ln(c/\epsilon)}{\ln\rho}. \tag{7}$$

TABLE II

APPROXIMATE ITERATIONS TO OBTAIN $\|x_k - x_\star\| \le \epsilon$ FOR GRADIENT OPTIMIZATION ALGORITHMS FOR LARGE κ. FOR EACH METHOD, $\tilde{c}$ IS A CONSTANT WHICH REMAINS BOUNDED AS $\kappa \to \infty$

Method	Iterations to converge
Gradient descent, $\alpha = 1/L$	$\ln(\tilde{c}/\epsilon)\,\kappa$
Gradient descent, $\alpha = 2/(L+m)$	$\ln(\tilde{c}/\epsilon)\,\kappa/2$
Nesterov's method	$\ln(\tilde{c}/\epsilon)\,2\sqrt{\kappa}$
Triple momentum method	$(\ln(\tilde{c}/\epsilon) + \ln(\sqrt{\kappa}))\,\sqrt{\kappa}$
Theoretical lower bound	$\ln(\tilde{c}/\epsilon)\,\sqrt{\kappa}/2$

For ill-conditioned problems in which the condition ratio is large, the convergence rate is approximately one so we can use the approximation $\ln(1+x) \approx x$ for small x to obtain

$$k_\epsilon \approx \frac{\ln(c/\epsilon)}{1-\rho}, \qquad \kappa \text{ large.} \tag{8}$$

This yields the approximate iterations to converge in Table II.

Nesterov's method is referred to as optimal since the number of iterations for his method to converge is proportional to the theoretical lower bound. For high desired accuracies (i.e., small ϵ), however, the TM method achieves a reduction by a factor of two over Nesterov's method and is within a factor of two of the lower bound.

Remark 1: For the TM method, the constant c in (6) depends on the condition number κ and can be large when κ is large. In particular, for $\kappa \to \infty$ we have $\delta = \mathcal{O}(\sqrt{\kappa})$ and $c = \mathcal{O}(\sqrt{\kappa})$. This produces the additional $\ln(\sqrt{\kappa})\sqrt{\kappa}$ term for the TM method in Table II. When ϵ is small, however, this term can be neglected. In other words, the TM method exploits a trade-off between the size of the constant and the corresponding convergence rate. Compared to Nesterov's method, the TM method has a faster rate ρ but a larger constant c.

III. ANALYSIS

In this section we prove the error bounds for the TM method in Theorem 1. To do this, we first give our main analysis theorem which can be used to prove linear convergence of a sequence with rate ρ after a given number of iterations.

Theorem 2 (Analysis): Let $x \in \ell_{2e}^n$, $x_\star \in \mathbb{R}^n$, and $k_0 \ge 0$. If there exists a sequence $q \in \ell_{2e}$ such that

$$\|x_{k+1} - x_\star\|^2 \le \rho^2\, \|x_k - x_\star\|^2 - q_k, \quad \forall k \ge k_0 \tag{9}$$

and

$$0 \le \sum_{j=0}^{k} \rho^{-2j}\, q_j, \quad \forall k \ge 0, \tag{10}$$

then x_k converges linearly to $x_\star$ with rate ρ after iteration k_0. In particular, we have $\|x_k - x_\star\| \le c\,\rho^k$ for all $k \ge k_0$ where

$$c = \left(\rho^{-2k_0}\, \|x_{k_0} - x_\star\|^2 + \sum_{j=0}^{k_0-1} \rho^{-2(j+1)} q_j\right)^{1/2}. \tag{11}$$

Proof: Suppose there exists a sequence $q \in \ell_{2e}$ which satisfies (9) and (10). Define the quanitity

$$\eta_k := \|x_{k+1} - x_\star\|^2 - \rho^2\, \|x_k - x_\star\|^2 + q_k. \tag{12}$$

From (9), we have $\eta_k \le 0$ for all $k \ge k_0$. Then we have the following telescoping sum,

$$0 \ge \sum_{j=k_0}^{k-1} \rho^{2(k-j-1)}\, \eta_j = \|x_k - x_\star\|^2 - \rho^{2(k-k_0)}\, \|x_{k_0} - x_\star\|^2 + \rho^{2(k-1)} \sum_{j=k_0}^{k-1} \rho^{-2j} q_j$$

for all $k \ge k_0$. From (10), we also have

$$-\sum_{j=0}^{k_0-1} \rho^{-2j}\, q_j \le \sum_{j=k_0}^{k-1} \rho^{-2j}\, q_j.$$

Combining these results gives the bound

$$\|x_k - x_\star\|^2 \le \rho^{2(k-k_0)}\, \|x_{k_0} - x_\star\|^2 + \rho^{2(k-1)} \sum_{j=0}^{k_0-1} \rho^{-2j} q_j$$

for $k \ge k_0$. Factoring out ρ^{2k} and taking the square root gives $\|x_k - x_\star\| \le c\,\rho^k$ for all $k \ge k_0$ where c is given by (11). ■

In order to use Theorem 2 to prove linear convergence, we need a sequence q_k that satisfies (10) and enables us to show that (9) holds. When the sequence x_k is generated by gradient algorithms applied to strongly convex functions, we can use the following lemma from [3] to generate the sequence q_k.

Lemma 1 [3, Lemma 10]: Suppose $f \in \mathcal{S}_{m,L}$ with $0 = \nabla f(x_\star)$. Define $p_r(y) := \nabla f(y) - r\,(y - x_\star)$. Given a sequence $y \in \ell_{2e}^n$, let

$$q_k = \begin{cases} -p_m(y_k)^T\big[p_L(y_k) - \rho^2\, p_L(y_{k-1})\big], & k \ge 1 \\ -p_m(y_0)^T p_L(y_0), & k = 0. \end{cases} \tag{13}$$

Then

$$0 \le \sum_{j=0}^{k} \rho^{-2j}\, q_j, \quad \forall k \ge 0. \tag{14}$$

We now prove that the sequence x_k produced by the TM method converges linearly to $x_\star$ with rate ρ using Theorem 2 along with the sequence q_k given in Lemma 1.

Proof of Theorem 1: Let x_k and y_k be the sequences generated by the TM method with initial conditions ξ_0 and ξ_{-1}, and let $k_0 = 1$. Let q_k be given by

$$q_k = \frac{1}{mL} \begin{cases} -p_m(y_k)^T\big[p_L(y_k) - \rho^2\, p_L(y_{k-1})\big], & k \ge 1 \\ -p_m(y_0)^T p_L(y_0), & k = 0 \end{cases}$$

where $p_r(y_k) = \nabla f(y_k) - r\,(y_k - x_\star)$. Then (10) holds from Lemma 1. Next, we show that (9) holds for all $k \ge 1$. The sequence η_k in (12) is

$$\eta_k = \|x_{k+1} - x_\star\|^2 - \rho^2\, \|x_k - x_\star\|^2 - \frac{1}{mL}\, p_m(y_k)^T\big[p_L(y_k) - \rho^2\, p_L(y_{k-1})\big] \tag{15}$$

for $k \geq 1$. From the definition of the TM method in (2), we can make the substitutions

$$\begin{aligned} x_{k+1} &\to (1+\delta)\xi_{k+1} - \delta\xi_k \\ x_k &\to (1+\delta)\xi_k - \delta\xi_{k-1} \\ y_k &\to (1+\gamma)\xi_k - \gamma\xi_{k-1} \\ y_{k-1} &\to (1+\gamma)\xi_{k-1} - \gamma\xi_{k-2} \\ \nabla f(y_k) &\to \big[-\xi_{k+1} + (1+\beta)\xi_k - \beta\xi_{k-1}\big]/\alpha \\ \nabla f(y_{k-1}) &\to \big[-\xi_k + (1+\beta)\xi_{k-1} - \beta\xi_{k-2}\big]/\alpha \end{aligned} \quad (16)$$

which gives η_k in terms of ξ_{k-2}, ξ_{k-1}, ξ_k, ξ_{k+1}, and $x_\star$. Substituting $\rho = 1 - \sqrt{m/L}$ and the TM method parameters in (3), it is straightforward to show that $\eta_k \equiv 0$ for all $\xi_{k-2}, \xi_{k-1}, \xi_k, \xi_{k+1}, x_\star \in \mathbb{R}^n$. (In fact, these parameters are the unique solution with $\rho \in [0, 1)$ to the equation $\eta_k \equiv 0$.) Then (9) is satisfied with equality for $k \geq 1$. Applying Theorem 2 gives the bound on the iterates in (4), and the bound on the function values in (5) follows since the gradient of f is Lipschitz continuous with Lipschitz constant L. ∎

Remark 2: It is often desired to have intuition about the design of optimization algorithms. One method of obtaining the TM method parameters is to define η_k as in (15), make the substitutions in (16), and then solve $0 \equiv \eta_k$. A more intuitive design process, however, can be obtained using integral quadratic constraints from robust control. We develop this approach in the Appendix.

IV. Simulations

To verify the TM method, we simulate the algorithm using smooth multidimensional piecewise objective functions similar to the heavy-ball counter-example in [3]. Let

$$f(x) = \sum_{i=1}^{p} g(a_i^T x - b_i) + \frac{m}{2}\|x\|^2 \quad (17)$$

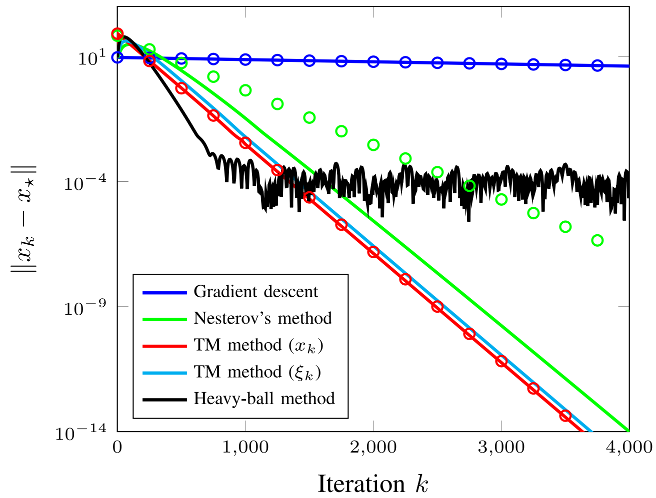
and

$$g(x) = \begin{cases} \frac{1}{2}x^2 e^{-r/x}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (18)$$

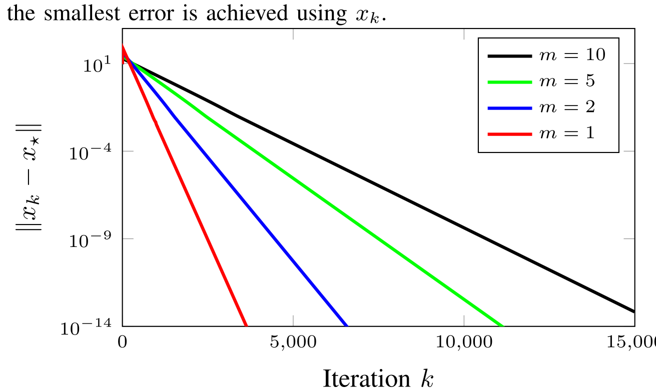
where $A = [a_1, \dots, a_p] \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^p$ with $\|A\| = \sqrt{L-m}$. Then $f \in \mathcal{S}_{m,L} \cap C^\infty$, i.e., f is m-strongly convex, its gradient is L-Lipschitz continuous, and it has continuous derivatives of all orders. We randomly generate the components of A and b from the normal distribution and then scale A so that $\|A\| = \sqrt{L-m}$.

In Fig. 2a, various methods are used to solve the same problem. Gradient descent is very slow due to the large condition number. The heavy-ball method has the fastest *local* convergence but contains stable limit cycles and does not converge globally. The TM method is proven to converge from any initial condition, and the proven rate is twice as fast as that of Nesterov's method. To verify Corollary 1, we also plot the error of ξ_k from the optimizer. While the rate is the same as that of x_k, the error using x_k as the output is smaller (note that ξ_k does not satisfy the same bound as x_k).

Since the parameter m is often unknown in practice, we also simulate the TM method where the parameters α, β, γ, and δ are designed using values of m which are larger than the strong convexity parameter of f; see Fig. 2b. The error still converges linearly in each case, although the convergence rate is slower if m is larger than the true value.



(a) Comparison of different methods. The circles indicate the proven error bounds (no bound is shown for the heavy-ball method since it is not globally convergent). For the TM method, the error using both x_k and ξ_k is shown; both sequences converge with the same rate, but the smallest error is achieved using x_k.



(b) The TM method is designed using the indicated values for m with $L = 10^4$. The convergence is linear in each case, although the rate is slower when m is larger than the true value.

Fig. 2. Simulation results using the objective function in (17) where $f \in \mathcal{S}_{1,10^4}$ with $n = 100$, $p = 5$, and $r = 10^{-6}$.

V. Conclusion

We have proposed a novel gradient-based algorithm for convex optimization. When $f \in \mathcal{S}_{m,L}$, the iterates converge linearly to the optimizer at rate $1 - \sqrt{m/L}$ from any initial condition. This is the fastest known convergence rate that has been proven for first-order algorithms which converge globally to the minimizer. For high levels of accuracy, the bound on the iteration complexity for our algorithm is half the known bound for Nesterov's method and within a factor of two of the theoretical lower bound in [2, Th. 2.1.13]. We gave a simple algebraic proof for the error bound of the TM method, and IQCs from robust control were used to motivate the design.

Appendix

Integral quadratic constraints (IQCs) are a powerful tool from robust control for analyzing interconnected dynamical systems which contain nonlinear components, including gradient-based optimization algorithms [3]. We now develop the IQC tools which will be used to give insight into the TM method.

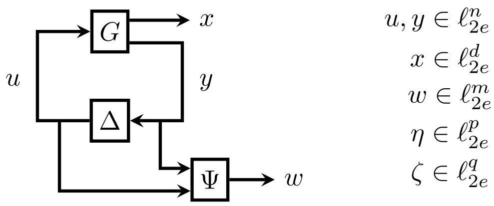


Fig. 3. The linear system G with state η is in feedback with the unknown function Δ, and the system output is x. The auxiliary system Ψ with state ζ filters u and y to produce an output w which satisfies the IQC.

In the robust control framework, we consider an unknown function $\Delta : \mathbb{R}^n \to \mathbb{R}^n$ in feedback with a known linear system $G : \mathbb{R}^n \to \mathbb{R}^d \times \mathbb{R}^n$, where G is given by the recursion

$$\begin{aligned} \eta_{k+1} &= A\eta_k + Bu_k, \quad \eta_0 \in \mathbb{R}^p \\ x_k &= C_1\eta_k \\ y_k &= C_2\eta_k. \end{aligned} \tag{19}$$

The internal state is $\eta \in \ell_{2e}^p$, the input is $u \in \ell_{2e}^n$, and the outputs are $x \in \ell_{2e}^d$ and $y \in \ell_{2e}^n$. The feedback is given by $u = \Delta(y)$ which produces the closed-loop system

$$\begin{aligned} \eta_{k+1} &= A\eta_k + B\Delta(C_2\eta_k), \quad \eta_0 \in \mathbb{R}^p \\ x_k &= C_1\eta_k. \end{aligned} \tag{20}$$

The system in (20) is difficult to analyze due to the unknown function Δ. The idea behind IQCs is to replace Δ with constraints that we know its input and output sequences must satisfy. If a result holds for any signals (y, u) which satisfy the constraints, then the result must also hold for the original system. To develop the constraints on (y, u), consider a linear system $\Psi : \ell_{2e}^n \times \ell_{2e}^n \to \ell_{2e}^m$ of the form

$$\begin{aligned} \zeta_{k+1} &= A_\Psi\zeta_k + B_\Psi^y y_k + B_\Psi^u u_k, \quad \zeta_0 = \zeta_\star \in \mathbb{R}^q \\ w_k &= C_\Psi\zeta_k + D_\Psi^y y_k + D_\Psi^u u_k \end{aligned} \tag{21}$$

where $\rho(A_\psi) < 1$ and $(\zeta_\star, w_\star, y_\star, u_\star)$ is the unique fixed-point of the system. This defines the map $w = \Psi(y, u)$ as shown in Figure 3.

We now define a ρ-IQC, which is a constraint on w which can be used to prove linear convergence with rate ρ.

Definition 3 (ρ-IQC, [3, Definition 3]): Suppose $u_\star, y_\star \in \mathbb{R}^n$ with $u_\star = \Delta(y_\star)$ and $y \in \ell_{2e}^n$ is an arbitrary square-summable sequence, i.e., $\sum_{k=0}^{\infty} \|y_k\|^2 < \infty$. Let $u = \Delta(y)$ and $w = \Psi(y, u)$. We say that $\Delta \in \text{IQC}(\Psi, M, u_\star, y_\star, \rho)$ if

$$0 \le \sum_{j=0}^{k} \rho^{-2j}(w_j - w_\star)^T M(w_j - w_\star), \quad \forall k \ge 0. \tag{22}$$

If $\Delta \in \text{IQC}(\Psi, M, u_\star, y_\star, \rho)$, then we can remove Δ from the block diagram and simply study the connection of the linear systems G and Ψ subject to the IQC constraint (22). For gradient-based optimization algorithms, we need an IQC which characterizes ∇f. The following lemma provides a useful class of IQCs which characterizes ∇f when $f \in \mathcal{S}_{m,L}$.

Lemma 2 [3, Lemma 10]: Suppose $f \in \mathcal{S}_{m,L}$ and $(u_\star, y_\star)$ is a reference point for the gradient of f, i.e., $u_\star = \nabla f(y_\star)$. Let $H(z) = \bar{\rho}^2/z$ with $\bar{\rho} \in [0, 1]$. Then $\nabla f \in \text{IQC}(\Psi, M, u_\star, y_\star, \rho)$ for any $\rho \in [\bar{\rho}, 1]$ where

$$\Psi = \begin{bmatrix} L(1-H) & -(1-H) \\ -m & 1 \end{bmatrix} \otimes I_n, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes I_n. \tag{23}$$

Remark 3: Lemmas 1 and 2 are related as follows. Using Ψ and M in (23) with $H(z) = \rho^2/z$, the sequence q_k in (13) is equal to

$$q_k = (w_k - w_\star)^T M(w_k - w_\star) \tag{24}$$

where w_k and $w_\star$ are the outputs of Ψ when the inputs are (y_k, u_k) and $(y_\star, u_\star)$, respectively. Condition (14) is then equivalent to the IQC condition in (22).

Now that we can characterize ∇f using an IQC, we would like conditions on G, Ψ, and M such that x_k converges linearly with rate ρ for any signal u such that the IQC condition in (22) is satisfied. Such conditions exist in both the time domain and the frequency domain [3], [9]–[12]. For example, the following theorem can be used to prove that the state of G_2 converges with rate ρ where G_2 is the transfer function from u to y.

Theorem 3 [9, Th. 2]: Let $G_2(\rho z) \in \mathcal{RH}_\infty^{m\times n}$ and let Δ be a bounded causal operator. Suppose that:

1) $\forall \tau \in [0, 1]$, the interconnection of G_2 and $\tau\Delta$ is well-posed.
2) $\forall \tau \in [0, 1]$, we have $\tau\Delta \in \text{IQC}(\Psi, M, u_\star, y_\star, \rho)$.
3) there exists $\epsilon > 0$ such that

$$\begin{bmatrix} G_2(z) \\ I \end{bmatrix}^* \Pi(z) \begin{bmatrix} G_2(z) \\ I \end{bmatrix} \preceq -\epsilon I, \quad \forall z \in \rho\mathbb{T} \tag{25}$$

where $\Pi(z) = \Psi(z)^* M \Psi(z)$.

Then the state of G_2 converges linearly with rate ρ.

The condition in (25) is a frequency-domain inequality (FDI) which must be satisfied at every point on $\rho\mathbb{T}$. This can be converted into a single linear matrix inequality (LMI) by applying the discrete-time KYP lemma [13]. This allows us to establish the following connection between Theorem 2 and the IQC framework: if we let x_k be the internal state of G, then conditions (9) and (10) in Theorem 2 are comparable to the FDI in (25) and the IQC condition in (22), respectively. Theorem 2, however, offers several advantages over the existing IQC framework; in particular, it

1) can be used to certify linear convergence after a finite number of iterations, and
2) can be applied to any sequence x_k, not just the internal state of G.

The conditions in Theorem 3 are not satisfied for the TM method, and therefore it cannot be used to prove convergence. However, we can gain insight into the design by relaxing the requirement that $\epsilon > 0$ in (25). Suppose $G_2(z) = g_2(z) \otimes I_n$. Then for the IQC in Lemma 2, the left-hand side of (25) simplifies to

$$-2\,\text{Re}\{(1 - H)\,(1 - L\,g_2)\,(1 - m\,g_2)^*\} \otimes I_n. \tag{26}$$

If we set $\epsilon = 0$ in (25), then this is equivalent to the following relaxed frequency-domain condition.

Condition 1 (Relaxed Frequency-Domain Condition):

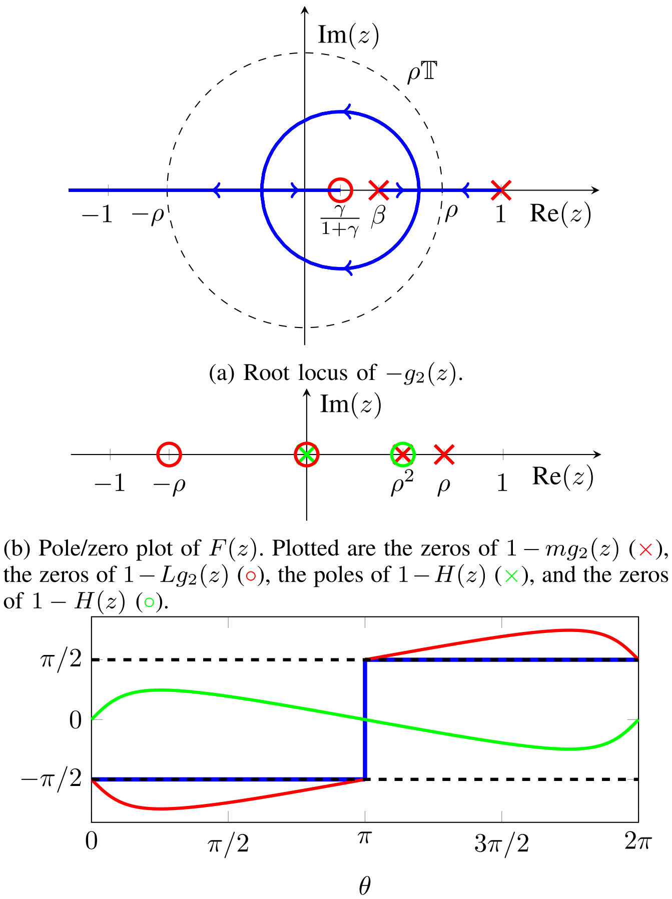
$$\angle F(z) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \quad \text{for all } z \in \rho\mathbb{T} \tag{27}$$

where

$$F(z) := \frac{1 - L\,g_2(z)}{1 - m\,g_2(z)}\big(1 - H(z)\big) \tag{28}$$

and $G_2(z) = g_2(z) \otimes I_n$.

We now describe how to design the system $g_2(z)$, the IQC parameter $H(z)$, and the convergence rate ρ to satisfy the



(a) Root locus of $-g_2(z)$.

(b) Pole/zero plot of $F(z)$. Plotted are the zeros of $1-mg_2(z)$ (×), the zeros of $1-Lg_2(z)$ (○), the poles of $1-H(z)$ (×), and the zeros of $1-H(z)$ (○).

(c) Phase plot of $(1-Lg_2(z))/(1-mg_2(z))$ (red), $1-H(z)$ (green), and $F(z)$ (blue) for $z=\rho e^{j\theta}$. Note that $\angle F(z) \in [-\pi/2, \pi/2]$ for all $z \in \rho\mathbb{T}$ as desired.

Fig. 4. Design of the TM method parameters (α, β, γ) and convergence rate ρ in Theorem 1.

relaxed frequency-domain condition. This procedure is then used to design the TM method.

1) Draw the root locus of $-g_2(z)$.
2) Draw a pole/zero plot of $F(z)$ as follows:
   a) The root locus poles at gain m are poles of $F(z)$.
   b) The root locus poles at gain L are zeros of $F(z)$.
   c) The poles and zeros of $1-H(z)$ are poles and zeros, respectively, of $F(z)$.
3) For all $z \in \rho\mathbb{T}$, calculate $\angle F(z)$ by summing the angles from z to the zeros and subtracting the angles from z to the poles. Condition 1 is satisfied if $\angle F(z) \in [-\pi/2, \pi/2]$ for all $z \in \rho\mathbb{T}$.

A. Design of the TM Method

To design the TM method using IQCs, we first formulate the problem in the robust control framework. We use the IQC in Lemma 2 to characterize ∇f when $f \in \mathcal{S}_{m,L}$. The gradient-based optimization algorithm in (2) is equivalent to the closed-loop system in (20) with $\Delta = \nabla f$, $\eta_0 = \begin{bmatrix}\xi_0^T & \xi_{-1}^T\end{bmatrix}^T$, $p = 2n$, and

$$G = \left[\begin{array}{c|c} A & B \\ \hline C_1 & 0 \\ C_2 & 0 \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)I_n & -\beta I_n & -\alpha I_n \\ I_n & 0_n & 0_n \\ \hline (1+\delta)I_n & -\delta I_n & 0_n \\ (1+\gamma)I_n & -\gamma I_n & 0_n \end{array}\right]. \tag{29}$$

The transfer function from u to y is $G_2(z) = g_2(z) \otimes I_n$ where

$$g_2(z) = -(1+\gamma)\alpha \frac{\left(z - \frac{\gamma}{1+\gamma}\right)}{(z-1)(z-\beta)}. \tag{30}$$

From the root locus of $-g_2(z)$ in Fig. 4a, it is clear that $(1-Lg_2(z))/(1-mg_2(z))$ will have negative phase for $\theta \in (0,\pi)$ and positive phase for $\theta \in (\pi, 2\pi)$ (see Fig. 4b). The phase of $1-H(z)$ is simply added to that of $(1-Lg_2(z))/(1-mg_2(z))$, so we want to choose $H(z)$ to have large positive phase in $(0,\pi)$ and large negative phase in $(\pi,2\pi)$. This is achieved using $H(z) = \rho^2/z$ (see Fig. 4b). Then $1-H(z)$ has a zero at $z=\rho^2$ and a pole at $z=0$. We design $g_2(z)$ to cancel out the pole and zero of $1-H(z)$ and to minimize the convergence rate ρ as follows:

- Design $1-mg_2(z)$ to have roots at $z=\rho^2$ (to cancel the zero of $1-H(z)$) and $z=\rho$.
- Design $1-Lg_2(z)$ to have roots at $z=0$ (to cancel the pole of $1-H(z)$) and $z=-\rho$.

These four conditions are used to solve for the parameters α, β, γ, and the convergence rate ρ. Since $g_2(z)$ does not depend on the parameter δ, it must be obtained in the time-domain by solving $0 \equiv \eta_k$ where η_k is defined in (15).

Remark 4: We choose $g_2(z)$ such that the roots of $1-mg_2(z)$ and $1-Lg_2(z)$ cancel the zero and pole, respectively, of $H(z)$. We can do this since the cancelled poles and zeros are inside $\rho\mathbb{T}$.

References

[1] 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.
[2] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. New York, NY, USA: Springer, 2004.
[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] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, "Global convergence of the heavy-ball method for convex optimization," in *Proc. Eur. Control Conf.*, Linz, Austria, Jul. 2015, pp. 310–315.
[5] Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach," *Math. Program.*, vol. 145, nos. 1–2, pp. 451–482, 2014.
[6] D. Kim and J. A. Fessler, "Optimized first-order methods for smooth convex minimization," *Math. Program.*, vol. 159, nos. 1–2, pp. 81–107, Sep. 2016.
[7] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to Nesterov's accelerated gradient descent," *arXiv:1506.08187 [math.OC]*, Jun. 2015.
[8] D. Drusvyatskiy, M. Fazel, and S. Roy, "An optimal first order method based on optimal quadratic averaging," *arXiv:1604.06543 [math.OC]*, Apr. 2016.
[9] R. Boczar, L. Lessard, and B. Recht, "Exponential convergence bounds using integral quadratic constraints," in *Proc. 54th IEEE Conf. Decis. Control*, Osaka, Japan, Dec. 2015, pp. 7516–7521.
[10] H. K. Khalil, *Nonlinear Systems*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[11] N. S. Ahmad, J. Carrasco, and W. P. Heath, "A less conservative LMI condition for stability of discrete-time systems with slope-restricted nonlinearities," *IEEE Trans. Autom. Control*, vol. 60, no. 6, pp. 1692–1697, Jun. 2015.
[12] P. Seiler, "Stability analysis with dissipation inequalities and integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 60, no. 6, pp. 1704–1709, Jun. 2015.
[13] A. Rantzer, "On the Kalman–Yakubovich–Popov lemma," *Syst. Control Lett.*, vol. 28, no. 1, pp. 7–10, Jun. 1996.