An Explicit Convergence Rate for Nesterov's Method from SDP

Sam Safavi
sam.safavi@tufts.edu

Bikash Joshi
bikash.joshi@imag.fr

Guilherme França
guifranca@jhu.edu

José Bento
jose.bento@bc.edu

***Abstract*—The framework of Integral Quadratic Constraints (IQC) introduced by Lessard et al. (2014) reduces the computation of upper bounds on the convergence rate of several optimization algorithms to semi-definite programming (SDP). In particular, this technique was applied to Nesterov's accelerated method (NAM). For quadratic functions, this SDP was explicitly solved leading to a new bound on the convergence rate of NAM, and for arbitrary strongly convex functions it was shown numerically that IQC can improve bounds from Nesterov (2004). Unfortunately, an explicit analytic solution to the SDP was not provided. In this paper, we provide such an analytical solution, obtaining a new general and explicit upper bound on the convergence rate of NAM, which we further optimize over its parameters. To the best of our knowledge, this is the best, and explicit, upper bound on the convergence rate of NAM for strongly convex functions.**

I. Introduction

Consider the problem

$$\min_{x\in\mathbb{R}^p} f(x) \tag{1}$$

under the following additional assumption, which holds throughout this paper.

Assumption 1. 1) *The function f is convex, closed and proper;*
2) *Let $S_d(m,L)$ be the set of functions $h:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ such that $m\|x-y\|^2 \le (\nabla h(x)-\nabla h(y))^T(x-y) \le L\|x-y\|^2$ for all $x,y\in\mathbb{R}^d$ where $0<m\le L<\infty$ and $\|\cdot\|$ denotes the Euclidean norm; We assume that $f\in S_p(m,L)$, i.e. f is strongly convex and ∇f is Lipschitz continuous.*

In this paper, we provide a new bound on the convergence rate of NAM when solving (1).

NAM has wide applications in machine learning. It is the base of the well-known FISTA algorithm largely used to solve problems arising in signal processing [1], and it was also extensively applied in compressed sensing, as for instance in [2], [3]. A trace norm regularization using NAM was proposed in [4], which has applications in multi-task learning, matrix classification and matrix completion. Even to train deep neural networks, it was shown that NAM with a careful initialization is able to achieve state-of-the-art accuracy [5].

NAM is parametrized by $\alpha>0$ and $\beta\ge 0$ and takes the form in Algorithm 1. We assume that α and β are fixed. A classical choice for these parameters is [6]

$$\alpha = 1/L, \qquad \beta = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1), \tag{2}$$

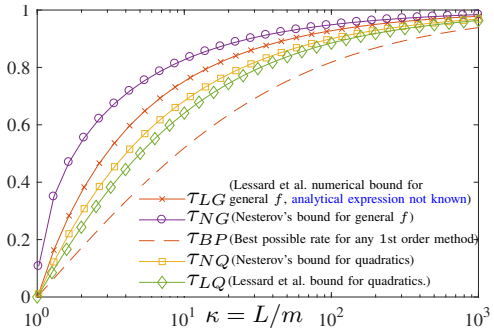

Fig. 1. Different known linear rate bounds for NAM.

where $\kappa = L/m$. We define an upper bound on the convergence rate of NAM, for fixed α, β and function f, as any $\tau\in[0,1]$ for which

$$\|x_t - x_*\| \le C\tau^t \|x_0 - x_*\|, \tag{3}$$

where $C>0$ is a constant, and x_* is a fixed point of Algorithm 1. Choosing α and β according to (2), [6] uses the technique of *estimate sequences* and obtains

$$\tau = \tau_{NG} \triangleq \sqrt{1-1/\sqrt{\kappa}}. \tag{4}$$

In addition, if f is quadratic, then

$$\tau = \tau_{NQ} \triangleq 1-1/\sqrt{\kappa}. \tag{5}$$

In [6] it was also shown that any first order method must obey

$$\tau \ge \tau_{BP} \triangleq 1-2/\sqrt{\kappa+1}. \tag{6}$$

Several recent works have revisited NAM and computed bounds on its convergence rate based on different techniques. Although these re-derivations have increased our understanding of NAM, and in some cases even inspiring new variations, they have not improved previous results. A partial exception is [7], where they reduce computing a bound on the rate of

Algorithm 1 Nesterov's accelerated method (parameters α, β)

1: Initialize x_0, x_1
2: **repeat**
3: $\quad y_t = (1+\beta)x_t - \beta x_{t-1}$
4: $\quad x_{t+1} = y_t - \alpha\nabla f(y_t)$
5: **until** stop criterion

convergence to finding solutions to a semi-definite programming (SDP) problem. This SDP has multiple solutions, each of which gives a bound on the convergence rate, some better than others. For quadratic functions, [7] explicitly solve this SDP, optimize the result over α and β, and obtain a new improved bound on the convergence rate of NAM with the tuning rules

$$\alpha = 4/(3L+m),\ \beta = (\sqrt{3\kappa+1}-2)/(\sqrt{3\kappa+1}+2),\\ \tau = \tau_{LQ} \triangleq 1 - 2/\sqrt{3\kappa+1}. \tag{7}$$

For general strongly convex functions, they numerically solve this SDP and obtain τ_{LG} as shown in Fig. 1. From the plot we see that the results from the IQC-framework improve on (4). However, no explicit and analytical solution to the SDP associated to NAM was provided. Even more discouraging is the fact that the only explicit solution obtained was for Gradient Descent (GD), yielding a previously known bound on convergence rate.

On the other hand, in our recent paper at ISIT 2016, [8], we show that it is possible to extract explicit solutions from the IQC-framework for non classical optimization algorithms and for general strongly convex functions. In particular, optimally tuning ADMM algorithm, we obtain a convergence rate that matches the τ_{BP}, the best possible for any first order methods.

The main contribution of this paper is to apply the IQC framework of [7] to obtain an explicit and new bound on the convergence rate of NAM. In particular, we derive an analytical solution to the corresponding SDP for which [7] only provides numerical solutions. To the best of our knowledge, our result is the best explicit bound for NAM and arbitrary strongly convex functions. It is also one of the only three explicit bounds obtained from the IQC-framework so far; others are for GD and ADMM.

II. Related Work

Several recent works have revisited NAM and computed bounds on its convergence rate based on different techniques. In addition to [7], the following works are relevant. [9] views NAM as a linear coupling between GD and Mirror Descent, and, for $f \in S_p(0,L)$, re-derives the previously known bound $f(x_t) - f(x_*) = \mathcal{O}\left(\frac{L}{t^2}\right)$, [10], with the choice $\alpha = \frac{C}{\sqrt{L}}$ and $\beta = \frac{1}{\alpha L+1}$, which is different from Nesterov's bound. This rate is not of the type (3) that we consider in this paper.

The work of [11] views (an adaptive version of) NAM with $\beta = \frac{t-1}{t+r-1}$ as the discretization of the second-order ODE $\frac{d^2x}{dt^2} + \frac{r}{t}\frac{dx}{dt} + \nabla f(x) = 0$. For $f \in S_p(m,L)$ and $2 \le \alpha \le 2r/3$ they obtain $f(x_t) - f(x_*) \le \frac{C(\alpha,r)}{t^\alpha}$. If $m > 0$, this leads to $\|x_t - x_*\| = \mathcal{O}\left(\frac{1}{t^{\alpha/2}}\right)$. Unfortunately, [11] show that their framework is incapable of providing linear convergence rates in general, which we know to hold for $f \in S_p(m,L)$.

The work of [12] does not give a bound for the fixed step-size NAM for a general smooth function but only for an adaptive NAM and a function $f \in S_p(0,L)$ that convex and quadratic. For such a function, and for $\beta = 1 - \frac{2}{t+1}$ and $\alpha = 1/L$, they show that $f(x_t) - f(x_*) = \mathcal{O}(\frac{L}{t^2})$.

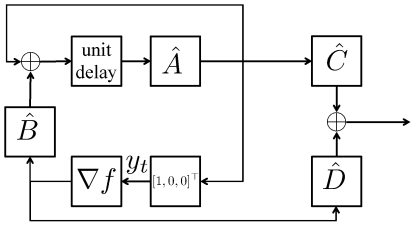

Fig. 2. The variables in NAM appear in a linear system inserted in a nonlinear feedback loop. Above, we show y_t only. $[1,0,0]^\top$ outputs the first component of its input. The system is more complex than NAM, and, in particular, the matrices $\hat{C}$ and $\hat{D}$ are used to probe it. [7] use the properties of the output of this probe to prove properties about the convergence of NAM as stated in Theorem 2.

The work of [13] focuses only on a convex quadratic function $f \in S_p(m,L)$, and obtain $\tau = 1 - 1/\sqrt{\kappa}$ for $\alpha = 1/L$ and $\beta = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, basically re-deriving (5).

Finally, [14] gives a possible geometric interpretation of why NAM accelerates convergence. For their NAM-type method, the result $\tau = \sqrt{1 - 1/\sqrt{\kappa}}$ is obtained, basically re-deriving (4).

III. Main results

We start recalling results from [7]. Algorithm 1 can be studied through a linear dynamical system involving the matrices[1]

$$\hat{A} = \begin{bmatrix} \beta+1 & -\beta & 0 \\ 1 & 0 & 0 \\ L(-\beta-1) & \beta L & 0 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} -\alpha \\ 0 \\ 1 \end{bmatrix}, \\ \hat{C} = \begin{bmatrix} L(\beta+1) & -L\beta & \rho^2 \\ -m(\beta+1) & m\beta & 0 \end{bmatrix}, \quad \hat{D} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \tag{8}$$

inserted in a nonlinear feedback loop, where the feedback gain is essentially ∇f. The constant $\rho > 0$ will be specified later. See Figure 2 for an illustration. The stability of this dynamical system is related to the convergence rate of Algorithm 1, which involves numerically solving a 4×4 semidefinite program.

Theorem 2 ([7]). *Let $\{x_t\}$ evolve according to Algorithm 1 for fixed $\alpha > 0$ and $\beta \ge 0$. Let x_* be a fixed point of the algorithm. Fix $0 < \rho \le \tau < 1$. If there exists a 3×3 matrix $P \succ 0$ and a constant $\lambda \ge 0$ such that*

$$\begin{bmatrix} \hat{A}^T P \hat{A} - \tau^2 P & \hat{A}^T P \hat{B} \\ \hat{B}^T P \hat{A} & \hat{B}^T P \hat{B} \end{bmatrix} + \lambda \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix}^T M \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix} \preceq 0, \tag{9}$$

where $M = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, then, for all $t \ge 0$, we have

$$\|x_t - x_*\| \le \sqrt{\kappa_P} C_0\, \tau^t, \tag{10}$$

where the constant $C_0 = \sqrt{\|x_1 - x_*\|^2 + \|x_0 - x_*\|^2}$, and $\kappa_P = \sigma_{\max}(P)/\sigma_{\min}(P)$ is the condition number of P.

Note that any fixed point x_* of Algorithm 1 satisfies the KKT conditions for problem (1), which due to strong convexity make x_* the unique minimizer. Thus, Theorem 2 enables

[1]The connection between Algorithm 1 and these matrices can only be formally established if we use $\hat{A} \otimes I_p$, $\hat{B} \otimes I_p$,$\hat{C} \otimes I_p$ and $\hat{D} \otimes I_p$, where $\otimes$ is the Kronecker product and $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix. Note, however, that Theorem 2 holds with the matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ exactly as specified in (8). See [7, Section 4.2] for more details.

us to find an explicit convergence rate $\tau = \tau(\alpha, \beta, L, m)$: just find P, λ, ρ and τ satisfying the conditions above.

Unfortunately, [7] does not give an explicit expression for τ as a function of κ, α and β. Our main result in this paper provides such an explicit formula when $\alpha = 1/L$. To arrive at this result, we first prove a series of intermediate steps.

Theorem 3. *Equation* (9) *holds if* $\beta > 0$, $\kappa > 1$, $\lambda = \alpha = L = 1$, $\rho = \tau > 0$, τ *is such that*

$$-4(-\kappa+1)^2\beta^2(-2+\omega) + \omega(\kappa - 1 + \kappa\omega)^2 - 4(-\kappa+1)\beta\omega(-3\kappa + 1 + \kappa\omega) = 0, \quad (11)$$

where $\omega = \tau^2$ *, and* $P = \left[\begin{smallmatrix} a & b & c \\ b & d & e \\ c & e & f \end{smallmatrix}\right]$*, where*

$$a = -\left(\tfrac{1}{\beta} + 2\right)\omega + 2(\beta+2) + \tfrac{\beta(s-1)}{\omega} - 2(\beta+1)s, \quad (12)$$

$$b = \tfrac{1}{2}\left((2\beta+1)(s-1) + \omega\right), \quad (13)$$

$$c = \beta - \tfrac{\omega(s+\omega-1)}{2\beta(s-1)} - (\beta+1)s - \omega + 1, \quad (14)$$

$$d = (1-s)\beta, \quad (15)$$

$$e = \omega - (1-s)\beta, \quad (16)$$

$$f = \tfrac{\omega^2}{\beta - \beta s}. \quad (17)$$

Note that ω *in* (12)-(17) *satisfies* (11) *and we defined* $s = \kappa^{-1}$.

Remark 4. *Note that* (11) *is a third degree polynomial in* ω *with real coefficients, which always has a real root. Moreover, all roots have a closed form expression. This defines* $\tau = \tau(\kappa, \beta)$ *through* $\tau^2 = \omega$.

Proof of Theorem 3. Let

$$H = \begin{bmatrix} H_1 & H_2 \\ H_2^\top & H_3 \end{bmatrix} \quad (18)$$

be the left hand side of (9) multiplied by -1, where H_1, H_2 and H_3 are 2×2 matrices and $H_2^\top$ denotes the transpose of H_2. To show that H is positive semidefinite we are going to use the following property of the Schur complement [15]: H is positive semidefinite if and only if

$$H_3 \succeq 0, \quad (19)$$

$$H_1 - H_2 H_3^\dagger H_2^\top \succeq 0, \quad (20)$$

$$(I - H_3 H_3^\dagger) H_2^\top = 0, \quad (21)$$

where $H_3^\dagger$ is the pseudoinverse of H_3 [16].

To check that conditions (19)–(21) hold, we first replace $\lambda = \alpha = L = 1$ and formulas (12)–(17) in H. Hence, for (19) we have

$$H_3 = \begin{bmatrix} \frac{\omega^3}{\beta - s\beta} & -\omega \\ -\omega & \frac{\beta - s\beta}{\omega} \end{bmatrix} \quad (22)$$

whose eigenvalues are 0 and $\frac{1}{\omega}\left(\beta(1-s) + \frac{\omega^4}{\beta(1-s)}\right)$. Both are nonnegative since $s = \kappa^{-1} < 1$, $\beta > 0$ and $\omega > 0$. Now we check (21). H_3 has no inverse but it has an explicit pseudoinverse given by

$$H_3^\dagger = \begin{bmatrix} -\frac{(s-1)\beta\omega^5}{(\omega^4 + (s-1)^2\beta^2)^2} & -\frac{(s-1)^2\beta^2\omega^3}{(\omega^4 + (s-1)^2\beta^2)^2} \\ -\frac{(s-1)^2\beta^2\omega^3}{(\omega^4 + (s-1)^2\beta^2)^2} & -\frac{(s-1)^3\beta^3\omega}{(\omega^4 + (s-1)^2\beta^2)^2} \end{bmatrix}. \quad (23)$$

Replacing this expression in the left hand side of (21) confirms that it holds true. Finally, we check (20). After a simple, but tedious, calculation one can obtain

$$H_1 - H_2 H_3^\dagger H_2^\top = \begin{bmatrix} \frac{-4\beta^2(s-1)^2(\omega-2) - 4\beta(s-1)\omega(s+\omega-3) + \omega(-s+\omega+1)^2}{4\beta(s-1)} & 0 \\ 0 & 0 \end{bmatrix}. \quad (24)$$

Let $\delta = \delta(\omega, s, \beta)$ be the numerator of the top-left element in the matrix above. A direct calculation shows that $\kappa^2\delta(\omega, \kappa^{-1}, \beta)$ is the left hand side of (11), which is zero by assumption. Hence, $H_1 - H_2 H_3^\dagger H_2^\top = 0$ and (20) is true. □

Let $\tau = \tau(\kappa, \beta)$ be the smallest (real) solution of (11) such that $\tau \in (0, 1)$. Our next theorem gives an expression for the choice of $\beta = \beta(\kappa)$ that minimizes $\tau(\kappa, \beta)$ for each $\kappa > 1$.

Theorem 5. *Let* $\beta(\kappa)$ *minimize* $\tau(\kappa, \beta)$*, for fixed* $\kappa > 1$*. We have*

$$\beta(\kappa) = \frac{2\kappa - \sqrt{2\kappa - 1} - 1}{2\left(\kappa + \sqrt{2\kappa - 1}\right)}, \quad (25)$$

$$\tau(\kappa, \beta(\kappa)) = \sqrt{1 - \frac{\sqrt{2\kappa - 1}}{\kappa}}. \quad (26)$$

Proof. Note that (11) is a quadratic polynomial in β. Its zeros are

$$\beta = \frac{x \pm \sqrt{y}}{z}, \quad (27)$$

where

$$x = (\kappa - 1)\omega(\kappa(\omega - 3) + 1), \quad (28)$$

$$y = 2(\kappa-1)^2(\omega-1)\omega\left(\kappa\left(\kappa(\omega-1)^2 - 2\right) + 1\right), \quad (29)$$

$$z = 2(\kappa-1)^2(\omega-2). \quad (30)$$

For each $\kappa > 1$, we want to find the smallest $\tau \in (0, 1)$ for which we still have real roots in the above equation. This is the same as finding the smallest $\omega \in (0, 1)$ for which $\left(\kappa\left(\kappa(\omega-1)^2 - 2\right) + 1\right)$, a quadratic function of ω, is nonnegative. This is easy to find, yielding

$$\omega = 1 - \frac{\sqrt{2\kappa - 1}}{\kappa}, \quad (31)$$

for which we have

$$\beta = \frac{x}{y} = \frac{\omega(\kappa(\omega-3)+1)}{2(\kappa-1)(\omega-2)} = -\frac{-2\kappa + \sqrt{2\kappa-1} + 1}{2\left(\kappa + \sqrt{2\kappa-1}\right)}. \quad (32)$$

□

Theorem 6. *If* τ *and* β *are chosen as* (25) *and* (26)*, respectively, and the entries in* P *according to* (12)–(17)*, then* $P \succ 0$.

Proof. Let P' be P with its rows and columns permuted such that the first, second and last row/column become the last, second and first row/column. Note that P' and P have the same spectrum. We are going to show that all the principal minors of P' are strictly positive, a necessary and sufficient

condition for positive definitiveness known as Sylvester's criterion [17].

Replacing (12)–(17) in P', the first minor is given by

$$f = \frac{\kappa\omega^2}{\beta(\kappa-1)} > 0. \tag{33}$$

The second minor is

$$\begin{vmatrix} d & e \\ e & f \end{vmatrix} = \frac{\beta(\kappa-1)(\beta(-\kappa)+\beta+2\kappa\omega)}{\kappa^2}, \tag{34}$$

whose sign is dictated by $\beta(-\kappa)+\beta+2\kappa\omega$ and which, by substituting (25)–(26), becomes

$$\beta(-\kappa)+\beta+2\kappa\omega = \frac{(\kappa-1)\left(2\kappa+\sqrt{2\kappa-1}-3\right)}{2\left(\kappa+\sqrt{2\kappa-1}\right)} > 0, \tag{35}$$

since $\kappa \geq 1$.

The third minor is just the determinant of P', which is

$$\begin{aligned} &\tfrac{1}{\omega}\beta^3\left(\tfrac{1}{\kappa}-1\right)^3(\omega-1)+\beta^2\left(\tfrac{1}{\kappa}-1\right)^2\left(\tfrac{1}{\kappa}+3\omega-5\right) \\ &+2\beta\left(\tfrac{1}{\kappa}-1\right)\omega\left(\tfrac{1}{\kappa}+\omega-3\right)-\tfrac{1}{2}\omega\left(-\tfrac{1}{\kappa}+\omega+1\right)^2. \end{aligned} \tag{36}$$

We can use (11) to simplify this expression to

$$\frac{\beta^2(\kappa-1)^2((\omega-1)(\beta(-\kappa)+\beta+\kappa\omega)+\omega)}{\kappa^3\omega}, \tag{37}$$

whose sign is dictated by $(\omega-1)(\beta(-\kappa)+\beta+\kappa\omega)+\omega$. If we substitute (25)–(26) we obtain

$$\begin{aligned} (\omega-1)(\beta(-\kappa)+\beta+\kappa\omega)+\omega &= \frac{(\kappa-1)\left(\sqrt{2\kappa-1}-1\right)}{2\kappa\left(\kappa+\sqrt{2\kappa-1}\right)} \\ &> 0 \end{aligned} \tag{38}$$

since $\kappa > 1$. □

We now provide our main result, which directly follows from our previous theorems and a simple rescaling argument.

Theorem 7. *Let $f \in S_p(m,L)$ and $\kappa = L/m \geq 1$. Consider Algorithm 1 to solve the optimization problem* (1)*. If $\alpha = \frac{1}{L}$ and $\beta = \frac{2\kappa-\sqrt{2\kappa-1}-1}{2\left(\kappa+\sqrt{2\kappa-1}\right)}$, then*

$$\|x_t - x_*\| \leq C_0 C_1\, \tau^t, \tag{39}$$

where $C_0 = \sqrt{\|x_1-x_*\|^2+\|x_0-x_*\|^2}$, $C_1 > 0$ is a function of κ, and

$$\tau = \sqrt{1-\frac{\sqrt{2\kappa-1}}{\kappa}}. \tag{40}$$

Proof. We can assume, without loss of generality, that $\kappa > 1$. The case $\kappa = 1$ follows by a continuity argument, applying a small quadratic perturbation to f and letting the perturbation converge to zero.

The convergence rate of Algorithm 1 on f with $\alpha = 1/L$ is the same as its convergence rate on $\hat{f} = f/L \in S_p(m,1)$ with $\alpha = 1$. In this setting, Theorem 3 and Theorem 6 tell us that the conditions to apply Theorem 2 hold for our choice of α and β. Furthermore, according to Theorem 5, for this choice of α and β, the convergence rate τ satisfies (40). □

IV. The pathway towards the proof

The reader might have noticed that our previous proofs amount to substituting expressions into conditions and subsequently checking that these conditions are satisfied. It is enlightening to explain how we obtained these expressions in the first place. Specifically, how did we obtain (12)–(17) from which all other formulas follow? In a nutshell, we built our ansatz based on numerical experimentation. Reveling this path might be useful for other researchers to use the IQC framework to derive explicit formulas for other algorithms as well.

First, we reduce the number of variables in the problem by setting $\lambda = 1$, $\rho = \tau$ and $\alpha = L = 1$.

Second, we fix $\beta > 0$ and $\kappa \in (0,1)$, and use a convex optimization solver to numerically find the smallest τ for which (9) is satisfied under the assumption that $P \succ 0$. Let H be the right hand side of (9) multiplied by -1. To find this τ, we start with $\tau = 0.5$ and check if the SDP

$$\min_P \ 1 \qquad \text{s.t. } H \succeq 0 \text{ and } P \succeq 0 \tag{41}$$

has a feasible solution[2]. In the affirmative case, we reduce τ, otherwise we increase τ. Notice that the eigenvalues of H increase monotonically with τ. Hence, we can use bisections to find the smallest possible τ in a few steps. After this procedure is done, we check if $P \succ 0$. If this does not hold, we try a different β and/or κ.

Third, we repeat this procedure for several pairs of (β,κ). For each pair, we obtain numerical values for P and H such that $H \succeq 0$ and $P \succ 0$ hold. From these numerical values, we try to identify some very simple properties that H or P might satisfy for all tested values of β and κ. Labeling the entries of P as in Theorem 3, the properties that we can easily guess based on our numerical experiments are the following:

1) Recall that $P = P^\top = \left[\begin{smallmatrix} a & b & c \\ b & d & e \\ c & e & f \end{smallmatrix}\right]$. Then,
$$e = \omega - d, \tag{42}$$
$$d = \beta(1-m). \tag{43}$$
2) Let Δ_i be the principal minor of H obtained by removing the ith row and column. We observe $\Delta_i = 0$ for $i = 1,\dots,4$;
3) Let $\Delta_{1,2;1,2}$ be the principal minor of H obtained from removing the 1st and 2nd column/row from H. We observe that $\Delta_{1,2;1,2} = 0$.

Fourth, we replace (42) and (43) into H and we solve the condition $\Delta_1 = 0$ for a. This leads to

$$a = \tfrac{1}{\omega}\left(-\beta+2c\omega-f\omega+\beta m+2\omega\right). \tag{44}$$

We substitute this expression into H and solve $\Delta_3 = 0$ for b, yielding

$$b = \tfrac{1}{2}\left((2\beta+1)(m-1)+\omega\right). \tag{45}$$

[2]Note that the standard formulation of convex optimization problems, and existing solvers, does not allow us to enforce $P \succ 0$. This is why we enforce $P \succeq 0$ and later check if $P \succ 0$.

Again, we substitute this expression in H and solve $\Delta_4 = 0$ for c, obtaining

$$c = \tfrac{1}{z}\left(x \pm \sqrt{y}\right), \tag{46}$$

where

$$\begin{aligned} x &= -2\beta(m-1)\omega((\beta+1)(m-1)-f) \\ &\quad -(2\beta+1)(m-1)\omega^2+\omega^3, \qquad (47)\\ y &= \omega(-4\beta^2(m-1)^2(\omega-2)-4\beta(m-1)\omega(m+\omega-3) \\ &\quad +\omega(-m+\omega+1)^2)(\beta f(m-1)+\omega^2), \qquad (48)\\ z &= 2\beta(m-1)\omega. \qquad (49) \end{aligned}$$

We substitute the expression for c with $+$ sign in H and solve $\Delta_{1,2;1,2} = 0$ for f. This leads to

$$f = -\frac{\omega^2}{\beta(m-1)}. \tag{50}$$

Finally, we eliminate f, d and c from equations (42), (44) and (46). This leads to (12)–(17), observing that $m = s = \kappa^{-1}$ when $L = 1$. Note that (11) can be obtained from (12)–(17) by forcing $H \succeq 0$ (see the proof of Theorem 3).

V. Numerical Results and Discussion

We first note that our optimal choice for β in (25) is numerically very close, but not equal, to Nestervo's choice in (2); see Figure 3 (left). Our convergence rate for NAM is almost indistinguishable to τ_{LG} in Figure 1, and it is indistinguishable from the curve obtained by running the Matlab code of [7] for the plot of τ_{LG} with our optimal choice of α and β. However, plotting τ_{LG} for the choice in (7) gives a numerical rate that is better than the one derived in this paper; see Figure 3 (right). This shows that we have not extracted the best possible convergence rate for NAM from the IQC framework. Indeed, we assumed that $\rho = \tau$ and $\alpha = 1/L$ which might be suboptimal. We did so because we were unable to find an ansatz without restricting α or ρ. There are too many free variables to perform closed form calculations, e.g. could not solve some of the resulting polynomial equations.

We know that any bound produced by the IQC-framework must be above or equal to τ_{LQ} in Figure 3. It is an important open question to know what is the best possible bound that the IQC-framework can produce. Can it reach τ_{LQ}?

VI. Conclusion and Future Work

We have derived a new, improved, and explicit convergence rate of Nesterov's accelerated method for strongly convex functions. Our numerical experiments using the IQC framework [7] show that our results can be further improved. Future work should include deriving better and explicit convergence rates using the IQC framework, and demonstrating that these cannot be improved. It would also be important to know if IQC allows us to prove the best possible upper bound on the convergence rate of Nesterov's method. To do so, one would have to produce a family of "bad" functions for which the convergence rate of Nesterov's method matches the rate obtained from IQC.

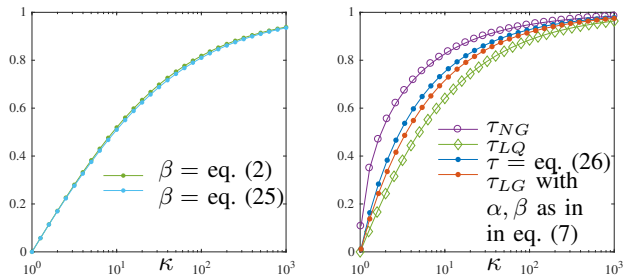

Fig. 3. *Left:* There is a very small difference between the standard choice for β given in (25) and our optimal choice of β in (2). *Right:* It is possible to obtain better rates than the one we derived in this paper if we choose α and β as in (7).

Acknowledgment

This work was partially funded by NIH/1U01AI124302 and NSF/IIS-1741129.

References

[1] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM. J. Imaging Sciences*, vol. 2, pp. 183–202, 2009.

[2] S. Becker, J. Bobin, and E. J. Candès, "NESTA: A Fast and Accurate First-Order Method for Sparse Recovery," *SIAM J. Imaging Sci.*, vol. 4, pp. 1–39, 2011.

[3] J. A. Tropp, J. N. Laska, and M. F. Duarte, "Beyond Nyquist: Efficient Sampling of Sparse Bandlimited Signals," *IEEE Transactions on Information Theory*, vol. 56, pp. 520–544, 2010.

[4] "An accelerated gradient method for trace norm minimization."

[5] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *International Conference on Machine Learning*, vol. 28, no. 3, pp. 1139–1147, 2013.

[6] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2013.

[7] L. Lessard, B. Recht, and A. Packard, "Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints," vol. 26, no. 1, pp. 57–95, 2016.

[8] G. França and J. Bento, "An Explicit Rate Bound for Over-Relaxed ADMM," *IEEE International Symposium on Information Theory (ISIT)*, pp. 2104–2108, 2016.

[9] Z. Allen-Zhu and L. Orecchia, "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent," 2014.

[10] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, pp. 127–152, 2005.

[11] W. Su, S. Boyd, and E. Candès, "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[12] N. Flammarion and F. Bach, "From Averaging to Acceleration, There is Only a Step-size," *Conference on Learning Theory*, vol. 40, 2015.

[13] Y. Arjevani, S. Shalev-Shwartz, and O. Shamir, "On lower and upper bounds in smooth and strongly convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 126, pp. 1–51, 2016.

[14] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to Nesterov's accelerated gradient descent," 2015.

[15] F. Zhang, *The Schur complement and its applications*. Springer Science & Business Media, 2006, vol. 4.

[16] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*. Springer Science & Business Media, 2003, vol. 15.

[17] C. D. Meyer, *Matrix analysis and applied linear algebra*. Siam, 2000, vol. 2.