An Explicit Rate Bound for the Over-Relaxed ADMM

Guilherme França
francag@bc.edu

José Bento
jose.bento@bc.edu

***Abstract*—The framework of Integral Quadratic Constraints of Lessard et al. (2014) reduces the computation of upper bounds on the convergence rate of several optimization algorithms to semi-definite programming (SDP). Followup work by Nishihara et al. (2015) applies this technique to the entire family of over-relaxed Alternating Direction Method of Multipliers (ADMM). Unfortunately, they only provide an explicit error bound for sufficiently large values of some of the parameters of the problem, leaving the computation for the general case as a numerical optimization problem. In this paper we provide an exact analytical solution to this SDP and obtain a general and explicit upper bound on the convergence rate of the entire family of over-relaxed ADMM. Furthermore, we demonstrate that it is not possible to extract from this SDP a general bound better than ours. We end with a few numerical illustrations of our result and a comparison between the convergence rate we obtain for the ADMM with known convergence rates for the Gradient Descent.**

I. Introduction

Consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x) + g(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^p$, $z \in \mathbb{R}^q$, $A \in \mathbb{R}^{r \times p}$, $B \in \mathbb{R}^{r \times q}$, and $c \in \mathbb{R}^r$ under the following additional assumption, which we assume throughout the paper.

Assumption 1.

1) *The functions f and g are 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*

$$m\|x-y\|^2 \leq \left(\nabla h(x) - \nabla h(y)\right)^T (x-y) \leq L\|x-y\|^2$$

for all $x, y \in \mathbb{R}^d$ where $0 < m \leq L < \infty$; We assume that $f \in S_p(m, L)$, in other words, f is strongly convex and ∇f is Lipschitz continuous; and that $g \in S_q(0, \infty)$;
3) *A is invertible and B has full column rank.*

In this paper we give an explicit convergence rate bound for a family of optimization schemes known as over-relaxed ADMM when applied to the optimization problem (1). This family is parametrized by $\alpha > 0$ and $\rho > 0$ and when applied to (1) takes the form in Algorithm 1. A classical choice of parameters is $\alpha = 1$ and $\rho = 1$. Several works have computed specific rate bounds for the ADMM under specific different regimes but a recent work by [2] allowed [1] to reduce the analysis of this entire family of solvers to finding solutions for a semi-definite programming problem. This SDP has multiple solutions and different solutions give different bounds on the convergence rate of the ADMM, some better than others. In their paper, [1] analyze this SDP numerically and also give one feasible solution to this SDP when $\kappa = (L/m)\kappa_A^2$ is sufficiently large, κ_A being the condition number of A. They further show, via a lower bound, that it is not possible to extract from this SDP a rate that is much better than the rate associated with their solution for large κ.

Algorithm 1 Family of Over-Relaxed ADMM schemes (parameters ρ, α)

```
Input: $f$, $g$, $A$, $B$, $c$;
Initialize $x_0, z_0, u_0$
repeat
    $x_{t+1} = \arg\min_x f(x) + \frac{\rho}{2}\|Ax + Bz_t - c + u_t\|^2$
    $z_{t+1} = \arg\min_z g(z) + \frac{\rho}{2}\|\alpha A x_{t+1} - (1-\alpha)Bz_t + Bz - \alpha c + u_t\|^2$
    $u_{t+1} = u_t + \alpha A x_{t+1} - (1-\alpha)Bz_t + Bz_{t+1} - \alpha c$
until stop criterion
```

An important problem remains open that we solve in this paper. Can we find a general explicit expression for the best[1] solution of this SDP? The answer is yes. As we explain later, our finding has both theoretical and practical interest.

II. Main results

We start by recalling the main result of [1] which is the starting point of our work. Based on the framework proposed in [2], it was later shown [1] that the iterative scheme of Algorithm 1 can be written as a dynamical system involving the matrices

$$\begin{aligned} \hat{A} &= \begin{bmatrix} 1 & \alpha-1 \\ 0 & 0 \end{bmatrix}, & \hat{B} &= \begin{bmatrix} \alpha & -1 \\ 0 & -1 \end{bmatrix}, \\ \hat{C}_1 &= \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}, & \hat{C}_2 &= \begin{bmatrix} 1 & \alpha-1 \\ 0 & 0 \end{bmatrix}, \\ \hat{D}_1 &= \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix}, & \hat{D}_2 &= \begin{bmatrix} \alpha & -1 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{2}$$

and the constants

$$\hat{m} = \frac{m}{\sigma_1^2(A)}, \qquad \hat{L} = \frac{L}{\sigma_p^2(A)}, \tag{3a}$$

$$\rho_0 = \rho(\hat{m}\hat{L})^{-1/2}, \qquad \kappa = \kappa_f \kappa_A^2. \tag{3b}$$

[1] "Best" in the sense that it gives the smallest rate bound.

Above, $\kappa_f = L/m$, $\sigma_1(A)$ ($\sigma_p(A)$) denotes the largest (smallest) singular value of the matrix A and $\kappa_A = \sigma_1(A)/\sigma_p(A)$ is the condition number of A; Throughout the paper, if M is a matrix κ_M denotes the condition number of M. Unless stated otherwise, throughout the paper we hold on to the definitions in (3).

The stability of this dynamical system is then related to the convergence rate of Algorithm 1 which in turn involves numerically solving a 4×4 semidefinite program as stated in following theorem.

Theorem 2 (See [1]). *Let the sequences $\{x_t\}$, $\{z_t\}$, and $\{u_t\}$ evolve according to Algorithm 1 with step size $\rho > 0$ and relaxation parameter $\alpha > 0$. Let $\varphi_t = [z_t, u_t]^T$ and φ_* be a fixed point of the algorithm. Fix $0 < \tau < 1$ and suppose there is a 2×2 matrix $P \succ 0$ and constants $\lambda_1, \lambda_2 \geq 0$ such that*

$$\begin{bmatrix} \hat{A}^T P \hat{A} - \tau^2 P & \hat{A}^T P \hat{B} \\ \hat{B}^T P \hat{A} & \hat{B}^T P \hat{B} \end{bmatrix} + \begin{bmatrix} \hat{C}_1 & \hat{D}_1 \\ \hat{C}_2 & \hat{D}_2 \end{bmatrix}^T \begin{bmatrix} \lambda_1 M_1 & 0 \\ 0 & \lambda_2 M_2 \end{bmatrix} \begin{bmatrix} \hat{C}_1 & \hat{D}_1 \\ \hat{C}_2 & \hat{D}_2 \end{bmatrix} \preceq 0 \quad (4)$$

where

$$M_1 = \begin{bmatrix} -2\rho_0^{-2} & \rho_0^{-1}(\kappa^{1/2} + \kappa^{-1/2}) \\ \rho_0^{-1}(\kappa^{1/2} + \kappa^{-1/2}) & -2 \end{bmatrix}, \quad (5)$$

$$M_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (6)$$

Then for all $t \geq 0$ we have

$$\|\varphi_t - \varphi_*\| \leq \kappa_B \sqrt{\kappa_P}\, \tau^t. \quad (7)$$

Notice that since A is non-singular, by step 6 in Algorithm 1, the rate bound τ also bounds $\|[x_t, z_t, u_t] - [x_*, z_*, u_*]\|$.

As already pointed out in [1], the weakness of Theorem 2 is that τ is not explicitly given as a function of the parameters involved in the problem, namely κ, ρ_0, and α. The factor κ_P in (7) is also not explicitly given. Therefore, for given values of these parameters one must perform a numerical search to find the minimal τ such that (4) is feasible. This in turn implies, for example, that to optimally tune the ADMM using this bound one might have perform this numerical search multiple times scanning the parameter space (α, ρ_0).

While from a practical point of view this may be enough for many purposes, this procedure can certainly introduce delays if, for example, (4) is used in an adaptive scheme where after every few iterations we estimate a local value for κ and then re-optimize α and ρ. Therefore, it is desirable to have an explicit expression for the smallest τ that Theorem 2 can give, from which the optimal values of the parameters follow. This expression is also desirable from a theoretical point of view. Our main goal in this paper is to complete the work initiated in [1], thus providing an explicit formula for the rate bound that the method of [2] can give for the over-relaxed ADMM.

Two of the most explicit bound rates that resemble the bound we give in this section are the ones found in [7] and [8]. The authors in [7] analyze the Douglas-Rachford splitting method, a scheme different but related to the one we analyze in this paper, for a problem similar to (1), and give a rate bound of $1 - \frac{\alpha}{1+\sqrt{\kappa_f}}$ where α is a step size and $\kappa_f = L/m$ where L and m bound the curvature of the objective function in the same sense as in Assumption 1. The authors in [8] apply the ADMM with $\alpha = 1$ and $\rho_0 = 1$ to the same problem as we do and give a rate bound of $1 - \frac{1}{\sqrt{\kappa}} + O(\frac{1}{\kappa})$, where, we recall, $\kappa = \kappa_f \kappa_A^2$.

We now state and prove our main results. Throughout the paper we often make use of the function

$$\chi(x) = \max(x, x^{-1}) \geq 1 \quad \text{for } x \in \mathbb{R} > 0. \quad (8)$$

Theorem 3. *For $0 < \alpha \leq 2$, $\kappa > 1$ and $\rho_0 > 0$, the following is an explicit feasible point of* (4) *with $\lambda_1, \lambda_2 \geq 0$, $P \succ 0$ and $0 < \tau < 1$.*

$$P = \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix}, \qquad \xi = -1 + \frac{\alpha(\chi(\rho_0)\sqrt{\kappa} - 1)}{1 - \alpha + \chi(\rho_0)\sqrt{\kappa}}, \quad (9)$$

$$\lambda_1 = \frac{\alpha \rho_0 \sqrt{\kappa}\,(1 - \alpha + \chi(\rho_0)\sqrt{\kappa})}{(\kappa - 1)\,(1 + \chi(\rho_0)\sqrt{\kappa})}, \quad (10)$$

$$\lambda_2 = 1 + \xi, \quad (11)$$

with

$$\tau = 1 - \frac{\alpha}{1 + \chi(\rho_0)\sqrt{\kappa}}. \quad (12)$$

Proof. First notice that since $\kappa > 1$ and $\chi(\rho_0) \geq 1$ we have that $\lambda_1, \lambda_2 \geq 0$ for the allowed range of parameters. Second, notice that the eigenvalues of P are $1 + \xi$ and $1 - \xi$ and since $\xi > -1$ we have that $P \succ 0$. Finally, consider the full matrix in the left hand side of (4) and let D_n denote an n^{th} principal minor. We will show through direct computation that $(-1)^n D_n \geq 0$ for all principal minors, which proves our claim.

Replacing (9)–(12) we have the matrix shown in equation (13). Note that it has vanishing determinant $D_0 = 0$. Let $J \subseteq \{1, 2, 3, 4\}$ and denote D_n^J the n^{th} principal minor obtained by deleting the rows and columns with indexes in J.

We consider the case $\rho_0 \geq 1$ first. The only *nonvanishing* principal minors are shown in equation (14). We obviously have (14a), (14b) ≥ 0 for the allowed range of parameters. For (14c) and (14d) we need to show that, for the allowed range of parameters, the concave 2nd order polynomial $w(\tilde{\kappa}) = 2(\alpha - 1)\tilde{\kappa} + (\alpha - 2)(1 + \tilde{\kappa}^2)\rho_0 - 2\rho_0^2\tilde{\kappa}$ is non-positive for $\tilde{\kappa} \equiv \sqrt{\kappa} > 1$. To do this, it suffices to show that the function and its first derivative are non-positive for $\tilde{\kappa} > 1$. We have $\partial_{\tilde{\kappa}} w(1) = w(1) = 2(1 + \rho_0)(\alpha - 1 - \rho_0) \leq 0$. Therefore $w(\tilde{\kappa}) \leq 0$ for $\tilde{\kappa} > 1$ implying that (14c), (14d) ≤ 0, as required. Analogously, for (14e) we only need to show that, for the allowed range of parameters, the 3rd order degree polynomial in $\tilde{\kappa} \equiv \sqrt{\kappa}$,

$$w(\tilde{\kappa}) = 2(\alpha - 1) + \rho_0\Big\{2(\alpha - 2)\tilde{\kappa} + \rho_0\{2 + 2\alpha(\tilde{\kappa}^2 - 1) - 4\tilde{\kappa}^2 + (\alpha - 2)(\tilde{\kappa}^2 - 1)\rho_0\tilde{\kappa}\}\Big\}, \quad (15)$$

which is the numerator in the fraction (14e), is non-positive for $\tilde{\kappa} > 1$. To do this, it suffices to show that the zeroth, first and second derivatives are non-positive. We have $w(1) =$

$$\left(\begin{array}{c|c|c|c} 1-\tau^2-\dfrac{2\lambda_1}{\rho_0^2} & \alpha-1-\xi\tau^2-\dfrac{2\lambda_1}{\rho_0^2} & \alpha-\dfrac{2\sqrt{\kappa}+\rho_0(1+\kappa)}{\rho_0^2\sqrt{\kappa}}\lambda_1 & 0 \\ \alpha-1-\xi\tau^2-\dfrac{2\lambda_1}{\rho_0^2} & (\alpha-1)^2-\tau^2-\dfrac{2\lambda_1}{\rho_0^2} & \alpha(\alpha-1)-\dfrac{2\sqrt{\kappa}+\rho_0(1+\kappa)}{\rho_0^2\sqrt{\kappa}}\lambda_1 & 0 \\ \alpha-\dfrac{2\sqrt{\kappa}+\rho_0(1+\kappa)}{\rho_0^2\sqrt{\kappa}}\lambda_1 & \alpha(\alpha-1)-\dfrac{2\sqrt{\kappa}+\rho_0(1+\kappa)}{\rho_0^2\sqrt{\kappa}}\lambda_1 & \alpha^2-\dfrac{2\rho_0^2\sqrt{\kappa}+2\sqrt{\kappa}+2\rho_0(1+\kappa)}{\rho_0^2\sqrt{\kappa}}\lambda_1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \tag{13}$$

$$D_2^{\{4,3\}} = \frac{2\alpha^2(2-\alpha)(\rho_0^2-1)\sqrt{\kappa}(1-\alpha+\rho_0\sqrt{\kappa})}{(\kappa-1)\rho_0(1+\rho_0\sqrt{\kappa})^3} \tag{14a}$$

$$D_2^{\{4,1\}} = D_2^{(4,3)}\cdot(1+\rho_0\sqrt{\kappa})^2 \tag{14b}$$

$$D_1^{\{4,3,2\}} = \alpha\cdot\frac{2(\alpha-1)\sqrt{\kappa}+(\alpha-2)(1+\kappa)\rho_0-2\rho_0^2\sqrt{\kappa}}{(\kappa-1)\rho_0(1+\rho_0\sqrt{\kappa})^2} \tag{14c}$$

$$D_1^{\{4,2,1\}} = D_1^{\{4,3,2\}}\cdot(1+\rho_0\sqrt{\kappa})^2 \tag{14d}$$

$$D_1^{\{4,3,1\}} = \alpha\sqrt{\kappa}\cdot\frac{2(\alpha-1)+\rho_0\big\{2(\alpha-2)\sqrt{\kappa}+\rho_0\big(2+2\alpha(\kappa-1)-4\kappa+(\alpha-2)(\kappa-1)\rho_0\sqrt{\kappa}\big)\big\}}{(\kappa-1)\rho_0(1+\rho_0\sqrt{\kappa})^2} \tag{14e}$$

$2(\alpha-1-\rho_0)(1+\rho_0) \le 0$, $\partial_{\tilde{\kappa}}w(1) = 2(\alpha-2)\rho_0(1+\rho_0)^2 \le 0$, and $\partial_{\tilde{\kappa}}^2 w(1) = 2(\alpha-2)\rho_0^2(2+3\rho_0) \le 0$. This implies that $w(\tilde{\kappa}) \le 0$ for $\tilde{\kappa} > 1$ and consequently (14e) ≤ 0. This concludes the proof for $\rho_0 \ge 1$.

For $\rho_0 < 1$ the analogous expressions to (14) are slightly different but the previous argument holds in exactly the same manner, thus we omit the details. □

In the following corollary, we allow $\kappa = 1$ but $0 < \alpha < 2$. It gives an explicit bound on the convergence rate of the over relaxed ADMM.

Corollary 4. *Consider the sequences $\{x_t\}$, $\{z_t\}$, and $\{u_t\}$, updated according to Algorithm 1 with step size $\rho > 0$, relaxation parameter $0 < \alpha < 2$ and for a problem with $\kappa \ge 1$. Let $\varphi_t = [z_t, u_t]^T$ and φ_* be a fixed point. Then the convergence rate of the over-relaxed ADMM obeys the following upper bound:*

$$\|\varphi_t - \varphi_*\| \le \kappa_B\sqrt{\chi(\eta)}\,\tau^t \tag{16}$$

with τ explicitly given by the formula (12) *and*

$$\eta = \frac{\alpha}{2-\alpha}\cdot\frac{\chi(\rho_0)\sqrt{\kappa}-1}{\chi(\rho_0)\sqrt{\kappa}+1}. \tag{17}$$

Proof. The proof for $\kappa > 1$ follows directly from Theorem 2 and Theorem 3. Indeed, all that we need to do is to compute κ_P in (7) for P as in Theorem 3. The two eigenvalues of P are $1-\xi$ and $1+\xi$ and the ratio of the largest to the smallest is precisely $\chi(\eta)$ where η is given in equation (17).

For $\kappa = 1$ the proof follows by continuity. First notice that, from one iteration to the next in Algorithm 1, $(x_{t+1}, z_{t+1}, u_{t+1})$ is a continuous function of (x_t, z_t, u_t, A) in a neighborhood of an invertible A if we assume everything else fixed (this can be derived from the properties of proximal operators, c.f. [5]). Therefore by the continuity of the composition of continuous functions, and assuming only A is free and everything else is fixed, $\|\varphi_t - \varphi_*\| = F(A)$ for some function that is continuous around a neighborhood of an invertible A. Now, add a small perturbation δA to A such that $\kappa_A > 1$. This perturbation makes $\kappa > 1$ and by the first part of this proof we can write that $F(A+\delta A) \le \kappa_B\sqrt{\chi(\eta+\delta\eta)}(\tau+\delta\tau)^t$, where $\delta\eta$ and $\delta\tau$ are themselves continuous functions of δA since both η and τ depend continuously on κ which in turn depends continuously on δA, around an invertible A. The theorem follows by letting $\delta A \to 0$ and using the fact that $\lim_{\substack{\delta A\to 0\\ \kappa_A>1}} F(A+\delta A) = F(A)$. □

The next result complements Theorem 3 by showing that the rate bound in equation (12) is the smallest one can get from the feasibility problem in Theorem 2.

Theorem 5. *If $0 < \alpha < 2$, $\rho_0 > 0$ and $\kappa \ge 1$, then the smallest τ for which one can find a feasible point of* (4) *is given by* (12).

Proof. The proof will follow by contradiction. Our counterexample follows [1] and [3]. Assume that for some $0 < \alpha < 2$, $\rho_0 > 0$ and $\kappa \ge 1$ it is possible to find a feasible solution with $\tau < \nu = 1 - \frac{\alpha}{1+\chi(\rho_0)\sqrt{\kappa}}$. Then, if we use the ADMM with this α and a $\rho = \rho_0\sqrt{\hat{m}\hat{L}}$ to solve any optimization problem with this same value of κ and satisfying Assumption 1 we have by Theorem 2 that $\|\varphi_t - \varphi_*\| \le C\tau^t$, where $\tau < \nu$ and $C > 0$ is some constant. In particular, if $\rho_0 \ge 1$, this bound on the error rate must hold if we try to solve a problem where $f(x) = \frac{1}{2}x^TQx$ and $g(z) = 0$, with $Q = \operatorname{diag}([m, L]) \in \mathbb{R}^{2\times 2}$, $A = I$, $B = -I$, and $c = 0$. Note that for this problem $\kappa_A = 1$, $\kappa = \kappa_f = L/m$, $\hat{m} = m$ and $\hat{L} = L$.

Applying Algorithm 1 to this problem yields

$$z_{t+1} = \big(I - \alpha(Q+I\rho)^{-1}Q\big)\,z_t. \tag{18}$$

If $z_{t=0}$ is in the direction of the smallest eigenvalue of Q, the error rate for z_t is,

$$1-\frac{\alpha}{1+\rho m^{-1}} = 1-\frac{\alpha}{1+\rho_0\sqrt{\kappa}}, \tag{19}$$

where in the second equality we replaced (3). But this means that the error rate for $\|\varphi_t - \varphi_*\|$ cannot be bounded by $\tau < \nu$ for $\rho_0 \geq 1$, which contradicts our original assumption.

The proof when $\rho_0 < 1$ is similar. We apply ADMM to the same problem as above but now with $A = \rho_0 I$ and the rest the same. Note that for this modified problem $\kappa_A = 1$, $\kappa = \kappa_f = L/m$, $\hat{m} = m/\rho_0^2$, $\hat{L} = L\rho_0^2$ and the ρ we choose for the ADMM is now $\rho = \sqrt{Lm}/\rho_0$ (while before it was $\rho = \rho_0\sqrt{Lm}$). □

Now we compare the rate bound of the ADMM with the rate bound of the gradient descent (GD) when we solve problem (1) with $B = I$. In what follows we use τ_{ADMM} and τ_{GD} when talking about rates of convergence for the ADMM and the GD respectively.

Before we state our result let us discuss how the GD behaves when we use it to solve this problem. To solve problem (1) using the GD with $B = I$ we reduce the problem to an unconstrained formulation by applying the GD to the function $F(z) = \tilde{f}(z) + g(z)$ where $\tilde{f}(z) = f(A^{-1}(c-z))$. We now assume that $F \in S_p(m_F, L_F)$ for some $0 < m_F \leq L_F < \infty$. The work of [6] gives an optimally tuned rate bound for the GD when applied to any objective function in $S_p(m_F, L_F)$. This rate is $1 - \frac{2}{1+\kappa_F}$ where $\kappa_F = L_F/m_F$. It is easy to see that, among all general bounds that only depend on κ_F, it is not possible to get a function smaller than this. Indeed, if the objective function is $x^T\text{diag}([m_F, L_F])x$ then the rate of convergence of the GD with step size β is given by the spectral radius of the matrix $I - \beta\text{diag}(\{m_F, L_F\})$ which is $\max\{|1-\beta L_F|, |1-\beta m_F|\}$ and which in turn has minimum value $1 - \frac{2}{1+\kappa_F}$ for $\beta = 2/(L_F + m_F)$. If $\mathcal{P}(\kappa_F)$ is the family of this unconstrained formulation of problem (1) with $B = I$ and $L_F/m_F = \kappa_F$, then we can summarize what we describe above as

$$\inf_{\beta} \sup_{\mathcal{P}(\kappa_F)} \tau_{\text{GD}} = 1 - \frac{2}{1+\kappa_F}. \tag{20}$$

In a similar way, if $\mathcal{P}(\kappa)$ is the family of problems of the form (1) with $B = I$, to be solved using Algorithm 1 under Assumption 1, where $f \in S_p(m, L)$ and $\kappa = L/m$, then Corollary 4 and the counterexample in the proof of Theorem 5 give us that

$$\inf_{\alpha,\rho_0} \sup_{\mathcal{P}(\kappa)} \tau_{\text{ADMM}} \leq \inf_{\alpha>2,\rho_0} \sup_{\mathcal{P}(\kappa)} \tau_{\text{ADMM}} = 1 - \frac{2}{1+\sqrt{\kappa}}, \tag{21}$$

where the last equality is obtained by setting $\alpha = 2$ and $\rho_0 = 1$ in equation (12).

The next theorem shows that the optimally tuned ADMM for worse-case problems has faster convergence rate than the optimally tunned GD for worse-case problems.

Theorem 6. *Let $\mathcal{P}(\kappa_F, \kappa)$ be the family of problems* (1) *with $B = I$ and under Assumption 1 such that $f \in S_p(m, L)$ with $L/m = \kappa$ and $F \in S_p(m_F, L_F)$ with $L_F/m_F = \kappa_F$, then*

$$\tau^*_{ADMM} \equiv \inf_{\alpha,\rho_0} \sup_{\mathcal{P}(\kappa_F,\kappa)} \tau_{ADMM} \leq \tau^*_{GD} \equiv \inf_{\beta} \sup_{\mathcal{P}(\kappa_F,\kappa)} \tau_{GD}. \tag{22}$$

More specifically,

$$\tau^*_{GD} \geq \frac{2\tau^\star_{ADMM}}{1+(\tau^\star_{ADMM})^2}. \tag{23}$$

Proof. First notice that (20) still holds if $\mathcal{P}(\kappa_F)$ is replaced by $\mathcal{P}(\kappa_F, \kappa)$ since the objective function used in the example given above (20) is also in $\mathcal{P}(\kappa_F, \kappa)$.

Second notice that, since $f \in S_p(m, L)$ and A is non-singular, we have that $\tilde{f} \in S_p(m_{\tilde{f}}, L_{\tilde{f}})$ for some $0 < m_{\tilde{f}} \leq L_{\tilde{f}} < \infty$. Thus, since $F = \tilde{f} + g \in S_p(m_F, L_F)$, we have that $g \in S_p(m_g, L_g)$ for some $0 \leq m_g \leq L_g < \infty$ (it might be that $m_g = 0$, i.e., g might not be strictly convex). Notice in addition that, without loss of generality, we can assume that $L_F \geq L_{\tilde{f}} + m_g$, $m_F \leq m_{\tilde{f}} + m_g$, $L_{\tilde{f}} = (\sigma_1(A^{-1}))^2 L_f$ and $m_{\tilde{f}} = (\sigma_p(A^{-1}))^2 m_f$. Therefore, if $F \in S_p(m_F, L_F)$ and $f \in S_p(m, L)$, then without loss of generality

$$\begin{aligned}\kappa_F = \frac{L_F}{m_K} \geq \frac{L_{\tilde{f}} + m_g}{m_{\tilde{f}} + m_g} \geq \frac{L_{\tilde{f}}}{m_{\tilde{f}}} = \frac{L_f(\sigma_1(A^{-1}))^2}{m_f(\sigma_p(A^{-1}))^2} \\ = \kappa_f(\kappa_{A^{-1}})^2 = \kappa_f(\kappa_A)^2 = \kappa. \end{aligned} \tag{24}$$

Finally, using the fact $\kappa_F \geq \kappa$ and equations (20) and (21) we can write

$$\begin{aligned}\inf_{\alpha,\rho_0} \sup_{\mathcal{P}(\kappa_F,\kappa)} \tau_{\text{ADMM}} \leq \inf_{\alpha,\rho_0} \sup_{\mathcal{P}(\kappa)} \tau_{\text{ADMM}} \leq 1 - \frac{2}{1+\sqrt{\kappa}} \\ \leq 1 - \frac{2}{1+\kappa} \leq 1 - \frac{2}{1+\kappa_F} = \inf_{\beta} \sup_{\mathcal{P}(\kappa_F,\kappa)} \tau_{\text{GD}}. \end{aligned} \tag{25}$$

Equation (23) follows from the fact that $\kappa_F \geq \kappa$ and the fact that $1 - \frac{2}{1+\sqrt{\kappa}} \leq 1 - \frac{2}{1+\kappa_F}$. □

III. Numerical Results

We now compare numerical solutions to the SDP in Theorem 2 with the exact formulas from Theorem 3. The numerical procedure was implemented in MATLAB using CVX and a binary search to find the minimal τ such that (4) is feasible. This is exactly the same procedure described in [1] and it works because the maximum eigenvalue of (4) decrease monotonically with τ. Figure 1 shows the rate bound τ against κ for several choices of parameters (α, ρ_0). The dots correspond to the numerical solutions and the solid lines correspond to the exact formula (12). Figure 2 compare the numerical values of λ_1 (circles) and λ_2 (squares) with the formulas (10) and (11) (solid lines). There is a perfect agreement between (9)–(12) and the numerical results, which strongly support Theorem 3 and Theorem 5.

The range $0 < \alpha < 1$ give worse convergence rates compared to $1 \leq \alpha < 2$. The best rate bound is attained with $\rho_0 = 1$, or equivalently $\rho = \sqrt{\hat{m}\hat{L}}$, and $\alpha = 2$. This is also evident from (12). Note, however, that (17) diverges when $\alpha \to 2$ so although the optimal rate bound, in the asymptotical sense, is $1 - \frac{2}{1+\chi(\rho_0)\sqrt{\kappa}}$, bound (16) suggests that in a practical setting with a maximum number of iterations it might be better to choose $\alpha < 2$.

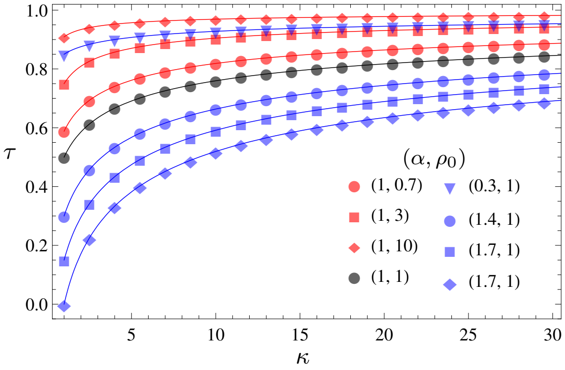

Fig. 1. Plot of τ versus κ for different values of parameters (α, ρ_0), as indicated in the legend. The dots correspond to the numerical solution to (4) while the solid curves are the exact formula (12). The best choice of parameters are $\rho_0 = 1$ and $\alpha = 2$. The convergence rate is improved with the choice $1 \le \alpha < 2$ compared to $0 < \alpha < 1$.

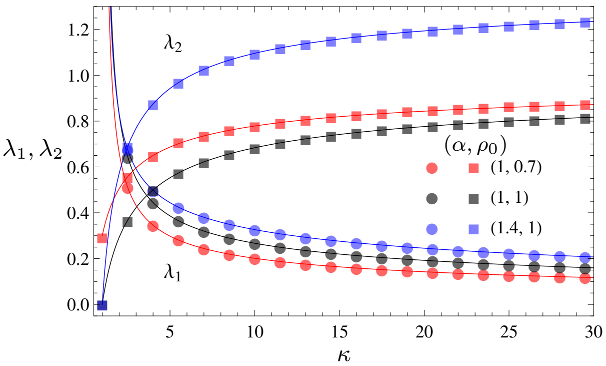

Fig. 2. We show λ_1 (circles) and λ_2 (squares) verus κ for some of the choices of parameters (α, ρ_0) in Figure 1. Note the exact match of numerical results with formulas (10) and (11) (solid lines).

Corollary 4 is valid only for $0 < \alpha < 2$ (for $\alpha > 2$, (12) can assume negative values). However, Theorem 2 does not impose any restriction on α and holds even for $\alpha > 2$ [1]. To explore the range $\alpha > 2$ we numerically solve (4) as shown in Figure 3. The dots correspond to the numerical solutions. The dashed blue line corresponds to (12) with $\alpha = 2$, and it is the boundary of the shaded region in which (12) can have negative values and is no longer valid. Although Theorem 3 does not hold for $\alpha > 2$, we deliberately included the solid lines representing (12) inside this region. Obviously, these curves do not match the numerical results.

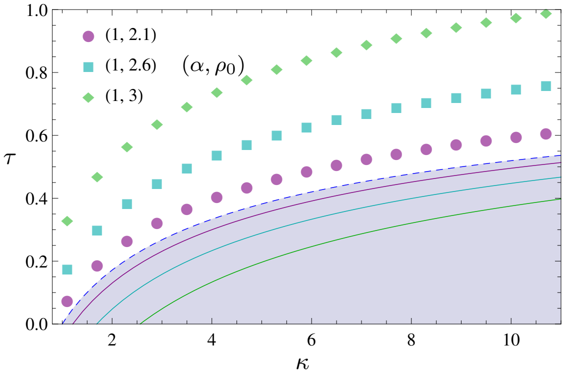

Fig. 3. Plot of τ versus κ for some values of (α, ρ_0) with $\alpha > 2$ and $\rho_0 = 1$. The dashed blue line corresponds to $\alpha = 2$ in formula (12). The shaded region contains curves given by (12) for values of α not allowed in Theorem 3. These numerical solutions with $\alpha > 2$ had to be restricted to a range $1 < \kappa \lesssim 11$. Moreover, notice that $\alpha > 2$ does not produce better convergence rates than $1 \le \alpha < 2$ through (12), which is valid for any $\kappa > 1$.

The first important remark is that, for a given $\alpha > 2$, we were unable to numerically find solutions for arbitrary $\kappa \ge 1$. For instance, for $\alpha = 2.6$ we can only stay roughly on the interval $1 < \kappa \lesssim 11$. The same behavior occurs for any $\alpha > 2$, and the range of κ becomes narrower as α increases. From the picture one can notice that $\tau = 1$ is actually attained with *finite* κ, while for (12) this never happens; it rather approaches $\tau \to 1^-$ as $\kappa \to \infty$. Therefore, although it is feasible to solve (4) with $\alpha > 2$, the solutions will be constrained to a small range of κ. The next question would be if Theorem 2 for $\alpha > 2$ could possibly give a better rate bound than Corollary 4 with $1 \le \alpha < 2$. We can see from the picture that this is probably not the case. We conclude that, as far as solutions to (4) are considered, there is no advantage in considering $\alpha > 2$ compared to (12) with $1 \le \alpha < 2$, and which holds for arbitrary $\kappa > 1$. It is an interesting problem to determine if proof techniques other than [2] can lead to good rate bounds for $\alpha > 2$.

IV. Conclusion

We introduced a new explicit rate bound for the entire family of over-relaxed ADMM. Our bound is the first of its kind and improves on [1] and [8]. In particular, the only explicit bound in [1] is a special case of our general explicit formula when κ is large. We also show that our bound is the best one can extract from the integral quadratic constrains framework of [2]. In [9] we find that $1 - 2/(1 + \sqrt{\kappa})$ bounds the convergence rate of any first order method on $S(m, L)$, $\kappa = m/L$, so we have also shown that the ADMM with $\alpha \to 2$ is close to being optimal on $S(m, L)$.

Although our analysis assumes that f is strongly convex, we can use a very-slightly modified ADMM algorithm to solve problem 1 when f is weakly convex using an idea of Elad Hazan explained in [2] Section 5.4.

References

[1] R. Nishihara, L. Lessard, B. Recht, A. Packard, M. I. Jordan, "A General Analysis of the Convergence of ADMM", *Int. Conf. on Machine Learning* 32 (2015), arXiv:1502.02009 [math.OC]

[2] L. Lessard, B. Recht, A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints" (2014), arXiv:1408.3595 [math.OC]

[3] E. Ghadimi, A. Teixeira, I. Shames, "Optimal Parameter Selection for the Alternating Direction Method of Multipliers (ADMM): Quadratic Problems", *IEEE Trans. on Automatic Control* 60 4 (2015)

[4] Y. Nesterov, "Introductory Lectures on Convex Optimization: A Basic Course", Kluwer Academic Publishers, Boston, MA, 2004

[5] N. Parikh, S. Boyd. "Proximal algorithms", *Foundations and Trends in optimization* 1.3 (2013): 123-231.

[6] Y. Nesterov. "Introductory lectures on convex optimization", Vol. 87 Springer Science and Business Media, 2004.

[7] G. Pontus, S. Boyd. "Diagonal scaling in Douglas-Rachford splitting and ADMM", *Decision and Control (CDC)*, 2014 IEEE 53rd Annual Conference on. IEEE, 2014.

[8] D. Wei, W. Yin. "On the global and linear convergence of the generalized alternating direction method of multipliers", *Journal of Scientific Computing* (2012): 1-28.

[9] Y. Nesterov. "Introductory lectures on convex optimization", *Springer Science & Business Media* (2004): 87.