On the Optimal Step-size Selection for the Alternating Direction Method of Multipliers ⋆

Euhanna Ghadimi, André Teixeira, Iman Shames and Mikael Johansson

School of Electrical Engineering, KTH, Stockholm, Sweden.
e-mail: {euhanna, andretei, imansh, mikaelj}@kth.se

Abstract: The alternating direction method of multipliers is a powerful technique for structured large-scale optimization that has recently found applications in a variety of fields including networked optimization, estimation, compressed sensing and multi-agent systems. While applications of this technique have received a lot of attention, there is a lack of theoretical support for how to set the algorithm parameters, and its step-size is typically tuned experimentally. In this paper we consider three different formulations of the algorithm and present explicit expressions for the step-size that minimizes the convergence rate. We also compare our method with one of the existing step-size selection techniques for consensus applications.

1. INTRODUCTION

The *alternating direction method of multipliers* (ADMM) is a powerful algorithm for solving structured convex optimization problems. Combining the strong convergence properties of the method of multipliers and the decomposability property of dual ascent, the method is particularly applicable to large-scale decision problems such as compressed sensing (Yang & Zhang 2011), image processing (Figueiredo & Bioucas-Dias 2010), regularized estimation (Wahlberg et al. 2012), and support vector machines (Forero et al. 2010). This broad array of applications has triggered recent attention in developing a better understanding of the theoretical properties of ADMM.

The origins of ADMM can be traced back to the alternating direction implicit (ADI) techniques for solving elliptic and parabolic partial difference equations. In the 70's, see Boyd et al. (2011) and references therein, ADMM was first introduced for solving optimization problems and enjoyed much attention in the following years. However, the main advantage of applying ADMM in solving optimization problems, its ability to deal with very large problem through its superior stability properties and its decomposability, remained largely untapped due to the lack of ubiquity of very large scale problems. Nevertheless, the technique has again raised to prominence in the last few years as there are many applications, *e.g.* in financial or biological data analysis, that are too large to be handled by generic optimization solvers.

Many large-scale problems can be cast as convex optimization problems. If the problem has favourable structure, then decomposition techniques such as primal and dual decomposition allows to distribute the computations on multiple processors. One then isolate subproblems that can be solved effectively, and coordinate these using gradient or sub-gradient methods. If problem parameters, such as Lipschitz constants and convexity parameters are known, then optimal step-sizes and associated convergence rates are well-known (Nesterov (2004)). Unfortunately, the convergence properties of the gradient method can be sensitive to the choice of the step-size, even to the point where poor parameters can lead to algorithm divergence (Ghadimi et al. 2011). The ADMM technique, on the other hand, converges for all values of a single tuning parameter. However, the parameter also influences the numerical conditioning and convergence speed of the method, and there is currently a lack of theoretical support for how to optimally tune this parameter. The aim of this paper is to contribute to a better understanding of the convergence properties of the ADMM method and to develop optimal step-size rules for some particular classes of problems.

The outline of this paper is as follows. In the next section we review the necessary background on the ADMM method. In Section 3 we study l_2-regularized quadratic programming and give explicit expressions for the optimal step-size that achieves the optimal convergence rate. We then shift our focus to the problem of achieving consensus in networks using ADMM in Section 4. We pose the problem for general graphs and provide closed-form solutions for the optimal step-size for the particular case of k-regular graphs. In Section 5 we briefly visit the problem of l_1-regularized quadratic programming and comment on equitable step-size selection policies. Numerical results are presented in Section 6. Concluding remarks and future directions are presented in the final section.

2. THE ADMM METHOD

The ADMM algorithm solves problems of the form

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

where f and g are convex functions, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$ and $c \in \mathbb{R}^p$; see Boyd et al. (2011) for a detailed review. Our focus on this paper is on the restricted class of problems of the form

⋆ This work was sponsored in part by the Swedish Foundation for Strategic Research, SSF, and the Swedish Research Council, VR.

$$\begin{aligned} &\text{minimize} \quad f(x)+g(z) \\ &\text{subject to } x-z=0. \end{aligned} \tag{1}$$

Relevant examples that can be put in this form are, *e.g.* regularized estimation, where f is the estimator loss and g is the regularization term, and various forms of networked optimization, see Erseghe et al. (2011) and (Boyd et al. 2011, § 7,8). The method is based on the *augmented Lagrangian*

$$L_\rho(x,z,\mu)=f(x)+g(z)+(\rho/2)\|x-z\|_2^2+\mu^T(x-z)$$

and performs sequential minimization of the x and z variables, followed by a dual variable update:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_x L_\rho(x,z^k,\mu^k) \\ z^{k+1} &= \operatorname*{argmin}_z L_\rho(x^{k+1},z,\mu^k) \\ \mu^{k+1} &= \mu^k+\rho(x^{k+1}-z^{k+1}). \end{aligned} \tag{2}$$

These iterations indicate that the method is particularly useful when the x- and z-minimizations can be carried out efficiently (*e.g.* admit closed-form expressions). One advantage of the method is that there is only one single algorithm parameter, ρ, and under rather mild conditions, the method can be shown to converge for all values of the parameter; see, e.g., Boyd et al. (2011), Mota et al. (2011). This is in contrast with, *e.g.* the gradient method where the iterates diverge if the step-size parameter is chosen too large. However, ρ has a direct impact on the convergence speed of the algorithm, and inadequate tuning of this parameter can render the method very slow. In the remaining parts of this paper, we will derive explicit expressions for the step-size that minimizes the convergence time for some particular classes of problems.

3. OPTIMAL CONVERGENCE SPEED FOR ℓ_2-REGULARIZED QUADRATIC MINIMIZATION

Regularized estimation problems

$$\text{minimize } f(x)+\frac{\delta}{2}\|x\|_p^q$$

are abound in statistics, machine learning, and control. In particular, ℓ_1-regularized estimation where $f(x)$ is quadratic and $p=q=1$, and *sum of norms* regularization, where $f(x)$ is quadratic, $p=2$ and $q=1$ have recently received significant attention.

Our initial result will focus on ℓ_2-regularized estimation, where $f(x)$ is quadratic and $p=q=2$, since the corresponding ADMM-iterations are linear and amendable to analysis. To this end, consider a problem in the form

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}x^\top Qx+q^\top x+\frac{\delta}{2}\|z\|_2^2 \\ &\text{subject to } x-z=0, \end{aligned} \tag{3}$$

where $Q\in\mathbb{S}_+^n$ is a positive definite $n\times n$ matrix, $x,q,z\in\mathbb{R}^n$ and $\delta\in\mathbb{R}_+$ is constant. The ADMM iterations read

$$\begin{aligned} x^{k+1} &= (Q+\rho I)^{-1}(\rho z^k-\mu^k-q) \\ z^{k+1} &= \frac{\mu^k+\rho x^{k+1}}{\delta+\rho} \\ \mu^{k+1} &= \mu^k+\rho(x^{k+1}-z^{k+1}). \end{aligned} \tag{4}$$

The z-update implies that $\mu^k=(\delta+\rho)z^{k+1}-\rho x^{k+1}$, that together with the μ-update gives

$$\mu^{k+1}=(\delta+\rho)z^{k+1}-\rho x^{k+1}+\rho(x^{k+1}-z^{k+1})=\delta z^{k+1}.$$

The iterations in (4) converge whenever the iterates on μ converges, *i.e.* when $x^k=z^k$. Hence, to study the convergence of (4) one can investigate how the errors associated with x^k or z^k vanish. Inserting the x-update into the z-update and using that $\mu^k=\delta z^k$, we find that

$$z^{k+1}=\frac{\delta I+\rho(\rho-\delta)(Q+\rho I)^{-1}}{\delta+\rho}z^k-\frac{\rho(Q+\rho I)^{-1}}{\delta+\rho}q.$$

Denote $e_{k+1}:=z^{k+1}-z^k$. Then

$$e_{k+1}=\frac{1}{\delta+\rho}\left(\delta I+\rho(\rho-\delta)\left(Q+\rho I\right)^{-1}\right)e_k, \tag{5}$$

and $E:=\frac{1}{\delta+\rho}\left(\delta I+\rho(\rho-\delta)\left(Q+\rho I\right)^{-1}\right)$. Hence, the convergence of (4) can be studied via the error dynamics (5). This allows us to state the following result.

Theorem 1. The iterations (4) converge for all values of $\rho>0$ and $\delta>0$. The optimal constant step-size $\rho^\star$ which minimizes the convergence rate is given by

$$\rho^\star=\begin{cases}\sqrt{\delta\lambda_1(Q)} & \text{if } \delta<\lambda_1(Q),\\ \sqrt{\delta\lambda_n(Q)} & \text{if } \delta>\lambda_n(Q),\\ \delta & \text{otherwise.}\end{cases} \tag{6}$$

Proof. The iterations (4) converge if and only if the spectral radius of the matrix E in (5) is less than one. Let $\lambda_i, i=1,\dots,n$ be the eigenvalues of Q. Then, the eigenvalues of E are given by

$$f(\rho,\lambda_i)=\frac{\delta+\frac{\rho(\rho-\delta)}{\lambda_i+\rho}}{\delta+\rho}=\frac{\rho^2+\lambda_i\delta}{\rho^2+(\lambda_i+\delta)\rho+\lambda_i\delta}. \tag{7}$$

Since $\lambda_i,\rho,\delta\in\mathbb{R}_+$, we conclude that $|f|<1$ which completes the first part of the proof. To find $\rho^\star$, note that

$$\rho^\star=\operatorname*{argmin}_\rho \max_i\{f(\rho,\lambda_i)\} \tag{8}$$

From the first equality in (7), $f(\rho,\lambda)$ is monotone decreasing in λ when $\rho>\delta$ and monotone increasing when $\rho<\delta$. Hence, when $\rho>\delta$,

$$\max_i f(\rho,\lambda_i)=f(\rho,\lambda_1)$$

By first-order optimality condition, $f(\rho,\lambda_1)$ is minimized by $\rho^\star=\sqrt{\lambda_1\delta}$. However, $\rho^\star>\delta$ only if $\lambda_1>\delta$. For $\delta\geq\lambda_1$, $0\leq(\rho-\delta)^2\leq(\rho-\delta)(\rho-\lambda_1)$ implies that

$$f(\rho,\lambda_1)\geq\frac{\rho^2+\lambda_1\delta}{\rho^2+(\lambda_1+\delta)\rho+\lambda_1\delta+(\rho-\delta)(\rho-\lambda_1)}=\frac{1}{2}$$

but $\rho=\delta$ attains $f(\delta,\lambda_1)=1/2$ and is hence optimal.

A similar argument applies for $\rho<\delta$, in which case

$$\max_i f(\rho,\lambda_i)=f(\rho,\lambda_n)$$

Similarly as above, when $\delta>\lambda_n$, $\rho^\star=\sqrt{\lambda_n\delta}$ is the optimal step-size. For $\delta\leq\lambda_n$, $0\leq(\delta-\rho)^2\leq(\lambda_n-\rho)(\delta-\rho)$ gives

$$f(\rho,\lambda_n)\geq\frac{1}{2}$$

hence $\rho=\delta$ is optimal. ■

Fig. 1 depicts the values of the optimal step-size $\rho^\star$ versus the penalty constant δ.

Corollary 2. For $\rho=\delta$,

$$\lambda_i(E)=1/2,\qquad i=1,\dots,n$$

and the convergence factor of the error dynamics (5) is independent of Q.

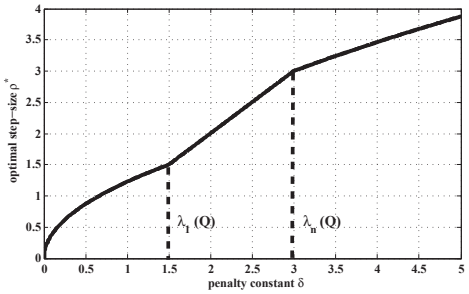


Fig. 1. Locus of $\rho^\star$ as a function of δ.

Remark 3. The aforementioned analysis also applies to the more general case with a cost function of the form $\frac{1}{2}\bar{x}^\top\bar{Q}\bar{x}+\bar{q}^\top\bar{x}+\frac{\delta}{2}\bar{z}^\top\bar{P}\bar{z}$ where $\bar{P}\in\mathbb{R}_+^{n\times n}$. One then first performs a change of variables of the form $z=\bar{P}^{1/2}\bar{z}$ that transforms the problem into the form (3) with $x=\bar{P}^{1/2}\bar{x}$, $q=\bar{P}^{-1/2}\bar{q}$, and $Q=\bar{P}^{-1/2}\bar{Q}\bar{P}^{-1/2}$, and everything follows as described earlier.

4. OPTIMAL CONVERGENCE RATE FOR CONSENSUS ON K-REGULAR GRAPHS

The ADMM method has also been used as a basis for distributed optimization and consensus algorithms on graphs. In this section, we develop optimal step-size rules for ADMM-based consensus and give explicit formulas for the optimal step-size for k-regular graphs.

Let $\mathcal{G}(\mathcal{V},\mathcal{E})$ be a connected undirected graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Each vertex $i\in\mathcal{V}$ represents an agent, and an edge $(i,j)\in\mathcal{E}$ means that agents i and j can exchange information. We let d_i denote the degree of vertex i, *i.e.* the number of edges incident on i. Each agent i holds a value y_i and it only coordinates with its neighbors $\mathcal{N}_i=\{j:(i,j)\in\mathcal{E}\}$ to compute the network-wide average $\bar{x}=\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}y_i$. Let us introduce auxiliary variables $z_{(ij)}\in\mathbb{R}$ for each $(i,j)\in\mathcal{E}$. The network-wide average is then the solution to the following optimization problem

$$\begin{aligned}&\text{minimize}\quad \frac{1}{2}\sum_{i\in\mathcal{V}}(x_i-y_i)^2\\ &\text{subject to } x_i=z_{(i,j)}\qquad \forall i\in\mathcal{V},\forall(i,j)\in\mathcal{E}.\end{aligned}\tag{9}$$

We will now use the ADMM algorithm to develop a distributed algorithm for finding the average consensus. To this end, the augmented Lagrangian is

$$\begin{aligned}L_\rho(x,z,\mu)=&\sum_{i\in\mathcal{V}}\frac{1}{2}(x_i-y_i)^2+\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i}\mu_{ij}(x_i-z_{(i,j)})\\&+\frac{\rho}{2}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i}(x_i-z_{(i,j)})^2,\end{aligned}\tag{10}$$

where μ_{ij} and ρ are the lagrange multiplier at node i associated with edge (i,j) and the constant step-size, respectively. Each node i, in addition to separately performing the minimization of (10) with respect to the primal variables x_i and $z_{(i,j)}$, executes the gradient ascent updates on the dual variables μ_{ij} as following

$$x_i^{k+1}=\frac{y_i-\sum_{j\in\mathcal{N}_i}\mu_{ij}^k+\rho\sum_{j\in\mathcal{N}_i}z_{(i,j)}^k}{1+\rho d_i},\tag{11}$$

$$z_{(i,j)}^{k+1}=\frac{x_i^{k+1}+x_j^{k+1}}{2}+\frac{\mu_{ij}^k+\mu_{ji}^k}{2\rho},\tag{12}$$

$$\mu_{ij}^{k+1}=\mu_{ij}^k+\rho(x_i^{k+1}-z_{(i,j)}^{k+1}).\tag{13}$$

Note that (11) and (12) are computed analytically by considering the first order optimality condition. The iteration can be simplified by noting that at optimality, (12) results in $\mu_{ij}^\star+\mu_{ji}^\star=0$. Thus, setting $\mu_{ij}^0+\mu_{ji}^0=0$ guarantees $\mu_{ij}^k+\mu_{ji}^k=0,\forall k>0$. Hence, after some simplifications, the above iterations read as

$$x_i^{k+1}=\frac{y_i-\frac{\rho}{2}\sum_{j\in\mathcal{N}_i}\sum_{t=1}^k x_i^t-x_j^t}{1+\rho d_i}+\frac{\frac{\rho}{2}\sum_{j\in\mathcal{N}_i}x_i^k+x_j^k}{1+\rho d_i}.$$

Or equivalently,

$$x^{k+1}=\Delta^{-1}\left[y+\frac{\rho}{2}\left((D+A)x^k-(D-A)\sum_{t=1}^k x^t\right)\right],\tag{14}$$

where $A\in\mathbb{R}^{n\times n}$ is the symmetric adjacency matrix associated with $\mathcal{G}$ and D is the diagonal degree matrix, i.e., $D=\mathrm{diag}(A\mathbf{1}_n)$, where $\mathbf{1}_n$ is a vector with n components equal to 1. Furthermore, $\Delta:=I+\rho D$. To study the stability of (14) we form the difference of two successive iterations $e^k:=x^{k+1}-x^k$ as following

$$e^k:=\Delta^{-1}\left[\frac{\rho}{2}(D+A)\Big(x^k-x^{k-1}\Big)-\frac{\rho}{2}(D-A)x^k\right].$$

Define $\tilde{x}^k:=\begin{bmatrix}x^{k+1}\\x^k\end{bmatrix}$ and form the following linear equality

$$\tilde{x}^k=M\tilde{x}^{k-1},\ M=\begin{bmatrix}I+\rho\Delta^{-1}A & -\frac{\rho}{2}\Delta^{-1}(D+A)\\ I & \mathbf{0}\end{bmatrix}.\tag{15}$$

One can check that any eigenvector of M given by $M\begin{bmatrix}u\\v\end{bmatrix}=\lambda\begin{bmatrix}u\\v\end{bmatrix}$ with nonzero $u,v\in\mathbb{C}^n$ should fulfill $u=\lambda v$. Replace u by λv and form the eigenvalue equality, as

$$\lambda^2 v-\lambda(I+\rho\Delta^{-1}A)v+\frac{\rho}{2}\Delta^{-1}(D+A)v=\mathbf{0}.\tag{16}$$

Multiplying both sides of (16) from the left by $v^\top\Delta$, we have

$$\lambda^2 v^\top\Delta v-\lambda v^\top(\Delta+\rho A)v+\frac{\rho}{2}v^\top(D+A)v=0.$$

Define $a_1(\rho,v):=v^\top(\Delta+\rho A)v, a_2(\rho,v):=v^\top\Delta v$ and $a_3(v):=v^\top(D+A)v\in\mathbb{R}$. It can be shown that the eigenvalues of M satisfy

$$\ell(\rho,v)=\frac{a_1\pm\sqrt{a_1^2-2\rho a_2 a_3}}{2a_2}:=\ell_r(\rho,v)+\jmath\ell_c(\rho,v),\tag{17}$$

where $\ell_r,\ \ell_c\in\mathbb{R}$. Next, we check the stability of (14).

Lemma 4. For all values of $\rho>0$ and $v\in\mathbb{C}^n$, $|\lambda_i|\le 1$, $i\in 1\dots 2n.$, where λ_i is the i-th eigenvalue of M and $|\lambda_1|\le|\lambda_i|\le|\lambda_{2n}|$. Moreover, vector $\mathbf{1}_{2n}$ is an eigenvector corresponding to $\lambda=1$.

Proof. From the fact that $\theta^\top M\theta\le|\lambda_{2n}|\theta^\top\theta$, with $\theta=\begin{bmatrix}\lambda v\\v\end{bmatrix}$, one can see that λ_{2n} is given by $\max_v|\ell|$ in (17).

If the value of (17) is real, i.e., $\ell_c(\rho, v) = 0$, then $|\ell| = \frac{|a_1|+\sqrt{a_1^2-2\rho a_2 a_3}}{2a_2}$. Note that $a_1 > 0$ (since $I+\rho(D+A) > 0$), so by replacing a_1, a_2, a_3 in (17), for the magnitude of the real eigenvalues of M denoted by $|\lambda_i^r|$, where $v = v_i$ such that $[\lambda_i^r v_i^\top v_i^\top]^\top$ is the corresponding eigenvector of λ_i^r, we obtain

$$2|\lambda_i^r| = 1 + \rho\frac{v^\top Av}{v^\top \Delta v} + \sqrt{1 + \left(\frac{\rho v^\top Av}{v^\top \Delta v}\right)^2 - 2\frac{\rho v^\top Dv}{v^\top \Delta v}}.$$

since $v^\top(D-A)v \geq 0$, one can replace $\rho v^\top Av$ with $\rho v^\top Dv$ under the square root in the above equation and find the following upper-bound:

$$2|\lambda_i^r| \leq 1 + \rho\frac{v^\top Av}{v^\top \Delta v} + \sqrt{1 + \left(\frac{\rho v^\top Dv}{v^\top \Delta v}\right)^2 - 2\frac{\rho v^\top Dv}{v^\top \Delta v}}$$
$$= 1 + \rho\frac{v^\top Av}{v^\top \Delta v} + |1 - \frac{\rho v^\top Dv}{v^\top \Delta v}| = 2 - \rho\frac{v^\top (D-A)v}{v^\top \Delta v} \leq 2.$$

On the other hand, when $\ell_c(\rho, v) \neq 0$, then the magnitude of eigenvalues of M are given by $|\lambda_i^c| = \sqrt{\frac{\rho v^\top (A+D)v}{2v^\top \Delta v}}$. One can check that $2v^\top \Delta v - \rho v^\top (A+D)v = v^\top (2I + \rho(D-A))v > 0$, hence $|\lambda_i^c| < 1$. To verify that $\lambda = 1$ is an eigenvalue of M, one can set $u = v = \mathbf{1}_n$ and conclude that $M\mathbf{1}_{2n} = \mathbf{1}_{2n}$. ■

Next, we pose our main result of this section and then we present required steps to prove it.

Theorem 5. For a k-regular graph, i.e., $d_i = k \geq 2 \forall i \in \mathcal{V}$, the optimal step-size which minimizes the convergence rate of (14) is given by

$$\rho^\star = \frac{1}{\sqrt{k^2 - l^2}} \tag{18}$$

where $l := \lambda_{n-1}(A)$.

We are interested in minimizing the magnitude of the second largest eigenvalue of M, i.e., $|\lambda_{2n-1}|$, which is equivalent to minimizing the convergence rate of (14). To this end one should maximize the value of $\frac{w^\top Mw}{w^\top w}$ where $w \in \mathbb{C}^{2n}$ and $w^\top \mathbf{1}_{2n} = 0$. The orthogonality condition is due to the fact that $\mathbf{1}_{2n}$ is the eigenvector corresponding to the maximum eigenvalue $\lambda = 1$ of M. To achieve this, we first define the function $s(\rho)$:

$$s(\rho) := \max_{v, v^\top \mathbf{1}_n = 0} |\frac{a_1 \pm \sqrt{a_1^2 - 2\rho a_2 a_3}}{2a_2}| \tag{19}$$

for $\rho \in \mathbb{R}_+$. For $|\lambda_{2n-1}|$ we have the following lemma.

Lemma 6. The magnitude of the second largest eigenvalue of M, i.e., $|\lambda_{2n-1}|$, is given by the following equation

$$|\lambda_{2n-1}| = \max\{s(\rho), \frac{\rho \sum_{i\in\mathcal{V}} d_i}{n + \rho \sum_{i\in\mathcal{V}} d_i}\}. \tag{20}$$

Proof. We know that all the eigenvalues of M satisfy (17). Consider $|\mathcal{V}| = n$ and set $v = \mathbf{1}_n$ then, $a_1 = n + 2\rho \sum_{i\in\mathcal{V}} d_i$, $a_2 = n + \rho \sum_{i\in\mathcal{V}} d_i$ and $a_3 = 2\sum_{i\in\mathcal{V}} d_i$. Replacing a_1, a_2 and a_3 in (17) leads to

$$\ell(\rho, \mathbf{1}_n) = \{1, \frac{\rho \sum_{i\in\mathcal{V}} d_i}{n + \rho \sum_{i\in\mathcal{V}} d_i}\}. \tag{21}$$

But $\lambda = 1$ is the simple maximum eigenvalue of M and we discard it. Still the second term of (21) might be the magnitude of the second largest eigenvalue of M. Note that any $v = \alpha\mathbf{1}_n$ for $\alpha \in \mathbb{R}$ leads to the same result as (21). Another possibility for the maximum magnitude of λ_{2n-1} is when we have $v^\top \mathbf{1}_n = 0$ in (17) and maximize it with respect to v. For such case (19) offers the maximum bound. Hence, (20) holds. ■

Thus, the optimal step-size $\rho^\star$ that minimizes $|\lambda_{n-1}|$ is

$$\rho^\star := \operatorname*{argmin}_\rho \max\{s(\rho), \frac{\rho \sum_{i\in\mathcal{V}} d_i}{n + \rho \sum_{i\in\mathcal{V}} d_i}\}. \tag{22}$$

Let $v_\perp \in \mathbb{C}^n$ and $v_\perp \mathbf{1}_n = 0$. The next result characterizes the behavior of $|\ell(\rho, v_\perp)|$ with respect to ρ for the case where the value of (17) is complex, i.e., $\ell_c(\rho, v_\perp) \neq 0$.

Lemma 7. Let $v_\perp \in \mathbb{C}^n$ such that $v_\perp \mathbf{1}_n = 0$. The magnitude of (17), $|\ell(\rho, v_\perp)|$, where $\ell_c(\rho, v_\perp) \neq 0$ is monotonically increasing with respect to ρ.

Proof. For $\ell_c(\rho, v_\perp) \neq 0$, $|\ell(\rho, v_\perp)| = \sqrt{\frac{\rho v_\perp^\top (A+D)v_\perp}{2v_\perp^\top \Delta v_\perp}}$. The derivative of $|\ell(\rho, v_\perp)|$ with respect to ρ is

$$\nabla_\rho |\ell| = \frac{1}{2}\sqrt{\frac{2v_\perp^\top \Delta v_\perp}{\rho v_\perp^\top (A+D)v_\perp}} \frac{(v_\perp^\top (A+D)v_\perp)(v_\perp^\top v_\perp)}{2(v_\perp^\top \Delta v_\perp)^2} \geq 0.$$

which proves that $|\ell(\rho, v_\perp)|$ is monotone increasing for $\ell_c(\rho, v_\perp) \neq 0$. ■

In the sequel, we restrict our results to k-regular graphs. In such graphs, $d_i = k \geq 2$, $\forall i \in \mathcal{V}$ and the adjacency matrix A has the largest eigenvalue $\lambda_n(A) = k$ associated with $\mathbf{1}_n$ as the eigenvector. For a k-regular graph, the following lemma characterizes the magnitude of the eigenvalues of M (17) is real.

Lemma 8. For a k-regular graph, if $\ell_c(\rho, v_\perp) = 0$, then the magnitude of (17), $|\ell(\rho, v_\perp)|$, is monotonically decreasing with respect to ρ.

Proof. When $\ell_c(\rho, v_\perp) = 0$, we have $|\ell(\rho, v_\perp)| = |\ell_r(\rho, v_\perp)|$. For a k-regular graph, we have $\Delta = (1+\rho k)I$ and $a_1 = v_\perp^\top((1+\rho k)I + A)v_\perp$, $a_2 = (1+\rho k)v_\perp^\top v_\perp$ and $a_3 = v_\perp^\top (A + kI)v_\perp$. By replacing a_1, a_2 and a_3 in (17) and taking into account that $\ell_c(\rho, v_\perp) = 0$, one obtains

$$2|\ell(\rho, v_\perp)| = 1 + \bar{\lambda} f(\rho) + \sqrt{1 + \bar{\lambda}^2 f^2(\rho) - 2kf(\rho)}$$

where $\bar{\lambda} := \frac{v_\perp^\top Av_\perp}{v_\perp^\top v_\perp}$ and $f(\rho) = \frac{\rho}{1+\rho k}$. Let $g(\rho, v_\perp) := 2|\ell(\rho, v_\perp)|$. Note that $\lambda_1(A) \leq \bar{\lambda} \leq \lambda_{n-1}(A) = l$. Taking the derivative of $g(\rho, v_\perp)$ with respect to ρ yields

$$\nabla_\rho g = f'(\rho)\left(\bar{\lambda} + (1 + \bar{\lambda}^2 f^2(\rho) - 2kf(\rho))^{-\frac{1}{2}}(\bar{\lambda}^2 f(\rho) - k)\right).$$

Since $f'(\rho) = \frac{1}{(1+\rho k)^2} > 0$, we can further simplify the above derivative and check its negativity:

$$\bar{\lambda} + (1 + \bar{\lambda}^2 f^2(\rho) - 2kf(\rho))^{-\frac{1}{2}}(\bar{\lambda}^2 f(\rho) - k) < 0.$$

By replacing $f(\rho)$ in the second term of the above inequality we have

$$\bar{\lambda} - (1 + \bar{\lambda}^2 f^2(\rho) - 2kf(\rho))^{-\frac{1}{2}}\left(\frac{k + \rho(k^2 - \bar{\lambda}^2)}{1+\rho k}\right)$$
$$< \bar{\lambda} - (1 - k\frac{\rho}{1+\rho k})^{-1}\left(\frac{k + \rho(k^2 - \bar{\lambda}^2)}{1+\rho k}\right)$$
$$= -(k - \bar{\lambda} + \rho(k^2 - \bar{\lambda}^2)) < 0.$$

Note that from the first to the second inequality we have replaced $\bar{\lambda}$ with k in the inverse square root term (since $\bar{\lambda} < k$). ■

Corollary 9. For a k-regular graph and a given vector $v_\perp \in \mathbb{C}^n$ and $v_\perp \mathbf{1}_n = 0$, the value of $\hat{\rho} = \underset{\rho}{\operatorname{argmin}} |\ell(\rho, v_\perp)|$ is obtained by setting the square root in (17) to be equal to 0, i.e., $a_1(\hat{\rho}, v_\perp)^2 - 2\hat{\rho} a_2(\hat{\rho}, v_\perp) a_3(v_\perp) = 0$.

In k-regular graphs and a given vector $v_\perp \mathbf{1}_n = 0$, Corollary 9 indicates that the minimum of $|\ell(\rho, v_\perp)|$ happens when $\ell(\rho, v_\perp)$ has multiplicity 2. Hence, if we set $D = kI$ and $\bar{\lambda} := \frac{v_\perp^\top A v_\perp}{v_\perp^\top v_\perp}$ and then replace them in $a_1(\hat{\rho}, v_\perp)^2 - 2\hat{\rho} a_2(\hat{\rho}, v_\perp) a_3(v_\perp) = 0$, we can find $\hat{\rho}$ as

$$\hat{\rho} := \underset{\rho}{\operatorname{argmin}} |\ell(\rho, v_\perp)| = \frac{1}{\sqrt{k^2 - \bar{\lambda}^2}}. \tag{23}$$

Additionally, $\hat{\rho}$ is the locus of minimal points of the magnitude of (17) for different values of $\bar{\lambda}$. In the sequel, we highlight the properties of $\hat{\rho}$.

Lemma 10. For k-regular graphs, $\hat{\rho}$ in (23) is monotone increasing with respect to $\bar{\lambda}$.

Proof. The result can be verified by checking that the derivative of (23) with respect to $\bar{\lambda} := \frac{v_\perp^\top A v_\perp}{v_\perp^\top v_\perp}$ is positive. ■

Lemma 11. For k-regular case, if the value of (17) is complex, i.e., $\ell_c(\rho, v_\perp) \neq 0$, then $|\ell(\rho, v_\perp)|$ is monotone increasing with respect to $\bar{\lambda} = \frac{v_\perp^\top A v_\perp}{v_\perp^\top v_\perp}$.

Proof. If $\ell_c(\rho, v_\perp) \neq 0$ in (17) then, $|\ell(\rho, v_\perp)| = \sqrt{\frac{\rho v_\perp^\top (A+D) v_\perp}{2 v_\perp^\top \Delta v_\perp}} = \sqrt{\frac{\rho(\bar{\lambda}+k)}{2(1+\rho k)}}$, which is monotone increasing with respect to $\bar{\lambda}$. ■

Now we are ready to prove the main result of this section.

Proof of Theorem 5. For a k-regular graph, in the light of Corollary 9, Lemma 10 and Lemma 11 we have

$$\rho_1^\star := \underset{\rho}{\operatorname{argmin}}\, s(\rho) = \underset{\rho}{\operatorname{argmin}} \max_{v, v^\top \mathbf{1}_n = 0} |\ell^k(\rho, v)| = \frac{1}{\sqrt{k^2 - l^2}}.$$

where $l := \lambda_{n-1}(A)$. Moreover, by replacing the value $\rho_1^\star$ and $v_\perp^\star := \underset{v^\top \mathbf{1}_n = 0}{\operatorname{argmax}} \frac{v^\top A v}{v^\top v}$ in (17) the magnitude $|\ell(\rho_1^\star, v_\perp^\star)|$ becomes

$$|\ell(\rho_1^\star, v_\perp^\star)| = \frac{1}{2} + \frac{\rho_1^\star l}{2(1 + \rho_1^\star k)}.$$

Moreover, from (21) we conclude that $|\ell(\rho, \mathbf{1}_n)| = \{1, \frac{\rho k}{1+\rho k}\}$. We use (22) to obtain the optimal step-size $\rho^\star := \underset{\rho}{\operatorname{argmin}} \max\{s(\rho), \frac{\rho k}{1+\rho k}\}$. Since $\rho_1^\star$ is the minimizer of $s(\rho)$ over ρ, if we show that $|\ell(\rho_1^\star, v_\perp^\star)| \geq |\ell(\rho_1^\star, \mathbf{1}_n)|$, then we have proved that $\rho^\star = \rho_1^\star$. By forming the aforementioned inequality we get

$$1 - \frac{\rho_1^\star (2k - l)}{1 + \rho_1^\star k} \geq 0.$$

Setting $\rho_1^\star = \frac{1}{\sqrt{k^2 - l^2}}$ in the above inequality we conclude $k - l \leq \sqrt{k^2 - l^2}$, and $l(l - k) < 0$. So we have $\rho^\star = \rho_1^\star = \frac{1}{\sqrt{k^2 - l^2}}$. ■

In what follows we discuss how $\rho^\star$ varies with the size of k-regular graphs and provide empirical rules for choosing a good step-size. Complete graphs with $n \geq 3$ nodes are k-regular with $k = n-1$ and $l = \lambda_{n-1}(A) = -1$. Thus, the optimal step-size is computed to be $\rho^\star = \frac{1}{\sqrt{n(n-2)}}$. Therefore, $\rho^\star$ tends to zero as the number of nodes increases.

Next we consider ring networks with n nodes and $k = 2$. For large number of nodes $\lambda_{n-1}(A)$ tends to $\lambda_n(A) = k$, i.e., $\frac{l}{k} \to 1$, and $\rho^\star$ grows significantly.

The above statements indicate that one should select small step-sizes for highly connected graphs and large step-sizes for very sparse graphs. Additionally, there exists a vast amount of results concerning the bounds on $l = \lambda_{n-1}(A)$ for generic k-regular graphs. Among all of them Quenell (1996) suggests the following lower bound

$$l > 2\sqrt{k-1} \cos(\frac{\pi}{r+1}) \tag{24}$$

where $k \geq 3$, $r = \lfloor \frac{\bar{d}}{2} \rfloor$ and $\bar{d}$ is the diameter of the graph. For large k-regular graphs Friedman (2004) specifies that $l < 2\sqrt{k-1} + \epsilon$. Applying these results to large k-regular networks one can then approximate $\lambda_{n-1}(A)$ by $2\sqrt{k-1}$ and use $\rho^\star = \frac{1}{k-2}$ for $k > 2$ as a reasonable guess for the optimal step-size.

5. ℓ_1-REGULARIZED LOSS MINIMIZATIONS

Consider the following quadratic loss minimization problem plus a ℓ_1 penalty function

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} x^\top Q x + q^\top x + \delta \|z\|_1 \\ \text{subject to} & x = z, \end{array} \tag{25}$$

where $Q \in \mathbb{S}_+^n$, $x, q, z \in \mathbb{R}^n$ and $\delta \in \mathbb{R}_+$ is constant. We formulate ADMM algorithm for this problem with the following change of variable $u = \frac{1}{\rho}\mu$. Remember ρ and μ are the ADMM constant step-size and Lagrangian variable corresponding to the equality constraint, respectively. Augmented Lagrangian can be written as

$$L_a(x, z, u) = \frac{1}{2} x^\top Q x + q^\top x + \delta \|z\|_1 + \frac{\rho}{2} \|x - z + u\|_2^2. \tag{26}$$

The algorithm becomes

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} \{\frac{1}{2} x^\top Q x + q^\top x + \frac{\rho}{2} \|x - z^k + u^k\|_2^2\} \\ z^{k+1} &= \underset{z}{\operatorname{argmin}} \{\delta \|z\|_1 + \frac{\rho}{2} \|x^{k+1} - z + u^k\|_2^2\} \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}. \end{aligned}$$

Note that z iterations are the proximity operator for the ℓ_1 norm. More precisely, we have

$$\operatorname{Prox}_h(x) = \underset{v}{\operatorname{argmin}} \left(h(v) + \frac{1}{2} \|v - x\|_2^2 \right).$$

By applying the above definition for the function $h(x) = \delta \|x\|_1$, we get a component-wise analytic solution for x which is called *soft thresholding* and is (see Boyd et al. (2011) for details)

$$\operatorname{Prox}_h(x)_i = \mathcal{S}_\delta(x) = \begin{cases} x_i - \delta & x_i \geq \delta \\ 0 & |x_i| \leq \delta \\ x_i + \delta & x_i \leq -\delta. \end{cases}$$

Hence, we obtain

$$\begin{aligned} x^{k+1} &= (Q+\rho I)^{-1}\left(-q+\rho(z^k-u^k)\right) \\ z^{k+1} &= \mathcal{S}_{\delta/\rho}(x^{k+1}+u^k) \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}. \end{aligned} \tag{27}$$

Iterations (27) are non-linear, and the rigorous analysis of this algorithm to obtain the optimal step-size $\rho^\star$ is identified as a future research problem. However, later in Section 6.3, we empirically offer some insights on different choices of the step-size ρ.

6. NUMERICAL EXAMPLES

In this section we conduct numerical examples to evaluate our step-size selection results.

6.1 ℓ_2-Regularized quadratic programming

Fig. 2 compares the convergence factor of ADMM for varying δ. We have considered two step-size selections: $\rho = \delta$ and $\rho = \rho^\star$ obtained from (6). The Q matrix is highly ill-conditioned with $\lambda_1(Q) \sim 41.3$ and $\lambda_n(Q) \sim 1.9 \times 10^3$. For comparison, the dashed-dotted curve shows the optimal convergence factor of the gradient iterations for the problem, *i.e.*

$$x_{k+1} = x_k - \alpha(Qx_k + q + \delta x_k),$$

where $\alpha < 2/\lambda_n(Q)$ is a constant step-size. For this problem, since the cost function is quadratic and its Hessian $f'' = Q + \delta I$ is bounded between $l = \lambda_1(Q) + \delta$ and $L = \lambda_n(Q) + \delta$, the optimal step-size is $\alpha^\star = \frac{2}{l+L}$ and the convergence factor is given by $q^\star = \frac{L-l}{L+l} = \frac{\lambda_n(Q)-\lambda_1(Q)}{\lambda_n(Q)+\lambda_1(Q)+2\delta}$ (Polyak (1987)). The figure illustrates the robust convergence properties of the ADMM method, and how it outperforms the gradient for small δ (ill-conditioned problem). We can also observe that the simple gradient method has better convergence factor as δ grows large (*i.e.* when regularization makes the overall problem well-conditioned).

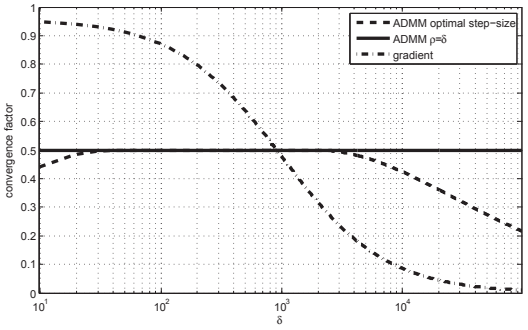
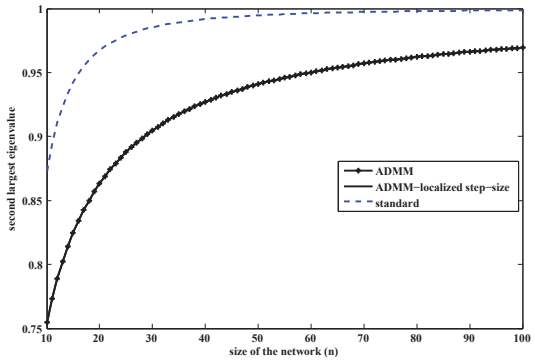


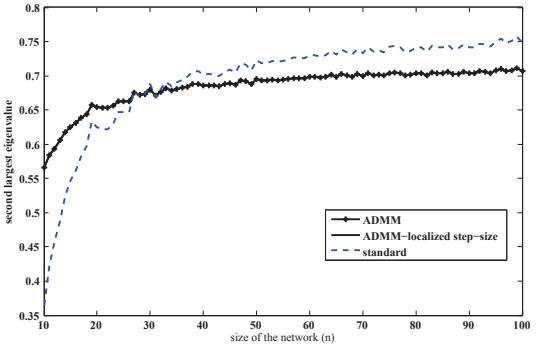
Fig. 2. Performance of ADMM for ℓ_2 regularized minimization.

6.2 Consensus on k-regular graphs

In this section we compare the convergence factor of the consensus iterations (11) with the optimal constant step-size $\rho^\star$ against the standard consensus algorithm presented in Xiao & Boyd (2004) and an alternative ADMM technique for consensus applications recently proposed in Erseghe et al. (2011). As an indicator of the convergence factor, we consider the second largest eigenvalue of the corresponding consensus matrices in linear iterations of the form $x(t+1) = Wx(t)$. Fig. 3 presents the second largest eigenvalue versus the number of nodes $n \in [10, 100]$ for randomly generated regular topologies.



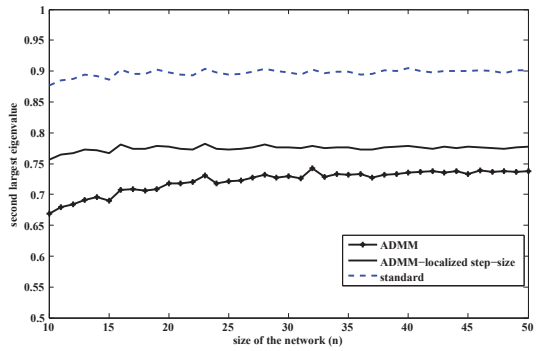
(a) 2-regular (ring)



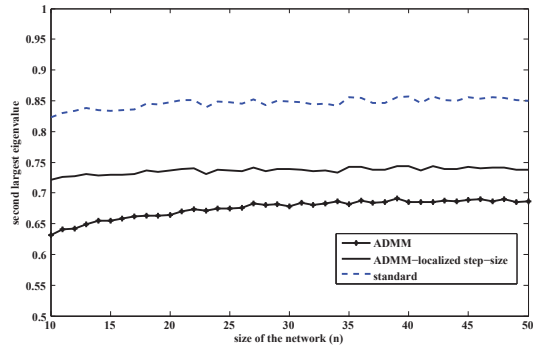
(b) 6-regular

Fig. 3. Performance comparison of optimal ADMM consensus algorithm for $\{2, 6\}$-regular graphs with standard algorithm and state of the art Erseghe et al. (2011) ADMM iterations with localized augmentation constant.

The algorithm corresponding to the solid curve captioned by ADMM is implemented via using $\rho^\star$ in (18) as the constant step-size, while the algorithm given in Erseghe et al. (2011) (annotated as ADMM-localized step-size) uses the optimal relaxed localized augmentation constants proposed by the respective authors. More precisely, this algorithm considers the symmetric matrix $C \in \mathbb{R}^{n\times n}$ to be the augmentation constant (step-size) matrix. Additionally, it is assumed that C is doubly stochastic. The authors in Erseghe et al. (2011) propose an optimal constant tuning parameter which minimizes the convergence factor of the algorithm. As for the standard consensus algorithm, we use *Metropolis-Hasting* (Xiao & Boyd (2004)) weight matrix as the augmentation matrix C in Erseghe et al. (2011). For a k-regular graph this matrix coincides with *constant weights* since both assign $\frac{1}{k}$ for each edge $(i, j) \in \mathcal{E}$ (see, Xiao & Boyd (2004)). This is a fair comparison since all the three algorithms need similar initialization steps. Although our proposed algorithm does not need a weight matrix, it uses the adjacency matrix A and degree matrix D of the graph, which contain the same amount of information available to the other two algorithms.



(a) $\epsilon = 0.3$

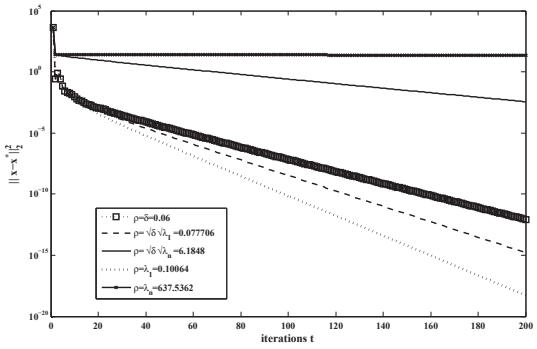


(b) $\epsilon = 0.7$

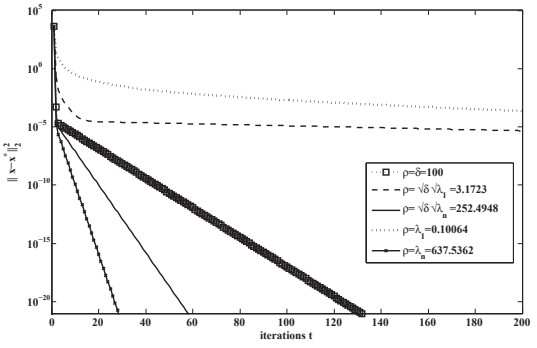
Fig. 4. Performance comparison of optimal ADMM consensus algorithm with standard algorithm and state of the art Erseghe et al. (2011) ADMM iterations with localized augmentation constant. The network of size $n = [10, 100]$ is randomly generated by Erdős-Rényi graphs with densities $\epsilon = \{0.3, 0.7\}$.

For each point in Fig. 3, the simulation is repeated for 10 random graphs with the same number of nodes and the average result is plotted. The plots show that both ADMM iterations offer faster convergence when compared to the standard case (specially for larger n). Furthermore, the convergence factor for the two ADMM alternatives are exactly equal in all the simulations. It is not surprising if we note that in k-regular graphs all the nodes have the same degree, hence they will have the same augmentation constant in the localized algorithm (Erseghe et al. (2011)) which, in turn, coincides with our optimal step-size.

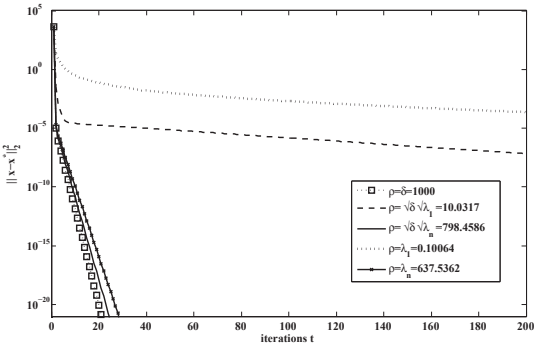
Fig. 4 depicts the performance comparison for the same algorithms for randomly generated Erdős-Rényi graphs, which may not be regular. According to this algorithm, each component (i, j) in the adjacency matrix A is set to 1 with probability $p = (1 + \epsilon)\frac{\log(n)}{n}$, where $\epsilon \in (0, 1)$ and n is the number of vertices. For the ADMM algorithm we calculate the constant step-size in (18) with k taken as the average degree of the graph. For each point of the plot, we have repeated the simulations for 100 times and the value of second largest eigenvalues are the mean value of corresponding random graphs with the same number of vertices and edge-density. Results show that in all the cases our sub-optimal constant step-size ADMM outperforms the method proposed by Erseghe et al. (2011) using the optimal localized step-sizes. These simulation results motivate us to investigate the problem of optimal step-size selection for general graphs as future work.



(a) $\delta < \lambda_1(Q)$



(b) $\lambda_1(Q) < \delta < \lambda_n(Q)$



(c) $\lambda_n(Q) < \delta$

Fig. 5. log scale of $\|x - x^\star\|_2^2$ versus the number of iterations t for the quadratic programming with ℓ_1 regularization. Different choices of penalty constant δ and step-size ρ are plotted.

6.3 ℓ_1-Regularized loss minimizations

Fig. 5 illustrates the convergence properties of ℓ_1-regularized ADMM algorithm for different choices of the step-size ρ. We simulate (25) with a randomly generated $Q \in \mathbb{S}_+^{50}$. The exhaustive iterations have been initialized to find the optimal point x in (27) for each setting of δ. Based on our initial simulation results, for the case where $\delta < \lambda_1(Q)$, $\rho = \lambda_1(Q)$ is a reasonable selection which accelerates the convergence of (27). Moreover, when $\lambda_1(Q) < \delta < \lambda_n(Q)$, $\rho = \lambda_n(Q)$ offers better performance while for $\delta > \lambda_n(Q)$ the proper step-size choice is $\rho = \delta$.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the optimal step-size selection for the alternating direction method of multipliers. In

particular, we investigated three different problem formulations: ℓ_2-regularized quadratic programming, ADMM-based consensus on k-regular graphs, and ℓ_1-regularized quadratic programming. We demonstrated both theoretically and numerically the optimal step-size selection for ℓ_2-regularized quadratic programming and consensus, and compared these with existing methods. Via simulations, we assessed suitable step constant for the ℓ_1-regularized estimation problem. As a future work, we plan to develop analytical results for the ℓ_1-regularized problem and extend our results for k-regular graphs to general graphs.

REFERENCES

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3 Issue: 1**, 1–122.

Erseghe, T., Zennaro, D., Dall'Anese, E. & Vangelista, L. (2011), 'Fast consensus by the alternating direction multipliers method', *Signal Processing, IEEE Transactions on* **59**, 5523 –5537.

Figueiredo, M. & Bioucas-Dias, J. (2010), 'Restoration of poissonian images using alternating direction optimization', *Image Processing, IEEE Transactions on* **19**(12), 3133 –3145.

Forero, P. A., Cano, A. & Giannakis, G. B. (2010), 'Consensus-based distributed support vector machines', *J. Mach. Learn. Res.* **99**, 1663–1707.

Friedman, J. (2004), 'A proof of alon's second eigenvalue conjecture and related problems', *CoRR, cs.DM/0405020.* .

Ghadimi, E., Johansson, M. & Shames, I. (2011), Accelerated gradient methods for networked optimization, *in* 'American Control Conference (ACC)'.

Mota, J. F. C., Xavier, J. M. F., Aguiar, P. M. Q. & Püschel, M. (2011), 'A Proof of Convergence For the Alternating Direction Method of Multipliers Applied to Polyhedral-Constrained Functions', *ArXiv e-prints* .

Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization: A Basic Course*, Springer-Verlag New York, LCC.

Polyak, B. (1987), *Introduction to Optimization*, ISBN 0-911575-14-6.

Quenell, G. (1996), 'Eigenvalue comparisons in graph theory', *Pacific J. Math* **176**, 443–461.

Wahlberg, B., Boyd, S., Annergren, M. & Wang, Y. (2012), An admm algorithm for a class of total variation regularized estimation problems, *in* '16th IFAC Symposium on System Identification'.

Xiao, L. & Boyd, S. (2004), 'Fast linear iterations for distributed averaging', *Systems and Control Letters* **53 Issue: 1**, 65–78.

Yang, J. & Zhang, Y. (2011), 'Alternating direction algorithms for l_1-problems in compressive sensing', *SIAM J. Sci. Comput.* **33**(1), 250–278.