Markov Chain Lifting and Distributed ADMM

Guilherme França*

Boston College, Computer Science Department and

Johns Hopkins University, Center for Imaging Science

José Bento†

Boston College, Computer Science Department

Abstract

The time to converge to the steady state of a finite Markov chain can be greatly reduced by a lifting operation, which creates a new Markov chain on an expanded state space. For a class of quadratic objectives, we show an analogous behavior where a distributed ADMM algorithm can be seen as a lifting of Gradient Descent algorithm. This provides a deep insight for its faster convergence rate under optimal parameter tuning. We conjecture that this gain is always present, as opposed to the lifting of a Markov chain which sometimes only provides a marginal speedup.

* guifranca@gmail.com

† jose.bento@bc.edu

I. INTRODUCTION

Let $\mathcal{M}$ and $\hat{\mathcal{M}}$ be two finite Markov chains with states $\mathcal{V}$ and $\hat{\mathcal{V}}$, of sizes $|\mathcal{V}| < |\hat{\mathcal{V}}|$, and with transition matrices M and $\hat{M}$, respectively. Let their stationary distributions be $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$. In some cases it is possible to use $\hat{\mathcal{M}}$ to sample from the stationary distribution of $\mathcal{M}$. A formal set of conditions under which this happen is known as *lifting*. We say that $\hat{\mathcal{M}}$ is a lifting of $\mathcal{M}$ if there is a row stochastic matrix $S \in \mathbb{R}^{|\hat{\mathcal{V}}|\times|\mathcal{V}|}$ with elements $S_{ij} \in \{0,1\}$ and a single nonvanishing element per line, where $\mathbf{1}_{|\mathcal{V}|} = S^\top \mathbf{1}_{|\hat{\mathcal{V}}|}$, and $\mathbf{1}_n$ is the all-ones n-dimensional vector, such that

$$\boldsymbol{\pi} = S^\top \hat{\boldsymbol{\pi}}, \qquad D_{\boldsymbol{\pi}} M = S^\top D_{\hat{\boldsymbol{\pi}}} \hat{M} S. \tag{1}$$

We denote $S^\top$ the transpose of S, and for any vector $\boldsymbol{v} \in \mathbb{R}^n$, $D_{\boldsymbol{v}} = \operatorname{diag}(v_1, \dots, v_n)$. Intuitively, $\hat{\mathcal{M}}$ contains copies of the states of $\mathcal{M}$ and transition probabilities between the extended sates $\hat{\mathcal{V}}$ such that it is possible to collapse $\hat{\mathcal{M}}$ onto $\mathcal{M}$. This is the meaning of relation (1). See Fig. 1 for an illustration. (We refer to [1] for more details on Markov chain lifting.)

The *mixing time* $\mathcal{H}$ is a measure of the time it takes for the distribution of a Markov chain $\mathcal{M}$ to approach stationarity. We follow the definitions of [1] but, up to multiplicative factors and slightly loser bounds, the reader can think of

$$\mathcal{H} = \min\{t : \max_{\{i, \boldsymbol{p}^0\}} |p_i^t - \pi_i| < 1/4\}, \tag{2}$$

where p_i^t is the probability of being on state i after t steps, starting from the initial distribution $\boldsymbol{p}^0$. Lifting is particularly useful when the mixing time $\hat{\mathcal{H}}$ of the lifted chain is much smaller than $\mathcal{H}$. There are several examples where $\hat{\mathcal{H}} \approx C\sqrt{\mathcal{H}}$, for some constant $C \in (0,1)$ which depends only on $\boldsymbol{\pi}$. However, there is a limit on how much speedup can be achieved. If $\mathcal{M}$ is irreducible, then $\hat{\mathcal{H}} \geq C\sqrt{\mathcal{H}}$. If $\mathcal{M}$ and $\hat{\mathcal{M}}$ are reversible, then the limitation is even stronger, $\hat{\mathcal{H}} \geq C\mathcal{H}$.

Consider the undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Let $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{V}|}$ with components z_i, and consider the quadratic problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^{|\mathcal{V}|}} \Big\{ f(\boldsymbol{z}) = \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} q_{ij}(z_i - z_j)^2 \Big\}. \tag{3}$$

We also write $q_{ij} = q_e$ for $e = (i,j) \in \mathcal{E}$. There is a connection between solving (3) by *Gradient Descent* (GD) algorithm and the evolution of a Markov chain. To see this,

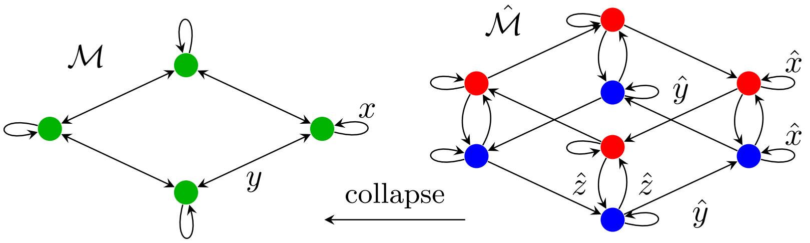


FIG. 1. The base Markov chain $\mathcal{M}$ over the cycle graph C_4 and its lifted Markov chain $\hat{\mathcal{M}}$ with duplicate states. Each state in $\mathcal{V}$ (green nodes) is expanded into two states in $\hat{\mathcal{V}}$ (blue and red nodes above each other).

consider $q_e = 1$ for simplicity. The GD iteration with step-size $\alpha > 0$ is given by

$$\boldsymbol{z}^{t+1} = (I - \alpha \nabla f)\, \boldsymbol{z}^t = (I - \alpha D_{\boldsymbol{d}}(I - W))\, \boldsymbol{z}^t \tag{4}$$

where $W = D_{\boldsymbol{d}}^{-1} A$ is the transition matrix of a random walk on $\mathcal{G}$, A is the adjacency matrix, and $D_{\boldsymbol{d}}$ is the degree matrix, where $\boldsymbol{d} = \text{diag}(d_1, \dots, d_{|\mathcal{V}|})$.

This connection is specially clear for d-regular graphs. Choosing $\alpha = 1/d$, equation (4) simplifies to $\boldsymbol{z}^{t+1} = W\boldsymbol{z}^t$. In particular, the *convergence rate* of GD is determined by the spectrum of W, which is connected to the mixing time of $\mathcal{M}$. More precisely, when W is irreducible and aperiodic, and denoting $\lambda_2(W)$ its second largest eigenvalue in absolute value, the mixing time of the Markov chain and the *convergence time* of GD are both equal to

$$\mathcal{H} = \frac{C}{\log(1/|\lambda_2|)} \approx \frac{C}{1 - |\lambda_2|} \tag{5}$$

where the constant C comes from the tolerance error, which in (2) is $1/4$. In the above approximation we assumed $\lambda_2 \approx 1$. Therefore, at least for GD, we can use the theory of Markov chains to analyze the convergence rate when solving optimization problems. For this example, and whenever there is linear convergence, the convergence rate τ and the convergence time $\mathcal{H}$ are related by $\tau^{\mathcal{H}} = \Theta(1)$. For an introduction on Markov chains, mixing times, and transition matrix eigenvalues, we refer the reader to [2].

The main goal of this paper is to extend the above connection to the *over-relaxed Alternating Direction Method of Multipliers* (ADMM) algorithm, and the concept of lifting will play an important role. Specifically, for problem (3), we show that a distributed implementation of over-relaxed ADMM can be seen as a lifting of GD, in the same way that $\hat{\mathcal{M}}$ is a lifting of the Markov chain $\mathcal{M}$. More precisely, there is a matrix M_A with stationary vector

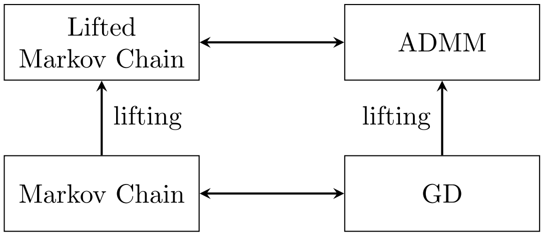


FIG. 2. GD is the analogue of a Markov chain, while distributed ADMM is the analogue of a lifted version of this Markov chain, which mixes faster.

v_A associated to distributed ADMM, and a matrix M_G with stationary vector v_G associated to GD, such that relation (1) is satisfied. This duality is summarized in Fig. 2. In some cases, M_A might have a few negative entries preventing it from being the transition matrix of a Markov chain. However, it always satisfies all the other properties of Markov matrices.

As explained in the example preceding (5), the convergence time of an algorithm can be related to the mixing time of a Markov chain. Let $\mathcal{H}_A$ be the convergence time of ADMM, and $\mathcal{H}_G$ the convergence time of GD. The lifting relation between both algorithms strongly suggest that, for problem (3) and optimally tuned parameters, ADMM is always faster than GD. Since lifting can speed mixing times of Markov chains up to a square root factor, we conjecture that the optimal convergence times $\mathcal{H}_A^\star$ and $\mathcal{H}_G^\star$ are related as

$$\mathcal{H}_A^* \leq C\sqrt{\mathcal{H}_G^*}, \tag{6}$$

where $C > 0$ is a universal constant. Moreover, we conjecture that (6) holds for *any* connected graph $\mathcal{G}$. Note that (6) is much stronger than the analogous relation for lifted of Markov chains, where for some graphs, e.g. with low conductance, the gain is marginal.

The outline of this paper is the following. After mentioning related works in Section II, we state our main results in Section III, which shows that distributed implementations of over-relaxed ADMM and GD obey the lifting relation (1). The proofs can be found in the Appendix. In Section IV we support conjecture (6) with numerical evidence. We present our final remarks in Section V.

II. RELATED WORK AND AN OPEN PROBLEM

We state conjecture (6) for the relatively simple problem (3), but, to the best of our knowledge, it cannot be resolved through the existing literature. We compare the *exact* *asymptotic convergence rates after optimal tuning* of ADMM and GD, while the majority of previous papers focus on upper bounding the *global convergence rate* of ADMM and, at best, optimize such an upper bound.

Furthermore, to obtain linear convergence, strong convexity is usually assumed [3], which does not hold for problem (3). Most results not requiring strong convexity focus on the convergence rate of the objective function, as opposed to this paper which focus on the convergence rate of the variables; see [4] for example.

Few papers consider a consensus problem with an objective function different than (3). For instance, [5] considers $f(\boldsymbol{z}) = \sum_{i\in\mathcal{V}} \sum \|z_i - c_i\|^2$, subject to $z_i = z_j$ if $(i,j) \in \mathcal{E}$, where $c_i > 0$ are constants. This problem is strongly convex and does not reduce to (3), and vice-versa. Other branch of research consider $f(\boldsymbol{z}) = \sum_i f_i(\boldsymbol{z})$ with ADMM iterations that are agnostic to whether or not $f_i(\boldsymbol{z})$ depends on a subset of the components of $\boldsymbol{z}$; see [6] and references therein. These are in contrast with our setting where decentralized ADMM is a message-passing algorithm [7], and the messages between agents i and j are only associated to the variables shared by functions f_i and f_j.

For quadratic problems, there are explicit results on the convergence rate and optimal parameters of ADMM [8–10]. However, their assumptions do not hold for the non strongly convex distributed problem considered in this paper. Moreover, there are very few results comparing the optimal convergence rate of ADMM as a function of the optimal convergence rate of GD. For a centralized setting, an explicit comparison is provided in [11], but it assumes strong convexity.

Finally, and most importantly, there is no prior result connecting GD and ADMM to lifted Markov chains. Lifted Markov chains were previously employed to speedup convergence time of distributed averaging and gossip algorithms [12–14], but these do not involve ADMM algorithm.

III. ADMM AS A LIFTING OF GRADIENT DESCENT

We now show that the lifting relation (1) holds when distributed implementations of over-relaxed ADMM and GD are applied to problem (3) defined over the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

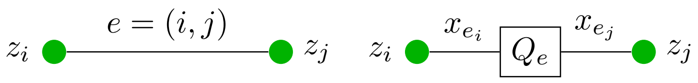
Let us introduce the extended set of variables $\boldsymbol{x} \in \mathbb{R}^{|\hat{\mathcal{E}}|}$, where

$$\hat{\mathcal{E}} = \{(e,i) : e \in \mathcal{E},\ i \in e,\ \text{and}\, i \in \mathcal{V}\}. \tag{7}$$

Note that $|\hat{\mathcal{E}}| = 2|\mathcal{E}|$. Each component of $\boldsymbol{x}$ is indexed by a pair $(e,i) \in \hat{\mathcal{E}}$. For simplicity we denote $e_i = (e,i)$. We can now write (3) as

$$\min_{\boldsymbol{x},\boldsymbol{z}} \left\{ f(\boldsymbol{x}) = \frac{1}{2} \sum_{e=(i,j)\in\mathcal{E}} q_e (x_{e_i} - x_{e_j})^2 \right\} \tag{8}$$

subject to $x_{e_i} = z_i$, $x_{e_j} = z_j$, for all $e=(i,j) \in \mathcal{E}$. The new variables are defined according to the following diagram:



Notice that we can also write $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top Q \boldsymbol{x}$ where Q is block diagonal, one block per edge $e=(i,j)$, in the form $Q_e = q_e \left(\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right)$. Let us define the matrix $S \in \mathbb{R}^{|\hat{\mathcal{E}}|\times|\mathcal{V}|}$ with components

$$S_{e_i,i} = S_{e_j,j} = \begin{cases} 1 & \text{if } e=(i,j)\in\mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The distributed over-relaxed ADMM is a first order method that operates on five variables: $\boldsymbol{x}$ and $\boldsymbol{z}$ defined above, and also $\boldsymbol{u}$, $\boldsymbol{m}$ and $\boldsymbol{n}$ introduced below. It depends on the relaxation parameter $\gamma \in (0,2)$, and several penalty parameters $\boldsymbol{\rho} \in \mathbb{R}^{|\hat{\mathcal{E}}|}$. The components of $\boldsymbol{\rho}$ are $\rho_{e_i} > 0$ for $e_i \in \hat{\mathcal{E}}$; see [7, 11] and also [15] for details on multiple ρ's. We can now write ADMM iterations as

$$\begin{aligned} \boldsymbol{x}^{t+1} &= A\boldsymbol{n}^t, \\ \boldsymbol{m}^{t+1} &= \gamma\boldsymbol{x}^{t+1} + \boldsymbol{u}^t, \\ \boldsymbol{s}^{t+1} &= (1-\gamma)\boldsymbol{s}^t + B\boldsymbol{m}^{t+1}, \\ \boldsymbol{u}^{t+1} &= \boldsymbol{u}^t + \gamma\boldsymbol{x}^{t+1} + (1-\gamma)\boldsymbol{s}^t - \boldsymbol{s}^{t+1}, \\ \boldsymbol{n}^{t+1} &= \boldsymbol{s}^{t+1} - \boldsymbol{u}^{t+1}, \end{aligned} \tag{10}$$

where $\boldsymbol{s}^t = S\boldsymbol{z}^t$, $B = S(S^\top D_{\boldsymbol{\rho}} S)^{-1} S^\top D_{\boldsymbol{\rho}}$, and $A = (I + D_{\boldsymbol{\rho}}^{-1} Q)^{-1}$. The next result shows that these iterations are equivalent to a linear system in $|\hat{\mathcal{E}}|$ dimensions. (The proofs of the following results are in the Appendix.)

Theorem 1 (Linear Evolution of ADMM). *Iterations* (10) *are equivalent to*

$$\boldsymbol{n}^{t+1} = T_A\,\boldsymbol{n}^t, \qquad T_A = I - \gamma(A + B - 2BA), \tag{11}$$

with $\boldsymbol{s}^t = B\boldsymbol{n}^t$ and $\boldsymbol{u}^t = -(I-B)\boldsymbol{n}^t$. All the variables of ADMM depend only on $\boldsymbol{n}^t$.

We can also write GD update $\boldsymbol{z}^{t+1} = (I - \alpha \nabla f)\boldsymbol{z}^t$ to problem (3) as

$$\boldsymbol{z}^{t+1} = T_G \boldsymbol{z}^t, \qquad T_G = I - \alpha S^\top Q S. \tag{12}$$

In the following, we establish lifting relations between distributed ADMM and GD in terms of matrices M_A and M_G, which are very closely related but not necessarily equal to T_A and T_G. They are defined as

$$M_G = (I - D_G)^{-1}(T_G - D_G), \tag{13}$$

$$M_A = (I - D_A)^{-1}(T_A - D_A), \tag{14}$$

where $D_G \neq I$ and $D_A \neq I$ are, for the moment, arbitrary diagonal matrices. Let us also introduce the vectors

$$\boldsymbol{v}_G = (I - D_G)\mathbf{1}, \tag{15}$$

$$\boldsymbol{v}_A = (I - D_A)\boldsymbol{\rho}. \tag{16}$$

As shown below, these matrices and vectors satisfy relation (1). Moreover, M_G can be interpreted as a probability transition matrix, and the rows of M_A sum up to one. We only lack the strict non-negativity of M_A, which in general is not a probability transition matrix. Thus, in general, we do not have a lifting between Markov chains, however, we still have lifting in the sense that M_A can be collapsed onto M_G according to (1).

Theorem 2. *For $(D_G)_{ii} < 1$ and sufficiently small α, M_G in* (13) *is a doubly stochastic matrix.*

Lemma 3. *The rows of M_G and M_A sum up to one, i.e. $M_G\mathbf{1} = \mathbf{1}$ and $M_A\mathbf{1} = \mathbf{1}$. Moreover, $\boldsymbol{v}_G^\top M_G = \boldsymbol{v}_G^\top$ and $\boldsymbol{v}_A^\top M_A = \boldsymbol{v}_A^\top$. These properties are shared with Markov matrices.*

Theorem 4 (ADMM as a Lifting of GD)**.** *M_A and M_G defined in* (13) *and* (14) *satisfy relation* (1)*, namely,*

$$\boldsymbol{v}_G = S^\top \boldsymbol{v}_A, \qquad D_{\boldsymbol{v}_G} M_G = S^\top D_{\boldsymbol{v}_A} M_A S, \tag{17}$$

provided D_G, D_A, α, γ, and $\boldsymbol{\rho}$ are related according to

$$S^\top D_{\boldsymbol{\rho}}(I - D_A)S = I - D_G, \tag{18}$$

$$\alpha = \frac{\gamma\, q_e \rho_{e_i,i}\, \rho_{e_j,j}}{\rho_{e_i,i}\, \rho_{e_j,j} + q_e\left(\rho_{e_i,i} + \rho_{e_j,j}\right)}, \tag{19}$$

for all $e = (i,j) \in \mathcal{E}$. Equation (19) *restricts the components of $\boldsymbol{\rho}$, and* (17) *is an equation for D_A and D_G.*

Theorem 5 (Negative Probabilities)**.** *There exists a graph $\mathcal{G}$ such that, for any diagonal matrix D_A, $\boldsymbol{\rho}$ and γ, the matrix M_A has at least one negative entry. Thus, in general, M_A is not a probability transition matrix.*

For concreteness, let us consider some explicit examples illustrating Theorem 4.

Regular Graphs. Let us consider the solution to equations (18) and (19) for d-regular graphs. Fix $q_e = 1$ and $\boldsymbol{\rho} = \rho\mathbf{1}$ for simplicity. Equation (18) is satisfied with

$$D_A = (1 - (\rho|\hat{\mathcal{E}}|)^{-1})I, \qquad D_G = (1 - |\mathcal{V}|^{-1})I, \tag{20}$$

since $d|\mathcal{V}| = |\hat{\mathcal{E}}| = 2|\mathcal{E}|$, while (19) requires

$$\alpha = \frac{\gamma\rho}{2+\rho}. \tag{21}$$

Notice that $(D_G)_{ii} < 1$ for all i, so choosing γ or ρ small enough we can make M_G positive. Moreover, the components of (15) and (16) are non-negative and sum up to one, i.e. $\boldsymbol{v}_G^\top\mathbf{1} = \boldsymbol{v}_A^\top\mathbf{1} = 1$, thus these vectors are stationary probability distributions of M_G and M_A.

Cycle Graph. Consider solving (3) over the 4-node cycle graph $\mathcal{G} = C_4$ shown in Fig. 1. By direct computation and upon using (20) we obtain

$$M_G = \begin{pmatrix} x & y & 0 & y \\ y & x & y & 0 \\ 0 & y & x & y \\ y & 0 & y & x \end{pmatrix}, \qquad M_A = \begin{pmatrix} \hat{x} & 0 & 0 & 0 & 0 & 0 & \hat{y} & \hat{z} \\ 0 & \hat{x} & \hat{z} & \hat{y} & 0 & 0 & 0 & 0 \\ \hat{y} & \hat{z} & \hat{x} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{x} & \hat{z} & \hat{y} & 0 & 0 \\ 0 & 0 & \hat{y} & \hat{z} & \hat{x} & 0 & 0 & 0 \\ 0 & 0 & \hat{y} & 0 & 0 & \hat{x} & \hat{z} & \hat{y} \\ 0 & 0 & 0 & 0 & \hat{y} & \hat{z} & \hat{x} & 0 \\ \hat{z} & \hat{y} & 0 & 0 & 0 & 0 & 0 & \hat{x} \end{pmatrix}, \tag{22}$$

where the probabilities of M_G are given by $x = 1 - 8\alpha$ and $x + 2y = 1$, and the probabilities of M_A are $\hat{x} = 1 - 4\gamma\rho$, $\hat{y} = 8\gamma\rho/(2+\rho)$ and $\hat{x} + \hat{y} + \hat{z} = 1$. The stationary probability vectors are $\boldsymbol{\pi} = \frac{1}{4}\mathbf{1}$ and $\hat{\boldsymbol{\pi}} = \frac{1}{8}\mathbf{1}$. Now (17) holds provided the parameters are related as (21). Moreover, in this particular case the matrix M_A is strictly non-negative, thus ADMM is a lifting of GD in the Markov chain sense.

Based on the above theorems we propose conjecture (6). The convergence rate τ is related to the convergence time, for instance $\mathcal{H} \sim (1-\tau)^{-1}$ if $\tau \approx 1$. Thus, let $\tau_G^\star$ and $\tau_A^\star$ be the optimal convergence rates of GD and ADMM, respectively. Then, at least for objective (3), and for any $\mathcal{G}$, we conjecture that there is some universal constant $C > 0$ such that

$$1 - \tau_A^\star \ge C\sqrt{1 - \tau_G^\star}. \tag{23}$$

IV. NUMERICAL EVIDENCE

For many graphs, we observe very few negative entries in M_A, which can be further reduced by adjusting the parameters $\boldsymbol{\rho}$ and γ. Nonetheless, in general, the lack of strict non-negativity of M_A prevents us from directly applying the theory of lifted Markov chains to prove (23). However, there is compelling numerical evidence to (23) as we now show.

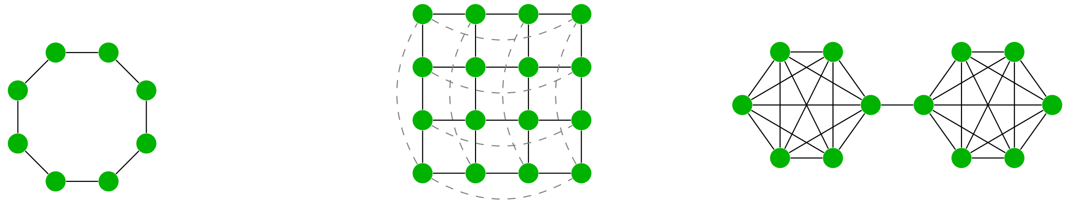
Consider a sequence of graphs $\{\mathcal{G}_n\}$, where $n = |\mathcal{V}|$, such that $\tau_G^\star \to 1$ and $\tau_A^\star \to 1$ as $n \to \infty$. Denote $\mathcal{R}_G(n) = (1 - \tau_G^\star)^{-1}$ and $\mathcal{R}_A(n) = (1 - \tau_A^\star)^{-1}$. We look for the smallest β such that $\mathcal{R}_A(n) \le C\,\mathcal{R}_G(n)^\beta$, for some $C > 0$, and all large enough n. If (23) was false, there would exist sequences $\{\mathcal{G}_n\}$ for which $\beta > 1/2$. For instance, if $\{\mathcal{G}_n\}$ have low conductance it is well-known that lifting does not speedup the mixing time and we could find $\beta = 1$.

To numerically find β, we plot

$$\hat{\beta}_1 = \frac{\log \mathcal{R}_A(n)}{\log \mathcal{R}_G(n)} \quad \text{and} \quad \hat{\beta}_2 = \frac{\mathcal{R}_G(n)}{\mathcal{R}_A(n)} \frac{\Delta \mathcal{R}_A(n)}{\Delta \mathcal{R}_G(n)} \tag{24}$$

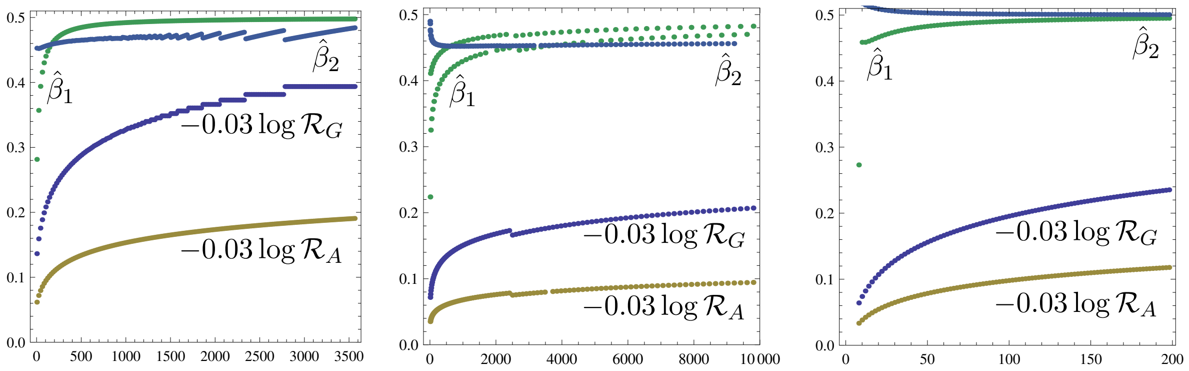
against n, where $\Delta h(n) = h(n+1) - h(n)$ for any function $h(n)$. The idea behind this is very simple. Let $f(x) = Cg(x)^\beta$, and $f, g \to \infty$ as $x \to \infty$. Then, $\log f / \log g \to \beta$ and also $\partial_x \log f / \partial_x \log g = (g\,\partial_x f)/(f\,\partial_x g) \to \beta$ as $x \to \infty$. Thus, we analyze (24) which are their discrete analogue. Given a graph $\mathcal{G}_n$, from (12) and (11) we numerically compute $\tau = \max_j\{|\lambda_j(T)| : |\lambda_j(T)| < 1\}$. The optimal convergence rates are thus given by $\tau_G^\star = \min_\alpha \tau_G$ and $\tau_A^\star = \min_{\{\gamma,\rho\}} \tau_A$, where we consider $\boldsymbol{\rho} = \rho\mathbf{1}$ with $\rho > 0$.

In Fig. 3 we show the three different graphs considered in the numerical analysis contained in the respective plots of Fig. 4. We show the values of (24) versus n, and also the curves $\log \mathcal{R}_A$ and $\log \mathcal{R}_G$ against n, which for visualization purposes are scaled by the factor -0.03. In Fig. 4a we see that (6), or equivalently (23), holds for the cycle graph $\mathcal{G} = C_n$. The same is true for the periodic grid, or torus grid graph $\mathcal{G} = T_n$, as shown in Fig. 4b. Surprisingly, as shown in Fig. 4c, we get the same square root speedup for a barbell graph, whose random



(a) (b) (c)

FIG. 3. We consider the following graphs for numerical analysis. (a) Cycle graph, $\mathcal{G} = C_n$. (b) Torus or periodic grid graph, $\mathcal{G} = T_n = C_{\sqrt{n}} \times C_{\sqrt{n}}$. (c) Barbel graph, obtained by connecting two complete graphs K_n by a bridge.



(a) (b) (c)

FIG. 4. Plot of (24) versus n, and also $\log \mathcal{R}_A$ and $\log \mathcal{R}_G$ versus n, scaled by -0.03 for visualization purposes only. The sequence of graphs $\{\mathcal{G}_n\}$ in each plot are of the types indicated in Fig. 3, in the same respective order. Notice that $\hat{\beta}_1, \hat{\beta}_2 \leq 1/2$, and $\hat{\beta}_1$ and $\hat{\beta}_2$ gets very close to $1/2$ for large n. (a) Cycle graph. (b) Torus grid graph. The two green curves occur because odd and even n behave differently. (c) Barbell graph. A Markov chain over this graph does not speedup via lifting, however, (23) still holds.

walk is known to not speedup via lifting. We find similar behavior for several other graphs but we omit these results due to the lack of space.

V. CONCLUSION

For a class of quadratic problems (3) we established a duality between lifted Markov chains and two important distributed optimization algorithms, GD and over-relaxed ADMM; see Fig. 2. We proved that precisely the same relation defining lifting of Markov chains (1), is satisfied between ADMM and GD. This is the content of Theorem 4. Although the lifting relation holds, in general, we cannot guarantee that the matrix M_A associated to ADMM is a probability transition matrix, since it might have a few negative entries. Therefore, in general, Theorem 4 is not a Markov chain lifting, but it is a lifting in a graph theoretical sense.

These negative entries actually make this parallel even more interesting since (6), or equivalently (23), do not violate theorems of lifted Markov chains where the square root improvement is a lower bound, thus the best possible, and in (6) it is an upper bound. For graphs with low conductance, the speedup given by Markov chain lifting is negligible. On the other hand, the lifting between ADMM and GD seems to always give the best possible speedup achieved by Markov chains, even for graphs with low conductance. This is numerically confirmed in Fig. 4c.

Due to the strong analogy with lifted Markov chains and numerical evidence, we conjectured the upper bound (6), or (23), which was well supported numerically. However, its formal explanation remains open.

Finally, although we considered a simple class of quadratic objectives, when close to the minimum the leading order term of more general convex functions is usually quadratic. In the cases where the dominant term is close to the form (3), the results presented in this paper should still hold. An attempt to prove our conjecture (23) is under investigation[1].

ACKNOWLEDGMENT

We would like to thank Andrea Montanari for his guidance and useful discussions. We also would like to thank the anonymous referees for careful reading of the manuscript and useful suggestions. This work was partially supported by an NIH/NIAID grant U01AI124302.

Appendix A: Proof of Main Results

In the main part of the paper we introduced the extended edge set $\hat{\mathcal{E}}$ which essentially duplicates the edges of the original graph, $|\hat{\mathcal{E}}| = 2|\mathcal{E}|$. This is the shortest route to state our results concisely but it complicates the notation in the following proofs. Therefore, we first introduce the notion of a factor graph for problem (3).

[1] Note added: soon after the acceptance of this paper, we found a proof of (23) for a class of quadratic problems. These results will be presented elsewhere.

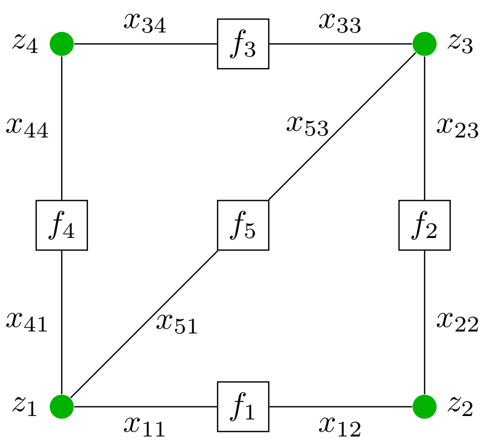


FIG. 5. Example of a factor graph $\bar{\mathcal{G}}$ for problem (3), and (8), where $\mathcal{G}$ is the complete graph K_4 with one edge removed.

1. Factor Graph

The factor graph $\bar{\mathcal{G}} = (\bar{\mathcal{F}}, \bar{\mathcal{V}}, \bar{\mathcal{E}})$ for problem (3) is a bipartite graph that summarizes how different variables are shared across different terms in the objective. This is illustrated in Fig. 5, where for this case (3) is given by

$$\begin{aligned} f(\boldsymbol{z}) = q^1(z_1 - z_2)^2 + q^2(z_2 - z_3)^2 + q^3(z_3 - z_4)^2 \\ + q^4(z_4 - z_1)^2 + q^5(z_4 - z_2)^2 \end{aligned} \tag{A1}$$

while (8) is given by

$$\begin{aligned} f(\boldsymbol{x}) = q^1(x_{12} - x_{11})^2 + q^2(x_{23} - x_{22})^2 \\ + q^3(x_{34} - x_{33})^2 + q^4(x_{41} - x_{44})^2 + q^5(x_{45} - x_{52})^2 \end{aligned} \tag{A2}$$

where $x_{11} = x_{41}$, $x_{12} = x_{22}$, $x_{23} = x_{33}$, and $x_{44} = x_{34}$.

The factor graph $\bar{\mathcal{G}}$ has two sets of vertices, $\bar{\mathcal{F}}$ and $\bar{\mathcal{V}}$. The circles in Fig. 5 represent the nodes in $\bar{\mathcal{V}} = \mathcal{V}$, and the squares represent the nodes in $\bar{\mathcal{F}} = \mathcal{E}$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the original graph. Note that each $a \in \bar{\mathcal{F}}$ is uniquely associated to one edge $e \in \mathcal{E}$, and uniquely associated to one term in the sum of the objective. In equation (8) we referred to each term as $f_e = q_e(x_{e_i} - x_{e_j})^2$, but now we refer to it by f_a. With a slightly abuse of notation we indiscriminately write $a \in \bar{\mathcal{F}}$ or $f_a \in \bar{\mathcal{F}}$. Each node $b \in \bar{\mathcal{V}}$ is uniquely associated to one node $i \in \mathcal{V}$, and uniquely associated to one component of $\boldsymbol{z}$. Before we referred to this variable by z_i, but now we refer to it by z_b, and indiscriminately write $b \in \bar{\mathcal{F}}$ or $z_b \in \bar{\mathcal{F}}$. Each edge

$(a,b) \in \bar{\mathcal{E}}$ must have $a \in \bar{\mathcal{F}}$ and $b \in \bar{\mathcal{V}}$, and its existence implies that the function f_a depends on variable z_b. Moreover, each edge $(a,b) \in \bar{\mathcal{E}}$ is also uniquely associated to one component of $\boldsymbol{x}$ in the equivalent formulation (8). In particular, if $a \in \bar{\mathcal{E}}$ is associated to $e \in \mathcal{E}$, and $b \in \bar{\mathcal{V}}$ is associated to $i \in \mathcal{V}$, then $(a,b) \in \bar{\mathcal{E}}$ is associated to x_{e_i}. Here, we denote x_{e_i} by x_{ab}. Thus, we can think of $\bar{\mathcal{E}}$ as being the same as $\hat{\mathcal{E}}$. Another way of thinking of $\bar{\mathcal{E}}$ and $\boldsymbol{x}$ is as follows. If $(a,b) \in \bar{\mathcal{E}}$ then $x_{ab} = z_b$ appears as a constraint in (8).

Let us introduce the neighbor set of a given node in $\bar{\mathcal{G}}$. For $a \in \bar{\mathcal{F}}$, the independent variables of f_a are in the set

$$N_a = \{b \in \bar{\mathcal{V}} : (a,b) \in \bar{\mathcal{G}}\}. \tag{A3}$$

Analogously, for $b \in \bar{\mathcal{V}}$, the functions that depend on z_b are in the set

$$N_b = \{a \in \bar{\mathcal{F}} : (a,b) \in \bar{\mathcal{G}}\}. \tag{A4}$$

In other words, $N_{\bullet}$ denotes the neighbors of either circle or square nodes in $\bar{\mathcal{G}}$. For $a \in \bar{\mathcal{F}}$ we define

$$I_a = \{e \in \bar{\mathcal{E}} : e \text{ is incident on } a\}. \tag{A5}$$

For $b \in \bar{\mathcal{V}}$ we define

$$I_b = \{e \in \bar{\mathcal{E}} : e \text{ is incident on } b\}. \tag{A6}$$

If we re-write problem (8) using this new notation, which indexes variables by the position they the take on $\bar{\mathcal{G}}$, the objective function takes the form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top Q \boldsymbol{x} = \frac{1}{2}\sum_{a \in \mathcal{F}} \boldsymbol{x}_a^\top Q^a \boldsymbol{x}_a \tag{A7}$$

where $Q \in \mathbb{R}^{\bar{\mathcal{E}} \times \bar{\mathcal{E}}}$ is block diagonal and each block, now indexed by $a \in \bar{\mathcal{F}}$, takes the form $Q^a = q^a \left(\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right)$, where $q^a > 0$, and $\boldsymbol{x}_a = (x_{ab}, x_{ac})^\top$ for $(a,b) \in \bar{\mathcal{E}}$ and $(a,c) \in \bar{\mathcal{E}}$. Here, q^a is the same as q_e in the main text. We also have the constraints $x_{ab} = x_{a'b} = z_b$ for each $a, a' \in N_b$ and $b \in \bar{\mathcal{V}}$. The row stochastic matrix S introduced in the ADMM iterations is now expressed as $S \in \mathbb{R}^{|\bar{\mathcal{E}}| \times |\bar{\mathcal{V}}|}$ and has a single 1 per row such that $S_{eb} = 1$ if and only if edge $e \in \bar{\mathcal{E}}$ is incident on $b \in \bar{\mathcal{V}}$. Notice that $S^\top S = D_{\boldsymbol{d}}$ is the degree matrix of the original graph $\mathcal{G}$.

With this notation at hands, we now proceed to the proofs.

2. Proof of Theorem 1

Recall that $B = S(S^\top D_\rho S)^{-1} S^\top D_\rho$, thus $B^2 = B$ is a projection operator, and $B^\perp = I - B$ its orthogonal complement. Consider updates (10). Substituting $\boldsymbol{x}^{t+1}$ and $\boldsymbol{m}^{t+1}$ into the other variables we obtain

$$\begin{pmatrix} I & 0 & 0 \\ I & I & 0 \\ -I & I & I \end{pmatrix} \begin{pmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{u}^{t+1} \\ \boldsymbol{n}^{t+1} \end{pmatrix} = \begin{pmatrix} (1-\gamma)I & B & \gamma BA \\ (1-\gamma)I & I & \gamma A \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{u}^{t} \\ \boldsymbol{n}^{t} \end{pmatrix}$$

which can be easily inverted yielding

$$\boldsymbol{s}^{t+1} = (1-\gamma)\boldsymbol{s}^t + B\boldsymbol{u}^t + \gamma BA\boldsymbol{n}^t, \tag{A8}$$

$$\boldsymbol{u}^{t+1} = B^\perp \boldsymbol{u}^t + \gamma B^\perp A \boldsymbol{n}^t, \tag{A9}$$

$$\boldsymbol{n}^{t+1} = (1-\gamma)\boldsymbol{s}^t + (B - B^\perp)\boldsymbol{u}^t + \gamma(B - B^\perp)A\boldsymbol{n}^t. \tag{A10}$$

Note the following important relations:

$$B\boldsymbol{n}^t = \boldsymbol{s}^t, \qquad B^\perp \boldsymbol{n}^t = -\boldsymbol{u}^t, \tag{A11}$$

$$B\boldsymbol{s}^t = \boldsymbol{s}^t, \qquad B^\perp \boldsymbol{s}^t = 0, \tag{A12}$$

$$B^\perp \boldsymbol{u}^t = \boldsymbol{u}^t, \qquad B\boldsymbol{u}^t = 0. \tag{A13}$$

Equation (A12) is a simple consequence of the definition of B, i.e.

$$B\,\boldsymbol{s}^t = S(S^\top D_{\boldsymbol{\rho}} S)^{-1}(S^\top D_{\boldsymbol{\rho}} S)\boldsymbol{z}^t = \boldsymbol{s}^t, \tag{A14}$$

which also implies $B^\perp \boldsymbol{s}^t = \mathbf{0}$. Since $BB^\perp = 0$, acting with B over (A9) implies $B\boldsymbol{u}^t = \mathbf{0}$ for every t, and also $B^\perp \boldsymbol{u}^t = \boldsymbol{u}^t$, which shows (A13). Now (A11) follows from these facts and the own definition $\boldsymbol{n}^t = \boldsymbol{s}^t - \boldsymbol{u}^t$. Finally, applying (A11) on (A10) we obtain $\boldsymbol{n}^{t+1} = T_A \boldsymbol{n}^t$ where $T_A = I - \gamma(A + B - 2BA)$.

3. Proof of Theorem 2

Write $Q = Q^+ + Q^-$ where Q^+ is diagonal and has only positive entries, and Q^- only has off-diagonal and negative entries. First, notice that $(S^\top Q^+ S)$ is also diagonal. Indeed, for $b \in \bar{\mathcal{V}}$ and $c \in \bar{\mathcal{V}}$, $(S^\top Q^+ S)_{bc} = \sum_{e \in \bar{\mathcal{E}}} S_{eb} Q^+_{ee} S_{ec} = \delta_{bc} \sum_{e \in I_b} Q^+_{ee}$ where δ is the Kronecker delta. By a similar argument, $S^\top Q^- S$ is off-diagonal. Hence, if $b \neq c$ we have

$$(T_G)_{bc} = -\alpha \sum_{e \in I_b} \sum_{e' \in I_c} Q^-_{ee'} \geq 0. \tag{A15}$$

Recall that $M_G = (I - D_G)^{-1}(T_G - D_G)$, where $D_G \neq I$ is diagonal. For M_G to be non-negative we first impose that $(D_G)_{bb} < 1$ for all $b \in \bar{\mathcal{V}}$. Then, since the off-diagonal elements of T_G are automatically positive by (A15), we just need to consider the diagonal elements of $T_G - D_G$. Thus we require that for every $b \in \bar{\mathcal{V}}$,

$$1 - \alpha \sum_{e \in I_b} Q_{ee} + (D_G)_{bb} \geq 0. \tag{A16}$$

Denoting $Q_{\max} = \max_{b \in \bar{\mathcal{V}}} \sum_{e \in I_b} Q_{ee}$ and $D_{G,\min}$ the smallest element of D_G, the matrix M_G will be non-negative provided $\alpha \leq (1 + D_{G,\min})/Q_{\max}$.

Notice that $S\mathbf{1}_{|\bar{\mathcal{V}}|} = \mathbf{1}_{|\bar{\mathcal{E}}|}$ and $Q\mathbf{1} = \mathbf{0}$. Thus $S^\top Q S \mathbf{1} = \mathbf{0}$, implying $T_G \mathbf{1} = \mathbf{1}$, and $\mathbf{1}^\top T_G = \mathbf{1}^\top$. From this we have $M_G \mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top M_G = \mathbf{1}^\top$, so all the rows and columns of M_G sum up to one.

4. Proof of Lemma 3

We proved above that M_G is a doubly stochastic matrix. Now let us consider M_A. Recall the definition of $B = S^\top (S^\top D_{\boldsymbol{\rho}} S)^{-1} S^\top D_{\boldsymbol{\rho}}$. Note that the action of B on a vector $\boldsymbol{v} \in \mathbb{R}^{|\bar{\mathcal{E}}|}$ is to take a weighted average of its components, namely, if $(a, b) \in \bar{\mathcal{E}}$ then

$$(B\boldsymbol{v})_{ab} = \frac{\sum_{c \in N_b} \rho_{cb} v_{cb}}{\sum_{c \in N_b} \rho_{cb}}. \tag{A17}$$

Therefore, $B\mathbf{1} = \mathbf{1}$. Recall that $Q\mathbf{1} = \mathbf{0}$, thus $A\mathbf{1} = \mathbf{1}$, where $A = (I + D_{\boldsymbol{\rho}}^{-1} Q)^{-1}$, which implies $T_A \mathbf{1} = \mathbf{1}$, and in turn $M_A \mathbf{1} = \mathbf{1}$. Now the other relations follow trivially.

5. Proof of Theorem 4

Due to the block diagonal structure of Q it is possible write A explicitly as

$$A = I - FQ, \tag{A18}$$

where F is a block diagonal matrix with $|\bar{\mathcal{F}}|$ blocks. Each block F^a, for $a \in \bar{\mathcal{F}}$, is of the form

$$F^a = \frac{q^a}{\rho_{ab}\rho_{ac} + q^a(\rho_{ab} + \rho_{ac})} \begin{pmatrix} \rho_{ac} & 0 \\ 0 & \rho_{ab} \end{pmatrix}, \tag{A19}$$

where $b, c \in N_a$. Now by the definition of B we have $S^\top D_{\boldsymbol{\rho}} B = S^\top D_{\boldsymbol{\rho}}$. Hence,

$$S^\top D_{\boldsymbol{v}_A} M_A S = S^\top D_{\boldsymbol{\rho}}(I - D_A)S - \gamma S^\top D_{\boldsymbol{\rho}} F Q S, \tag{A20}$$

$$D_{\boldsymbol{v}_G} M_G = (I - D_G) - \alpha S^\top Q S. \tag{A21}$$

Equating the first term of (A20) to the first term of (A21), and also the second terms to each other, on using (A18) we obtain

$$S^\top D_{\boldsymbol{\rho}}(I - D_A)S = I - D_G, \tag{A22}$$

$$\alpha = \frac{\gamma\, q^a \rho_{ab}\rho_{ac}}{\rho_{ab}\rho_{ac} + q^a(\rho_{ab} + \rho_{bc})}, \tag{A23}$$

where (A23) must hold for all $a \in \bar{\mathcal{F}}$ and $b, c \in N_a$. This gives the second equality in (17) together with relations (18) and (19). Finally, since diagonal matrices commute, $S^\top \boldsymbol{v}_A = S^\top (I - D_A) D_{\boldsymbol{\rho}} S \mathbf{1}_{|\bar{\mathcal{V}}|} = (I - D_G)\mathbf{1}_{|\bar{\mathcal{V}}|} = \boldsymbol{v}_G$, which gives the first relation in (17).

6. Proof of Theorem 5

It suffices to show one example with at least one negative entry. Let $\mathcal{G}$ be the complete graph K_4 with one edge removed, as shown in Fig. 5. By direct inspection one finds the following sub-matrix of T_A:

$$T^{(S)} = \begin{pmatrix} (T_A)_{21} & (T_A)_{24} \\ (T_A)_{31} & (T_A)_{34} \end{pmatrix} \tag{A24}$$

whose elements are explicitly given by

$$(T_A)_{21} = \frac{\gamma\, \rho_{11}(\rho_{12} - \rho_{22})}{(\rho_{12} + \rho_{22})(\rho_{11} + \rho_{12} + \rho_{11}\rho_{12})}, \tag{A25}$$

$$(T_A)_{24} = \frac{2\gamma}{(\rho_{12} + \rho_{22})\left(1 + \rho_{22}^{-1} + \rho_{23}^{-1}\right)}, \tag{A26}$$

$$(T_A)_{31} = \frac{2\gamma}{(\rho_{12} + \rho_{22})\left(1 + \rho_{11}^{-1} + \rho_{12}^{-1}\right)}, \tag{A27}$$

$$(T_A)_{34} = -\frac{\gamma\, \rho_{23}(\rho_{12} - \rho_{22})}{(\rho_{12} + \rho_{22})(\rho_{22} + \rho_{23} + \rho_{22}\rho_{23})}. \tag{A28}$$

First notice that subtracting D_A from T_A does not affect $T^{(S)}$. Now recall that all components of $\boldsymbol{\rho}$ must be strictly positive. The elements (A25) and (A28) have opposite signs, so one of them is negative. Since (A26) and (A27) are both positive, one cannot remove the negative entries of an entire row of T_A by multiplying T_A by the diagonal matrix $(I - D_A)^{-1}$. Therefore, $M_A = (I - D_A)^{-1}(T_A - D_A)$ has at least one negative entry.

---

[1] F. Chen, L. Lovász, and L. Pak. Lifting Markov Chains to Speed up Mixing. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 275–281, 1999.

[2] J. Norris. Markov Chains. Cambridge University Press, Cambridge, 1998.

[3] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the Linear Convergence of the ADMM in Decentralized Consensus Optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[4] D. Davis and W. Yin. Convergence Rate Analysis of Several Splitting Schemes. *arXiv preprint arXiv:1406.4834*, 2014.

[5] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista. Fast Consensus by the Alternating Direction Multipliers Method. *IEEE Transactions on Signal Processing*, 59(11):5523–5537, 2011.

[6] E. Wei and A. Ozdaglar. Distributed Alternating Direction Method of Multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.

[7] N. Derbinsky, J. Bento, V. Elser, and J. Yedidia. An Improved Three-Weight Message Passing Algorithm. arXiv:1305.1961v1 [cs.AI], 2013.

[8] André Teixeira, Euhanna Ghadimi, Iman Shames, Henrik Sandberg, and Mikael Johansson. Optimal Scaling of the ADMM Algorithm for Distributed Quadratic Programming. In *52nd IEEE Conference on Decision and Control*, pages 6868–6873. IEEE, 2013.

[9] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson. Optimal Parameter Selection for the Alternating Direction Method of Multipliers (ADMM): Quadratic Problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2015.

[10] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem. Explicit Convergence Rate of a Distributed Alternating Direction Method of Multipliers. *IEEE Transactions on Automatic Control*, 61(4):892–904, 2016.

[11] G. França and J. Bento. An Explicit Rate Bound for Over-Relaxed ADMM. In *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016*, pages 2104–2108, 2016.

[12] K. Jung, D. Shah, and J. Shin. Fast Gossip Through Lifted Markov Chains. In *Proc. Allerton Conf. on Comm., Control, and Computing, Urbana-Champaign, IL*, 2007.

[13] W. Li, H. Dai, and Y. Zhang. Location-Aided Fast Distributed Consensus in Wireless Networks. *IEEE Transactions on Information Theory*, 56(12):6208–6227, 2010.

[14] K. Jung, D. Shah, and J. Shin. Distributed averaging via lifted markov chains. *IEEE Transactions on Information Theory*, 56(1):634–647, 2010.

[15] J. Bento, N. Derbinsky, J. Alonso-Mora, and J. Yedidia. A Message-Passing Algorithm for Multi-Agent Trajectory Planning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 521–529. Curran Associates, Inc., 2013.