Markov chain lifting and distributed ADMM

Guilherme França
Boston College, Department of Computer Science
Johns Hopkins University, Center for Imaging Science
guifranca@gmail.com

José Bento
Boston College, Department of Computer Science
jose.bento@bc.edu

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 whereby a distributed ADMM algorithm can be seen as a lifting of Gradient Descent. This provides a deep insight for its faster convergence rate under optimal parameter tuning. We conjecture that this gain is always present, contrary to when lifting a Markov chain where sometimes we only obtain a marginal speedup.

1 Introduction

Let $\mathcal{M}$ and $\hat{\mathcal{M}}$ be two finite Markov chains with states $\mathcal{V}$ and $\hat{\mathcal{V}}$, of sizes $|\hat{\mathcal{V}}| > |\mathcal{V}|$, and with transition matrices M and $\hat{M}$, respectively. Let their stationary distributions be $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$. Sometimes 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* [1]. We say that $\hat{\mathcal{M}}$ is a lifting of $\mathcal{M}$ if there is a row stochastic matrix $S \in \mathbb{R}^{|\hat{\mathcal{V}}|\times|\mathcal{V}|}$, where $\mathbf{1}_{\hat{\mathcal{V}}}^\top S = \mathbf{1}_{\mathcal{V}}^\top$ and $\mathbf{1}$ is the all-ones 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 $\boldsymbol{v} \in \mathbb{R}^n$, $D_{\boldsymbol{v}} = \mathrm{diag}(v_1, \ldots, v_n)$.

Lifting is very useful when the mixing time $\hat{\mathcal{H}}$ of the lifted chain $\hat{\mathcal{M}}$ is much smaller than the mixing time $\mathcal{H}$ of the original chain $\mathcal{M}$[1]. There are several examples where $\hat{\mathcal{H}} \approx C\sqrt{\mathcal{H}}$, for some constant $C > 0$. However, there is a limit on how much speedup lifting can achieve [1]. 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 $\hat{\mathcal{H}} \geq C\mathcal{H}$.

On the other hand, there is a well known relationship between Gradient Descent (GD) and Markov chains. Consider the problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^{|\mathcal{V}|}} \left\{ f(\boldsymbol{z}) = \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} (z_j - z_j)^2 \right\} \tag{2}$$

defined over the *undirected* and *connected* graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. The GD iteration is a function of the probability transition matrix M associated to the random walk on $\mathcal{G}$:

$$\boldsymbol{z}^{t+1} = [I - \alpha \nabla f]\, \boldsymbol{z}^t = [I - \alpha \mathcal{D}(I - M)]\, \boldsymbol{z}^t, \tag{3}$$

[1] We follow the definitions of [1] but up to multiplicative factors and slightly loser bounds, one can think of mixing time as $\mathcal{H} = \min\{t : \max_{\{i,\boldsymbol{p}^0\}} |p_i^t - \pi_i| < 1/4\}$, where p_i^t is the probability of being on state i after t steps elapse from the initial distribution $\boldsymbol{p}^0$.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

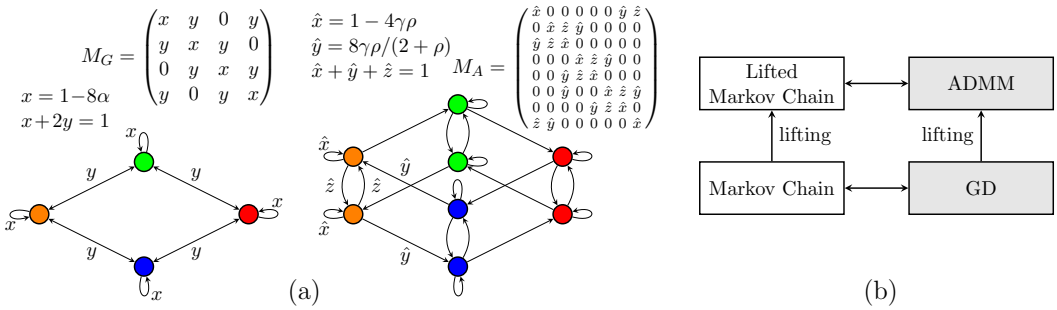


Figure 1: (a) For the 4-node ring the evolution of GD and ADMM is related to probability transition matrices M_G and M_A, where the later is a lifting of the former. Here α is the step-size of GD and (γ, ρ) the parameters of ADMM. (b) We propose that this holds in general, where ADMM is the analogous of a lifted Markov chain while GD is the analogous of the original Markov chain.

where $\alpha > 0$ is the step-size and $\mathcal{D} = D_{\boldsymbol{d}}$ is the degree matrix of $\mathcal{G}$, where $\boldsymbol{d} = \mathrm{diag}(d_1, \ldots, d_{|\mathcal{V}|})$ and d_i is the degree of node $i \in \mathcal{V}$.

The above connection is specially clear for d-regular graphs. Choosing $\alpha = 1/d$, equation (3) simplifies to $\boldsymbol{z}^{t+1} = M\boldsymbol{z}^t$. If M is irreducible and aperiodic, denoting $\lambda_2(M)$ its second largest eigenvalue in absolute value, then both the mixing time of a Markov chain on $\mathcal{G}$ and the time for GD to reach a given accuracy are equal to

$$\mathcal{H} = \frac{C}{\log(1/|\lambda_2|)} \approx \frac{C}{1-|\lambda_2|} \tag{4}$$

for some constant C, and for $\lambda_2 \approx 1$ in the second equality.

Let us anticipate our proposal by a concrete example. Consider solving (2) over the 4-node ring graph with two algorithms, GD and the distributed *Alternating Direction Method of Multipliers* (ADMM). Both are first-order methods and the objective is quadratic, so they can be written as linear systems with matrices T_G and T_A, respectively. Surprisingly, there are Markov matrices M_G and M_A, very closely related to T_G and T_A, where M_A is a lifting of M_G! The situation is illustrated in Figure 1 (a). In this case, we have the lifting equations (1) satisfied with $\boldsymbol{\pi} = \frac{1}{4}\mathbf{1}$, $\hat{\boldsymbol{\pi}} = \frac{1}{8}\mathbf{1}$, and $\alpha = \frac{\gamma\rho}{2+\rho}$.

We conjecture that this connection holds more generally, as described in Figure 1 (b). First, we show that for problems like (2) GD and ADMM satisfy an analogous relation to (1), although M_A might have few negative entries. Therefore, ADMM can be seen as a "lifting" of GD. Second, since lifting can speed mixing times up to a square root factor, we conjecture that the optimal convergence time $\mathcal{H}_A^\star$ of ADMM is related to the optimal convergence time $\mathcal{H}_G^\star$ of GD as follows:

Conjecture 1 (ADMM lifting speedup)**.** *For problems like* (2) *over any connected graph $\mathcal{G}$, there is a universal constant $C > 0$ such that*

$$\mathcal{H}_A^* \leq C\sqrt{\mathcal{H}_G^*}. \tag{5}$$

Note that this is a much stronger statement than for lifted Markov chains, where for some graphs the gain in speed is only marginal. We support this conjecture with numerical evidence. Due to lack of space, our proofs are included as supplementary material. We end the paper with related works.

2 ADMM as a lifting of Gradient Descent

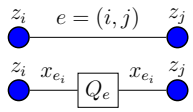
In this section we show that the lifting relations (1) hold for the following generalization of (2):

$$\min_{\boldsymbol{z}\in\mathbb{R}^{|\mathcal{V}|}} \Big\{ f(\boldsymbol{z}) = \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} q_{ij}(z_i - z_j)^2 \Big\}, \tag{6}$$

where $q_{ij} = q_e$ and $e = (i, j)$. Let us introduce the extended set of variables $\boldsymbol{x} \in \mathbb{R}^{|\hat{\mathcal{E}}|}$, where $\hat{\mathcal{E}} = \{(e, i) : e \in \mathcal{E},\ i \in e, \text{ and } i \in \mathcal{V}\}$. 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) by e_i. Now (6) can be written as

$$\min_{\boldsymbol{x},\boldsymbol{z}} \ \Big\{ f(\boldsymbol{x}) = \frac{1}{2} \sum_{e=(i,j)\in\mathcal{E}} q_e (x_{e_i} - x_{e_j})^2 \Big\} \quad \text{subject to } x_{e_i} = z_i,\, x_{e_j} = z_j,\, \forall\, e=(i,j)\in\mathcal{E} \tag{7}$$



Notice that $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top Q \boldsymbol{x}$ where Q is block diagonal, one block per edge $e=(i,j)$, in the form $Q_e = q_e \left(\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right)$. The diagram in (7) explains the new variables introduced. We define the matrix $S \in \mathbb{R}^{|\hat{\mathcal{E}}|\times|\mathcal{V}|}$ with components such that $S_{e_i,i} = S_{e_j,j} = 1$ for all $e=(i,j)\in\mathcal{E}$ and zero otherwise.

The distributed over-relaxed ADMM is a first order method that operates in five variables: $\boldsymbol{x}$ and $\boldsymbol{z}$, defined in (7), but also $\boldsymbol{u}$, $\boldsymbol{m}$ and $\boldsymbol{n}$, introduced below. It depends on the relaxation parameter $\gamma \in (0,2)$ and several 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 [2, 3], and also [4] for more details on multiple ρ's). We can now write ADMM iterations as

$$\begin{aligned} &\boldsymbol{x}^{t+1} = A\boldsymbol{n}^t, \qquad \boldsymbol{m}^{t+1} = \gamma\boldsymbol{x}^{t+1} + \boldsymbol{u}^t, \qquad \boldsymbol{s}^{t+1} = (1-\gamma)\boldsymbol{s}^t + B\boldsymbol{m}^{t+1}, \\ &\boldsymbol{u}^{t+1} = \boldsymbol{u}^t + \gamma\boldsymbol{x}^{t+1} + (1-\gamma)\boldsymbol{s}^t - \boldsymbol{s}^{t+1}, \qquad \boldsymbol{n}^{t+1} = \boldsymbol{s}^{t+1} - \boldsymbol{u}^{t+1}, \end{aligned} \tag{8}$$

where $\boldsymbol{s}^t = S\boldsymbol{z}^t$, $B = S(S^\top D_{\boldsymbol{\rho}} S)^{-1} S^\top D_{\boldsymbol{\rho}}$, and $A = (I + D_{\boldsymbol{\rho}}^{-1} Q)^{-1}$.

Theorem 2 (Linear evolution for ADMM)**.** *Iterations* (8) *are equivalent to*

$$\boldsymbol{n}^{t+1} = T_A\, \boldsymbol{n}^t \qquad \textit{where} \qquad T_A = I - \gamma(A + B - 2BA), \tag{9}$$

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

We can also generalize GD rule (3) to problem (6) as

$$\boldsymbol{z}^{t+1} = T_G \boldsymbol{z}^t \qquad \text{where} \qquad T_G = I - \alpha S^\top Q S. \tag{10}$$

In the following we establish lifting relations between 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

$$M_G = (I - D_G)^{-1}(T_G - D_G), \qquad \boldsymbol{v}_G = (I - D_G)\mathbf{1}, \tag{11}$$

$$M_A = (I - D_A)^{-1}(T_A - D_A), \qquad \boldsymbol{v}_A = (I - D_A)\boldsymbol{\rho}, \tag{12}$$

where $D_G \neq I$ and $D_A \neq I$ are arbitrary diagonal matrices. We also introduced two vectors.

We demonstrate below that M_G and M_A satisfy (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. Therefore, in general, we do not have a lifting between Markov chains. Some of the proof techniques we use are standard, and we include them as supplementary material.

Theorem 3. *For $(D_G)_{ii} < 1$, and sufficiently small α, M_G in* (11) *is a doubly stochastic matrix.*

Lemma 4. *The rows of M_G and M_A sum up to one: $M_G\mathbf{1} = \mathbf{1}$ 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 are properties shared with Markov matrices (see Section 1).*

Theorem 5 (ADMM as a lifting of GD)**.** *M_A and M_G satisfy relation* (1)*, namely,*

$$\boldsymbol{v}_G = S^\top \boldsymbol{v}_A, \qquad D_{\boldsymbol{v}_G} M_G = S^\top D_{\boldsymbol{v}_A} M_A S, \tag{13}$$

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

$$S^\top D_{\boldsymbol{\rho}}(I - D_A)S = I - D_G, \qquad \alpha = \frac{\gamma\, q_e \rho_{e_i,i}\rho_{e_j,j}}{\rho_{e_i,i}\rho_{e_j,j} + q_e\left(\rho_{e_i,i} + \rho_{e_j,j}\right)}, \tag{14}$$

for all $e=(i,j)\in\mathcal{E}$. The last equation restricts the components of $\boldsymbol{\rho}$.

Theorem 6 (Negative probabilities)**.** *There exists a graph $\mathcal{G}$ such that, for any diagonal matrix D_A, $\boldsymbol{\rho}$ and γ, M_A has at least one negative entry. Thus in general M_A is not a transition matrix.*

Remark 7 (Regular graphs)**.** *As shown in Figure 1, for the n-node ring we have true lifted Markov chain since M_A is non-negative. Now consider d-regular graphs. We fix $q_e = 1$ for simplicity, and $\boldsymbol{\rho} = \rho\mathbf{1}$. Equation* (14)*-left is satisfied by $D_A = (1 - (\rho|\hat{\mathcal{E}}|)^{-1})I$ and $D_G = (1 - |\mathcal{V}|^{-1})I$, since $d|\mathcal{V}| = |\hat{\mathcal{E}}| = 2|\mathcal{E}|$. Equation* (14)*-right imposes $\alpha = \gamma\rho/(2+\rho)$. Notice that $(D_G)_{ii} < 1$ for all i, thus choosing γ or ρ small enough we can make M_G positive. Moreover, $\boldsymbol{v}_G^\top\mathbf{1} = \boldsymbol{v}_A^\top\mathbf{1} = 1$, and all components of $\boldsymbol{v}_G$ and $\boldsymbol{v}_A$ are non-negative, hence these are stationary probability distributions of M_G and M_A. Thus, except for a few negative entries, M_A is a Markov chain lifting of M_G.*

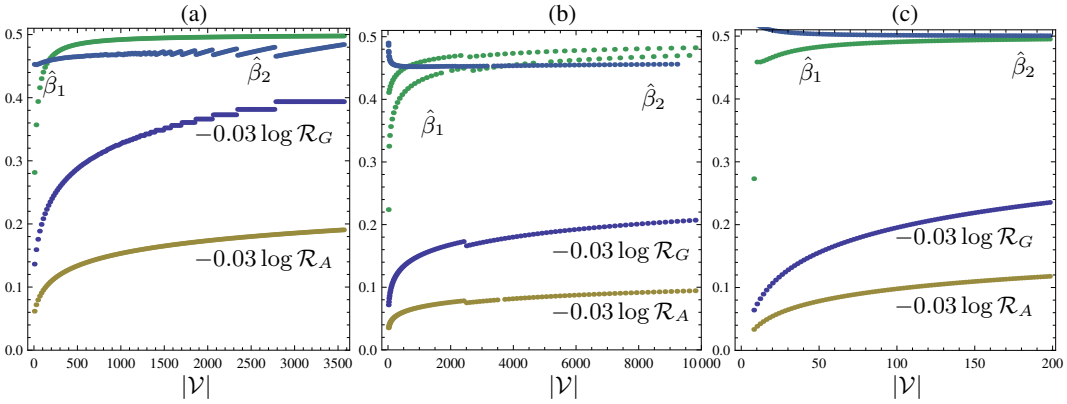


Figure 2: Convergence rate of ADMM and GD for different graphs. (a) Chain; (b) Periodic 2D grid. The two green curves occur because odd and even $|\mathcal{V}|$ behave differently; (c) Barbell graph.

Based on the above results we propose Conjecture 1, which is equivalent to the following statement. Let $\tau_G^\star$ and $\tau_A^\star$ be the optimal convergence rates of GD and ADMM, respectively. At least for objective (7), and for any $\mathcal{G}$, there is some universal constant $C > 0$ such that

$$1 - \tau_A^\star \geq C\sqrt{1 - \tau_G^\star} \tag{15}$$

when $\tau_G^\star$ and $\tau_A^\star$ are close to 1. To see this equivalence, just recall equation (4).

3 Numerical evidence

For many graphs, we observe very few negative entries in T_A and M_A. These can be reduced even more by adjusting $\boldsymbol{\rho}$ and γ. Nonetheless, in general, the lack of strict non-negativity in M_A prevents us from directly applying the theory of lifted Markov chains to relate ADMM and GD.

However, there is compelling numerical evidence to Conjecture 1, or equivalently (15). Consider a sequence $\{\mathcal{G}_n\}$ of graphs, 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 $\mathcal{R}_A(n) \leq C\,\mathcal{R}_G(n)^\beta$, for some $C > 0$ and all n large enough. Would (15) be false, there would exist sequences $\{\mathcal{G}_n\}$ for which $\beta \neq 1/2$. For instance, if $\{\mathcal{G}_n\}$ have low conductance, lifting does not speedup the mixing time of random walks on $\{\mathcal{G}_n\}$, and we could find $\beta = 1$.

To find β we plot $\hat{\beta}_1 = \frac{\log \mathcal{R}_A(n)}{\log \mathcal{R}_G(n)}$ and $\hat{\beta}_2 = \frac{\mathcal{R}_G(n)}{\mathcal{R}_A(n)} \frac{\Delta \mathcal{R}_A(n)}{\Delta \mathcal{R}_G(n)}$ as functions of n, where for any function $h(n)$, $\Delta h(n) = h(n+1) - h(n)$. The idea is simple. Let $f(x) = Cg(x)^\beta$ and $f, g \to \infty$ as $x \to \infty$. Then, $\frac{\log f}{\log g} \to \beta$ and also $\frac{\partial_x \log f}{\partial_x \log g} = \frac{g}{f}\frac{\partial_x f}{\partial_x g} \to \beta$, as $x \to \infty$. We thus numerically analyze their discrete analogues. Given $\mathcal{G}_n$, from (10) and (9) we compute $\tau = \max_j\{|\lambda_j(T)| : |\lambda_j(T)| < 1\}$. The optimal convergence rates are then given by $\tau_G^\star = \min_\alpha \tau_G$, and $\tau_A^\star = \min_{\{\gamma,\rho\}} \tau_A$ where $\boldsymbol{\rho} = \rho\mathbf{1}$. In Figure 2 we see that (5) holds for three very different types of graphs. Surprisingly, we get the same $\sqrt{\cdot}$ speedup for a barbell graph, which is known to not speedup random walks via lifting. We find similar behavior for several other graphs, but we omit these results due to lack of space.

4 Related work, conclusion, and an open problem

We state our Conjecture 1 for a relatively simple problem but, to the best of our knowledge, we cannot resolve it through existing literature on GD or ADMM. Our conjecture compares the *exact asymptotic optimal rates* of convergence of ADMM and GD. On the contrary, most literature on ADMM focus on upper bounding global convergence rate and, at best, optimize these upper bounds. Furthermore, to get linear convergence rates, strong convexity is usually assumed, which does not hold for our problem; see e.g. [5]. Most papers not requiring strong convexity, focus on the convergence rate of the objective and not on the convergence rate of the variables like us; see e.g. [6]. Some works consider a consensus problem which is related but different from ours. They use the objective $f(\boldsymbol{z}) = \sum_{i\in\mathcal{V}} \sum \|z_i - c_i\|^2$ subject to $z_i = z_j$ if $(i,j) \in \mathcal{E}$ where $c_i > 0$ are constants [7]. A branch of research considers $f(\boldsymbol{z}) = \sum_i f_i(\boldsymbol{z})$ and ADMM iterations that are agnostic to whether or not $f_i(\boldsymbol{z})$ depends on a subset of the components of $\boldsymbol{z}$; see e.g. [8]. This contrasts with our setting, where we can interpret ADMM's scheme as a message-passing algorithm where the messages between

agents i and j are only associated to the variables shared by f_i and f_j [3]. The works closest to help resolve our conjecture are [9, 10] on optimally tunning ADMM for quadratic problems, and also [11] that contains explicit rates of convergence. Nonetheless, their theorems' assumptions do not hold for our non-strongly-convex distributed problem. Very few works express the optimal convergence rate of ADMM as a function of the optimal convergence rate of GD. For example [2] does it, but assumes strong convexity. Finally, and most importantly, no prior work has connected GD and ADMM with Markov chain lifting, although some works on non-ADMM-based distributed averaging and gossip algorithms have made use of lifting to speed up convergence time [12, 13, 14].

References

[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] 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.

[3] N. Derbinsky, J. Bento, V. Elser, and J. Yedidia. An improved three-weight message passing algorithm. arXiv:1305.1961v1 [cs.AI], 2013.

[4] 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.

[5] 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.

[6] D. Davis and W. Yin. Convergence rate analysis of several splitting schemes. *arXiv preprint arXiv:1406.4834*, 2014.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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.

5 Supplementary material for "Markov chain lifting and distributed ADMM"

In the main part of the paper, we introduced the extended set $\hat{\mathcal{E}}$ which essentially duplicates all edges of the original graph, $|\hat{\mathcal{E}}| = 2|\mathcal{E}|$; see discussion before (7). This is the shortest route to state our results concisely, but it complicates the notation in the following proofs. Therefore, in this section we introduce the notion of a factor graph for problem (6).

The factor-graph $\bar{\mathcal{G}} = (\bar{\mathcal{F}}, \bar{\mathcal{V}}, \bar{\mathcal{E}})$ for problem (6) is a bipartite graph that summarizes how different variables are shared across different terms in the objective. This is illustrated in Figure 3.

The factor-graph $\bar{\mathcal{G}}$ has two sets of vertices, $\bar{\mathcal{F}}$ and $\bar{\mathcal{V}}$. The blue circles represent the nodes in $\bar{\mathcal{V}} = \mathcal{V}$, and the red 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. Before, we referred to this function by f_e but now we refer to it by f_a. With a slight 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 in $\boldsymbol{z}$. 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 (7). 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 $\bar{\mathcal{E}}$ has being $\hat{\mathcal{E}}$. 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 problem (7).

Let us introduce the neighbor set of a given node in $\bar{\mathcal{G}}$. For $a \in \bar{\mathcal{F}}$, the variables that f_a depends on are in the set $N_a = \{b \in \bar{\mathcal{V}} : (a, b) \in \bar{\mathcal{G}}\}$. Analogously, for $b \in \bar{\mathcal{V}}$, the functions that depend on z_b are in the set $N_b = \{a \in \bar{\mathcal{F}} : (a, b) \in \bar{\mathcal{G}}\}$. 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 $I_a = \{e \in \bar{\mathcal{E}} : e \text{ is incident on } a\}$. For $b \in \bar{\mathcal{V}}$ we define $I_b = \{e \in \bar{\mathcal{E}} : e \text{ is incident on } b\}$.

If we re-write problem (7) using the new notation, which indexes variables by the position they the take in $\bar{\mathcal{G}}$, the objective takes the form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top Q \boldsymbol{x} = \frac{1}{2}\sum_{a \in \bar{\mathcal{F}}} \boldsymbol{x}_a^\top Q^a \boldsymbol{x}_a \tag{16}$$

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 $Q^a = q^a \left(\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right)$ ($q^a > 0$), and $\boldsymbol{x}_a = (x_{ab}, x_{ac})^\top$ for $(a, b), (a, c) \in \bar{\mathcal{E}}$. 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 $S_{eb} = 1$ if and only if edge $e \in \bar{\mathcal{E}}$ is incident on $b \in \bar{\mathcal{V}}$ in the factor-graph. Notice that $S^\top S = \mathcal{D}$ is the degree matrix of the original graph $\mathcal{G}$.

Proof of Theorem 2. 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 (8). Substituting $\boldsymbol{x}^{t+1}$ and $\boldsymbol{m}^{t+1}$ into the other variables we obtain

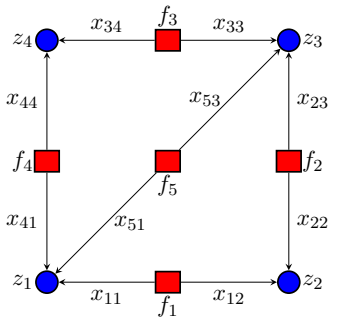
$$\begin{pmatrix} I & 0 & 0 \\ I & I & 0 \\ -I & I & I \end{pmatrix} \begin{pmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{u}^{t+1} \\ \boldsymbol{n}^{t+1} \end{pmatrix} = \begin{pmatrix} (1-\gamma)I & B & \gamma BA \\ (1-\gamma)I & I & \gamma A \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{u}^{t} \\ \boldsymbol{n}^{t} \end{pmatrix} \tag{17}$$

which can be easily inverted yielding

$$\boldsymbol{s}^{t+1} = (1-\gamma)\boldsymbol{s}^t + B\boldsymbol{u}^t + \gamma BA\boldsymbol{n}^t, \tag{18}$$

$$\boldsymbol{u}^{t+1} = B^\perp \boldsymbol{u}^t + \gamma B^\perp A \boldsymbol{n}^t, \tag{19}$$

$$\boldsymbol{n}^{t+1} = (1-\gamma)\boldsymbol{s}^t + (B - B^\perp)\boldsymbol{u}^t + \gamma(B - B^\perp)A\boldsymbol{n}^t. \tag{20}$$



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

$$f(\boldsymbol{x}) = q^1(x_{12} - x_{11})^2 + q^2(x_{23} - x_{22})^2 + q^3(x_{34} - x_{33})^2 + q^1(x_{41} - x_{44})^2 + q^1(x_{45} - x_{52})^2$$
with $x_{11} = x_{41}$, $x_{12} = x_{22}$, $x_{23} = x_{33}$, $x_{44} = x_{34}$

Figure 3: Example of a factor graph $\bar{\mathcal{G}}$ for problem (6) and (7), where $\mathcal{G}$ is the complete graph K_4 with one edge removed.

Note the following important relations:

$$B\boldsymbol{n}^t = \boldsymbol{s}^t, \qquad B^\perp \boldsymbol{n}^t = -\boldsymbol{u}^t, \tag{21}$$

$$B\boldsymbol{s}^t = \boldsymbol{s}^t, \qquad B^\perp \boldsymbol{s}^t = 0, \tag{22}$$

$$B^\perp \boldsymbol{u}^t = \boldsymbol{u}^t, \qquad B\boldsymbol{u}^t = 0. \tag{23}$$

The relations (22) follow simply by definition, $B\,\boldsymbol{s}^t = S(S^\top D_{\boldsymbol{\rho}} S)^{-1}(S^\top D_{\boldsymbol{\rho}} S)\boldsymbol{z}^t = \boldsymbol{s}^t$, which also implies $B^\perp \boldsymbol{s}^t = \mathbf{0}$. Since $BB^\perp = 0$, acting with B over (19) implies $B\boldsymbol{u}^t = \mathbf{0}$ for every t, and also $B^\perp \boldsymbol{u}^t = \boldsymbol{u}^t$. Thus we have shown (22) and (23). Now (21) follows simply by using these facts and the own definition $\boldsymbol{n}^t = \boldsymbol{s}^t - \boldsymbol{u}^t$. Finally, applying the relations (21) on (20) we obtain $\boldsymbol{n}^{t+1} = T_A \boldsymbol{n}^t$ where $T_A = I - \gamma(A + B - 2BA)$. □

Proof of Theorem 3. 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 $b, c \in \bar{\mathcal{V}}$, $(S^\top Q^+ S)_{bc} = \sum_{e \in \bar{\mathcal{E}}} S_{eb} Q^+_{ee} S_{ec} = \delta_{bc} \sum_{e \in I_b} Q^+_{ee}$ where δ is the Kronecker delta. By a similar argument, $S^\top Q^- S$ is off-diagonal. Hence, if $b, c \in \bar{\mathcal{V}}$ and $b \neq c$,

$$(T_G)_{bc} = -\alpha \sum_{e \in I_b} \sum_{e' \in I_c} Q^-_{ee'} \geq 0. \tag{24}$$

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 (24), we just need to consider the diagonal elements of $T_G - D_G$. Thus we require that for every $b \in \bar{\mathcal{V}}$,

$$1 - \alpha \sum_{e \in I_b} Q_{ee} + (D_G)_{bb} \geq 0. \tag{25}$$

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 $\alpha \leq (1 + D_{G,\min})/Q_{\max}$.

Now 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. □

Proof of Lemma 4. 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

$$(B\boldsymbol{v})_{ab} = \frac{\sum_{c \in N_b} \rho_{cb} v_{cb}}{\sum_{c \in N_b} \rho_{cb}}. \tag{26}$$

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. □

Proof of Theorem 5. Due to the block diagonal structure of Q it is possible write A explicitly as

$$A = I - FQ, \tag{27}$$

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

$$F^a = \frac{q^a}{\rho_{ab}\rho_{ac} + q^a(\rho_{ab} + \rho_{ac})} \begin{pmatrix} \rho_{ac} & 0 \\ 0 & \rho_{ab} \end{pmatrix}, \tag{28}$$

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,

$$S^\top D_{\boldsymbol{v}_A} M_A S = S^\top D_{\boldsymbol{\rho}}(I - D_A)S - \gamma S^\top D_{\boldsymbol{\rho}} FQS, \tag{29}$$

$$D_{\boldsymbol{v}_G} M_G = (I - D_G) - \alpha S^\top QS, \tag{30}$$

Equating the first term of (29) to the first term of (30), and also the second terms to each other, on using (27) we obtain

$$S^\top D_{\boldsymbol{\rho}}(I - D_A)S = I - D_G, \qquad \alpha = \frac{\gamma\, q^a \rho_{ab}\rho_{ac}}{\rho_{ab}\rho_{ac} + q^a(\rho_{ab} + \rho_{bc})}, \tag{31}$$

where the second equality above must hold for all $a \in \bar{\mathcal{F}}$ and $b, c \in N_a$. These give the second equality in (13) and relations (14). 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 (13). □

Proof of Theorem 6. 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 Figure 3. By direct inspection, one finds the following sub-matrix of T_A:

$$\begin{pmatrix} (T_A)_{21} & (T_A)_{24} \\ (T_A)_{31} & (T_A)_{34} \end{pmatrix} = \gamma \begin{pmatrix} \frac{\rho_{11}(\rho_{12}-\rho_{22})}{(\rho_{12}+\rho_{22})(\rho_{11}+\rho_{12}+\rho_{11}\rho_{12})} & \frac{2}{(\rho_{12}+\rho_{22})\left(1+\rho_{22}^{-1}+\rho_{23}^{-1}\right)} \\ \frac{2}{(\rho_{12}+\rho_{22})\left(1+\rho_{11}^{-1}+\rho_{12}^{-1}\right)} & \frac{-\rho_{23}(\rho_{12}-\rho_{22})}{(\rho_{12}+\rho_{22})(\rho_{22}+\rho_{23}+\rho_{22}\rho_{23})} \end{pmatrix}. \tag{32}$$

First notice that subtracting D_A from T_A does not affect this sub-matrix. Now recall that all components of $\boldsymbol{\rho}$ must be strictly positive. The elements $(T_A)_{21}$ and $(T_A)_{34}$ have opposite signs, so one of them is negative. Since $(T_A)_{24}$ and $(T_A)_{31}$ 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. □