Distributed optimization for multi-task learning via nuclear-norm approximation

David Mateos-Núñez [*] **Jorge Cortés** [*]

[*] *Department of Mechanical and Aerospace Engineering, University of California, San Diego, USA (e-mail: {dmateosn,cortes}@ucsd.edu).*

Abstract: We exploit a variational characterization of the nuclear norm to extend the framework of distributed convex optimization to machine learning problems that focus on the sparsity of the aggregate solution. We propose two distributed dynamics that can be used for multi-task feature learning and recommender systems in scenarios with more tasks or users than features. Our first dynamics tackles a convex minimization on local decision variables subject to agreement on a set of local auxiliary matrices. Our second dynamics employs a saddle-point reformulation through Fenchel conjugation of quadratic forms, avoiding the computation of the inverse of the local matrices. We show the correctness of both coordination algorithms using a general analytical framework developed in our previous work that combines distributed optimization and subgradient methods for saddle-point problems.

Keywords: Distributed optimization; multi-task learning; nuclear norm; matrix completion

1. INTRODUCTION

Motivated by applications in machine learning, this paper considers the design of distributed algorithmic solutions to problems that involve the joint minimization over a set of local variables of a sum of convex functions together with a regularizing term that favors sparsity patterns in the resulting aggregate solution. Our framework can be seen as a generalization of distributed convex optimization problems that employ the nuclear norm as a regularization technique to capture sparsity patterns in the data.

Literature review: The increasing body of literature on cooperative strategies for distributed convex optimization, see (Nedic and Ozdaglar, 2009; Boyd et al., 2011; Zhu and Martínez, 2012; Gharesifard and Cortés, 2014) and references therein, renders itself naturally to large-scale problems like distributed estimation in sensor networks or distributed label feedback in machine learning. Data is usually geographically distributed and often private, all of which favor cooperative fusion of local models to exploit the network decentralized resources such as automatic data collection, computation capabilities, and limited communication bandwidth. These problems consider a sum of convex functions subject to an agreement constraint in their arguments. The key observation here is that oftentimes a global decision vector, or global parameter, needs to be replaced by local parameter vectors that are coupled in a more flexible way than agreement to capture patterns in the decentralized data. In particular, the nuclear norm of the matrix composed of the local parameter vector across the network promotes low-rank solutions and as such is less rigid than the agreement constraint.

Mathematical models that use a low-rank matrix estimate are key in applications such as recommender systems through matrix completion (Candès and Recht, 2009), dimension reduction in multivariate regression (Yuan and Lin, 2007), and multi-task feature learning (Ando and Zhang, 2005; Argyriou et al., 2006, 2008). The basic underlying structure is the same: an estimate of a matrix that is *assumed or postulated* to be low rank. While the rank function is nonconvex, it turns out that the nuclear norm, defined as the one norm of the vector of singular values, is the convex surrogate of the rank function (Fazel, 2002). When used as a regularization in optimization problems, the nuclear norm promotes a low-rank solution and in some cases it even allows to recover the exact low-rank solution (Candès and Tao, 2010; Recht et al., 2010). The applications of nuclear norm regularization described above have inspired research in parallel computation following the model of stochastic gradient descent (Recht and Ré, 2013), but these developments emphasize the parallel aspect alone, rather than other aspects such as geographically distributed data, communication bandwidth, and privacy. Other strategies to address the problem that focus neither on the parallel aspect, nor in the distributed aspect, but instead try to overcome the nonsmooth nature of the nuclear norm, use techniques such as approximate singular value decompositions (Woolfe et al., 2008; Witten and Candès, 2015); coordinate descent and subspace selection (Dudík et al., 2012; Hsieh and Olsen, 2014); and successive over-relaxation (Wen et al., 2012), which is again related to coordinate descent. Finally, the technical analysis here builds on our recent work (Mateos-Núñez and Cortés, 2015) which develops a general analytical framework combining distributed optimization and subgradient methods for saddle-point problems.

Statement of contributions: We motivate the nuclear norm regularization in two problems that can benefit from distributed strategies: multi-task feature learning and matrix completion. Then we introduce two distributed formulations of the resulting optimization problems: a separable convex minimization, and a separable saddle-point problem, and we make the presentation systematic as to the automatic derivation of distributed coordination algorithms. After introducing each formulation, we show the existence of critical points that solve the original problem and also present the corresponding distributed subgradient dynamics. To the best of our knowledge, the subgradient

[⋆] The authors are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, USA, {dmateosn,cortes}@ucsd.edu.

saddle-point method proposed in the second case is a novel coordination algorithm even in its centralized version and we argue its advantages and general application to each of the motivational problems. For both families of distributed strategies, we show the convergence guarantees using the results in (Mateos-Núñez and Cortés, 2015). In our conclusions, we describe how our systematic treatment of the nuclear norm in distributed optimization opens the way to the design of additional novel strategies. The convergence results are illustrated in a simulation example of low-rank matrix completion. All the proofs are omitted for reasons of space and will be presented elsewhere.

2. PRELIMINARIES

We present some preliminaries on matrix norms, graph theory, and variational characterizations of the nuclear norm.

Notational conventions. We let $\mathbb{R}^n$ be the n-dimensional Euclidean space, $\mathrm{I}_n \in \mathbb{R}^{n\times n}$ the identity matrix in $\mathbb{R}^n$, and e_i the ith column of I_n. Given a vector $v \in \mathbb{R}^n$, we denote its one-norm by $\|v\|_1 = \sum_{i=1}^n |v_i|$ and its Euclidean norm (or two-norm) by $\|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$. Given a matrix $A \in \mathbb{R}^{n\times m}$, we denote its $L_{2,1}$-norm by $\|A\|_{2,1} := \|(\|a_1\|_2, \dots, \|a_m\|_2)\|_1$, which is the one-norm of the vector of two-norms of the columns of A. We denote the nuclear norm (or trace norm) by $\|A\|_* = \operatorname{trace}(\sqrt{A^\top A})$. This coincides with the sum of the singular values of A, $\|A\|_* = \sum_{i=1}^{\min\{n,m\}} \sigma_i$. We denote the Frobenius norm by $\|A\|_{\mathcal{F}} = \sqrt{\operatorname{trace}(A^\top A)} = \sqrt{\operatorname{trace}(AA^\top)} = \sqrt{\sum_{i=1}^{\min\{n,m\}} \sigma_i^2}$. Note that for any $A \in \mathbb{R}^{m\times n}$ with rank r, the nuclear norm and the Frobenius norm are related by

$$\|A\|_* \le \sqrt{r}\|A\|_{\mathcal{F}} \le \sqrt{\min\{n,m\}}\|A\|_{\mathcal{F}}. \tag{1}$$

We denote by $A^\dagger$ the Moore-Penrose pseudoinverse of A and by $\mathcal{C}(A)$ its column space, i.e., the vector space generated by the columns of A. The sets $\mathbb{S}^d, \mathbb{S}^d_{\succeq 0}, \mathbb{O}^d \subseteq \mathbb{R}^{d\times d}$ represent, respectively the symmetric, positive semidefinite, and orthogonal matrices. The following sets play a central role in our optimization problems. For any $c, r \in \mathbb{R}_{>0}$, let

$$\mathfrak{D}(c,r) := \{D \in \mathbb{S}^d_{\succeq 0} \,:\, D \succeq c\mathrm{I},\ \|D\|_{\mathcal{F}} \le r\}, \tag{2a}$$

$$\varDelta(c) := \{D \in \mathbb{S}^d_{\succeq 0} \,:\, D \succeq c\mathrm{I},\ \operatorname{trace}(D) \le 1\}. \tag{2b}$$

We refer to these sets as *reduced ice-cream* and *reduced spectraplex*, resp., based on the fact that they correspond to the intersection of the *reduced* cone $\{D \in \mathbb{S}^d : D \succeq c\mathrm{I}_d\} \subseteq \mathbb{S}^d_{\succeq 0}$ with the ball given by the Frobenius norm and with the trace constraint, resp. Given a closed convex set $\mathcal{C}$, we define the orthogonal projection onto $\mathcal{C}$ by

$$\mathcal{P}_{\mathcal{C}}(x) \in \arg\min_{x' \in S} \|x - x'\|_2.$$

A vector $\xi_x \in \mathbb{R}^n$ is a subgradient of a convex function $f : \mathcal{C} \to \mathbb{R}$ at $x \in \mathcal{C}$ if $f(y) - f(x) \ge \xi_x^\top (y - x)$, for all $y \in \mathcal{C}$. We denote by $\partial f(x)$ the set of all such subgradients.

Graph theory. We review basic notions from graph theory following the exposition in (Bullo et al., 2009). A (weighted) digraph $\mathcal{G} := (\mathcal{I}, \mathcal{E}, \mathsf{A})$ is a triplet where $\mathcal{I} := \{1, \dots, N\}$ is the vertex set, $\mathcal{E} \subseteq \mathcal{I} \times \mathcal{I}$ is the edge set, and $\mathsf{A} \in \mathbb{R}^{N\times N}_{\ge 0}$ is the weighted adjacency matrix with the property that $\mathsf{a}_{ij} := A_{ij} > 0$ if and only if $(i,j) \in \mathcal{E}$. Given $\mathcal{G}_1 = (\mathcal{I}, \mathcal{E}_1, \mathsf{A}_1)$ and $\mathcal{G}_2 = (\mathcal{I}, \mathcal{E}_2, \mathsf{A}_2)$, their union is the digraph $\mathcal{G}_1 \cup \mathcal{G}_2 = (\mathcal{I}, \mathcal{E}_1 \cup \mathcal{E}_2, \mathsf{A}_1 + \mathsf{A}_2)$. A path is an ordered sequence of vertices such that any pair of vertices appearing consecutively is an edge. A digraph is strongly connected if there is a path between any pair of distinct vertices. A sequence of digraphs $\{\mathcal{G}_t := (\mathcal{I}, \mathcal{E}_t, \mathsf{A}_t)\}_{t\ge 1}$ is δ-nondegenerate, for $\delta \in \mathbb{R}_{>0}$, if the weights are uniformly bounded away from zero by δ whenever positive, i.e., for each $t \in \mathbb{Z}_{\ge 1}$, $\mathsf{a}_{ij,t} := (\mathsf{A}_t)_{ij} > \delta$ whenever $\mathsf{a}_{ij,t} > 0$. A sequence $\{\mathcal{G}_t\}_{t\ge 1}$ is B-jointly connected, for $B \in \mathbb{Z}_{\ge 1}$, if for each $k \in \mathbb{Z}_{\ge 1}$, the digraph $\mathcal{G}_{kB} \cup \dots \cup \mathcal{G}_{(k+1)B-1}$ is strongly connected. The weighted out-degree and in-degree of $i \in \mathcal{I}$ are, respectively, $d_{\text{out}}(i) := \sum_{j=1}^N \mathsf{a}_{ij}$ and $d_{\text{in}}(i) := \sum_{j=1}^N \mathsf{a}_{ji}$. A digraph is weight-balanced if $d_{\text{out}}(i) = d_{\text{in}}(i)$ for all $i \in \mathcal{I}$.

Variational characterizations of the nuclear norm. The following characterizations of the nuclear norm play a key role in our forthcoming distributed formulations,

$$2\|W\|_* = \min_{\substack{D \in \mathbb{S}^d_{\succeq 0} \\ \mathcal{C}(W) \subseteq \mathcal{C}(D)}} \operatorname{trace}\left(D^\dagger WW^\top\right) + \operatorname{trace}(D), \tag{3a}$$

$$\|W\|_*^2 = \min_{\substack{D \in \mathbb{S}^d_{\succeq 0},\, \operatorname{trace}(D) \le 1 \\ \mathcal{C}(W) \subseteq \mathcal{C}(D)}} \operatorname{trace}\left(D^\dagger WW^\top\right). \tag{3b}$$

Defining $C := WW^\top$, the minimizers are, respectively,

$$D_1^* := \sqrt{C} \quad \text{and} \quad D_2^* := \frac{\sqrt{C}}{\operatorname{trace}(\sqrt{C})}. \tag{4}$$

A proof sketch of the latter can be found in (Argyriou et al., 2006, Thm 4.1). A different proof, valid when C is positive definite, can also be found in (Argyriou et al., 2008, Appendix A). Adding the penalty $\epsilon \operatorname{trace}(D^\dagger)$ in either minimization, and factoring out $D^\dagger$, gives $C_\epsilon = WW^\top + \epsilon \mathrm{I}_d$ in the formula for the optimizers (4). The optimal values then change according to

$$\begin{aligned} \operatorname{trace}\left(\sqrt{WW^\top + \epsilon \mathrm{I}_d}\right) &= \operatorname{trace}\left(\sqrt{[W|\sqrt{\epsilon}\mathrm{I}_d][W|\sqrt{\epsilon}\mathrm{I}_d]^\top}\right) \\ &= \|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*, \end{aligned}$$

which is the nuclear norm of the block matrix comprised of W and $\sqrt{\epsilon}\mathrm{I}_d$. Also, for any $W \in \mathbb{R}^{d\times N}$, one has

$$\|W\|_* = \min_{U \in \mathbb{O}^d} \|W^\top U\|_{2,1}. \tag{5}$$

This result can be found in the proof of (Argyriou et al., 2006, Thm 4.1). (This reference uses the notation $\|\cdot\|_{2,1}$ interchanging columns and rows.)

3. OPTIMIZATION WITH NUCLEAR NORM REGULARIZATION

We are interested in developing distributed coordination algorithms to solve the optimization problem

$$\min_{\substack{w_i \in \mathcal{W}, \\ i \in \{1,\dots,N\}}} \sum_{i=1}^N f_i(w_i) + \gamma \Omega(W), \tag{6}$$

where $\mathcal{W} \subseteq \mathbb{R}^d$ is a closed convex set; the matrix $W \in \mathbb{R}^{d\times N}$ aggregates the vectors $\{w_i\}_{i=1}^N$ as columns, i.e., $W := [w_1| \dots |w_N]$; each function $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex; $\gamma \in \mathbb{R}_{>0}$ is a design parameter; and $\Omega : \mathbb{R}^{d\times N} \to$ is a joint regularizer to promote solutions with low rank or other sparsity patterns. We next motivate the distributed optimization problem with nuclear-norm regularization.

3.1 Multi-task feature learning

In data-driven optimization problems each function f_i often codifies the loss incurred by the vector of weighting parameters w_i with respect to a set of n_i data points $\{p_j, y_j\}_{j=1}^{n_i}$. As such, this loss can be called *residual* or *margin*, depending on whether we are considering regression or classification problems. The work (Argyriou et al., 2008) exploits the relation (5) as follows. For a given $W \in \mathbb{R}^{d\times N}$, the following regularizer is used,

$$\Omega(W) = \min_{\substack{U\in\mathbb{O}^d,\, A\in\mathbb{R}^{d\times N}\\ W=UA}} \|A^\top\|_{2,\,1}$$
$$= \min_{U\in\mathbb{O}^d} \|W^\top U\|_{2,\,1} = \|W\|_*.$$

This minimization promotes a *dictionary* matrix U of orthonormal columns such that the columns of W are sparse linear combinations of them. The latter is achieved through $\|A^\top\|_{2,\,1}$, which 'favors' rows of small size because the one-norm is the convex surrogate of the zero-norm, or number of nonzero elements. This offers an interesting perspective on minimization problems that are convex on the *product* UA, with $U \in \mathbb{O}^d$, and have a penalty term $\|A^\top\|_{2,\,1}$. As pointed by Argyriou et al. (2008), the above characterization enables a convex reformulation on the matrix variable $W = UA$.

3.2 Matrix completion for recommender systems

The estimation of a low-rank matrix from a set of entries, or matrix completion, see, e.g., (Mazumder et al., 2010), also fits naturally in the framework of (6) with nuclear-norm regularization. This is because the nuclear norm is the convex surrogate of the rank function (Fazel, 2002). Let $Z \in \mathbb{R}^{d\times N}$ be a low-rank matrix of unknown rank for which only a few entries per column are known. The goal is then to determine a matrix W that minimizes the Frobenius norm across the revealed entries while keeping small the nuclear norm,

$$\min_{\substack{w_i\in\mathcal{W},\\ i\in\{1,\dots,N\}}} \sum_{i=1}^{N}\sum_{j\in\Upsilon_i} (W_{ji} - Z_{ji})^2 + \gamma\|W\|_* \tag{7}$$

where, for each $i \in \{1, \dots, N\}$,

$$\Upsilon_i := \{j \in \{1, \dots, d\} \,:\, Z_{ji} \text{ is a revealed entry of } Z\}.$$

3.3 A case for distributed optimization

The optimization problem (6) can be formulated as a convex and separable minimization when the joint regularizer is $\|\cdot\|_*$ or $\|\cdot\|_*^2$ using the characterizations (3a) or (3b). Assuming that a minimum exists, we can write

$$\min_{W\in\mathbb{R}^{d\times N}} \sum_{i=1}^{N} f_i(w_i) + \gamma\|W\|_*^2$$
$$= \min_{\substack{W\in\mathbb{R}^{d\times N}\\ D\in\mathbb{S}^d_{\succeq 0},\, \mathrm{trace}(D)\le 1\\ w_i\in\mathcal{C}(D),\forall i}} \sum_{i=1}^{N} f_i(w_i) + \gamma\sum_{i=1}^{N} w_i^\top D^\dagger w_i\,.$$
$$= \min_{\substack{w_i\in\mathcal{W},\forall i\\ D_i\in\mathbb{S}^d_{\succeq 0},\, \mathrm{trace}(D_i)\le 1,\,\forall i\\ w_i\in\mathcal{C}(D_i),\,\forall i\\ D_i=D_j,\,\forall i,j}} \sum_{i=1}^{N} f_i(w_i) + \gamma\sum_{i=1}^{N} w_i^\top D_i^\dagger w_i\,, \tag{8}$$

and similarly for $\Omega(W) = 2\|W\|_*$ replacing the constraint $\mathrm{trace}(D) \le 1$ by the penalty functions $\gamma\sum_{i=1}^{N}\frac{1}{N}\,\mathrm{trace}(D_i)$. When $d \ll N$, it is reasonable to design distributed strategies that use local gradient descent and consensus to solve this problem because the objective can be split across a network of agents, and the only coupling constraint is the agreement on the matrix arguments, $D_i = D_j$ for each i, j, whose dimensions do not grow with the network size. The condition $d \ll N$ in multi-task feature learning implies that there are far less features than tasks or users (for instance, there are less diseases or symptoms than people). The same observation applies to matrix completion in collaborative filtering where the rows represent features and the columns represent users.

However, the design of distributed strategies to solve (8) raises the following challenges,

(i) The constraint set $\{w \in \mathbb{R}^d, D \in \mathbb{S}^d_{\succeq 0} \,:\, w \in \mathcal{C}(D)\}$ is convex but not closed, which is a difficulty when designing a projection among the local variables. Note that for any fixed matrix D_i, one could project w_i onto $\mathcal{C}(D_i)$ by computing $D_i D_i^\dagger w$, but this projection is state-dependent.

(ii) The computation of $D_i^\dagger$ is a concern because D_i might be rank deficient and the pseudoinverse might be *discontinuous* when the rank of D_i changes.

We avoid these difficulties by enforcing the solution to be within a margin of the boundary of the positive semidefinite cone. This is achieved by considering an approximate regularization that we introduce in Section 4.1. Our first dynamics solves the *nuclear-norm regularization as a separable minimization with agreement constraint.* Even with (ii) addressed, an additional challenge involves the efficient computation of the inverse:

- Iterative algorithms involving the computation of D^{-1} are computationally expensive and potentially lead to numerical instabilities.

We eliminate the necessity of computing D^{-1} altogether in Section 4.2 by transforming the convex minimization into a saddle-point problem. This transformation is general and does not require the approximate treatment of the nuclear norm regularization in Section 4.1. Our second dynamics solves the *nuclear-norm regularization as a separable min-max problem with agreement constraint.*

4. DISTRIBUTED COORDINATION ALGORITHMS

Here we address the three challenges outlined in Section 3 to solve the optimization problem (8). In the forthcoming discussion, we present two reformulations of this problem and two distributed coordination algorithms to solve them.

4.1 Nuclear norm approximate regularization

In relation to the first two challenges outlined above, note that the optimal values D_1^* and D_2^* in (4) for the variational characterizations of $\|\cdot\|_*$ and $\|\cdot\|_*^2$ are in general positive semidefinite. To enforce these optimal values to be in the interior of the positive semidefinite cone, following the technique in (Argyriou et al., 2008, Sec. 4), we consider an approximate problem by introducing in (8) the barrier function $\epsilon\,\mathrm{trace}(D^\dagger)$ for some $\epsilon \in \mathbb{R}_{>0}$. We next justify how the optimizer of the approximate problem, which depends on ϵ, is farther than some *margin* from the boundary of $\mathbb{S}^d_{\succ 0}$ (in turn, this fact allows to insert in our optimization problem a dummy constraint of the form $D \succeq c\mathrm{I}$, where c is what we refer to as the margin). For $\Omega_\epsilon(W) = 2\|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*$, this is easy to see because, in view of (4),

$$D_{1,\epsilon}^* := \sqrt{WW^\top + \epsilon\mathrm{I}_d} \succeq \sqrt{\epsilon}\mathrm{I}_d.$$

For $\Omega_\epsilon(W) = \|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*^2$, we need more care and we offer next a result using the notation for the reduced spectraplex defined in Section 2.

Lemma 4.1. **(Dummy constraint for ϵ-approximate regularization under $\Omega(W) = \|W\|_*^2$):** Let $W \in \mathbb{R}^{d\times N}$ be any matrix whose columns have two-norm bounded by r_w. Then

$$D_{2,\epsilon}^* := \frac{\sqrt{WW^\top + \epsilon \mathrm{I}_d}}{\mathrm{trace}(\sqrt{WW^\top + \epsilon \mathrm{I}_d})} \tag{9}$$

is the optimizer of both

$$\min_{\substack{D\in\mathbb{S}^d_{\succeq 0},\, \mathrm{trace}(D)\leq 1,\\ \mathcal{C}(W)\subseteq\mathcal{C}(D)}} \mathrm{trace}\left(D^\dagger(WW^\top + \epsilon\,\mathrm{I})\right) \tag{10}$$

and

$$\min_{D\in\Delta(c_\epsilon)} \mathrm{trace}\left(D^\dagger(WW^\top + \epsilon\,\mathrm{I})\right)$$

(attaining the optimal value $\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*^2$), where the margin c_ϵ of the reduced spectraplex $\Delta(c_\epsilon)$ is

$$c_\epsilon := \frac{\sqrt{\epsilon}}{\sqrt{d}\sqrt{Nr_w^2 + \epsilon\, d}}\,. \tag{11}$$

Furthermore, c_ϵ in (11) satisfies $c_\epsilon \leq 1/d$ for any $\epsilon, r_w \in \mathbb{R}_{>0}$. Hence, $\Delta(c_\epsilon)$ is nonempty for any $\epsilon, r_w \in \mathbb{R}_{>0}$.

As a result, when we add the barrier terms $\sum_{i=1}^N \frac{\epsilon}{N}\,\mathrm{trace}(D_i^\dagger)$ to the optimization in (8), the constraints $D_i \in \mathbb{S}^d_{\succeq 0}$ and $w_i \in \mathcal{C}(D_i)$ can be replaced by $D_i \succeq c_\epsilon \mathrm{I}_d$. Hence, the variational characterization of $\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*^2$ can be written over the compact domain $\Delta(c_\epsilon)$. Alternatively, in the case of $2\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*$, we saw above that we can use the constraint $D_i \succeq \sqrt{\epsilon}\mathrm{I}_d$ to achieve the same effect. However, because the trace constraint is now absent, we construct a compact domain containing the optimal value $D_{1,\epsilon}^*$ by introducing one more dummy constraint $\|D_i\|_{\mathcal{F}} \leq r_\epsilon$, with

$$r_\epsilon := \sqrt{N}r_w + \sqrt{\epsilon d}. \tag{12}$$

This, together with the constraint $D_i \succeq \sqrt{\epsilon}\mathrm{I}_d$, yields the compact domain given by the reduced ice-cream $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$. The derivation is similar to the proof of Lemma 4.1; here we compute an upper bound as opposed to a lower bound. In both cases, we use the fact that the columns of W are contained in the ball $\bar{\mathcal{B}}(0, r_w) \subseteq \mathbb{R}^d$.

The following results summarizes our discussion above.

Corollary 4.2. **(Separable minimization with agreement constraint):** Let $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$ and define c_ϵ as in (11). Then

$$\min_{W\in\mathbb{R}^{d\times N}} \sum_{i=1}^N f_i(w_i) + \gamma\Omega_\epsilon(W), \tag{13}$$

with $\Omega_\epsilon(W) = \|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*^2$ is equal to

$$\min_{\substack{w_i\in\mathcal{W},\forall i,\\ D_i\in\Delta(c_\epsilon),\,\forall i,\\ D_i=D_j,\,\forall i,j}} \sum_{i=1}^N f_i(w_i) + \gamma\sum_{i=1}^N \left(w_i^\top D_i^{-1} w_i + \tfrac{\epsilon}{N}\,\mathrm{trace}(D_i^{-1})\right). \tag{14}$$

The analogous result is valid for $\Omega_\epsilon(W) = 2\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*$ replacing $\Delta(c_\epsilon)$ by $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$ and including the penalty functions $\gamma\sum_{i=1}^N \frac{1}{N}\,\mathrm{trace}(D_i)$.

In both cases of Corollary 4.2, Weierstrass' Theorem guarantees that the minimum is reached since we are minimizing a continuous function over a compact set. This leads to our first candidate dynamics.

Distributed subgradient dynamics for nuclear optimization. Our first coordination algorithm for the distributed optimization with nuclear norm (13) is a subgradient algorithm with proportional feedback on the disagreement on the matrix variables:

$$\begin{aligned}
\hat{w}_i(k+1) &= w_i(k) - \eta_k\big(g_i(k) + 2\gamma D_i(k)^{-1} w_i(k)\big),\\
\hat{D}_i(k+1) &= D_i(k) - \eta_k\gamma\Big(- D_i^{-1}(k) w_i(k) w_i(k)^\top D_i^{-1}(k)\\
&\qquad + \tfrac{\alpha}{N}\mathrm{I}_d - \tfrac{\epsilon}{N} D_i^{-2}(k)\Big) + \sigma\sum_{j=1}^N \mathsf{a}_{ij,t}(D_j(k) - D_i(k)),\\
w_i(k+1) &= \mathcal{P}_{\mathcal{W}}(\hat{w}_i(k+1)),\\
D_i(k+1) &= \mathcal{P}_{\mathcal{D}}(\hat{D}_i(k+1)),
\end{aligned} \tag{15}$$

where $g_i(k) \in \partial f_i(w_i(k))$, for each $i \in \{1, \ldots, N\}$, and $\mathcal{P}_{\mathcal{W}}(\cdot)$ and $\mathcal{P}_{\mathcal{D}}(\cdot)$ denote the projections onto the compact convex sets $\mathcal{W}$ and $\mathcal{D}$. This notation allows us to consider both approximate regularizers: for the case $2\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*$, the trace acts as a penalty, i.e., $\alpha = 1$, and the domain is $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$; for the case $\|[\,W\,|\,\sqrt{\epsilon}\,\mathrm{I}_d\,]\|_*^2$, the trace acts as a constraint, i.e., $\alpha = 0$, and $\mathcal{D} = \Delta(c_\epsilon)$.

4.2 Separable saddle-point formulation

In the previous section we have written the optimization (13) with approximate nuclear norm regularization as a separable convex optimization with an agreement constraint on auxiliary local matrices. Here we derive an equivalent min-max problem that is also separable and has the advantage of enabling iterative distributed strategies that avoid the computation of the inverse of the local matrices. To achieve this aim, the next result expresses the quadratic forms $w^\top D^\dagger w$ and $\mathrm{trace}(D^\dagger) = \sum_{j=1}^d e_j^\top D^\dagger e_j$ as the maximum of concave functions in additional auxiliary variables. We write these expressions using Fenchel conjugacy of quadratic forms, and in doing this, we avoid the need to compute the pseudoinverse of D.

Proposition 4.3. **(Min-max formulation via Fenchel conjugacy):** For $i \in \{1, \ldots, N\}$ and $\alpha \in \mathbb{R}_{\geq 0}$, let $F_i : \mathcal{W}\times \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ be defined by

$$\begin{aligned}
F_i\left(w, D, x, Y\right) &:= f_i(w) + \gamma\,\mathrm{trace}\left(D(-xx^\top - \tfrac{\epsilon}{N}YY^\top)\right)\\
&\quad - 2\gamma w^\top x - 2\gamma\frac{\epsilon}{N}\,\mathrm{trace}(Y) + \frac{\alpha}{N}\,\mathrm{trace}(D).
\end{aligned} \tag{16}$$

Then, the following two optimizations are equivalent

$$\begin{aligned}
&\min_{\substack{D\in\mathbb{S}^d_{\succeq 0},\\ w\in\mathcal{W}\cap\mathcal{C}(D)}} f_i(w) + \gamma\big(w^\top D^\dagger w + \tfrac{\epsilon}{N}\,\mathrm{trace}(D^\dagger) + \tfrac{\alpha}{N}\,\mathrm{trace}(D)\big)\\
&= \min_{w\in\mathcal{W},\, D\in\mathbb{R}^{d\times d}} \ \sup_{x\in\mathbb{R}^d, Y\in\mathbb{R}^{d\times d}} F_i(w, D, x, Y).
\end{aligned} \tag{17}$$

Moreover, the minimization on the right does not change with the addition of the constraints $D \in \mathbb{S}^d_{\succeq 0}$ and $w \in \mathcal{C}(D)$ (which allows to replace the operator sup by max).

The function $w^\top D^\dagger w$ is jointly convex in the convex domain $\{w \in \mathcal{W},\, D \in \mathbb{S}^d_{\succeq 0} : w \in \mathcal{C}(D)\}$ because it is a point-wise maximum of linear functions indexed by x. (The function is also proper but not closed because the domain is not closed). The same considerations apply adding the constraint $\mathrm{trace}(D) \leq 1$. We are now ready to show the main equivalence between optimization problems.

Corollary 4.4. **(Separable min-max problem with agreement constraint):**. The optimization (13) with $\Omega_\epsilon(W) = \|[\, W \,|\, \sqrt{\epsilon}\,\mathrm{I}_d \,]\|_*^2$ is equivalent to

$$\min_{\substack{w_i \in \mathcal{W},\, D_i \in \mathbb{R}^{d\times d}, \\ \mathrm{trace}(D_i) \le 1,\, \forall i, \\ D_i = D_j\ \forall i,j}} \ \sup_{\substack{x_i \in \mathbb{R}^d,\, \forall i \\ Y_i \in \mathbb{R}^{d\times d},\, \forall i}} \ \sum_{i=1}^N F_i(w_i, D_i, x_i, Y_i)\,, \tag{18}$$

without the penalty on the trace in F_i (i.e., $\alpha = 0$) for each $i \in \{1, \dots, N\}$. As long as c_ϵ is given by (11) and $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$, the constraints $D_i \in \mathit{\Delta}(c_\epsilon)$ are not necessary, but including them allows to replace the operator sup by max. An analogous result holds for $\Omega_\epsilon(W) = 2\|[\, W \,|\, \sqrt{\epsilon}\,\mathrm{I}_d \,]\|_*$ when, instead of the trace constraints, one has the penalty terms $\sum_{i=1}^N \frac{1}{N}\,\mathrm{trace}(D_i)$ (i.e., $\alpha = 1$). In this case, as long as r_ϵ is given by (12) and $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$, the constraints $D_i \in \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$ are not necessary.

Next we state the existence of a saddle-point for the convex -concave formulation of the ϵ-approximate minimization. Define $F : \mathcal{W}^N \times \mathit{\Delta}(c_\epsilon) \times (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N \to \mathbb{R}$ as

$$F(\boldsymbol{w}, D, \boldsymbol{x}, \boldsymbol{Y}) := \sum_{i=1}^N F_i(w_i, D, x_i, Y_i), \tag{19}$$

where $\boldsymbol{w} := (w_1, \dots w_N)$, $\boldsymbol{x} := (x_1, \dots x_N)$, $\boldsymbol{Y} := (Y_1, \dots Y_N)$.

Proposition 4.5. **(Existence of saddle points):** For $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$ and $\mathcal{D}$ equal to either $\mathit{\Delta}(c_\epsilon)$ or $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, the set of saddle points of F on $\mathcal{W}^N \times \mathcal{D} \times (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N$ is nonempty and compact, and, as a consequence,

$$\begin{aligned} &\max_{x_i \in \mathbb{R}^d,\, Y_i \in \mathbb{R}^{d\times d},\, \forall i} \ \min_{w_i \in \mathcal{W},\, \forall i,\, D \in \mathit{\Delta}(c_\epsilon)} \sum_{i=1}^N F_i(w_i, D, x_i, Y_i) \\ = &\min_{w_i \in \mathcal{W},\, \forall i,\, D \in \mathit{\Delta}(c_\epsilon)} \ \max_{x_i \in \mathbb{R}^d,\, Y_i \in \mathbb{R}^{d\times d},\, \forall i} \sum_{i=1}^N F_i(w_i, D, x_i, Y_i). \end{aligned}$$

(The agreement constraints $D_i = D_j$ for all $i, j \in \{1, \dots, N\}$ are implicit because the existence of saddle-points is established within those agreement constraints.)

The above leads us to our second candidate dynamics.

Distributed saddle-point dynamics for nuclear optimization. Our second coordination algorithm for the distributed optimization with nuclear norm (13) is a saddle-point subgradient dynamics with proportional feedback on the disagreement of a subset of the variables:

$$\begin{aligned} w_i(k+1) &= \mathcal{P}_{\mathcal{W}}\big(w_i(k) - \eta_k\big(g_i(k) - 2\gamma x_i(k)\big)\big), \\ D_i(k+1) &= \mathcal{P}_{\mathcal{D}}\Big(D_i(k) - \eta_k\gamma\big(-x_i x_i^\top - \tfrac{\epsilon}{N} Y_i Y_i^\top + \tfrac{\alpha}{N}\mathrm{I}_d\big) \\ &\qquad + \sigma \sum_{j=1}^N \mathsf{a}_{ij,t}(D_j(k) - D_i(k))\Big), \\ x_i(k+1) &= x_i(k) + \eta_k\gamma\big(-2D_i x_i(k) - 2w_i(k)\big), \\ Y_i(k+1) &= Y_i(k) + \eta_k\gamma\big(-\frac{2\epsilon}{N} D_i(k) Y_i(k) - \frac{2\epsilon}{N}\mathrm{I}_d\big), \end{aligned} \tag{20}$$

where $g_i(k) \in \partial f_i(w_i(k))$, for each $i \in \{1, \dots, N\}$, and $\mathcal{P}_{\mathcal{W}}(\cdot)$ and $\mathcal{P}_{\mathcal{D}}(\cdot)$ denote the projections onto the compact convex sets $\mathcal{W}$ and $\mathcal{D}$. For the case of the regularizer $2\|[\, W \,|\, \sqrt{\epsilon}\,\mathrm{I}_d \,]\|_*$ we set $\alpha = 1$ and $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, and for the regularizer $\|[\, W \,|\, \sqrt{\epsilon}\,\mathrm{I}_d \,]\|_*^2$, we set $\alpha = 0$ and $\mathcal{D} = \mathit{\Delta}(c_\epsilon)$.

5. CONVERGENCE ANALYSIS

The convergence result of the distributed strategies (15) and (20) follows from the analysis framework developed in Mateos-Núñez and Cortés (2015), as we outline next.

Theorem 5.1. **(Convergence of the coordination algorithms** (15) **and** (20)**):** Let the convex compact set $\mathcal{W} \subseteq \mathbb{R}^d$ be contained in $\bar{\mathcal{B}}(0, r_w)$ and let the bounds c_ϵ and r_ϵ be defined as in (11) and (12). Assume that each dynamics evolves over a sequence $\{\mathcal{G}_t\}_{t\ge 1}$ of B-jointly connected, δ-nondegenerate, weight-balanced digraphs with uniformly bounded Laplacian eigenvalues. Let σ be as follows: for any $\tilde{\delta}' \in (0,1)$, let $\tilde{\delta} := \min\big\{\, \tilde{\delta}',\ (1-\tilde{\delta}')\frac{\delta}{d_{\max}} \,\big\}$, where $d_{\max} := \max\big\{\, d_{\mathrm{out,t}}(j)\ :\ j \in \mathcal{I},\ t \in \mathbb{Z}_{\ge 1} \,\big\}$, and choose

$$\sigma \in \Big[\, \frac{\tilde{\delta}}{\delta},\ \frac{1-\tilde{\delta}}{d_{\max}} \,\Big].$$

Assume also that the learning rates are taken according to the doubling trick: for $m = 0, 1, 2, \dots, \lceil \log_2 t \rceil$, fix $\eta_s = \frac{1}{\sqrt{2^m}}$ in each period of 2^m rounds $s = 2^m, \dots, 2^{m+1} - 1$. Then both the dynamics (15) and (20) converge to an optimizer of (13). The evaluation error with respect to any minimum of (14), or with respect to any saddle point of the convex-concave function (18), is proportional to $1/\sqrt{t}$.

6. SIMULATION EXAMPLE

Here we illustrate the performance of the distributed saddle-point algorithm (20) on a matrix completion problem, cf. Section 3.2. The matrix $Z \in \mathbb{R}^{8\times 20}$ has rank 2 and each agent is assigned a column. From each column, only 5 entries have been revealed, and with this partial information, and without knowledge about the rank of Z, the agents execute the coordination algorithm (20) to solve the optimization (7). In this application each local function $f_i(w_i) = \sum_{j \in \Upsilon_i} (W_{ji} - Z_{ji})^2$ is not strongly convex, but just convex, in line with the hypotheses of Theorem 5.1. Figure 1 illustrates the matrix fitting error, the evolution of the network cost function, and the disagreement of the local auxiliary matrices.

7. CONCLUSIONS

We have considered a class of optimization problems that involve the joint minimization over a set of local variables of a sum of convex functions together with a regularizing term that favors sparsity patterns in the resulting aggregate solution. Particular instances of these optimization problems include multi-task feature learning and matrix completion. We have exploited the separability property of a variational characterization of the nuclear norm to design two types of provably-correct distributed coordination algorithms. Our analysis relies on the body of work on distributed convex optimization and saddle-point dynamics. To the best of our knowledge, the proposed coordination algorithms are the first distributed dynamics for convex optimization with nuclear-norm regularization. Future work will include the use of Fenchel duality in place of Fenchel conjugacy, the treatment of other barrier functions like the logarithm of the determinant, and the extension to applications with chordal sparsity.

ACKNOWLEDGEMENTS

The first author thanks Ashish Cherukuri and Karan Sikka for stimulating discussions, and Mariza Suárez for her

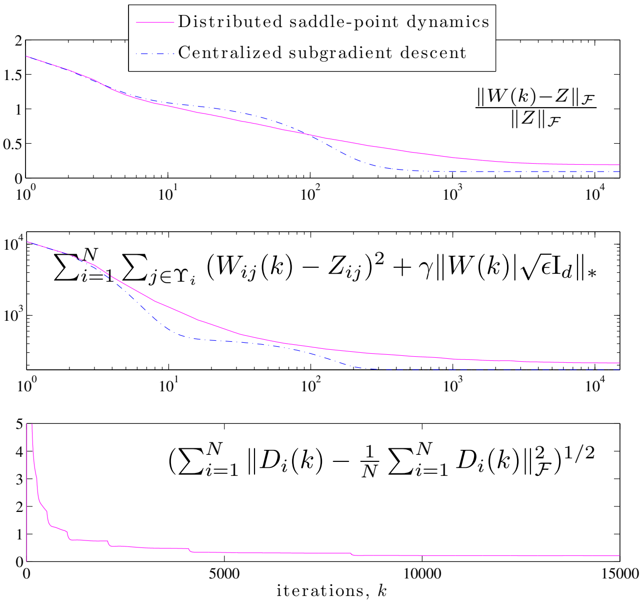

Fig. 1. **Matrix fitting error, evolution of network cost function, and disagreement of local matrices.** Here we represent the evolution of algorithm (20) (magenta solid line). The comparison is made with respect to a standard subgradient descent algorithm (blue dashed line) with constant gradient stepsize equal to 0.1. (The subgradient of the nuclear norm employed therein takes the form $U_r V_r^\top \in \partial\|W(k)\|_*$, where $U_r \Sigma_r V_r^\top$ is the reduced singular value decomposition of $W(k)$.) The optimization parameter weighting the nuclear norm is $\gamma = 2$, and the parameter of the approximate regularization is $\epsilon = 10^{-3}$. We use as constraint set $\mathcal{W} = \bar{\mathcal{B}}(0, r_w)$ with $r_w = 800$. In the distributed algorithm, the constraint set for the auxiliary matrices is $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, the consensus stepsize is $\sigma = 0.5$, and the communication topology is a ring connecting the 20 agents. Our algorithm is slower because it halves the learning rates (subgradient stepsizes) according to the doubling trick. This is necessary for asymptotic convergence in Theorem 5.1, in sharp contrast with standard (centralized) gradient descent that uses constant subgradient stepsize. The third plot shows the disagreement among the auxiliary matrices for our distributed algorithm. For decreasing learning rates, which is our case, the disagreement is guaranteed to converge to zero.

encouragement since the inception of these ideas. This work was partially supported by NSF Award CMMI-1300272.

REFERENCES

Ando, R.K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11), 1817–1853.

Argyriou, A., Evgeniou, T., and Pontil, M. (2006). Multi-task feature learning. In *Advances in Neural Information Processing Systems*, volume 19, 41–48.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3), 243–272.

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

Bullo, F., Cortés, J., and Martínez, S. (2009). *Distributed Control of Robotic Networks*. Applied Mathematics Series. Princeton University Press. Electronically available at http://coordinationbook.info.

Candès, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717–772.

Candès, E.J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2053–2080.

Dudík, M., Harchaoui, Z., and Malick, J. (2012). Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, 327–336. JMLR Workshop and Conference Proceedings.

Fazel, M. (2002). *Matrix rank minimization with applications*. Ph.D. thesis, Stanford University.

Gharesifard, B. and Cortés, J. (2014). Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3), 781–786.

Hsieh, C.J. and Olsen, P.A. (2014). Nuclear norm minimization via active subspace selection. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32. JMLR Workshop and Conference Proceedings.

Mateos-Núñez, D. and Cortés, J. (2015). Distributed subgradient methods for saddle-point problems. In *IEEE Conf. on Decision and Control*. Osaka, Japan. To appear.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11, 2287–2322.

Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1), 48–61.

Recht, B., Fazel, M., and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3), 471–501.

Recht, B. and Ré, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2), 201–226.

Wen, Z., Yin, W., and Zhang, Y. (2012). Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4), 333–361.

Witten, R. and Candès, E. (2015). Randomized algorithms for low-rank matrix factorizations: Sharp performance bounds. *Algorithmica*, 72(1), 264–281.

Woolfe, F., Liberty, E., Rokhlin, V., and Tygert, M. (2008). A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3), 335–366.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, 68(1), 49–57.

Zhu, M. and Martínez, S. (2012). On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, 57(1), 151–164.