Reweighted Nuclear Norm Minimization with Application to System Identification

Karthik Mohan and Maryam Fazel

***Abstract*—The matrix rank minimization problem consists of finding a matrix of minimum rank that satisfies given convex constraints. It is NP-hard in general and has applications in control, system identification, and machine learning. Reweighted trace minimization has been considered as an iterative heuristic for this problem. In this paper, we analyze the convergence of this iterative heuristic, showing that the difference between successive iterates tends to zero. Then, after reformulating the heuristic as *reweighted nuclear norm minimization*, we propose an efficient gradient-based implementation that takes advantage of the new formulation and opens the way to solving large-scale problems. We apply this algorithm to the problem of low-order system identification from input-output data. Numerical examples demonstrate that the reweighted nuclear norm minimization makes model order selection easier and results in lower order models compared to nuclear norm minimization without weights.**

I. Introduction

A. Background

The matrix rank minimization problem consists of finding a matrix of minimum rank that satisfies given convex constraints, i.e.,

$$\begin{array}{ll}\text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & X \in C,\end{array} \tag{1}$$

where $X \in \mathbb{R}^{m\times n}$ is the optimization variable and C is a convex set. When C is described by affine equality constraints, (1) is the matrix extension of the popular sparse signal recovery problem in *compressed sensing*. The rank minimization problem arises in a diverse set of fields, where notions of order, complexity, or dimension are expressed by means of the rank of an appropriate matrix. Applications include system identification, low-order controller design, collaborative filtering in machine learning, and Euclidean embedding problems (see [14] and references therein). Problem (1) is in general NP-hard. A common convex heuristic [7] replaces rank with the *nuclear norm* (also known as the Schatten 1-norm or trace norm) of the matrix, denoted by $\|X\|_* = \sum_i^r \sigma_i(X)$ where $\sigma_i(X)$ are the singular values and $r = \operatorname{rank}(X)$. The heuristic solves the convex problem

$$\begin{array}{ll}\text{minimize} & \|X\|_* \\ \text{subject to} & X \in C.\end{array} \tag{2}$$

This heuristic and its variants have lately received a lot of interest. One reason is the recent progress in both the theory and the algorithms for this heuristic. Several conditions have been studied that guarantee that the heuristic yields an exact solution when the constraints are linear equalities (e.g., [14]). Development of various classes of algorithms for this heuristic that exploit specific problem structure is also an active research area (e.g., [17]). Finally, new applications have initiated interest in special cases of the rank minimization problem, e.g., the low-rank matrix completion problem [4] arising in machine learning. We also mention that matrix rank and nuclear norm minimization have a natural connection to vector sparsity and ℓ_1-norm: the former reduces to the latter if the matrix variable is taken to be diagonal.

A variation on this basic heuristic that helps reduce the rank of the solution further, is to use a weighted objective (see [3], [6] for the vector version, and [8] for the matrix version). In this paper we study the *reweighted trace* heuristic, which is based on using a nonconvex surrogate function for the rank and solving the resulting problem locally via a sequence of convex problems. First, note that problem (1) can also be expressed in a positive semidefinite form [9]:

$$\begin{array}{ll}\text{minimize} & \operatorname{rank}(Y) + \operatorname{rank}(Z) \\ \text{subject to} & \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0,\ X \in C,\end{array} \tag{3}$$

where X, $Y \in \mathbb{R}^{m\times m}$, and $Z \in \mathbb{R}^{n\times n}$ are the optimization variables. Then, replacing rank with trace, we obtain a semidefinite programming problem that is equivalent to (2) [7]. The heuristic given in [8] replaces the rank of positive semidefinite matrices Y, Z by a surrogate function as follows:

$$\begin{array}{ll}\text{minimize} & \log\det(Y+\delta I) + \log\det(Z+\delta I) \\ \text{subject to} & \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0,\ X \in C,\end{array} \tag{4}$$

where $\delta > 0$ is a small regularization constant. Problem (4) can be solved locally by iterative linearization of the objective. The kth step of this algorithm solves the problem

$$\begin{array}{ll}\text{minimize} & \operatorname{Tr}\left(Y^k+\delta I\right)^{-1} Y + \operatorname{Tr}\left(Z^k+\delta I\right)^{-1} Z \\ \text{subject to} & \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0,\ X \in C,\end{array} \tag{5}$$

to get $X^{k+1}, Y^{k+1}, Z^{k+1}$. Throughout this paper, we refer to (5) as the *reweighted trace heuristic* (*RTH*). We often take Y^0 and Z^0 to be identity, thus the first iteration of the algorithm will simply minimize $\operatorname{Tr} Y + \operatorname{Tr} Z$.

B. Summary of results

We examine the convergence of the reweighted trace heuristic, showing that the difference between the successive iterates of the heuristic tends to zero.

Electrical Engineering Department, University of Washington, Seattle. Email: karna@u.washington.edu, mfazel@u.washington.edu.

Research funded in part by NSF CAREER grant ECCS-0847077.

We give an example of other concave functions that can be used as a surrogate for rank in (3), giving rise to other heuristics that have similar convergence properties.

RTH as given in (5) would require solving a semidefinite program (SDP) at each iteration. We reformulate the iterations in terms of the matrix nuclear norm, to which several first-order gradient-based methods could be applied efficiently. We apply the reweighted nuclear norm heuristic to a classic system identification problem, finding a low order system from input-output data. We use the gradient projection method to implement the heuristic efficiently. We give a numerical example (from the system identification database [11]) and show that the reweighted nuclear norm heuristic gives a clearer description of the model order and results in lower order models compared to a simple subspace method and also the un-weighted or nuclear norm minimization as in (2).

C. Related work

The papers [3], [6] study reweighted ℓ_1 minimization as a heuristic to find the sparsest vector in a convex set (a special case of the rank minimization problem). It is shown that this heuristic works better in practice than ℓ_1 minimization [3]. *RTH* was first proposed in [8]. While this paper shows that the objective function value converges, it does not discuss if the iterates themselves converge or if the difference between the successive iterates converge. That the difference between iterates goes to zero was shown for the reweighted ℓ_1 heuristic in [6], but the proof there does not extend to the matrix case.

The paper [17] gives an interior point method for nuclear norm minimization, and applies it for system identification as an alternative approach to subspace based identification methods (see, e.g., [12], [5]). It shows that nuclear norm minimization determines the lowest system order better than existing methods. The interior point implementation is more efficient than generic SDP solvers, however it does not scale as well as the first-order implementation we discuss here.

II. Convergence of the reweighted trace heuristic

Let $X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{m\times m}, Z \in \mathbb{R}^{n\times n}$. Define the function $g : \mathbb{R}^{m\times n} \times \mathbb{S}^m_+ \times \mathbb{S}^n_+ \to \mathbb{R}$ as $g(X,Y,Z) = \log\det(Y+\delta I) + \log\det(Z+\delta I)$. Note that $g(X,Y,Z)$ is a continuously differentiable function over its domain. The gradient of g is given by

$$\nabla g(X,Y,Z) = (0, (Y+\delta I)^{-1}, (Z+\delta I)^{-1}).$$

Let $\bar{W} = (W_1, W_2, W_3)$, $\bar{X} = (X,Y,Z)$ be two points in the domain of g . Define the inner product on the cross product space, $\mathbb{R}^{m\times n}\times\mathbb{S}^m_+\times\mathbb{S}^n_+$ as $\left\langle \bar{W}, \bar{X}\right\rangle_c = \langle W_1, X\rangle + \langle W_2, Y\rangle + \langle W_3, Z\rangle$, where $\langle X, Y\rangle$ is the standard inner product between two matrices. Since g is strictly concave on its domain, we have

$$g(\bar{X}) < g(\bar{W}) + \left\langle \nabla g(\bar{W}), \bar{X} - \bar{W}\right\rangle_c \;:=\; h(\bar{X}, \bar{W})$$

for all $\bar{W} \neq \bar{X} \in \bar{D}$, where

$$\bar{D} = \{(X,Y,Z) : \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0, X \in C\}.$$

Since $g(\bar{X}) < h(\bar{X}, \bar{W}),\ \forall\ \bar{W} \neq \bar{X} \in \bar{D}$ and $g(\bar{X}) = h(\bar{X}, \bar{X}),\ \forall \bar{X} \in \bar{D}$, the function h majorizes g. Let $\bar{X}^k$ denote (X^k, Y^k, Z^k). *RTH* is thus a *Majorization-Minimization (MM)* algorithm:

$$\begin{aligned}\bar{X}^{k+1} &= \arg\min_{\bar{X}\in\bar{D}} h(\bar{X}, \bar{X^k}) = \arg\min_{\bar{X}\in\bar{D}} \left\langle \nabla g(\bar{X}^k), \bar{X}\right\rangle_c \\ &= \arg\min_{\bar{X}\in\bar{D}} \mathrm{Tr}\,(Y^k+\delta I)^{-1} Y + \mathrm{Tr}\,(Z^k+\delta I)^{-1} Z \quad (6)\end{aligned}$$

Note that if $\bar{X}^k \neq \bar{X}^{k+1}$,

$$g(\bar{X}^{k+1}) < h(\bar{X}^{k+1}, \bar{X}^k) \leq h(\bar{X}^k, \bar{X}^k) = g(\bar{X}^k). \quad (7)$$

We first define a stationary point of a function before giving the theorem on convergence.

Definition II.1. *x is a stationary point of a continuously differentiable function f over a set D if $x \in \arg\min_{y\in D} \nabla f(x)^T y$.*

Theorem II.2. *Every convergent subsequence of the reweighted trace heuristic (5) converges to a stationary point of g over $\bar{D}$. Further the norm of the difference between succesive iterates tends to zero, $\|\bar{X}^{k+1} - \bar{X}^k\|_F \to 0$.*

Proof: Let $\bar{X}^1 = (X^1, Y^1, Z^1)$ be the solution to the nuclear norm minimization problem (2). The set $\{\bar{X} | g(\bar{X}) \leq g(\bar{X}^1)\}$ is bounded, because $\|Y\|_F, \|Z\|_F \to \infty$ implies $g(\bar{X}) \to \infty$. Also, if $\|X\|_F \to \infty$, then $\|Y\|_F \times \|Z\|_F \to \infty$ (since the block matrix is positive semidefinite), hence $g(\bar{X}) \to \infty$. Thus, $\{\bar{X} | g(\bar{X}) \leq g(\bar{X}^1)\}$ is a compact set. Therefore the iterates of *RTH* are bounded (since $g(\bar{X}^{k+1}) \leq g(\bar{X}^k),\ \forall k \geq 1$). Therefore, the sequence $\{\bar{X}^k\}$ has a convergent subsequence. Let $\{\bar{X}^{n_i}\}, i = 1, 2, \ldots$ denote this subsequence with a limit $\bar{X}^*$. Let $\bar{X}^{n_{i+1}} \to \bar{X}^a$. Assume that $\bar{X}^* \neq \bar{X}^a$. Now, (7) with the fact that g is continuously differentiable implies that

$$g(\bar{X}^*) = \lim g(\bar{X}^{n_i}) > \lim g(\bar{X}^{n_{i+1}}) = g(\bar{X}^a). \quad (8)$$

(7) also implies that $g(\bar{X}^{i+1}) \leq g(\bar{X}^i)\ \forall i$. Since g is bounded below, $\{g(\bar{X}^i)\}$ converges and

$$\lim g(\bar{X}^i) = \lim g(\bar{X}^{n_i}) = g(\bar{X}^*) = g(\bar{X}^{n_{i+1}}) = g(\bar{X}^a) \quad (9)$$

But (9) contradicts (8), hence $\bar{X}^* = \bar{X}^a$ and by definition, this implies that $\bar{X}^*$ is a stationary point. Now, assume that there exists a subsequence and a $\delta > 0$ such that $\|\bar{X}^{n_i} - \bar{X}^{n_{i+1}}\|_F > \delta,\ i = 1, 2, \ldots$. Let $\{\bar{X}^{n_i}\} \to \bar{X}^*$ and $\{\bar{X}^{n_{i+1}}\} \to X^a \neq \bar{X}^*$. Then, by a similar argument as above we arrive at a contradiction. Thus, $\bar{X}^* = \bar{X}^a$ and it holds that $\|\bar{X}^{k+1} - \bar{X}^k\|_F \to 0$. ■

A. Convergence through conditional gradient method

For a generic problem with a continuously differentiable objective function $f : \mathbb{R}^n \to \mathbb{R}$ and a convex, compact constraint set C, the *conditional gradient method* [1] yields the iterates $\bar{x}^k = \arg\min_{x\in C} \left\langle \nabla f(x^k), x \right\rangle$, $x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k)$. If we use a the so-called limited minimization rule to pick the step-size α^k, we get $\alpha^k = 1$ since the objective function f is concave. Thus, the heuristic (5) is the same as the conditional gradient method applied to the problem (4). It is known (e.g., [1]) that every cluster point of the conditional gradient iterates is a stationary point of f over the set C, and we obtain the first part of the convergence result in the previous theorem.

B. Other surrogate functions

We considered the concave surrogate function $\log\det(X)$ in (4). We can similarly use other concave surrogates such as $-\operatorname{Tr}(X^{-1})$ (see e.g. [2] for proof of concavity) and apply the conditional gradient method or the Majorization-minimization algorithm to obtain other heuristics for which the same convergence results hold. For example, using the surrogate function $-\operatorname{Tr}(X^{-1})$ yields the following heuristic:

$$\bar{X}^{k+1} = \arg\min_{\bar{X}\in\bar{D}} \operatorname{Tr}\left(Y^k + \delta I\right)^{-2} Y + \operatorname{Tr}\left(Z^k + \delta I\right)^{-2} Z \quad (10)$$

with $\bar{X}$ as defined in section 2. Comparing the performance of these heuristics is a direction for future work.

This section establishes the convergence of difference of successive iterates of *RTH* and shows that the limit point of every convergent subsequence is a stationary point. However, we can't say if the trace heuristic can achieve the global minimum of (4). We initialize the weighted iterations with the solution of nuclear norm minimization (2), thus *RTH* can be thought of as improving on the solution of the nuclear norm heuristic.

III. Reweighted Nuclear norm heuristic

Recall from (5) that in the kth iteration of the reweighted trace heuristic we solve the following problem:

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}\left(Y^k + \delta I\right)^{-1} Y + \operatorname{Tr}\left(Z^k + \delta I\right)^{-1} Z \\ \text{subject to} & \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0,\ X \in C. \end{array} \quad (11)$$

In this section, we reformulate this problem as a (reweighted) nuclear norm minimization problem. This reformulation is algorithmically beneficial: it allows us to exploit the properties of the nuclear norm and the problem structure to obtain an efficient first-order algorithm for solving the heuristic.

Let $W_1^k = \left(Y^k + \delta I\right)^{-\frac{1}{2}}$ and $W_2^k = \left(Z^k + \delta I\right)^{-\frac{1}{2}}$. Since W_1^k, W_2^k are positive definite for any feasible Y^k, Z^k and $\delta > 0$, the constraint in (11) is equivalent to

$$\left[\begin{array}{cc} W_1^k & 0 \\ 0 & W_2^k \end{array}\right] \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \left[\begin{array}{cc} W_2^k & 0 \\ 0 & W_2^k \end{array}\right] \geq 0.$$

Thus, problem (11) is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(\operatorname{Tr} W_1^k Y W_1^k + \operatorname{Tr} W_2^k Z W_2^k) \\ \text{subject to} & \left[\begin{array}{cc} W_1^k Y W_2^k & W_1^k X W_2^k \\ W_2^k X^T W_1^k & W_2^k Z W_2^k \end{array}\right] \geq 0 \\ & X \in C. \end{array} \quad (12)$$

Using the following characterization of the nuclear norm (see e.g. [7], [14]),

$$\begin{array}{rl} \|X\|_* = & \frac{1}{2}\min(\operatorname{Tr} Y + \operatorname{Tr} Z) \\ & \text{subject to } \left[\begin{array}{cc} Y & X \\ X^T & Z \end{array}\right] \geq 0, \end{array}$$

we can write problem (12) as

$$\begin{array}{ll} \text{minimize} & \|W_1^k X W_2^k\|_* \\ & X \in C, \end{array} \quad (13)$$

which is a (weighted) nuclear norm minimization. Once the optimal solution X^{k+1} is found, the weights W_1^{k+1} and W_2^{k+1} are updated as follows. Let $W_1^k X^{k+1} W_2^k = U\Sigma V^T$ be the reduced singular value decomposition of $W_1^k X^{k+1} W_2^k$, where $U \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ and $V \in \mathbb{R}^{n\times r}$. It can be checked [14] that the optimal Y^{k+1} and Z^{k+1} in (12) are given by

$$\begin{aligned} Y^{k+1} &= \left(W_1^k\right)^{-1} U\Sigma U^T \left(W_1^k\right)^{-1}, \\ Z^{k+1} &= \left(W_2^k\right)^{-1} V\Sigma V^T \left(W_2^k\right)^{-1}, \end{aligned} \quad (14)$$

so the weights can be updated as

$$\begin{aligned} W_1^{k+1} &= \left(Y^{k+1} + \delta I\right)^{-1/2}, \\ W_2^{k+1} &= \left(Z^{k+1} + \delta I\right)^{-1/2}. \end{aligned} \quad (15)$$

The update equations (14),(15) together with (13) describe the *reweighted nuclear norm heuristic*. If the set C in (13) is described by convex constraints $f_i(X) \leq 0,\ i = 1, \ldots, m$, we can write the problem in the regularized form

$$X^{k+1} = \arg\min \sum_i \lambda_i f_i(X) + \|W_1^k X W_2^k\|_* \quad (16)$$

with W_1^k, W_2^k defined above, and a suitable choice of λ_i.

We note that if in addition to $X \in C$ we have the constraint that X be positive semidefinite, then the *reweighted nuclear norm heuristic* reduces to,

$$X^{k+1} = \arg\min_{X\in C, X\succeq 0} \operatorname{Tr}\left(X^k + \delta I\right)^{-1} X. \quad (17)$$

In the next section, we apply this *regularized reweighted nuclear norm heuristic* (that we abbreviate as *RRNH*) to a system identification problem using an efficient first-order method.

IV. Efficient implementation of the RRNH for System identification

Consider the problem of identifying a discrete-time, linear time-invariant state-space model,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

given a set of inputs $u(t) \in \mathbb{R}^m$ and noisy measured outputs $y_{meas}(t) \in \mathbb{R}^p$, for $t = 0, 1, \ldots, N-1$. Here $x(t) \in \mathbb{R}^n$ is the state of the system at time t, and n is the order of the model. We would like to find the matrices A, B, C, D, the initial state $x(0)$, and the lowest possible order n that satisfy $y(t) \approx y_{meas}(t)$. Let $\hat{Y} = [y(0), \ldots, y(N-1)] \in \mathbb{R}^{p \times N}$, $Y_{meas} = [y_{meas}(0), \ldots, y_{meas}(N-1)] \in \mathbb{R}^{p \times N}$, $\hat{U} = [u(0), \ldots, u(N-1)] \in \mathbb{R}^{m \times N}$. Define the linear operator H_r as follows:

$$H_r(\hat{Y}) = \begin{bmatrix} y(0) & y(1) & \ldots & y(N-r-1) \\ y(1) & y(2) & \ldots & y(N-r) \\ \vdots & \vdots & & \vdots \\ y(r) & y(r+1) & \ldots & y(N-1) \end{bmatrix}, \quad (18)$$

which is a block-Hankel matrix, with $\hat{Y}$ defined as earlier. The adjoint operator, $H_r^*(W)$, is as below:

$$H_r^*(W) = H_r^* \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1,N-r} \\ w_{21} & w_{22} & \ldots & w_{2,N-r} \\ \vdots & \vdots & & \vdots \\ w_{r+1,1} & \ldots & \ldots & w_{r+1,N-r} \end{pmatrix}$$
$$= \begin{bmatrix} w_{11} & w_{21}+w_{12} & w_{31}+w_{22}+w_{13} & \ldots & w_{r+1,N-r} \end{bmatrix}$$

Define $X = [x(0), x(1), \ldots, x(N-r-1)]$, $U = H_r(\hat{U})$, $Y = H_r(\hat{Y})$ with $\hat{X}, \hat{U}$, and $\hat{Y}$ as defined earlier, and let

$$G = \begin{bmatrix} C^T & (CA)^T & \ldots & (CA^r)^T \end{bmatrix}^T,$$

$$F = \begin{bmatrix} D & 0 & 0 & \ldots & 0 \\ CB & D & 0 & \ldots & 0 \\ CAB & CB & D & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{r-1}B & CA^{r-2}B & CA^{r-3}B & \ldots & D \end{bmatrix}.$$

It is easy to see that $Y = GX + FU$, and thus $YU^\perp = GXU^\perp$, where $U^\perp \in \mathbb{R}^{N-r+1 \times q}$ is a set of orthonormal basis vectors for null-space of U. If X has a rank n and there is no rank cancellation in $XU^\perp$, one can find the system order from the rank of $YU^\perp$ (see, e.g., [17], [16] for more details). Liu et al [16], [17] propose a nuclear norm heuristic for minimizing the rank of $YU^\perp$ as

$$\text{minimize}_{\hat{Y}} \|H_r(\hat{Y})U^\perp\|_* + \frac{\lambda}{2}\|\hat{Y} - Y_{\text{meas}}\|_F^2, \quad (19)$$

with λ a positive parameter. We apply the *RRNH* to find a minimum order system giving the following iterative minimization:

$$\hat{Y}^{k+1} = \arg\min_{\hat{Y}} \frac{\lambda}{2}\|\hat{Y} - Y_{meas}\|_F^2 + \|W_1^k H_r(\hat{Y})U^\perp W_2^k\|_*, \quad (20)$$

where we define $f = \frac{1}{2}\|\hat{Y} - Y_{meas}\|_F^2$ in (16) and W_1^k, W_2^k are as in (13). Once we obtain an optimal $\hat{Y}$, we compute the rank of $H_r(\hat{Y})U^\perp$ by looking at its singular values. The thumbrule we use to obtain the rank is the number of singular values after which there is a sharp drop (differentiating the significant singular values from the non-significant ones). If we don't observe a sharp drop, the thumbrule used is to choose the rank of the matrix to be the number of singular values that are within 0.1 percent of the largest singular value (as in ([17]). Once we identify the rank of $YU^\perp$, the matrices A, B, C, D of the LTI state-space model can be estimated as detailed in section 5 of [17].

A. Problem Reformulation

To solve (20) efficiently, we reformulate it by making use of the structure of the regularized constraint in (19) and the fact that the nuclear norm is the dual of the spectral norm,

$$\|Y\|_* = \max_Z \{\langle Y, Z\rangle : \|Z\|_F \leq 1\}. \quad (21)$$

Using (21) the primal problem in (20) at the kth iteration can be formulated as (note that switching Z to $-Z$ does not change (21)):

$$\min_y \max_{Z: Z^T Z \leq I} \frac{\lambda}{2}\|\hat{Y} - Y_{meas}\|_F^2 - \langle Z, W_1^k H_r(\hat{Y})U^\perp W_2^k\rangle \quad (22)$$

The dual problem corresponding to (22) is obtained by interchanging the $\min$ and $\max$ in the primal as follows:

$$\max_{Z: Z^T Z \leq I} \min_y \frac{\lambda}{2}\|\hat{Y} - \bar{Y}\|_F^2 - \langle Z, W_1^k H_r(\hat{Y})U^\perp W_2^k\rangle$$

Define the operator $\Phi_k : \mathbb{R}^{p \times N} \rightarrow \mathbb{R}^{(r+1)p \times (N-r)}, k = 0, 1, 2, \ldots$, with $\Phi_k(\hat{Y}) = W_1^k H_r(\hat{Y})U^\perp W_2^k$. It is easy to check that the adjoint operator $\Phi_k^* : \mathbb{R}^{(r+1)p \times (N-r)} \rightarrow \mathbb{R}^{p \times N}$ is given by $\Phi_k^*(Z) = H_r^*(W_1^k Z W_2^k {U^\perp}^T)$. The dual problem can now be reframed as:

$$\max_{Z: Z^T Z \leq I} \min_{\hat{Y}} \frac{\lambda}{2}\|\hat{Y} - Y_{meas}\|_F^2 - \langle Z, \Phi_k(\hat{Y})\rangle \quad (23)$$

Minimizing over $\hat{Y}$, the optimality conditions give

$$\lambda(\hat{Y} - Y_{\text{meas}}) - \Phi_k^*(Z) = 0 \quad (24)$$

Note that the primal (22) is a convex problem and obeys Slater's conditions, hence the duality gap between (22) and (23) is zero. Thus the primal optimal solution can be obtained from the dual optimal solution, which is the basis for the implementation described later. Substituting $\hat{Y}$ from (24) back into (23), the dual problem reduces to:

$$\min_{Z: Z^T Z \leq I} \frac{1}{2\lambda}\|\Phi_k^*(Z)\|_F^2 + \langle Y_{\text{meas}}, \Phi_k^*(Z)\rangle \quad (25)$$

The dual objective is scaled by λ so that the objective is independent of it:

$$\min_{Z: \lambda^2 (Z^T Z) \leq I} \frac{1}{2}\|\Phi_k^*(Z)\|_F^2 + \langle Y_{\text{meas}}, \Phi_k^*(Z)\rangle \quad (26)$$

The *RRNH* for the System Identification application using an efficient first order method (i.e., the Gradient projection method applied to the dual, see e.g. [15]) can be summarized as follows:

1) Set $k = 0$. Initialize $W_1^0 = I, W_2^0 = I$.

2) Solve the dual problem (26) using the gradient projection algorithm, obtain Z^{k+1}.

3) Obtain $\hat{Y}^{k+1} = Y_{\text{meas}} + \frac{1}{\lambda}\Phi_k^*(\lambda Z^{k+1})$ (using 24).

4) Let $Y^{k+1} = H_r(\hat{Y}^{k+1})$, let $U\Sigma V^T$ be the reduced SVD of $\bar{Y}^{k+1} = W_1^k Y^{k+1} W_2^k$. Set

$$W_1^{k+1} = ((W_1^k)^{-1} U\Sigma U^T (W_1^k)^{-1} + \delta I)^{-1/2},$$
$$W_2^{k+1} = ((W_2^k)^{-1} V\Sigma V^T (W_2^k)^{-1} + \delta I)^{-1/2}.$$

5) Stop if termination criterion is satisfied, else set $k = k + 1$ and go to step 2.

Define $D^k(Z) = \frac{1}{2}\|\Phi_k^*(Z)\|_F^2 + \langle Y_{\text{meas}}, \Phi_k^*(Z)\rangle$ to be the dual objective in (26). Step 2 of the above algorithm applies the gradient projection method to solve the dual (26). We note that the projection method works well when the step size in the gradient-descent step of the method is chosen to be inversely proportional to the lipschitz constant of $D^k(Z)$(see e.g. [15]). An estimate of the Lipschitz constant of ∇D^k can be obtained as $L^k = r\lambda_{max}^2(W_1^k)\lambda_{max}^2(W_2^k)$ with the details given in the Appendix.

B. Numerical results

In [16], Liu et al mention that the main advantage of nuclear norm technique is that it makes the selection of an appropriate model order easier. We present an example, where we show that the *RRNH* improves on the nuclear norm technique for model order selection. We apply the *RRNH* implementation (algorithm described at the end of subsection B) to one of the data sets (96-006, [11]) available from (`http://homes.esat.kuleuven.be/ smc/daisy/`). The parameter r, which is the number of row-blocks in the block Hankel matrices $H_r(\hat{U})$, $Y = H_r(\hat{Y})$, is chosen so that the number of rows is greater than the expected system order. We choose an r sufficiently large, i.e. $r : rp = 60$, where p is the size of the output of the system. The parameter λ is chosen to give approximately the smallest identification error when *RRNH* is run for one iteration (i.e. just the nuclear norm heuristic as in (19)). The identification error is given by

$$e_I = \left(\frac{\|Y_{\text{measi}} - \tilde{Y}\|_F^2}{\|Y_{\text{measi}} - \bar{Y}_I\|_F^2}\right), \tag{27}$$

where $\tilde{Y} = [\tilde{y}(0), \ldots, \tilde{y}(N_I - 1)]$ denotes the output of the identified state-space model, and $\bar{Y}_I$ has each of its N_I columns equal to $Y_{\text{measi}}\mathbf{1}$. $Y_{\text{measi}} \in \mathbb{R}^{p\times N_I}$ denotes the first N_I output measurements. Similarly the validation error,e_V, can be obtained by replacing N_I by N_V in the computations in (27). The number of data points used for the identification experiment is $N_I = 150$ and for computing validation error is $N_V = 400$. The trade-off curve between the identification error,validation error and nuclear norm for is shown in Fig. 1. We pick $\lambda = 6$, that corresponds with approximately the smallest identification error of 0.0725 as indicated in Fig. 1. The normalized singular values of $YU^\perp$ (obtained by setting maximum singular value to 1) using just nuclear norm heuristic as in (19) are shown in Fig. 2. As can be seen from Fig. 2, there is no sharp drop in singular values that would clearly indicate the rank of $YU^\perp$, therefore we use the thumbrule described earlier to obtain the rank (and order of the system) to be 6. The termination criterion we use for *RRNH* is to stop after 4 iterations since we observe empirically that there is no significant change in the optimized variable after 4 iterations. For the dual-gradient method used in each iteration of *RRNH*, the termination criterion used is such that the number of iterations, $Q = \min(Q_1, Q_2)$, where Q_1 is the number of iterations for the duality gap to fall below a tolerance of 10^{-4} and $Q_2 = 4000$. Fig. 3 shows the results of *RRNH* for $\lambda = 6$ and different values of the parameter δ. The identification error and validation error were obtained as 0.0691 and 0.1154 respectively, which is comparable to the errors (0.069 and 0.12 respectively) obtained for this data set in [17]. The parameter δ, which is used as a regularization term in the weights, W_1^k, W_2^k seems to have an influence on the singular values of $H_r(\hat{Y})U^\perp$ and thus its rank. We observe empirically that as λ increases, smaller values of

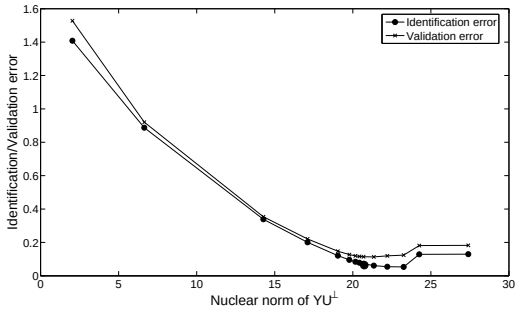

Fig. 1. Trade off curve between Identification error and Nuclear norm for the data-set. The bigger dot on the identification error curve corresponds to a very small identification error of 0.0725 and corresponds to a $\lambda = 6$

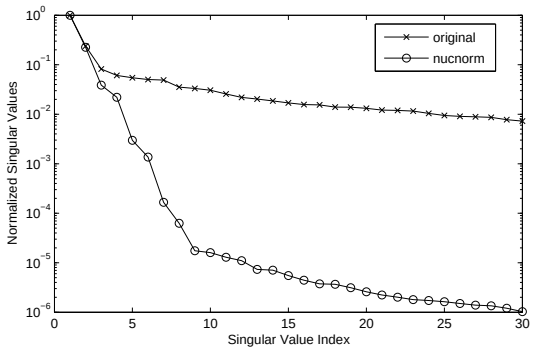

Fig. 2. Normalized singular values of $YU^\perp$ with $Y = H_r(\hat{Y})$ obtained through nuclear norm heuristic for the data-set. A plot of the normalized singular values of $Y_{\text{measi}}U^\perp$ (original) is also shown.

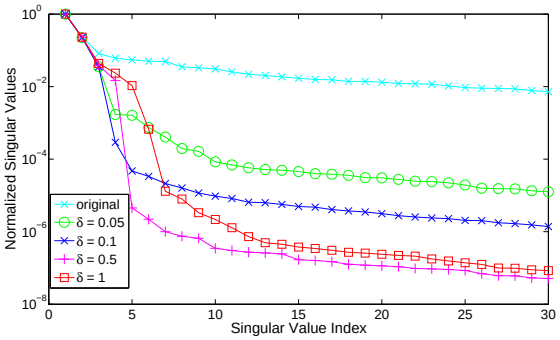

Fig. 3. Normalized singular values of $YU^{\perp}$ with $Y = H_r(\hat{Y})$ obtained through regularized reweighted nuclear norm heuristic(RRNH) for the data-set for different values of δ. A plot of the normalized singular values of $Y_{\text{measi}}U^{\perp}$ (original) is also shown.

δ give a clearer rank description for $H_r(\hat{Y})U^{\perp}$. As can be seen from Fig. 3,$\delta = 0.5$ gives a clearer description of the rank, i.e., equal to 4. Smaller values of δ (less than 0.01) don't seem to provide a clear description of the rank and this observation mirrors the observations made in [3] about the choice of ϵ for which the *iterative weighted* ℓ_1 *algorithm* recovers sparse solutions. Thus, for the data-set considered, we obtain a reduction in model order (from 6 to 4) as well as a much clearer description of model order by using *RRNH* as compared to the nuclear norm heuristic.

V. Conclusions

We explored the convergence properties of the *reweighed trace heuristic*, showing that the difference between the successive iterates of this heuristic goes to zero, and that every convergent subsequence converges to a stationary point of the concave surrogate function. We gave a reformulation of this heuristic as the *reweighted nuclear norm heuristic* (*RNH*), which allows for efficient and scalable implementation through first-order gradient methods such as the gradient projection method and conditional gradient method, as compared to the reweighted trace formulation which requires solving an SDP at each iteration. We apply the *RRNH* to a System Identification application and show that the *RRNH* provides a clearer description of the matrix rank (and hence system order) through a sharp fall in singular values in the singular value plot of $H_r(\hat{Y})U^{\perp}$. We also observe that the *RRNH* gives a lower system order as compared to *nuclear norm heuristic* (without weighting) for the data set considered, with the identification and validation errors comparable to those obtained for this data set in [17]. We observe empirically that as λ increases, smaller values of δ give a clearer rank description for $H_r(\hat{Y})U^{\perp}$. It would be useful to understand precisely how δ plays a role in providing a clear rank description as λ varies. We mentioned that the *RRNH* allows for an efficient implementation of the reweighted trace heuristic. It would be useful to quantify the efficiency and scalability of *RRNH* and compare it with the *nuclear norm heuristic* implemented using the interior point method detailed in [17].

Acknowledgements

We would like to thank Paul Tseng for very helpful discussions and for his implementation of the dual gradient projection method, which we adapted for our algorithm.

Appendix

Estimate of Lipschitz constant. For any $Z_1, Z_2 \in \mathbb{R}^{rp\times q}$,

$$\|\nabla D^k(Z_1) - \nabla D^k(Z_2)\|_F = \|\Phi_k\Phi_k^*(Z_1 - Z_2)\|_F \quad (28)$$

Note that $H_rH_r^*$ and $\Phi_k\Phi_k^*$ are self-adjoint operators. $\Phi_k\Phi_k^*$ is a compact operator since its range is finite dimensional, and it is positive since $\langle Z, \Phi_k\Phi_k^*(Z)\rangle = \langle \Phi_k^*(Z), \Phi_k^*(Z)\rangle = \|\Phi_k^*(Z)\|_F^2 \geq 0 \ \forall Z$. We apply the Rayleigh-Ritz method for self-adjoint, positive, compact operators [13] to express the maximum eigenvalue of $\Phi_k\Phi_k^*$ as $\lambda_{\max}(\Phi_k\Phi_k^*) = \sup_{W:\|W\|_F=1}\langle W, \Phi_k\Phi_k^*(W)\rangle$ and get

$$\frac{\|\Phi_k\Phi_k^*(Z_1 - Z_2)\|_F^2}{\|Z_1 - Z_2\|_F^2} \leq \sup_W \frac{\|\Phi_k\Phi_k^*(W)\|_F^2}{\|W\|_F^2} = \lambda_{\max}^2(\Phi_k\Phi_k^*)$$

Thus an upper bound on $\lambda_{\max}$ can be used to find an estimate L^k of the Lipschitz constant of the gradient of the dual objective,∇D^k. We obtain an estimate of $\lambda_{\max}(\Phi_k\Phi_k^*)$ below. From (19), it is easy to see by using the properties of norms that $\|H_r^*(X)\|_F^2 \leq r\|X\|_F^2$. Also by using the properties of trace, we have $\|\Phi_k^*(Z)\|_F^2 \leq r\langle W^kZV^k, W^kZV^k\rangle$. Let $A = (W_2^k)^2$, with eigenvalues $\rho_1^2 \geq \rho_2^2 \geq \ldots \rho_q^2$, and let $\gamma_1^2 \geq \gamma_2^2 \geq \ldots \gamma_{rp}^2$ be the eigenvalues of $(W_1^k)^2$. Using Von Neumann's Trace inequality (see e.g. [10]), it can be shown that $\langle W_1^kZW_2^k, W_1^kZW_2^k\rangle \leq {\rho_1}^2{\gamma_1}^2\|Z\|_F^2$. Thus we have

$$\begin{aligned}\frac{\|\Phi\Phi^*(Z_1 - Z_2)\|_F^2}{\|Z_1 - Z_2\|_F^2} &\leq \lambda_{\max}(\Phi\Phi^*)^2 \\ &= \left(\sup_W \frac{\|\Phi^*(W)\|_F^2}{\|W\|_F^2}\right)^2 \\ &\leq r^2{\rho_1}^4{\gamma_1}^4 \end{aligned} \quad (29)$$

Thus, $L^k = r{\rho_1}^2{\gamma_1}^2$(upper-bound on $\lambda_{\max}(\Phi\Phi^*)$) is an estimate of the lipschitz constant of ∇D^k.

References

[1] D.P. Bertsekas. *Nonlinear Programming*. 1999.
[2] J. Brinkhuis, Z. Luo, and S. Zhang. Matrix convex functions with applications to weighted centres for semidefinite programming. 2005.
[3] E.J. Candes, M.B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.
[4] E.J.Candes and B.Recht. Exact matrix completion via convex optimization.
[5] L Ljung. *System Identification - Theory for the User*. 2002.
[6] M.S. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs.
[7] M.Fazel, H.Hindi, and S.Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings American Control Conference*, 2001.
[8] M.Fazel, H.Hindi, and S.Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings American Control Conference*, 2003.
[9] M.Fazel, H.Hindi, and S.Boyd. Rank minimization and applications in system theory. In *Proceedings American Control Conference*, 2004.
[10] L. Mirsky. A trace inequality of john von neumann. *Monatshefte fur Mathematik*, 1975.

[11] B. Moor, P. Gersem, B. Schutter, and W. Favoreel. Daisy: A database for the identification of systems.

[12] B.D. Moor, M. Moonen, L.Vandenberghe, and J. Vandewalle. A geometrical approach for the identification of state space models with the singular value decomposition. In *Proceedings of the 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988.

[13] A.W. Naylor and G.R. Sell. *Linear Operator Theory in Engineering and Science*. 1982.

[14] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. Accepted for publication, *SIAM Review*.

[15] T.K.Pong, P. Tseng, S. Ji, and J. Ye. Trace norm regularization: Formulations, algorithms, and mult-task learning. 2009.

[16] Z.Liu and L.Vandenberghe. Semidefinite programming methods for system realization and identification. In *Proceedings Conference on Decision and Control*, 2009.

[17] Z.Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *Submitted to SIAM Journal on Matrix Analysis and Applications*, 2009.