**21st International Symposium on**
**Mathematical Theory of Networks and Systems**
**July 7-11, 2014. Groningen, The Netherlands**

# Low rank approximations in adaptive modeling approaches.

Kristiaan Pelckmans*, Uppsala University, SE, kp@it.uu.se

***Abstract*— This work investigates the use of low nuclear norm matrices in methods of recursive identification and adaptive modelling. The key idea is to use a Hankel matrix associated to a low-dimensional system, and to keep it as low-rank as possible during adaption. In this way, one can guarantee the power of the predictions made by this strategy, in case the true system can be approximated well as a low-dimensional system. This result is obtained by using the convex technique of the nuclear-norm projection. Proper formulation leads to a strategy with high predictive power, while guaranteeing a low-dimensional minimal realisation of the models which are built along the process. This work extends the work of Fazel et. al. [FHB01], [RFP10] and Liu et al. [LV09] to a context of online learning. An example is given towards the task of fault detection.**

Methods of recursive identification [Lju99], adaptive filtering [Say08], stochastic approximation [KY03] optimisation algorithms [BV04], and online learning [CBL06] all provide different perspectives to the problem of sequential modelling.

The proposed scheme can be seen as an implementation of recursive Subspace Identification [VODM94], since it yields at every time step $t$ an estimate $(A_t, B_t, C_t)$ based on only in- and output measurements. However, recursive forms of the original approach of subspace identification go along different lines [LGV00], [OK02], [MLL04], [MBL08], namely by recovering the range-space of the extended observability matrix. However, formal analyses of such recursive schemes are not very common till date [Knu01], [CP04], [ML07].

The present work however starts from a different optimality criterion, going along the lines as set out in the theory of online learning [CBL06]. The criterium is given as

$$\min_{A} \max_{\{u_t, y_t\}_t} \sum_{t=1}^{n} (y_t - \hat{y}_t(A))^2, \tag{1}$$

where $\hat{y}_t(A)$ is the prediction made by the algorithm $A$, by using all previously available information. In other words, $\hat{y}_t(A)$, is based on the model which is estimated thus far. This means that one is not really interested in *recovering (identification)* of the LTI underlying the data, but in the total cost which is accumulated when traversing the data using the estimate thus far. It is a happy coincidence for both to coincide, being a consequence of the squared loss and the proper stochastic assumptions (time-invariant). Note that in the present setting, a *worst case* approach is taken, rather than making the usual stochastic assumptions on the involved signals.

Lets formalise the setting. At first, consider the SISO case represented as a filter $h$. The filter $h$ is modelled as follows:

$$\begin{aligned} y_t + \epsilon_t = u_t * h + \epsilon_t &= \sum_{\tau=0}^{d-1} u_{t-\tau} h_\tau + \epsilon_t \\ &= \mathrm{tr}\left(\mathcal{H}_d(h) U_d^T(t)\right) + \epsilon_t, \end{aligned} \tag{2}$$

where $\{\epsilon_t\}_t$ are output errors, and

$$\mathcal{H}_d(h) = \begin{bmatrix} h_0 & h_1 & h_2 & \dots & h_{d-1} \\ h_1 & h_2 & h_3 & & \\ h_2 & h_3 & h_4 & & \\ \vdots & & & \ddots & \\ h_{d-1} & & & \dots & h_{2d-1} \end{bmatrix}. \tag{3}$$

Note that this matrix is of the same rank as the order of the LTI system represented by $h$. And $U_d(t)$ is defined as the symmetrical matrix

$$U_d(t) = \begin{bmatrix} u_t & \frac{u_{t-1}}{2} & \frac{u_{t-2}}{3} & \dots & \frac{u_{t-d+1}}{d} \\ \frac{u_{t-1}}{2} & \frac{u_{t-2}}{3} & \frac{u_{t-3}}{4} & & 0 \\ \frac{u_{t-2}}{3} & \frac{u_{t-3}}{4} & \frac{u_{t-4}}{5} & & 0 \\ \vdots & & & \ddots & \\ \frac{u_{t-d+1}}{d} & 0 & 0 & \dots & 0 \end{bmatrix}. \tag{4}$$

Some elementary matrix algebra shows that

$$\|U_d(t)\|_F^2 = \sum_{\tau=0}^{d} \frac{\tau u_{t-\tau}^2}{\tau^2} \leq R^2 \ln(d+1). \tag{5}$$

where one has $u_t^2 \leq R^2$ for any $t$. Now the question becomes how to learn $H_d(h)$ from given measurements $\{U_d(t)\}_t$.

Consider the more general case of a MIMO a system described in state-space form as

$$\begin{cases} x_t = A x_{t-1} + B u_t \\ y_t = C x_t + \epsilon_t \end{cases} \quad \forall\ t = 1, 2, \dots, \tag{6}$$

where $m$ is the (minimal) dimension of the state sequence $\{x_t \in \mathbb{R}^m\}_t$ where we assume that $x_0 = 0_m \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times n_u}, C \in \mathbb{R}^{n_y \times m}$ are given.

$$y_t \approx \mathrm{tr}\left(\mathcal{H}_d(A, B, C) U_d(t)\right) + \epsilon_t \tag{7}$$

with equality for large enough $d$, and where we defined $\mathcal{H}_d(A, B, C)$

$$= \begin{bmatrix} CB & CAB & CA^2B & & CA^{d-1}B \\ CA^1B & CA^3B & & & \vdots \\ CA^2B & & & & \\ \vdots & & & \ddots & \\ CA^{d-1}B & CA^dB & \dots & & CA^{2d-1}B \end{bmatrix} = \mathcal{O}\mathcal{C}, \tag{8}$$

*This work was supported in part by Swedish Research Council under contract 621-2007-6364.

and $U_d(t)$ is as defined in eq. (4). Stochastic Approximation (SA) applied to this case implements the following recursion for $t = 1, 2, \ldots$

$$\mathcal{H}^{t+1} = \mathcal{H}^t + \gamma_t \left(y_t - \operatorname{tr}\left(\mathcal{H}^t U_d^T(t)\right)\right) U_d^T(t), \qquad (9)$$

where $H^0 = 0_d 0_d^T$ and $\gamma_t = \frac{\nu}{t}$ for a given $\nu > 0$. The proximal gradient descent algorithm however implements the following scheme

$$\mathcal{H}^{t+1} = \operatorname{prox}_{\gamma_t, \|\cdot\|_*}\left(\mathcal{H}^t + \gamma_t \left(y_t - \operatorname{tr}\left(\mathcal{H}^t U_d^T(t)\right)\right) U_d^T(t)\right) \qquad (10)$$

where

$$\operatorname{prox}_{\gamma, \|\cdot\|_*}(H) = \arg\min_Z \frac{1}{2\gamma}\|Z - H\|_F^2 + \|Z\|_*, \qquad (11)$$

or equivalently

$$\operatorname{prox}_{\gamma, \|\cdot\|_*}(H) = P(H) T_\gamma(\Sigma(H)) Q^T(H), \qquad (12)$$

where we let the SVD of a matrix $H$ be defined as $H = P(H)\Sigma(H)Q^T(H)$, and $T_\gamma$ is defined as

$$T_\gamma(\sigma_i(H)) = \begin{cases} \sigma_i(H) - \gamma & \sigma_i(H) > \gamma \\ \sigma_i(H) 0 & -\gamma \leq \sigma_i(H) \leq \gamma \\ \sigma_i(H) + \gamma & \sigma_i(H) < -\gamma, \end{cases} \qquad (13)$$

and $T_\gamma(\Sigma(H)) = \operatorname{diag}\left(T_\gamma(\sigma_1(H)), \ldots, T_\gamma(\sigma_n(H))\right)$ as usual. Such schemes are extensively investigated in the context of bi-objective function minimisation in a context of optimisation algorithms. The presented Proximal Gradient Descent (PGD) algorithm is now part of modern textbooks on optimisation [NN04]. Its analysis and the choice of its step-length $\nu$ are well-studied, see e.g. [Pau08]. This line of investigations gave rise to even faster approaches as accelerated first order methods, see e.g. [BT09], [TY10].

However, experiments indicate that a slightly different scheme performs better. Rather than mapping the estimate at every iteration though the prix map, and iterating further with the newly obtained estimate, the prox map is used only to generate the prediction. That is, the recursion proceeds as a traditional gradient descent scheme, with the additional property that the resulting estimate is mapped using $T(\gamma)$ in order to generate the current prediction. Formally, given $\{\gamma_t > 0\}_t$ then $\forall t = 1, 2, 3, \ldots$

$$\begin{cases} \hat{y}_t = \operatorname{tr}\left(\operatorname{prox}_{\gamma_t, \|\cdot\|_*}(\mathcal{H}^{t-1}, \gamma_{t-1}) U_d^T(t)\right) \\ \mathcal{H}^t = \mathcal{H}^{t-1} + \gamma_t (y_t - \hat{y}_t) U_d^T(t), \end{cases} \qquad (14)$$

and $\mathcal{H}^0 = 0_d 0_d^T$.

This line of representing LTI systems, and the use of its low-rank structure can be employed in various ways

- (Adaptive Filtering) Estimating LTI systems or filters from data belongs to the core of the methodology of adaptive signal processing, see e.g. [Say08].
- (Adaptive Control) Our interest is mostly in adaptive control, where availability of the (changing) system matrix is assumed to design an optimal control strategy for the system at hand.
- (Fault Detection) An interesting third application is found in the task of fault detection or signal detection. This application is worked out in the manuscript [Pel13].

This presentation presents the formal and experimental arguments supporting the various statements and design decisions as sketched here. Hereto, we both draw on (1) techniques of optimisation with proximal gradient descent [NN04] and low-rank matrix completion as in [RFP10], [LV09], as well as on (2) experimental results.

## REFERENCES

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BV04] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Press, 2004.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CP04] Alessandro Chiuso and Giorgio Picci. The asymptotic variance of subspace estimates. *Journal of Econometrics*, 118(1):257–291, 2004.

[FHB01] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[Knu01] Torben Knudsen. Consistency analysis of subspace identification methods based on a linear regression approach. *Automatica*, 37(1):81–89, 2001.

[KY03] H.J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer Verlag, 2003.

[LGV00] Marco Lovera, Tony Gustafsson, and Michel Verhaegen. Recursive subspace identification of linear and non-linear wiener state-space models. *Automatica*, 36(11):1639–1650, 2000.

[Lju99] L. Ljung. *System identification*. Wiley Online Library, 1999.

[LV09] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235, 2009.

[MBL08] Guillaume Mercère, Laurent Bako, and Stéphane Lecœuche. Propagator-based methods for recursive subspace model identification. *Signal Processing*, 88(3):468–491, 2008.

[ML07] Guillaume Mercère and Marco Lovera. Convergence analysis of instrumental variable recursive subspace identification algorithms. *Automatica*, 43(8):1377–1386, 2007.

[MLL04] Guillaume Mercère, Stéphane Lecoeuche, and Marco Lovera. Recursive subspace identification based on instrumental variable unconstrained quadratic optimization. *International Journal of Adaptive Control and Signal Processing*, 18(9-10):771–797, 2004.

[NN04] Yurii Nesterov and I?U E Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[OK02] Hiroshi Oku and Hidenori Kimura. Recursive 4sid algorithms using gradient type subspace tracking. *Automatica*, 38(6):1035–1043, 2002.

[Pau08] Tseng Paul. On accelerated proximal gradient methods for convex-concave optimization. 2008.

[Pel13] Kristiaan Pelckmans. Towards an online, non-stochastic approach to fault detection. *Submitted*, 2013.

[RFP10] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[Say08] A.H. Sayed. *Adaptive filters*. Wiley-IEEE Press, 2008.

[TY10] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[VODM94] Peter Van Overschee and Bart De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.