KALMAN FILTERING IN STOCHASTIC GRADIENT ALGORITHMS: CONSTRUCTION OF A STOPPING RULE

Barbara Bittner[*]*, Luc Pronzato*

Laboratoire I3S
Université de Nice-Sophia Antipolis — CNRS
Bât. Euclide, Les Algorithmes
2000 route des Lucioles, BP 121
06903 Sophia Antipolis cedex, France

ABSTRACT

Stochastic gradient algorithms are widely used in signal processing. Whereas stopping rules for deterministic descent algorithms can easily be constructed, using for instance the norm of the gradient of the objective function, the situation is more complicated for stochastic methods since the gradient needs first to be estimated. We show how a simple Kalman filter can be used to estimate the gradient, with some associated confidence, and thus construct a stopping rule for the algorithm. The construction is illustrated by a simple example. The filter might also be used to estimate the Hessian, which would open the way to a possible acceleration of the algorithm. Such developments are briefly discussed.

1. INTRODUCTION

We want to minimize the function $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, applying the stochastic gradient method. We assume that $f(\mathbf{x})$ is bounded from below (and thus has at least one local minimum). For each value of $\mathbf{x}_t$ proposed by the algorithm, we observe ∇g_t, a noisy realization of $\nabla f_t = \nabla f(\mathbf{x}_t)$, the gradient of f at $\mathbf{x}_t$. The idea is to use the information present during the progress of the algorithm to estimate the value of the true gradient ∇f_t and construct a stopping rule.

We assume that

$$\nabla g_t = \nabla f_t + \mathbf{w}_t \tag{1}$$

where $(\mathbf{w}_t)$ corresponds to a sequence of independent random vectors with $\mathbb{E}(\mathbf{w}_t) = \vec{0}$ and covariance $\mathbf{W}_t$.

One iteration of the stochastic gradient algorithm, see e.g. [1, 2] is described by

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \lambda_t \nabla g_t \tag{2}$$

where the step-length λ_t satisfies (i) $\lambda_t > 0$, (ii) $\sum \lambda_t = \infty$ and (iii) $\sum \lambda_t^2 < \infty$.

Assume that $f(\mathbf{x})$ can be replaced by its quadratic approximation at $\mathbf{x}_t$, this gives

$$\nabla f_{t+1} = \nabla f_t + \mathbf{H}_t \Delta \mathbf{x}_{t+1} \tag{3}$$

where $\mathbf{H}_t$ is the Hessian of f at $\mathbf{x}_t$, a $d \times d$ symmetric matrix, and $\Delta \mathbf{x}_{t+1} = \mathbf{x}_{t+1} - \mathbf{x}_t$. For f smooth enough, this quadratic approximation will become more and more valid as $\mathbf{x}_t$ gets closer to the (a) minimizing point. Assume, moreover, that the Hessian is nearly constant so that

$$\mathbf{H}_{t+1} = \mathbf{H}_t \,. \tag{4}$$

Again, this assumption will become more and more reasonable closer to the optimum.

The idea is to consider (3,4) as the evolution equation, and (1) as the observation equation for a linear dynamical system, with state given by ∇f_t and $\mathbf{H}_t$ and input $\Delta \mathbf{x}_{t+1}$. We can then re-write (3,4) and (1) in a standard state-space representation, and use Kalman filtering to estimate the part ∇f_t of the state (we shall see in Section 2.2 that the part $\mathbf{H}_t$ is not observable). Notice that the system is driven by the optimisation algorithm through (2), but this (nonlinear) control is separated from the estimation: the algorithm runs independently from the filter. Connecting them opens some perspectives briefly discussed in Section 4.

2. A STATE-SPACE REPRESENTATION AND FILTER

We replace the Hessian $\mathbf{H}_t \in \mathbb{R}^{d \times d}$ by the vector $\mathbf{h}_t \in \mathbb{R}^{m \times 1}$ with $m = d(d+1)/2$ and (when omitting the index t for the elements $h_{ij}^{(t)}$ of the Hessian $\mathbf{H}_t$ to facilitate notation)

$$\begin{aligned} \mathbf{h}_t \;=\; & (h_{11}\; h_{22} \cdots h_{dd}\; h_{12}\; h_{13} \cdots h_{1d} \\ & h_{23}\; h_{24} \cdots h_{2d} \cdots h_{(d-1)d})^{\mathrm{T}} \,. \end{aligned} \tag{5}$$

[*]On leave from the Department of Medical Statistics at the University of Vienna.

We thus obtain the state-space representation

$$\begin{pmatrix} \nabla f_{t+1} \\ \mathbf{h}_{t+1} \end{pmatrix} = \mathbf{A}_t \begin{pmatrix} \nabla f_t \\ \mathbf{h}_t \end{pmatrix} \quad (6)$$

$$\nabla g_t = \mathbf{C}_t \begin{pmatrix} \nabla f_t \\ \mathbf{h}_t \end{pmatrix} + \mathbf{w}_t \quad (7)$$

with, as already mentioned, $\mathbf{w}_t$ the measurement noise satisfying $\mathbb{E}(\mathbf{w}_t) = \vec{0}$ and $\mathbf{W}_t = \mathbb{E}(\mathbf{w}_t \mathbf{w}_t^{\mathrm{T}})$, and

$$\mathbf{A}_t = \begin{pmatrix} \mathbf{I}_d & \mathbf{X}_{t+1} \\ 0_{m\times d} & \mathbf{I}_m \end{pmatrix}$$
$$\mathbf{C}_t = \mathbf{C} = \begin{pmatrix} \mathbf{I}_d & 0_{d\times m} \end{pmatrix} .$$

(The symbol $0_{d\times m}$ stands for the $d \times m$ matrix of zeros, $\mathbf{I}_d$ for the $d \times d$ identity matrix.) The matrix $\mathbf{X}_{t+1} \in \mathbb{R}^{d\times m}$ is constructed to fulfill the condition

$$\mathbf{X}_{t+1}\mathbf{h}_t = \mathbf{H}_t \Delta \mathbf{x}_{t+1}$$

when $\mathbf{h}_t$ is defined as in Equation (5).

We can now apply a Kalman filter on the system (6,7). The estimated values of ∇f and $\mathbf{h}$ at iteration t will be denoted $\hat{\nabla} f_{t|t}$ and $\hat{\mathbf{h}}_{t|t}$ respectively, while $\hat{\nabla} f_{t+1|t}$ and $\hat{\mathbf{h}}_{t+1|t}$ will denote their prediction at iteration $t+1$. Using (6), we obtain

$$\begin{pmatrix} \hat{\nabla} f_{t+1|t} \\ \hat{\mathbf{h}}_{t+1|t} \end{pmatrix} = \mathbf{A}_t \begin{pmatrix} \hat{\nabla} f_{t|t} \\ \hat{\mathbf{h}}_{t|t} \end{pmatrix}$$
$$\mathbf{P}_{t+1|t} = \mathbf{A}_t \mathbf{P}_{t|t} \mathbf{A}_t^{\mathrm{T}} ,$$

with $\mathbf{P}_{t|t}$ and $\mathbf{P}_{t+1|t}$ respectively the covariances of the estimation error of $[\nabla f_t, \mathbf{h}_t]$ and prediction error at iteration $t+1$.

The observation equation (7) gives

$$\begin{aligned} \begin{pmatrix} \hat{\nabla} f_{t+1|t+1} \\ \hat{\mathbf{h}}_{t+1|t+1} \end{pmatrix} &= \begin{pmatrix} \hat{\nabla} f_{t+1|t} \\ \hat{\mathbf{h}}_{t+1|t} \end{pmatrix} \\ &\quad + \mathbf{K}_{t+1} \left(\nabla g_t - \mathbf{C} \begin{pmatrix} \hat{\nabla} f_{t+1|t} \\ \hat{\mathbf{h}}_{t+1|t} \end{pmatrix} \right) \\ &= \begin{pmatrix} \hat{\nabla} f_{t+1|t} \\ \hat{\mathbf{h}}_{t+1|t} \end{pmatrix} + \mathbf{K}_{t+1} \left(\nabla g_t - \hat{\nabla} f_{t+1|t} \right) \\ \mathbf{P}_{t+1|t+1} &= \mathbf{P}_{t+1|t} - \mathbf{K}_{t+1} \mathbf{C} \mathbf{P}_{t+1|t} , \end{aligned} \quad (8)$$

with $\mathbf{K}_{t+1}$ the gain of the filter,

$$\mathbf{K}_{t+1} = \mathbf{P}_{t+1|t} \mathbf{C}^{\mathrm{T}} (\mathbf{W}_{t+1} + \mathbf{C}\mathbf{P}_{t+1|t}\mathbf{C}^{\mathrm{T}})^{-1} . \quad (9)$$

2.1. Evolution of the covariance of the prediction error

We decompose the covariance matrix $\mathbf{P}_{s|t}$, $s \in \{t, t+1\}$, into four sub-matrices that correspond to the partition of the state into gradient and Hessian,

$$\mathbf{P}_{s|t} = \begin{pmatrix} \mathbf{G}_{s|t} & \mathbf{E}_{s|t} \\ \mathbf{E}_{s|t}^{\mathrm{T}} & \mathbf{F}_{s|t} \end{pmatrix} ,$$

with $\mathbf{E}_{s|t} \in \mathbb{R}^{d\times m}$, $\mathbf{F}_{s|t} \in \mathbb{R}^{m\times m}$, and $\mathbf{G}_{s|t} \in \mathbb{R}^{d\times d}$.

The evolution of the covariance of the prediction error then satisfies

$$\begin{aligned} \mathbf{G}_{t+1|t} &= \mathbf{G}_{t|t} + \mathbf{X}_{t+1}\mathbf{E}_{t|t}^{\mathrm{T}} + \mathbf{E}_{t|t}\mathbf{X}_{t+1}^{\mathrm{T}} \\ &\quad + \mathbf{X}_{t+1}\mathbf{F}_{t|t}\mathbf{X}_{t+1}^{\mathrm{T}} \\ \mathbf{E}_{t+1|t} &= \mathbf{E}_{t|t} + \mathbf{X}_{t+1}\mathbf{F}_{t|t} \\ \mathbf{F}_{t+1|t} &= \mathbf{F}_{t|t} \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}_{t+1|t+1} &= \begin{pmatrix} \mathbf{G}_{t+1|t} & \mathbf{E}_{t+1|t} \\ \mathbf{E}_{t+1|t}^{\mathrm{T}} & \mathbf{F}_{t+1|t} \end{pmatrix} \\ &\quad - \underbrace{\begin{pmatrix} \mathbf{G}_{t+1|t} \\ \mathbf{E}_{t+1|t}^{\mathrm{T}} \end{pmatrix} \left(\mathbf{W}_{t+1} + \mathbf{G}_{t+1|t} \right)^{-1}}_{\mathbf{K}_{t+1}} \\ &\quad \times \begin{pmatrix} \mathbf{G}_{t+1|t} & \mathbf{E}_{t+1|t} \end{pmatrix} \end{aligned}$$

Notice that the part $\mathbf{F}_{t|t}$ of the covariance matrix concerning the Hessian does not change during the step from $\mathbf{P}_{t|t}$ to $\mathbf{P}_{t+1|t}$. Further, it can only decrease during the step from $\mathbf{P}_{t+1|t}$ to $\mathbf{P}_{t+1|t+1}$.

When the variance of the measurement noise, $\mathbf{W}_{t+1}$, is large the Kalman gain $\mathbf{K}_{t+1}$ is small and the decrease of the covariance of the prediction error slow, see (9). On the contrary, for a small $\mathbf{W}_{t+1}$ the Kalman gain is big and the covariance of the prediction error quickly decreases. Equation (9) can also be manipulated as follows, see e.g. [3]: multiply first by $(\mathbf{W}_{t+1} + \mathbf{C}\mathbf{P}_{t+1|t}\mathbf{C}^{\mathrm{T}})$ and then by $\mathbf{W}_{t+1}^{-1}$ from the right, and plug in equation (8), this gives

$$\begin{aligned} \mathbf{K}_{t+1} &= (\mathbf{P}_{t+1|t} - \mathbf{K}_{t+1}\mathbf{C}\mathbf{P}_{t+1|t})\mathbf{C}^{\mathrm{T}}\mathbf{W}_{t+1}^{-1} \\ &= \mathbf{P}_{t+1|t+1}\mathbf{C}^{\mathrm{T}}\mathbf{W}_{t+1}^{-1} . \end{aligned}$$

When the covariance of the prediction error is decomposed above, we obtain

$$\mathbf{K}_{t+1} = \begin{pmatrix} \mathbf{G}_{t+1|t+1} \\ \mathbf{E}_{t+1|t+1}^{\mathrm{T}} \end{pmatrix} \mathbf{W}_{t+1}^{-1} ,$$

and we can see that the larger the variance of the estimation error for ∇f_t and the covariance of the errors for ∇f_t and $\mathbf{h}_t$, the larger the Kalman gain.

2.2. Observability

The observability matrix $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ at iteration t is defined by

$$\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)} = \begin{pmatrix} \mathbf{C}_t \\ \mathbf{C}_{t+1}\mathbf{A}_t \\ \mathbf{C}_{t+2}\mathbf{A}_{t+1}\mathbf{A}_t \\ \vdots \\ \mathbf{C}_{t+d+m-1}\mathbf{A}_{t+d+m-2}\dots\mathbf{A}_t \end{pmatrix}$$

and, for the system to be observable this $d(d+m)\times(d+m)$ matrix must have full rank $d+m$, see e.g. [4]. Here $\mathbf{C}$ does not vary with t, which gives

$$\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)} = \begin{pmatrix} \mathbf{I}_d & 0_{d\times m} \\ \mathbf{I}_d & \mathbf{X}_{t+1} \\ \mathbf{I}_d & \mathbf{X}_{t+1}+\mathbf{X}_{t+2} \\ \vdots & \vdots \\ \mathbf{I}_d & \sum_{j=t+1}^{t+d+m-1}\mathbf{X}_j \end{pmatrix} .$$

It can easily be checked that any sum of matrices $\mathbf{X}_t$ is telescopic and therefore the elements of any matrix $\sum_{j=t+1}^{n}\mathbf{X}_t$ will either be zero or $[\mathbf{x}_{n+1}]_i - [\mathbf{x}_t]_i$ for some $i\in\{1,\ldots,d\}$, where $[\mathbf{x}_k]_i$ stands for the i-th component of the vector $\mathbf{x}_k$. Whether or not the matrix $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ has rank $d+m$ thus depends on the matrices $\mathbf{X}_t$ and hence on the differences between the vectors $\mathbf{x}_t$. Since the iteration (2) is converging (the stochastic gradient algorithm converges to a local solution), the difference between the vectors $\mathbf{x}_t$ is tending to zero and the rank of $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ is decreasing towards d,

$$\lim_{t\to\infty}\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)} = \begin{pmatrix} \mathbf{I}_d & 0_{d\times m} \\ \vdots & \vdots \\ \mathbf{I}_d & 0_{d\times m} \end{pmatrix} .$$

Since the first d columns of the matrix $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ are stationary and non null, the system remains *partially observable*, that is, the *gradient components* of the state are observable. Also, it is *observable at the beginning* since the rank of $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ is $d+m$ before the algorithm converges.

2.3. A stopping criterion for stochastic gradient

The filter constructed above gives an estimate $\hat{\nabla}f_{t|t}$ of the gradient of f at iteration t of the algorithm and a covariance matrix $\mathbf{G}_{t|t}$ for the estimation error. A natural stopping criterion is then given by

$$\hat{\nabla}f_{t|t}^{\mathrm{T}}\hat{\nabla}f_{t|t} + \mathrm{trace}\mathbf{G}_{t|t} < \varepsilon\,. \qquad (10)$$

3. EXAMPLE

We take $f(\mathbf{x}) = (\mathbf{x}-\mathbf{a})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{a})$, with

$$\mathbf{M} = \begin{pmatrix} 2 & 0.2 \\ 0.2 & 1 \end{pmatrix}\,,\quad \mathbf{a} = (3\ \ 3)^{\mathrm{T}}\,.$$

The measurement noise $\mathbf{w}_t$ is normal $\mathcal{N}(\vec{0},\sigma^2\,\mathbf{I}_2)$ with $\sigma = 0.3$. The stochastic gradient algorithm is initialized at $\mathbf{x}_0 = \vec{0}$ and the stopping rule is given by $\varepsilon = 5\times 10^{-3}$ in (10).

The filter can be initialized as follows. Take an arbitrary initial estimate of $\mathbf{H}$, here $\hat{\mathbf{H}}_0 = 10\mathbf{I}_m$, with a large associated covariance matrix, here $\mathbf{F}_{0|0} = 100\mathbf{I}_m$, to account for a lack of confidence in $\hat{\mathbf{H}}_0$. The estimation of ∇f_0 can be based on ∇g_0, and we can take $\hat{\nabla}f_{0|0} = \nabla g_0$ with $\mathbf{F}_{0|0} = \mathbf{W}$ and $\mathbf{E}_{0|0} = 0_{d\times m}$. Numerical experiments show that the behavior (stopping time t_s when (10) is satisfied) is not sensitive to this initialization. One may thus simply take $\mathbf{P}_{0|0} = 100\mathbf{I}_{d+m}$ with a rather arbitrary value for $\hat{\nabla}f_{0|0}$.

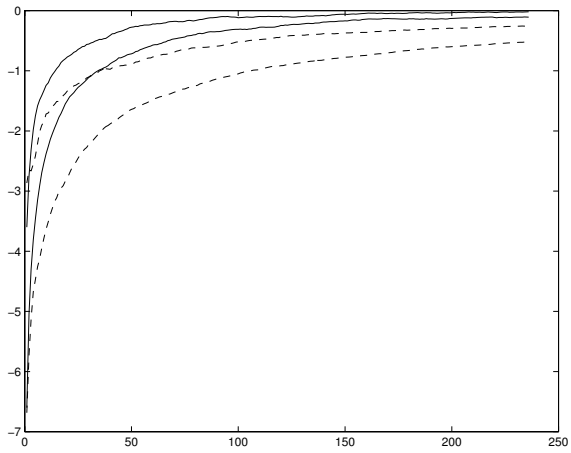

Fig. 1. Evolution of the empirical mean $\tilde{\nabla}f_t$ (11, dashed line) and exact gradient (full line) as functions of t.

We compare the evolution of the filtered estimates $\hat{\nabla}f_{t|t}$ with the true values ∇f_t and also consider the naive estimates

$$\tilde{\nabla}f_t = \frac{1}{t+1}\sum_{i=0}^{t}\nabla g_t\,, \qquad (11)$$

which corresponds to the empirical mean of the observed gradients along the trajectory imposed by the optimisation algorithm.

Figures 1 and 2 respectively present the evolutions of $\tilde{\nabla}f_t, \nabla f_t$ and $\hat{\nabla}f_{t|t}, \nabla f_t$.

Kalman filtering is well known to be robust with respect to mis-specifications of the noise characteristics. This is illustrated by Figure 3, where the noise $\mathbf{w}_t$ satisfies

$$\mathbf{w}_t = \mathbf{v}_t^1 + \mathbf{v}_t^2\,,$$

with $\mathbf{v}_t^1$ normal $\mathcal{N}(\vec{0},\sigma^2\,\mathbf{I}_2)$, $\sigma = 0.3$, and $\mathbf{v}_t^2$ uniformly distributed in $[-0.2,0.2]^2$ whereas the filter uses the covariance matrix $\mathbf{W} = (1/10)\sigma^2\,\mathbf{I}_2$.

4. FURTHER DEVELOPMENTS

We have shown how a Kalman filter could be used to define a stopping rule in a stochastic gradient algorithm. However, the benefit of the implementation of this filter may seem questionable if one observes that restricting the computation of the empirical mean estimate $\tilde{\nabla}f_t$ given by (11) to a

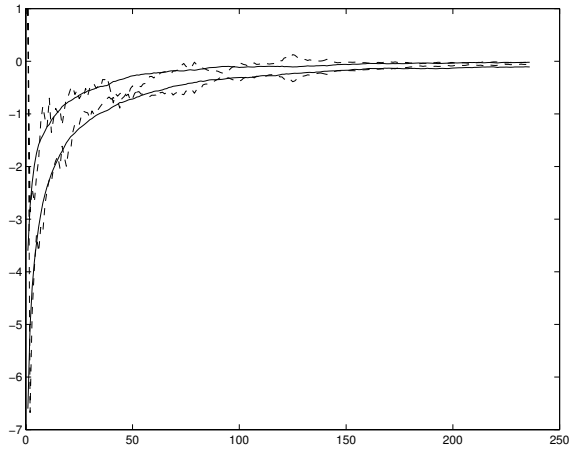

Fig. 2. Evolution of the filtered estimate $\hat{\nabla} f_{t|t}$ (dashed line) and exact gradient (full line) as functions of t.

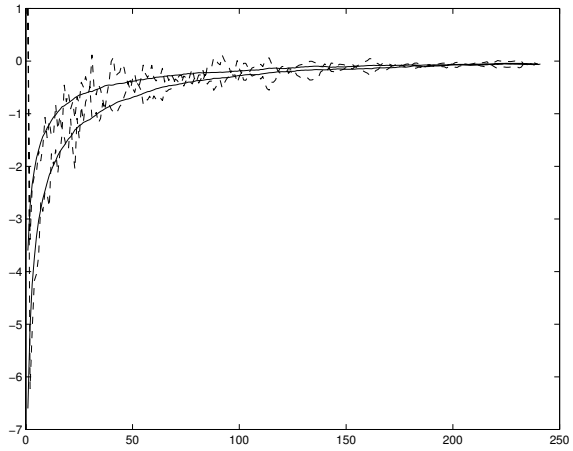

Fig. 3. Evolution of the filtered estimate $\hat{\nabla} f_{t|t}$ (dashed line) and exact gradient (full line) as functions of t, with misspecification of the noise characteristics.

sliding window of suitable fixed length T, that is,

$$\tilde{\nabla} f_t = \frac{1}{T+1} \sum_{i=t-T}^{t} \nabla g_t \, , t \geq T \, ,$$

can be expected to also yield a reasonable stopping rule. The implementation of the filter will thus be of real interest only if it provides a more useful information than the stopping rule (10). This is discussed below. Also, we assumed that the Hessian $\mathbf{H}_t$ was constant. Although Kalman filtering is robust to modelling errors, it would be more reasonable to replace (4) by

$$\mathbf{H}_{t+1} = \mathbf{H}_t + \mathbf{V}_t$$

in case f is not quadratic, where $\mathbf{V}_t$ represents some process noise. In this case, since the observability matrix $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$ tends to become singular as t increases, one may expect numerical difficulties due to the increase of the matrix $\mathbf{F}_{t|t}$. A possible way to overcome these difficulties would be to waste some iterations in terms of optimisation of f, and use them to help the estimation of $\mathbf{H}_t$; that is, take control of the $\Delta \mathbf{x}_t$ variable, and thus of $\mathbf{X}_t$, to keep control of the rank of $\mathbf{O}_{(\mathbf{A},\mathbf{C})}{}^{(t)}$.

The resulting algorithm would then choose $\mathbf{x}_t$ in order to fulfill two objectives: minimize f and help the estimation of its important characteristics, gradient and Hessian. The important benefit would then be the possible acceleration of the algorithm due to the knowledge of $\mathbf{H}$. Designing such an algorithm with dual features (optimisation and estimation) is closely connected to dual control. The approach developed in [5] in another context might then reveal appropriate.

5. REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Math. Stat.*, pp. 400–407, 1951.

[2] H.J. Kushner and G.G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer, Heidelberg, 1997.

[3] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*, Springer, Heidelberg, 1997.

[4] G. Bornard, F. Celle-Couenne, and G. Gilles, "Observability and observers," in *Nonlinear Systems: Modeling and Estimation*, A.J. Fossard and D. Normand-Cyrot, Eds., chapter 5, pp. 173–216. Chapman & Hall, London, 1993.

[5] L. Pronzato, "Adaptive optimisation and D-optimum experimental design," *Annals of Statistics*, vol. 28, no. 6, pp. 1743–1761, 2000.