Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

M. Kárný*, K. Dedecius

Approximate Bayesian Recursive Estimation

On Approximation Errors

No. 2317 January 26, 2012

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: +420 286892337, Fax: +420 266052274, Url: http://www.utia.cas.cz,
E-mail: *school@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the institute.

Abstract

Adaptive systems rely on recursive estimation of a firmly bounded complexity. As a rule, they have to use an approximation of the posterior probability density function (pdf), which comprises unreduced information about the estimated parameter. In recursive setting, the latest *approximate* pdf is updated using the learnt system model and the newest data and then approximated. The fact that approximation errors may accumulate over time course is mostly neglected in the estimator design and, at most, checked ex post. The paper inspects this problem and concludes that a sort of forgetting (flattening) is an indispensable part of approximate recursive estimation algorithms. The conclusion results from Bayesian paradigm complemented by the minimum cross-entropy (also known as Kullback-Leibler divergence, KLD) principle. Claims of the paper are illustrated on approximate recursive estimation of the mode and scaling factor of Cauchy pdf.

Keywords: approximate estimation; adaptive systems; recursive estimation; Kullback-Leibler divergence; forgetting

1. Introduction

Model-based adaptive control [3], computer intensive single-pass data processing [10] and various practical applications [1] strongly rely on recursive estimation. Mostly, the exact recursive estimation is infeasible and a sort of approximation coping with computational complexity is used [7]. Without a special care, approximation errors may accumulate to the extent that spoils the estimation. Stochastic approximations [4] represent dominating tool used for *analysis*, whether a specific estimator suffers this problem or not. *Design* of estimators avoiding it is less developed and mostly relies on stochastic Lyapunov stability theory [20]. It depends significantly on a difficult choice of Lyapunov function. Both analysis and design mostly focus on a point estimation.

Since the estimates often serve a subsequent dynamic decision making, the Bayesian estimation, exploiting the pdf of unknown parameter, became an essential tool [5]. Inspection of the approximation-errors influence has been neglected in this context. The collection of papers [13, 14, 15, 16, 17] is an exception that analyses schemes without accumulation of approximation errors. The papers conclude that this accumulation is avoided if and only if values of a finite collection of fixed linear functionals acting on logarithms of the posterior pdfs are used for construction of the approximate posterior pdf. The estimator design then reduces to the choice of functionals, whose values serve as information-bearing constraints used in approximation. This useful class of statistics is, however, too narrow to include many cases of practical interest. Thus, it is desirable to inspect approximate recursive Bayesian estimation employing statistics that lead to approximations with non-zero errors caused by the recursive treatment but prevent their accumulation.

The discussed problem is wide-spread and mostly ignored. Pointing to its existence and proposing a possible way to overcome it form the core of this brief paper. The problem is addressed from the Bayesian viewpoint. The formulation respects the ignorance of the exact pdf (Radon-Nikodým derivative with respect to a dominating measure [22]) to be approximated. Since the recursively stored information about this exact pdf is inevitably partial, the minimum KLD principle [23] serves for its completion. The considered completion recovers the use a common "naive" approximate recursive estimation when applied to a *forgotten* approximate pdf.

Section 2 formulates the problem. Core Section 3 provides its Bayesian solution employing a variable forgetting factor. Its simple data-based choice is proposed in Section 4. An example illustrating claims of the paper is in Section 5. Section 6 contains closing comments.

2. Addressed Problem

A parametric model $\mathsf{m}_t(\Theta)$ describes system output y_t in discrete time $t \in \{1, 2, \ldots\}$. The model is a pdf of y_t conditioned on the prior information, on the past measured outputs $y_1, \ldots, y_{t-1}$, on the past and current applied inputs $u_1, \ldots, u_t$ and on an unknown parameter Θ belonging to a subset $\Theta^\star$ of a finite dimensional space. At time $t-1$, the *full* information conditions the parameter Θ through the conditional *exact* pdf $\mathsf{f}_{t-1} = \mathsf{f}_{t-1}(\Theta)$. This pdf evolves according to the Bayes' rule

$$\mathsf{f}_t(\Theta) \propto \mathsf{f}_{t-1}(\Theta)\mathsf{m}_t(\Theta) \quad \text{for all} \quad \Theta \in \Theta^\star, \tag{1}$$

where $\propto$ means equality up to normalisation. This form is valid under the natural conditions of control [21], asserting that Θ is unknown to the input source. The recursion (1) starts from a user-supplied prior pdf $\mathsf{f}_0 = \mathsf{f}_0(\Theta)$ describing the prior information.

In the considered situation, the exact pdf f_t is too complex to be handled. Therefore, it is replaced by an *approximate* pdf $\hat{\mathsf{f}}_t$. The pdf $\hat{\mathsf{f}}_t$ is a projection of f_t on a user-selected set $\hat{\mathsf{f}}_t^\star$ of feasible pdfs. In [6], it was shown that the *optimal* pdf ${}^{O}\hat{\mathsf{f}}_t \in \hat{\mathsf{f}}_t^\star$ approximating optimally the exact pdf f_t is to be a KLD-minimiser $\mathsf{D}(\mathsf{f}_t||\hat{\mathsf{f}})$ [19]

$$
{}^{O}\hat{\mathsf{f}}_t \in \arg\min_{\hat{\mathsf{f}}\in\hat{\mathsf{f}}_t^\star} \mathsf{D}(\mathsf{f}_t||\hat{\mathsf{f}}) = \arg\min_{\hat{\mathsf{f}}\in\hat{\mathsf{f}}_t^\star} \int_{\Theta^\star} \mathsf{f}_t(\Theta)\ln\left(\frac{\mathsf{f}_t(\Theta)}{\hat{\mathsf{f}}(\Theta)}\right)\mathsf{d}\Theta. \tag{2}
$$

Since the direct use of (2) with f_t evolving according to (1) is prevented by the problem definition, the recursive evaluation *without an additional error* should instead evolve the optimal pdf ${}^{O}\hat{\mathsf{f}}_t$, i.e., to realise the map

$$
\left(\mathsf{m}_t(\Theta),\ {}^{O}\hat{\mathsf{f}}_{t-1}(\Theta)\right) \to {}^{O}\hat{\mathsf{f}}_t(\Theta),\ \forall\Theta\in\Theta^\star. \tag{3}
$$

The papers [13, 14, 15, 16, 17] mentioned in Introduction have shown that such a construction is possible if and only if the sets $\hat{\mathsf{f}}_t^\star$ are delimited by a finite collection of values of linear time-invariant functionals $\mathcal{F}_j(\ln(\mathsf{f}_t))$ fulfilling $\mathcal{F}_j(1) = 0$, $j = 1,\ldots,J$. However, this is not the case of most commonly stored statistics, for instance, the mean and covariance values in unscented approximation [12], the likelihood values on a variable (e.g., Monte Carlo generated) grid [9], statistics determining finite Gaussian mixtures with a fixed number of components [2], statistics yielded by variational Bayes [24] etc.

Due to the non-commutativity of the Bayes rule (1) and projections determined by the discussed techniques, the optimal recursive approximation (3) is not reachable. Consequently, instead of ${}^{O}\hat{\mathsf{f}}_t$, only an approximate pdf $\hat{\mathsf{f}}_t \in \hat{\mathsf{f}}_t^\star$ can be evolved. Then (3) is replaced by

$$
\left(\mathsf{m}_t(\Theta),\ \hat{\mathsf{f}}_{t-1}(\Theta)\right) \to \hat{\mathsf{f}}_t(\Theta),\ \forall\Theta\in\Theta^\star. \tag{4}
$$

It is mostly constructed in the following *naive* way:

$$
\begin{array}{ll}
\text{Define} & \tilde{\mathsf{f}}_t(\Theta) \propto \hat{\mathsf{f}}_{t-1}(\Theta)\mathsf{m}_t(\Theta),\ \forall\Theta\in\Theta^\star, \\
\text{Find} & \hat{\mathsf{f}}_t \in \arg\min\limits_{\hat{\mathsf{f}}\in\hat{\mathsf{f}}_t^\star} \mathsf{D}(\tilde{\mathsf{f}}_t||\hat{\mathsf{f}}),
\end{array} \tag{5}
$$

often with other proximity measures than the KLD considered. Let us stress that the optimal but infeasible projection (3) with the definition (2) would be

$$
\begin{array}{ll}
\text{Define} & \breve{\mathsf{f}}_t(\Theta) \propto \mathsf{f}_{t-1}(\Theta)\mathsf{m}_t(\Theta),\ \forall\Theta\in\Theta^\star, \\
\text{Find} & {}^{O}\hat{\mathsf{f}}_t \in \arg\min\limits_{\hat{\mathsf{f}}\in\hat{\mathsf{f}}_t^\star} \mathsf{D}(\breve{\mathsf{f}}_t||\hat{\mathsf{f}}).
\end{array} \tag{6}
$$

Here the question arises how to construct the map (4) respecting $\hat{\mathsf{f}}_{t-1} \neq \mathsf{f}_{t-1}$ or, in other words, how to modify the naive recursive approximate estimation (5) so that the approximation-errors accumulation is counteracted.

3. Solution

At time $t-1$, the approximate pdf $\hat{\mathsf{f}}_{t-1}$ represents the available information about the exact pdf f_{t-1}. It differs both from the optimal approximate pdf ${}^{O}\hat{\mathsf{f}}_{t-1}$ and from the unknown exact pdf f_{t-1}. The already cited result [6] implies that the approximate pdf $\hat{\mathsf{f}}_{t-1}$ is acceptable if and only if there is a finite, ideally small, non-negative β_{t-1} such that

$$\mathsf{D}(\mathsf{f}_{t-1}||\hat{\mathsf{f}}_{t-1}) \leq \beta_{t-1}. \tag{7}$$

The axiomatically justified minimum KLD principle [23] recommends to replace the unknown f_{t-1} by a pdf ${}^{\lambda}\mathsf{f}_{t-1}$ with the smallest KLD on a pdf representing the information before processing the information contained in $\hat{\mathsf{f}}_{t-1}$ and β_{t-1}. The prior pdf $\mathsf{f}_0(\Theta)$ is a natural descriptor of such (vague) information. The choice of ${}^{\lambda}\mathsf{f}_{t-1}$ is made in $\mathsf{f}^{\star}_{t-1}$ containing pdfs f_{t-1} on $\Theta^{\star}$ meeting (7). Thus, the minimum KLD principle recommends the choice

$${}^{\lambda}\mathsf{f}_{t-1} \in \arg\min_{\mathsf{f}_{t-1}\in\mathsf{f}^{\star}_{t-1}} \mathsf{D}(\mathsf{f}_{t-1}||\mathsf{f}_0). \tag{8}$$

The Kuhn-Tucker optimality conditions [11] provide directly the solution of this task. It is determined by a factor λ_{t-1} (motivating the notation ${}^{\lambda}\mathsf{f}_{t-1}$)

$$\lambda_{t-1} = (1+\beta_{t-1})^{-1} \in [0,1] \tag{9}$$

and has the form ${}^{\lambda}\mathsf{f}_{t-1} \propto \mathsf{f}_0^{(1-\lambda_{t-1})}\, \hat{\mathsf{f}}_{t-1}^{\lambda_{t-1}}$ where

$$\begin{array}{lcl} \lambda_{t-1} & = & 0 \quad \text{if} \quad \mathsf{D}(\mathsf{f}_0||\hat{\mathsf{f}}_{t-1}) < \beta_{t-1} \\ \lambda_{t-1} & & \text{solves} \quad \mathsf{D}({}^{\lambda}\mathsf{f}_{t-1}||\hat{\mathsf{f}}_{t-1}) = \beta_{t-1} \text{ otherwise.} \end{array} \tag{10}$$

${}^{\lambda}\mathsf{f}_{t-1}$ comprises *all information* about Θ in the remembered elements $\mathsf{f}_0, \hat{\mathsf{f}}_{t-1}$, β_{t-1}, thus it is legitimate to identify ${}^{\lambda}\mathsf{f}_{t-1}$ with f_{t-1}. Therefore, we can propagate it via the Bayes rule (1), and repeat the procedure for all t:

$$\begin{array}{ll} \text{Define} & \tilde{\mathsf{f}}_t \propto \mathsf{f}_0^{(1-\lambda_{t-1})}\, \hat{\mathsf{f}}_{t-1}^{\lambda_{t-1}} \mathsf{m}_t \\ \text{Find} & \hat{\mathsf{f}}_t \in \arg\min_{\hat{\mathsf{f}}\in\hat{\mathsf{f}}^{\star}_t} \mathsf{D}(\tilde{\mathsf{f}}_t||\hat{\mathsf{f}}), \end{array} \tag{11}$$

i.e., the naive way (5) recovers after the use of a stabilised forgetting [18] applied to the approximate pdf $\hat{\mathsf{f}}_{t-1}$.

Many estimation methods employ similar techniques to avoid accumulation of errors, see, e.g., the Monte Carlo methods in recursive parameters estimation [8]. The above arguments only justify the need of a technique of this type and recommend its form (10). Note that the naive way (5) coincides with (11) for $\lambda_{t-1} = 1$. Then, Equation (9) implies $\beta_{t-1} = 0$, which induces $\hat{\mathsf{f}}_{t-1} = \mathsf{f}_{t-1}$. This indicates that the derived extension (11) reduces smoothly to the naive way (5) which is adequate only in non-recursive (one-step) estimation. Obviously, the limit case $\lambda_{t-1} = 1$ is rather exceptional.

4. Data-based Choice of $\boldsymbol{\lambda_t}$

Prior knowledge of β_{t-1} and thus of λ_{t-1} in (10) can hardly be supposed. Their dependence on time and data makes the estimation of λ_{t-1} difficult. However, the recursive nature of the approximate estimator (11) implies that an incorrectly chosen λ_{t-1} only increases the approximation error, which is counteracted by forgetting applied in the subsequent estimation steps. Thus, even an extremely simple guess of λ_{t-1} is expected to serve the purpose. This conjecture, whose validity is experimentally supported and illustrated in Section 5, led us to the following use of the standard Bayesian hypotheses testing:

- Hypotheses $H_k: \ \lambda_t = \lambda_k \in [0, 1]$, $k = 1, 2$, $\lambda_1 \neq \lambda_2$, are formulated.
- The approximate recursive estimation (11) is run in parallel for $\lambda_t = \lambda_k$ yielding the pdfs $\hat{\mathsf{f}}_{t;k}$, for $k = 1, 2$.
- The values of the predictors $\int_{\Theta^\star} {}^{\lambda_k}\hat{\mathsf{f}}_{t-1}(\Theta)\mathsf{m}_t(\Theta)\,\mathsf{d}\Theta$ are evaluated during the projection step and used for incrementing log-likelihoods $l_{t-1;k}$ of H_k, $k = 1, 2$.
- The hypotheses are undecided until $\Delta l_t = |l_{t;1} - l_{t;2}|$ crosses a threshold $h \in [3, 7]$ implying that the probability $(1 + \exp(l_{t;2} - l_{t;1}))^{-1}$ of H_1 is close to 1 or 0.
- The better (even not winning yet) H_k provides parameter estimates exploited by the supported adaptive system.
- The poorer of H_k is discarded when $\Delta l_t > h$.

- A new hypothesis is constructed to replace the discarded one. It uses the latest ${}^{\lambda}\hat{\mathsf{f}}_{t-1}$ of the winning hypothesis as the pdf to be updated and a newly selected λ generated in $[0,1]$. At present, the middle between the winning λ_k and an end-point of a sub-interval of $[0,1]$ is chosen. The right end is chosen if the straight line through the log-likelihood values increases, the left end otherwise.

This algorithm can be improved in many directions: i) more hypotheses can be used; ii) the search for a new λ_k can emulate a derivative-free maximisation of log-likelihood normalised by the number of undecided steps; iii) probabilities of hypotheses can be evaluated on a fixed λ grid and combined with a second-level forgetting [25]. Various tested approaches led to similar results.

5. Illustrative Example

This section illustrates the discussed drawback of the naive way and the advantage of the proposed technique using even the simple choice of the factor λ_k.

Experiment setup Mutually independent scalar uncontrolled system outputs y_t were generated from the Cauchy distribution $\mathsf{m}_t(\Theta) \propto \sqrt{r}(1 + (y_t - \mu)^2/r)^{-1}$ with the mode $\mu = 1$ and scaling factor $r = 0.1$.

The approximate pdfs $\hat{\mathsf{f}}_t$ of the unknown parameter $\Theta = (\mu, r)$ were searched in the normal inverse-gamma (NiΓ) class. For it, as for any member of pdfs conjugated to exponential family [5], the approximation (11) reduces to moment matching. The moments were evaluated by a straightforward Monte-Carlo procedure with 500 samples generated from $\hat{\mathsf{f}}_{t-1}$. This intentionally high number of samples suppresses possible side-effects of numerical integration.

The recursive estimation run exactly according to (11) and with λ_k chosen according to Section 4. The NiΓ prior pdf $\hat{f}_0$ had zero expectation of μ with unit variance; expectation of r equal 10^{-4} and the number of degrees of freedom equal 3. The initial forgetting factor $\lambda_{0;k} \in \{0.7, 0.8\}$ were allowed to change within the range $[2/3, 1]$, the threshold $h = 5$ was used.

Commented results The naive way (5) was compared with its proposed modification (11). Selected sample statistics presented in Table 1 for $T = 500$ and $T = 5000$ output samples provide numerical performance indicators. Furthermore, longer simulation runs were performed confirming stability of the method. Each time instant, the hypothesis with a higher likelihood supplied parameter estimates compared with the true value.

	naive T = 500	**proposed** T = 500	**naive** T = 5000	**proposed** T = 5000
mean	0.1463	0.0111	0.1281	0.0072
median	0.1348	0.0080	0.1260	0.0077
std	0.0310	0.1198	0.0115	0.0616
rmse	0.1495	0.1203	0.1286	0.0620

Table 1: Sample statistics of difference between modes of $\hat{\mathsf{f}}_t$ and μ.

Table 1 shows: i) root mean square error (*rmse*) is smaller for the proposed technique; ii) biases, visible in *mean* values, are predominantly responsible for rmse of the naive way, which sticks at biased value even in long run (cf. small standard deviation, *std*); iii) biases of the proposed modification are significantly smaller than those offered by the naive way; iv) smaller biases are reached at the cost of a larger standard deviation *std* (this volatility reflects that the estimation respects its approximate nature) but altogether they lead to smaller rmse; v) closeness of mean and median in respective columns indicates symmetry of distributions of the inspected deviations.

Note that the algorithm performance is influenced by "tuning knobs", which are fixed in the presented experiments. Unreported evaluations lead to the conclusions on their influence on results: very weak for the threshold $h \in [3, 7]$; negligible for the initial values of forgetting factors; weak for statistics of the prior pdf f_0; mild for the number of Monte Carlo samples needed for the projection and prediction. This indicates a solid robustness of the method with respect to its initial setting.

6. Concluding Remark

This brief paper indicates that a real need exists for constructive ways counteracting the accumulation of approximation errors in a range of approximation techniques including, e.g., unscented transformations, recursive Monte Carlo methods, variational Bayes etc. The problem is especially urgent in parameter estimation, where consequences are not damped by a nontrivial stable state evolution, but the discussed errors surely degrades quality of various filters, too. The proposed solution provides a way to visible improvements.

Acknowledgement

The research has been supported by MŠMT 1M0572 and GAČR 102/08/0567.

Bibliography

[1] H.K. Alaei, K. Salahshoor, and H.K. Alaei. Model predictive control of distillation column based recursive parameter estimation method using HYSYS simulation. *Intelligent Computing and Cognitive Informatics, International Conference on*, 0 (2010) 308–311.

[2] D.L. Alspach and H.W. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximation. *IEEE Tran. on Automatic Control*, 17(4) (1972) 439–448.

[3] K.J. Astrom and B. Wittenmark. *Adaptive Control*. Addison-Wesley, Massachusetts, (1989).

[4] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, Berlin, (1990).

[5] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, (1985).

[6] J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3) (1979) 686–690.

[7] F. Daum. Nonlinear filters: beyond the Kalman filter. *Aerospace and Electronic Systems Magazine, IEEE*, 20(8) (2005) 57–69.

[8] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, (2001).

[9] A. Doucet and V. B. Tadizic. Parameter estimation in general state-space models using particle methods. *J. Annals of the Institute of Statistical Mathematics*, 55(2) (2003) 409–422.

[10] D. Hand, H. Mannila, and P. Smyth. *Principles of data mining.* MIT Press, (2001).

[11] R. Horst and H. Tuy. *Global Optimization.* Springer, (1996). 727 pp.

[12] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. A new approach for the nonlinear transformation of means and covariances in linear filters. *IEEE Tran. on Automatic Control*, 5(3) (2000) 477–482.

[13] R. Kulhavý. A Bayes-closed approximation of recursive nonlinear estimation. *Int. J. Adaptive Control and Signal Processing*, 4 (1990) 271–285.

[14] R. Kulhavý. Recursive Bayesian estimation under memory limitations. *Kybernetika*, 26 (1990) 1–20.

[15] R. Kulhavý. Recursive nonlinear estimation: A geometric approach. *Automatica*, 26(3) (1990) 545–555.

[16] R. Kulhavý. Can approximate Bayesian estimation be consistent with the ideal solution? In *Proc. of the 12th IFAC World Congress*, volume 4, pages 225–228, Sydney, Australia, 1993.

[17] R. Kulhavý. Implementation of Bayesian parameter estimation in adaptive control and signal processing. *The Statistician*, 42 (1993) 471–482.

[18] R. Kulhavý and M. B. Zarrop. On a general concept of forgetting. *Int. J. of Control*, 58(4) (1993) 905–924.

[19] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 (1951) 79–87.

[20] H.J. Kushner. *Stochastic stability and control.* Academic Press, (1967).

[21] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[22] M.M. Rao. *Measure Theory and Integration.* John Wiley, New York, (1987).

[23] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Tran. on Information Theory*, 26(1) (1980) 26–37.

[24] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, (2005).

[25] A. Votava. Choice of optional parameters determining partial forgetting. Technical report, Faculty of Mathematics and Physics, Charles University, Prague, 2010. master thesis.