ON SELF-ASSESSMENT OF PROFICIENCY OF AUTONOMOUS SYSTEMS

Petar M. Djurić$^{(1)}$ *and Pau Closas*$^{(2)}$

(1) Dept. of Electrical & Computer Engineering, Stony Brook University, Stony Brook, NY (USA)
(2) Dept. of Electrical & Computer Engineering, Northeastern University, Boston, MA (USA)

ABSTRACT

In this paper we propose a probabilistic framework for proficiency self-assessment of autonomous systems. We define proficiency as a mathematical concept, i.e., as a metric that depends on a variety of factors. This concept allows for assessment of the degree of completion of a given task by a system. We provide the rationale behind the proposed concept and its forms for various settings. Further, we present motivating examples with details of evaluation of the proficiency. We anticipate that our definition of proficiency is a step forward toward achieving "self-awareness" of autonomous systems.

Index Terms— Self-assessment, autonomous systems, sequential inference, proficiency.

1. INTRODUCTION

In statistical inference, the problem of determining in real-time whether a method run by a system performs well has often been addressed in the literature. Most of the time, the emphasis has been on the ability of the system to estimate unknown parameters or to select a model for the observed measurements. In principle, however, none of the adopted metrics for measuring the performance of the system (e.g., mean square errors of parameter estimates, statistics of model residuals, or probabilities of correct model selections) are appropriate. The main reason for this is that in real world settings, we neither know the parameters nor the model. A better approach is to use the ability of the method to predict new observations, which is the strategy we adopt in this paper.

In the context of sequential methods for model validation, there are some notable approaches that are only theoretically justified under the assumption of linear and Gaussian systems. These approaches are typically based on concepts of residual analysis of processed time series, where, if the model is correct the residual sequences are white noise processes. These procedures are prompt to fail in more general nonlinear/non-Gaussian systems, where the assumed whiteness does not hold. Attempts to go beyond simple cases can be found in [1], where the focus was on model validation. In practice, the used models that describe observations are often only approximately accurate, and then the question is how robust they are to deviations from the made assumptions under which they hold. We have investigated this issue in a preliminary work in [2]. In this context, other developments in classical Bayesian non-sequential inference include [3]. Relevant to our concept of proficiency is [4], where convergence results of predictive distributions obtained by sequential Monte Carlo methods are discussed. The main idea is to use the predictive distribution to validate the ability of a filter to predict new observations and build statistically meaningful tests to assess whether the prediction is valid or not.

In the literature, there has been interest in proficiency self-assessment (PSA) and some progress has been made in machine learning, information theory, and artificial intelligence. Bayesian optimization is at the core of many approaches [5]. A relevant information-theoretic work discussing model complexity is [6], where predictive abilities of studied systems are exploited. Similarly, [7, 8, 9] elaborate on prediction and self-healing systems for evolutionary robotics and artificial intelligence. For instance, in [10] the problem of autonomous robot navigation is addressed from a machine learning perspective using neural networks.

Another relevant set of research efforts on self-assessment is found in the literature on human metacognition [11] and the use of self-assessment in artificial intelligence [12, 13, 14, 15, 16]. Self-diagnosis in the context of wireless communication networks can be found in [17, 18].

The main contribution of this paper is the proposal of a new metric for self-assessment of proficiency in autonomous systems. The foundations and the rationale of the proposed PSA are provided in Section 2. In Section 3 we address approaches to computing the PSA. Sections 4 through 6 provide examples of use in particular setups that are relevant to many applications. Finally, Section 7 concludes the paper with final remarks.

2. FOUNDATIONS OF PROFICIENCY SELF-ASSESSMENT

In engineering and science, we describe observed data by models. Let us denote generically a considered model by $\mathcal{M}$.

PMD has been supported by NSF under Awards CCF-1618999 and CNS-1642965. PC has been supported by NSF under Awards CNS-1815349 and ECCS-1845833.

One way of assessing models is by their predictive power. For simplicity, assume that the model $\mathcal{M}$ is used to predict scalar observations, i.e.,

$$\widehat{y}_t = g_{\mathcal{M}}(y_{1:t-1}), \tag{1}$$

where t is a discrete time index, and the notation $g_{\mathcal{M}}(y_{1:t-1})$ means that the model $\mathcal{M}$ uses previous data $y_{1:t-1}$ and some mathematical function $g_{\mathcal{M}}(\cdot)$ to predict y_t, where the predicted value is $\widehat{y}_t$. Typically, in that case, one would use metrics, e.g., like $\mathbb{E}(\widehat{y}_t - y_t)^2$ to compare models and evaluate how well they perform. It will be beneficial if we can obtain bounds on how well models can perform.

In estimation theory, the most common benchmark for assessing estimators is the Cramér-Rao bound (CRB). This bound provides the smallest achievable variance of an unbiased estimators. In Bayesian theory, the equivalent to the CRB is the Bayesian CRB [19]. This bound is the lowest possible minimum mean square error (MMSE) that can be obtained in estimating a random parameter. Again, for simplicity, consider a scalar random parameter θ. The following theorem states the Bayesian CRB [19]:

Theorem: Let θ be a random variable, y an observation vector, and $p_{y,\theta}(y,\theta)$ their joint distribution. Under some mild conditions, the mean square error of any estimate $\widehat{\theta}(y)$ based on the measurements y satisfies the inequality:

$$\begin{aligned}\mathbb{E}(\widehat{\theta}(y) - \theta)^2) &\geq \left(\mathbb{E}\left[\left(\frac{\partial \ln p_{y,\theta}(y,\theta)}{\partial \theta}\right)^2\right]\right)^{-1} \\ &= \left(-\mathbb{E}\left[\frac{\partial^2 \ln p_{y,\theta}(y,\theta)}{\partial \theta^2}\right]\right)^{-1}. \end{aligned} \tag{2}$$

In practice, in most of the cases the use of the bound is to estimate how well an estimator performs. One first obtains estimates of the parameters, then computes their MSEs, and finally compares the MSEs to the Bayesian CRB. This process, however, is not realistic simply because the ground truth (the true values of the parameters) is unknown. The parameter values are only known in simulation settings; if they were known, their estimation would not be needed. In our work, we want to use the Bayesian CRB to define mathematically the concept of proficiency of a model without the need to know the true values of the model parameters and without access to any observations.

We define the proficiency of a model $\mathcal{M}$ at time t, $\mathcal{P}_t(\mathcal{M})$, by

$$\begin{aligned}\mathcal{P}_t(\mathcal{M}) &= \mathbb{E}\left[\left(\frac{\partial \ln p(y_t, y_{1:t-1}|\mathcal{M})}{\partial y_t}\right)^2\right] \\ &= -\mathbb{E}\left[\frac{\partial^2 \ln p(y_t, y_{1:t-1}|\mathcal{M})}{\partial y_t^2}\right], \end{aligned} \tag{3}$$

where the above expectations are with respect to $p(y_t, y_{1:t-1}|\mathcal{M})$, the joint distribution of the observation that is predicted, y_t, and the previous observations, $y_{1:t-1}$. We have

$$\mathcal{P}_t(\mathcal{M}) \geq \mathbb{E}\left[\left(y_t - g_{\mathcal{M}}(y_{1:t-1})\right)^2\right]^{-1}. \tag{4}$$

In the above definition, we basically treat the next observation as a random parameter and then exploit the theorem for the Bayesian CRB to define the proficiency of a model. It goes without saying that in Bayesian theory, there is no difference in the treatment of parameters and observations.

We note that the proficiency of a model satisfies $\mathcal{P}_t(\mathcal{M}) \geq 0$, and that $\mathcal{P}_t(\mathcal{M})$ is not bounded from above. We refer to this type of PSA as *a priori PSA*. A simple example of analytically obtaining the a priori proficiency $\mathcal{P}_t(\mathcal{M})$ is given in Section 4.

2.1. Conditioning of PSA on observed data

Here we propose a measure of proficiency that is not defined by averaging of all possible (seen and unseen data) but only on unseen data. It is given by

$$\begin{aligned}\mathcal{P}_t(\mathcal{M}|y_{1:t-1}) &= \mathbb{E}\left[\left(\frac{\partial \ln p(y_t|y_{1:t-1},\mathcal{M})}{\partial y_t}\right)^2\right] \quad (5)\\ &= -\mathbb{E}\left[\frac{\partial^2 \ln p(y_t|y_{1:t-1},\mathcal{M})}{\partial y_t^2}\right], \quad (6)\end{aligned}$$

where the expectation is over the predictive distribution according to the model $\mathcal{M}$. We can rewrite (6) by

$$\mathcal{P}_t(\mathcal{M}|y_{1:t-1}) \geq \left[\int (y_t - g_{\mathcal{M}}(y_{1:t-1}))^2 p(y_t|y_{1:t-1},\mathcal{M}) \mathrm{d}y_t\right]^{-1}, \tag{7}$$

where $g_{\mathcal{M}}(y_{1:t-1})$ is a predicted value of the next observation y_t obtained by the model $\mathcal{M}$, and $p(y_t|y_{1:t-1},\mathcal{M})$ is the predictive distribution of y_t. We call this type of PSA *in situ PSA*. The computation of in situ proficiency based on an autoregressive model of order one is given in Section 5.

2.2. Extensions to settings with state-space models

The proposed definition of proficiency can readily be extended to dynamic processes represented by state-space models. We recall that state-space models are comprised of two processes, one hidden (also referred to as a state process) and another that is observed. The objective is to estimate the state from the observed process. In mathematical terms, we express the model by

$$x_t = g(x_{t-1}, u_t), \tag{8}$$
$$y_t = h(x_t, v_t), \tag{9}$$

where x_t is the state at time t, y_t is the observation, and u_t and v_t are random perturbations. The conditional proficiency of a system is defined as in (6). The computation of the proficiency is discussed further below. The work in this area is somewhat related to the Bayesian CRB presented in [20]. In Section 6, we show how we can compute the proficiency of a Kalman filter.

3. COMPUTATIONAL APPROACHES TO PSA

In most cases, the analytical evaluation of the system's proficiency would not be possible. Instead, one has to resort to computational methods. Consider the conditional proficiency given by (6). Recall that the expectation is with respect to the predictive distribution $p(y_t|y_{1:t-1}, \mathcal{M})$. The expectation in (6) can then readily be computed by drawing samples according to

$$y_t^{(m)} \sim p(y_t|y_{1:t-1}, \mathcal{M}), \tag{10}$$

where $y_t^{(m)}$ is the mth drawn sample of y_t. The generation of these samples can be either done directly (if possible), by applying a Markov chain Monte Carlo scheme [21], or by an importance sampling method [22]. When we deal with nonlinear state-space models, a natural candidate for computing the desired expectations is by way of particle filtering [23]. In fact, particle filtering has already been used in related applications where the objective has been to assess if an underlying model of the data should be rejected or not [24], if the data contains outliers [25], or if the data experience a change of model and one needs to change the number of particles to track the hidden states [4].

4. PROFICIENCY AS A FUNCTION OF A PRIORI KNOWLEDGE

Let us assume that observed data are generated by a model of the form

$$\mathcal{M}: \quad y_t = a + w_t, \tag{11}$$

where $t = \{1, 2, \dots\}$, $a \sim \mathcal{N}(a_0, \sigma_a^2)$ and $w_t \sim \mathcal{N}(0, \sigma_w^2)$. Thus, the model represents a constant signal embedded in noise. The value of the signal is a and the noise w_t is zero-mean Gaussian. The prior knowledge about the constant signal is modeled by a Gaussian distribution $\mathcal{N}(a_0, \sigma_a^2)$.

Suppose next that y_1 is observed and one wants, given $\mathcal{M}$, to predict y_2. What is the best possible performance of any predictor given the available information (y_1, $a \sim \mathcal{N}(a_0, \sigma_a^2)$, and $w_t \sim \mathcal{N}(0, \sigma_w^2)$)?

We use (2), which now becomes

$$\mathcal{P}_2(\mathcal{M}) = \left(-\mathbb{E}\left[\frac{\partial^2 \ln p(y_2, y_1|\mathcal{M})}{\partial y_2^2}\right]\right). \tag{12}$$

We need the expression for $p(y_2, y_1|\mathcal{M})$, and we obtain it by

$$p(y_2, y_1|\mathcal{M}) = \int_{-\infty}^{\infty} p(y_2, y_1|a, \mathcal{M})p(a|\mathcal{M})\mathrm{d}a. \tag{13}$$

where $p(y_2, y_1|a, \mathcal{M}) = \mathcal{N}(\mu, \sigma_w^2 I)$, where $\mu = [a_0 \ a_0]^\top$, and I is the 2×2 identity matrix. It can be shown that upon integrating out a, we have

$$p(y_2, y_1|\mathcal{M}) = \mathcal{N}(\mu, \Sigma), \tag{14}$$

where

$$\Sigma = \begin{bmatrix} \sigma_w^2 + \sigma_a^2 & \sigma_a^2 \\ \sigma_a & \sigma_w^2 + \sigma_a^2 \end{bmatrix}. \tag{15}$$

Next, we can readily find that

$$\frac{\partial^2 \ln p(y_2, y_1|\mathcal{M})}{\partial y_2^2} = \frac{\sigma_w^2 + \sigma_a^2}{\sigma_w^4 + 2\sigma_w^2\sigma_a^2}. \tag{16}$$

This expression is not a function of the observations, and therefore

$$\mathcal{P}_2(\mathcal{M}) = \frac{\sigma_w^2 + \sigma_a^2}{\sigma_w^4 + 2\sigma_w^2\sigma_a^2}. \tag{17}$$

We note that when we have perfect knowledge of a ($\sigma_a^2 = 0$), according to the model we do not learn anything from y_1 and $\mathcal{P}_2(\mathcal{M}) = 1/\sigma_w^2$. When we have no knowledge of a and express this by $\sigma_a^2 >> \sigma_w^2$, we have that $\mathcal{P}_2(\mathcal{M}) \approx 1/(2\sigma_w^2)$. In fact, we can write

$$\lim_{\sigma_a^2 \to \infty} \mathcal{P}_2(\mathcal{M}) = \frac{1}{2\sigma_w^2}, \tag{18}$$

and not surprisingly, $\mathcal{P}_2(\mathcal{M})$ reduces with the increase of σ_w^2.

Next suppose that we have three observations y_1, y_2, and y_3. With the same line of reasoning, we obtain

$$\mathcal{P}_3(\mathcal{M}) = \frac{\sigma_w^2 + 2\sigma_a^2}{\sigma_w^4 + 3\sigma_w^2\sigma_a^2}. \tag{19}$$

For as long as $\sigma_w^2 > 0, \sigma_a^2 > 0$, the inequality $\mathcal{P}_2(\mathcal{M}) < \mathcal{P}_3(\mathcal{M})$ holds, that is, the proficiency of the system with three samples is always better than the proficiency based on two samples.

Now, we assume that we have n observations. Then

$$\mathcal{P}_n(\mathcal{M}) = \frac{\sigma_w^2 + (n-1)\sigma_a^2}{\sigma_w^4 + n\sigma_w^2\sigma_a^2}. \tag{20}$$

It is immediately clear that when the number of samples n tends to infinity, we have $\lim_{n\to\infty} \mathcal{P}_n(\mathcal{M}) = 1/\sigma_w^2$. This results shows that the maximum possible proficiency is given by the inverse of the variance of the noise in the system. In Fig. 1, we see how the proficiency depends on our knowledge of a (measured by σ_a^2) for three different numbers of observed data and $\sigma_w^2 = 1$. We observe that when the number of samples is 10, the proficiency is not affected much by σ_a^2. The reason is simple: it is expected that the system will learn the unknown a from the 9 previous samples and use it successfully to predict y_{10}. In Fig. 2, we see the dependence of the proficiency on the number of samples for $\sigma_w^2 = 1$ and $\sigma_a^2 = 10$. We observe that the proficiency grows quickly for about the first 10 samples, and then it slows down. For large values of n it gets close to the theoretical limit of the model. $1/\sigma_w^2$.

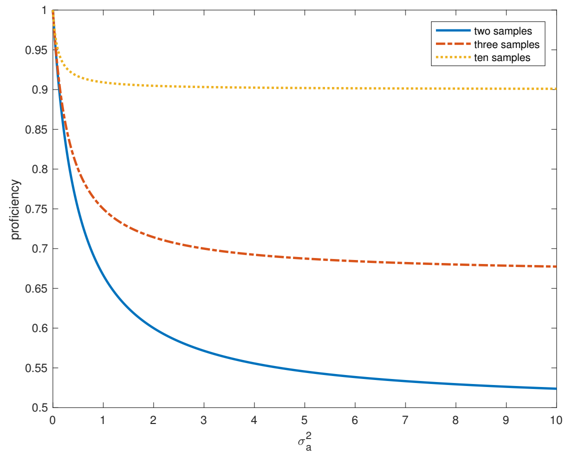

Fig. 1. Proficiency of the model in (11) as a function of σ_a^2.

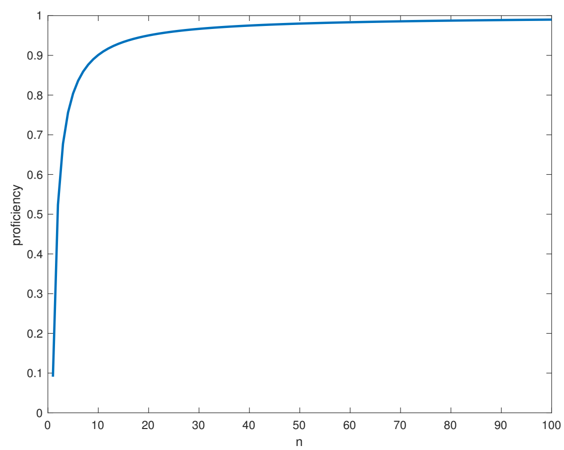

Fig. 2. Proficiency of the model in (11) as a function of the number of samples.

5. PROFICIENCY FOR AUTOREGRESSIVE MODELS

Let the generating model of the data be

$$y_t = ay_{t-1} + w_t, \tag{21}$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2)$. We want to find the in-situ PSA, of the model given $y_{1:t-1}$. It is not difficult to show that

$$\mathcal{P}_t(\mathcal{M}|y_{1:t-1}) = \frac{\sigma_w^2 + h_{t-2}^\top h_{t-2}\sigma_a^2}{\sigma_w^4 + h_{t-1}^\top h_{t-1}\sigma_w^2\sigma_a^2}, \tag{22}$$

where $h_{t-1} = [y_{t-1} \;\; y_{t-2} \cdots \;\; y_1]^\top$. This proficiency is clearly a function of the history of observed data $y_{1:t-1}$. One can also readily see that as $t \to \infty$, we have

$$\lim_{t\to\infty} \mathcal{P}_t(\mathcal{M}|y_{1:t-1}) = \frac{1}{\sigma_w^2}. \tag{23}$$

We note that now $\mathcal{P}_t(\mathcal{M}|y_{1:t-1})$ is not necessarily greater than $\mathcal{P}_{t-1}(\mathcal{M}|y_{1:t-2})$. The reason is that the proficiency depends on past data. However, the proficiency in general increases with time and asymptotically reaches the bound $1/\sigma_w^2$. We observe that this bound is reached when one has perfect knowledge of a.

6. PROFICIENCY IN LINEAR/GAUSSIAN STATE SPACES

Consider the linear state-space model with scalar processes x_t and y_t,

$$x_t = ax_{t-1} + u_t \tag{24}$$
$$y_t = bx_t + v_t, \tag{25}$$

where a and b are known scalars and u_t and v_t are Gaussian noises with known variances σ_u^2 and σ_v^2. Then it is not difficult to show that the proficiency of the system based on the model given by (24) and (25) is given by

$$\mathcal{P}_t(\mathcal{M}|y_{1:t-1}) = \frac{1}{b^2\sigma_{t|t-1}^2 + \sigma_v^2}, \tag{26}$$

where $\sigma_{t|t-1}^2$ is the variance of the predicted state $\widehat{x}_{t|t-1}$ and obtained from

$$\sigma_{t|t-1}^2 = a^2\sigma_{t-1|t-1}^2 + \sigma_u^2, \tag{27}$$

with $\sigma_{t-1|t-1}^2$ being the variance of the filtered state, which is also recursively obtained.

7. CONCLUSIONS

This paper introduced a metric for assessing the degree of proficiency of a system in completing a task in an autonomous manner. The most important factor affecting the capabilities and limitations of a system are the uncertainties within which the system operates. This is the reason why the introduced probabilistic formulation of proficiency self-assessment provides a natural framework. Our metric is based on Bayesian theory, and it quantifies the smallest uncertainty in predicting future observations of a system that operates under a given model. The definition of proficiency allows for several relevant extensions. This work also provides a general methodology for self-assessing the proficiency. We showed on several examples how the proficiency based on given models varies with the number of observations and the parameters of the model, thereby providing insights about the proposed metric. One methodology for computing the assessment is sequential in nature in that the systems reassess their abilities as they acquire new observations. The approach can be further generalized to include information about the environment in which the systems operate.

8. REFERENCES

[1] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadic, "Particle methods for change detection, system identification, and control," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 423–438, 2004.

[2] P. M. Djurić, P. Closas, M. F. Bugallo, and J. Míguez, "Evaluation of a method's robustness," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 3598–3601.

[3] M. Bayarri and J. Morales, "Bayesian measures of surprise for outlier detection," *Journal of Statistical Planning and Inference*, vol. 111, no. 1, pp. 3–22, 2003.

[4] V. Elvira, J. Míguez, and P. M. Djurić, "Adapting the number of particles in sequential Monte Carlo methods through an online scheme for convergence assessment," *IEEE Transactions on Signal Processing*, vol. 65, no. 7, pp. 1781–1794, 2017.

[5] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[6] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Stat. Meth.)*, vol. 64, no. 4, pp. 583–639, 2002.

[7] A. K. Seth, "The cybernetic Bayesian brain," in *Open Mind*. Open MIND. Frankfurt am Main: MIND Group, 2014.

[8] C. Schneider, A. D. Barker, and S. A. Dobson, "Evaluating unsupervised fault detection in self-healing systems using stochastic primitives," *EAI Endorsed Transactions on Self-Adaptive Systems*, 2015.

[9] C. Schneider, A. Barker, and S. Dobson, "Autonomous fault detection in self-healing systems using restricted boltzmann machines," *arXiv preprint arXiv:1501.01501*, 2015.

[10] A. Jauffret, N. Cuperlier, P. Gaussier, and P. Tarroux, "From self-assessment to frustration, a small step toward autonomy in robotic navigation," *Frontiers in Neurorobotics*, vol. 7, p. 16, 2013.

[11] M. T. Cox, "Metareasoning, monitoring, and self-explanation," *Metareasoning: Thinking about thinking*, pp. 131–149, 2011.

[12] M. T. Cox, T. Oates, and D. Perlis, "Toward an integrated metacognitive architecture," in *AAAI Fall Symposium: Advances in Cognitive Systems*, 2011.

[13] K. M. MBalé and D. Josyula, "Integrating metacognition into artificial agents," in *AAAI 2013 Fall Symposium Series*, 2013, pp. 55–62.

[14] M. T. Cox, "Goal-driven autonomy and question-based problem recognition," in *Second Annual Conference on Advances in Cognitive Systems 2013, Poster Collection*, 2013, pp. 29–45.

[15] M. F. Caro, D. P. Josyula, M. T. Cox, and J. A. Jiménez, "Design and validation of a metamodel for metacognition support in artificial intelligent systems," *Biologically Inspired Cognitive Arch.*, vol. 9, pp. 82–104, 2014.

[16] S. M. Fleming and N. D. Daw, "Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation," *Psychological review*, vol. 124, no. 1, p. 91, 2017.

[17] V. Volotka, "Methods of self-diagnosing in telecommunication networks," in *Second International Scientific-Practical Conference on Problems of Infocommunications Science and Technology*, 2015, pp. 131–134.

[18] H. J. Steinhauer, A. Karlsson, G. Mathiason, and T. Helldin, "Root-cause localization using restricted boltzmann machines," in *19th International Conference on Information Fusion*. IEEE, 2016, pp. 248–255.

[19] H. L. Van Trees and K. L. Bell, *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*, 2007, vol. 10.

[20] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1386–1396, 1998.

[21] F. Liang, C. Liu, and R. Carroll, *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. John Wiley & Sons, 2011, vol. 714.

[22] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djurić, "Adaptive importance sampling: The past, the present, and the future," *IEEE Sig. Proc. Mag.*, vol. 34, no. 4, pp. 60–79, 2017.

[23] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, "Particle filtering," *IEEE Sig. Proc. Mag.*, vol. 20, no. 5, pp. 19–38, 2003.

[24] P. M. Djurić and J. Míguez, "Assessment of nonlinear dynamic models by Kolmogorov–Smirnov statistics," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5069–5079, 2010.

[25] C. S. Maiz, E. M. Molanes-Lopez, J. Miguez, and P. M. Djurić, "A particle filtering scheme for processing time series corrupted by outliers," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4611–4627, 2012.