Learning Adaptive Parameter Policies for Nonlinear Bayesian Filtering

Ondřej Straka
European Centre of Excellence NTIS,
University of West Bohemia in Pilsen, Czech Republic
E-mail: straka30@kky.zcu.cz

Felipe Giraldo-Grueso, and Renato Zanetti
Department of Aerospace Engineering and Engineering Mechanics
The University of Texas at Austin, Austin, Texas 78712
Email: {fgiraldo, renato}@utexas.edu

***Abstract*—Algorithms for Bayesian state estimation of nonlinear systems inevitably introduce approximation errors. These algorithms depend on parameters that influence the accuracy of the numerical approximations used. The parameters include, for example, the number of particles, scaling parameters, and the number of iterations in iterative computations. Typically, these parameters are fixed or adjusted heuristically, although the approximation accuracy can change over time with the local degree of nonlinearity and uncertainty. The approximation errors introduced at a time step propagate through subsequent updates, affecting the accuracy, consistency, and robustness of future estimates. This paper presents adaptive parameter selection in nonlinear Bayesian filtering as a sequential decision-making problem, where parameters influence not only the immediate estimation outcome but also the future estimates. The decision-making problem is addressed using reinforcement learning to learn adaptive parameter policies for nonlinear Bayesian filters. Experiments with the unscented Kalman filter and stochastic integration filter demonstrate that the learned policies improve both estimate quality and consistency.**

***Index Terms*—Bayesian estimation; parameter adaptation; Markov decision problem, reinforcement learning; Gaussian assumed filters**

I. Introduction

State estimation for stochastic dynamical systems is traditionally addressed using the Bayesian framework [1], leading to Bayesian recursive relations (BRRs) that enable the calculation of the posterior probability density function (PDF) of the state conditioned on previous available measurements. For linear systems with Gaussian noise, the Kalman filter (KF) provides an analytic recursive solution. Such an analytic solution is available only for a few special cases, usually involving linear systems. For nonlinear systems, a wide range of approximations have been developed, including the extended KF (EKF), unscented KF (UKF) [2], stochastic integration filter (SIF) [3], ensemble-based filters [4]–[6], Gaussian sum filters [7], and particle filters (PFs) [8].

Despite their success, these methods share a common structural property: the state posterior PDF update rule is fixed. Given a previous state posterior PDF and a new observation, the estimator applies a predetermined update operator derived from local approximations or sampling schemes. The update operator is often parametrized by a set of scalar parameters, sampling distribution [9], or grid parameters [10]. Their specification is either fixed in time or adapted based on current measurements [11]–[14] to achieve optimal quality or the smallest possible approximation error with respect to the true posterior PDF. Optimality guarantees, where available, rely on restrictive assumptions such as linearity, Gaussianity, or correct model specification. In practice, these assumptions are often violated. Nonlinear measurement functions, heavy-tailed noise, and model mismatch often result in estimator inconsistency, divergence, or irreversible information loss, such as the premature collapse of multimodal posteriors.

In recent years, machine learning (ML) techniques, particularly reinforcement learning (RL), have been applied to improve the performance of classical estimators. Representative approaches include ML-based adaptation of filter parameters [15], [16], RL-based adaptation of process and measurement noise covariance matrices [17], [18], direct learning of Kalman gains or correction terms [19], neural augmentation of prediction or update steps [20], or end-to-end neural filters trained to output state estimates [21].

These methods have demonstrated empirical improvements in some scenarios, especially under unknown or time-varying noise statistics. However, they retain the underlying estimator structure: the posterior PDF update mechanism remains fixed, and learning is confined to parameter tuning within that structure. These approaches are unable to express non-myopic estimation strategies and cannot explicitly reason about the long-term consequences of posterior probability distribution updates. Improvements are typically measured in terms of instantaneous or short-horizon error metrics such as mean squared error or its robust counterpart correntropy [22], rather than long-term estimator reliability or stability.

In control theory, dynamic programming [23], and RL [24] are standard tools for handling long-horizon decision-making under uncertainty [25]. In contrast, estimation is almost exclusively treated as an inference problem, despite its inherently sequential nature. A small body of theoretical work has explored connections between Bayesian filtering and optimal control in posterior space, but these insights have rarely been exploited to design practical estimation algorithms [26]. In particular, there is no general framework that formulates posterior PDF updates as decisions, optimizes estimator behavior over

The work was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic under project ROBOPROX - Robotics and Advanced Industrial Production CZ.02.01.01/00/22_008/0004590.

long horizons, and explains classical filters as special cases of optimal policies.

This paper aims to take a first step toward formulating the estimator as a decision maker. While keeping the filter update rule fixed, the filter parameters are optimized non-myopically to improve filter consistency. The problem is first formulated for an arbitrary parameter of any nonlinear Bayesian filter. Then, it is tailored to the adaptation of the UKF scaling parameter and the number of SIF iterations.

The paper is structured as follows: Section II provides the formulation of the state estimation problem, introduces the Bayesian and optimization approaches, and discusses approximate solutions. Section III then formulates the Bayesian state estimation as a Markov decision problem (MDP) and shows its myopic property. Section IV then shows that parameter adaptation can be formulated as a non-myopic MDP and presents its solution using the RL framework. Application of the RL framework to parameter adaptation in the Gaussian assumed filter is shown in Section V. Numerical illustration is presented in Section VI and concluding remarks are drawn in Section VII.

II. State estimation problem

A. System model

Consider a discrete-time stochastic system described by a nonlinear state-space model

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k) + \mathbf{w}_k, \tag{1a}$$

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k, \tag{1b}$$

where $\mathbf{x}_k \in \mathbb{R}^{n_x}$ and $\mathbf{z}_k \in \mathbb{R}^{n_z}$ represent the immeasurable state of the system and the available measurement at time instant $k = 0, 1, \ldots$, respectively. The functions $\mathbf{f}_k : \mathbb{R}^{n_x} \mapsto \mathbb{R}^{n_x}$ and $\mathbf{h}_k : \mathbb{R}^{n_x} \mapsto \mathbb{R}^{n_z}$ are assumed known. The state noise $\mathbf{w}_k \in \mathbb{R}^{n_x}$ and measurement noise $\mathbf{v}_k \in \mathbb{R}^{n_z}$ are described by known PDFs $p_{\mathbf{w}_k}$ and $p_{\mathbf{v}_k}$, respectively. The initial state $\mathbf{x}_0$ is given by known PDF $p_{\mathbf{x}_0}$. Both noises are assumed to be white, mutually independent, and independent of the initial state.

The model (1) can be expressed using the transition PDF (2a) and measurement PDF (2b)

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k) = p_{\mathbf{w}_k}(\mathbf{x}_{k+1} - \mathbf{f}_k(\mathbf{x}_k)), \tag{2a}$$

$$p(\mathbf{z}_k|\mathbf{x}_k) = p_{\mathbf{v}_k}(\mathbf{z}_k - \mathbf{h}_k(\mathbf{x}_k)). \tag{2b}$$

B. Bayesian state estimation

The goal of state estimation in the Bayesian framework is to infer the posterior PDF $p(\mathbf{x}_k|\mathbf{z}^k)$ of the state $\mathbf{x}_k$ given all the measurements available up to time k, denoted as $\mathbf{z}^k \coloneqq [\mathbf{z}_1^\intercal, \mathbf{z}_2^\intercal, \cdots, \mathbf{z}_k^\intercal]^\intercal$. The posterior PDF is computed by the BRRs consisting of the Bayes equation (3a) and the Chapman-Kolmogorov equation (CKE) (3b)

$$p(\mathbf{x}_k|\mathbf{z}^k) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}^{k-1})}{\int p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}^{k-1})\mathrm{d}\mathbf{x}_k} \tag{3a}$$

$$p(\mathbf{x}_k|\mathbf{z}^{k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}^{k-1})\mathrm{d}\mathbf{x}_{k-1}. \tag{3b}$$

The initial condition for the BRRs is $p(\mathbf{x}_0|\mathbf{z}^0) = p(\mathbf{x}_0)$. The calculation of the BRRs thus involves alternating the filtering step (3a) and the prediction step (3b). Usually, an approximate solution has to be used to obtain the filtering PDF $p(\mathbf{x}_k|\mathbf{z}^k)$ and the predictive PDF $p(\mathbf{x}_k|\mathbf{z}^{k-1})$.

C. State estimation as an optimization problem

State estimation can also be formulated as a search for a point estimate of $\mathbf{x}_k$ given the available information $\mathbf{z}^k$. The estimate is denoted as $\hat{\mathbf{x}}_k(\mathbf{z}^k)$. If the criterion $J(\tilde{\mathbf{x}}_k)$ to be minimized with $\tilde{\mathbf{x}}_k(\mathbf{z}^k) \coloneqq \mathbf{x}_k - \hat{\mathbf{x}}_k(\mathbf{z}^k)$ representing the estimate error is the mean-squared error (MSE)

$$J_{\text{MSE}}(\tilde{\mathbf{x}}_k) = \mathsf{E}\left[(\tilde{\mathbf{x}}_k)^\intercal \tilde{\mathbf{x}}_k\right], \tag{4}$$

the optimum estimate is the conditional mean $\hat{\mathbf{x}}(\mathbf{z}^k) = \mathsf{E}[\mathbf{x}_k|\mathbf{z}^k]$. Replacing the square of the L2-norm in (4) with the L1-norm yields the median as the optimal solution. Additional alternatives are the correntropy criterion [22] and the Huber cost function [27].

D. Approximate solutions

An analytic solution to (3) can be obtained only for a few special combinations of the model (1) and noise PDFs. Mostly, an approximate solution must be searched for. Assuming the Gaussian joint PDF $p(\mathbf{z}_k, \mathbf{x}_k|\mathbf{z}^{k-1})$ in the numerator of (3a) leads to the Gaussian-assumed filters such as the cubature KF (CKF) [1] or SIF [3]. Assuming a Gaussian mixture distributed joint PDF leads to the Gaussian mixture filter [7] and assuming the joint PDF being Student-t leads to the Student-t filter [28]. Since the paper demonstrates parameter adaptation using Gaussian-assumed filters, its generic algorithm is described in Algorithm 1 for the zero-mean Gaussian noises $\mathbf{w}_k$ and $\mathbf{v}_k$ with covariance matrices $\mathbf{Q}_k$ and $\mathbf{R}_k$, respectively, and Gaussian initial condition $p(\mathbf{x}_0) = \mathcal{N}\{\mathbf{x}_0; \bar{\mathbf{x}}_0, \mathbf{P}_0\}$. The algorithm calculates the mean and covariance matrix of the prediction PDF $p(\mathbf{x}_{k+1}|\mathbf{z}^k) = \mathcal{N}\{\mathbf{x}_{k+1}; \hat{\mathbf{x}}_{k+1|k}, \mathbf{P}^{\mathbf{xx}}_{k+1|k}\}$ and filtering PDF $p(\mathbf{x}_{k+1}|\mathbf{z}^{k+1}) = \mathcal{N}\{\mathbf{x}_{k+1}; \hat{\mathbf{x}}_{k+1|k+1}, \mathbf{P}^{\mathbf{xx}}_{k+1|k+1}\}$

III. State estimation as a Markov decision problem

When formulating the state estimation as a sequential optimization problem, for all time instants $k \geq 0$, we aim to derive an estimator $\hat{\mathbf{x}}_k(\mathbf{z}^k)$ based on the available information. To do this, we introduce a loss function $L(\mathbf{x}_k, \hat{\mathbf{x}}_k(\mathbf{x}^k))$ and seek to minimize the cumulative expected loss defined as follows

$$J = \lim_{T\to\infty} \mathsf{E}_{\mathbf{x}^T, \mathbf{z}^T}\left[\sum_{k=0}^{T} \gamma^k L(\mathbf{x}_k, \hat{\mathbf{x}}_k(\mathbf{z}^k))\right], \tag{8}$$

where $\gamma \in (0, 1]$ is the discount factor.

Since the state $\mathbf{x}_k$ is not available, a reformulation is introduced, replacing the unknown state $\mathbf{x}_k$ by an information state $\mathbf{s}_k$ [29]. The information state $\mathbf{s}_k$ is defined as the conditional PDF

$$\mathbf{s}_k \coloneqq p(\mathbf{x}_k|\mathbf{z}^k). \tag{9}$$

Algorithm 1: Gaussian assumed filter

Initialization:
Set $k \leftarrow 0$
Define the initial condition $p(\mathbf{x}_0|\mathbf{z}^0) = \mathcal{N}\{\mathbf{x}_0; \hat{\mathbf{x}}_{0|0}, \mathbf{P}^{\mathbf{xx}}_{0|0}\}$, with $\hat{\mathbf{x}}_{0|0} = \bar{\mathbf{x}}_0$ and $\mathbf{P}^{\mathbf{xx}}_{0|0} = \mathbf{P}_0$
while *new measurement* $\mathbf{z}_{k+1}$ *available* **do**

Predict:
Compute the state prediction mean and covariance:

$$\hat{\mathbf{x}}_{k+1|k} = \mathsf{E}[\mathbf{f}_k(\mathbf{x}_k)|\mathbf{z}^k]$$
$$\mathbf{P}^{\mathbf{xx}}_{k+1|k} = \mathsf{E}[(\mathbf{f}_k(\mathbf{x}_k) - \hat{\mathbf{x}}_{k+1|k})(\cdot)^\intercal|\mathbf{z}^k] + \mathbf{Q}_k$$

Update:
Compute the measurement prediction mean and covariance:

$$\hat{\mathbf{z}}_{k+1|k} = \mathsf{E}[\mathbf{h}_{k+1}(\mathbf{x}_{k+1})|\mathbf{z}^k] \tag{5}$$
$$\mathbf{P}^{\mathbf{zz}}_{k+1|k} = \mathsf{E}[(\mathbf{h}_{k+1}(\mathbf{x}_{k+1}) - \hat{\mathbf{z}}_{k+1|k})(\cdot)^\intercal|\mathbf{z}^k] + \mathbf{R}_{k+1} \tag{6}$$
$$\mathbf{P}^{\mathbf{xz}}_{k+1|k} = \mathsf{E}[(\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1|k})(\mathbf{h}_{k+1}(\mathbf{x}_{k+1}) - \hat{\mathbf{z}}_{k+1|k})^\intercal|\mathbf{z}^k] \tag{7}$$

Compute the gain $\mathbf{K}_{k+1} = \mathbf{P}^{\mathbf{xz}}_{k+1|k}(\mathbf{P}^{\mathbf{zz}}_{k+1|k})^{-1}$
Update mean using $\tilde{\mathbf{z}}_{k+1|k} \coloneqq (\mathbf{z}_{k+1} - \hat{\mathbf{z}}_{k+1|k})$ as

$$\hat{\mathbf{x}}_{k+1|k+1} = \hat{\mathbf{x}}_{k+1|k} + \mathbf{K}_{k+1}\tilde{\mathbf{z}}_{k+1|k}$$

Update covariance:

$$\mathbf{P}^{\mathbf{xx}}_{k+1|k+1} = \mathbf{P}^{\mathbf{xx}}_{k+1|k} - \mathbf{K}_{k+1}\mathbf{P}^{\mathbf{zz}}_{k+1|k}\mathbf{K}^\intercal_{k+1}$$

$k \leftarrow k + 1$

Note: $(\mathbf{a})(\cdot)^\intercal$ denotes $(\mathbf{a})(\mathbf{a})^\intercal$.

The information state $\mathbf{s}_k$ dynamics is given by

$$\mathbf{s}_{k+1} = \mathbf{\Phi}_{\text{BRR}}(\mathbf{s}_k, \mathbf{z}_{k+1}), \tag{10}$$

where $\mathbf{\Phi}_{\text{BRR}}$ is the Bayesian update operator composed of (3a) and (3b). From (10), it follows that the information state process is Markov. Due to the random nature of the measurement, the transition of the information state is stochastic.

Using the information state, the sequential optimization problem with the cumulative loss (8) can be reformulated as

$$J = \lim_{T\to\infty} \mathsf{E}_{\mathbf{z}^T}\left[\sum_{k=0}^{T} \gamma^k \bar{L}(\mathbf{s}_k, \hat{\mathbf{x}}_k)\right], \tag{11}$$

where

$$\bar{L}(\mathbf{s}_k, \hat{\mathbf{x}}_k) = \mathsf{E}_{\mathbf{x}_k|\mathbf{z}^k}[L(\mathbf{x}_k, \hat{\mathbf{x}}_k)] = \int_{\mathbb{R}^{n_x}} L(\mathbf{x}_k, \hat{\mathbf{x}}_k)\mathbf{s}_k(\mathbf{x}_k)\mathrm{d}\mathbf{x}_k. \tag{12}$$

The solution $\hat{\mathbf{x}}_k(\mathbf{z}^k)$ to the original problem (8) now becomes $\hat{\mathbf{x}}_k(\mathbf{s}_k)$. The solution to the reformulated problem is identical to that of the original problem [29]. This problem is called MDP, more specifically, the partially observable Markov decision problem (POMDP) due to the unavailability of $\mathbf{x}_k$.

For this reformulated problem, the Bellman recursion [23] has the form

$$V(\mathbf{s}_k) = \min_{\hat{\mathbf{x}}_k}\left[\bar{L}(\mathbf{s}_k, \hat{\mathbf{x}}_k) + \gamma\mathsf{E}\left[V\left(\mathbf{\Phi}_{\text{BRR}}(\mathbf{s}_k, \mathbf{z}_{k+1})\right)\right]\right], \tag{13}$$

where $V : \mathcal{S} \to \mathbb{R}$ is the Bellman (value) function. The optimum estimate is then

$$\hat{\mathbf{x}}_k(\mathbf{s}_k) = \arg\min_{\hat{\mathbf{x}}_k}\left[\bar{L}(\mathbf{s}_k, \hat{\mathbf{x}}_k) + \gamma\mathsf{E}\left[V\left(\mathbf{\Phi}_{\text{BRR}}(\mathbf{s}_k, \mathbf{z}_{k+1})\right)\right]\right]. \tag{14}$$

Since the second term in the RHS of (14) does not depend on $\hat{\mathbf{x}}_k$, the estimator becomes myopic

$$\hat{\mathbf{x}}_k(\mathbf{s}_k) = \arg\min \bar{L}(\mathbf{s}_k, \hat{\mathbf{x}}_k) \tag{15}$$

and no multi-step coupling exists. The optimum estimate then corresponds to that computed using the standard non-MDP optimization procedure.

IV. Parameter Adaptation as MDP

For most nonlinear or non-Gaussian models, the Bayesian update (3) is intractable, necessitating an approximation. The approximate posterior generated by an estimator is typically characterized by a set of parameters $\boldsymbol{\theta} \in \Theta$ utilized by that estimator. The quality of the approximation at time k then depends on previous decisions (values of $\boldsymbol{\theta}_t$, $t \leq k$), and similarly, the current parameters $\boldsymbol{\theta}_k$ influence future approximations. The problem of parameter adaptation will now be formulated as an MDP, as in the previous section. However, instead of looking for an estimate $\hat{\mathbf{x}}_k$, we will look for parameters $\boldsymbol{\theta}_k$ assuming a fixed structure of the estimator (e.g., the Gaussian-assumed filter such as the UKF).

A. Specification of information state, parameters, and cost

As the posterior PDF is not available, the information state is assembled from the estimator's internal representation of the state PDF, such as the predictive state mean and covariance, sigma points, or a set of particles. The new information state will be denoted as $\bar{\mathbf{s}}_k \in \bar{\mathcal{S}}$ to distinguish it from $\mathbf{s}_k$. Then, the information state $\bar{\mathbf{s}}_k$ dynamics can be written as[1]

$$\bar{\mathbf{s}}_{k+1} = \mathbf{\Phi}_{\text{filter}}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k). \tag{16}$$

In the MDP context, the estimator's parameters now serve as actions that affect the future information state. They are generated by a *policy* $\pi : \bar{\mathcal{S}} \times \mathbb{R}^{n_z} \to \Theta$ that assigns a parameter $\boldsymbol{\theta}_k$ to each decision situation at time k characterized by the information state and current measurement:

$$\boldsymbol{\theta}_k \sim \pi(\bar{\mathbf{s}}_k, \mathbf{z}_k). \tag{17}$$

The parameters $\boldsymbol{\theta}_k$ may include the number of PF samples, the model selection criteria, or the UKF scaling parameter.

The parameters $\boldsymbol{\theta}_k$ enter the cost $\bar{L}$ through the estimate $\hat{\mathbf{x}}_k(\boldsymbol{\theta}_k)$ $\bar{L}(\bar{\mathbf{s}}_k, \hat{\mathbf{x}}_k(\boldsymbol{\theta}_k))$. The specification of such a cost may be tricky, as the reference value for the estimate $\hat{\mathbf{x}}_x(\boldsymbol{\theta}_k)$ is calculated by the same estimator. A more convenient form of the cost function is defined in terms of the measurement-related quantities such as the measurement $\mathbf{z}_k$, its prediction depending on the parameters $\hat{\mathbf{z}}_{k|k-1}(\boldsymbol{\theta}_k)$, and the covariance $\mathbf{P}^{\mathbf{zz}}_{k|k-1}(\boldsymbol{\theta}_k)$. Such cost can be denoted

[1]Note that, in contrast to (10), the information state $\bar{\mathbf{s}}_{k+1}$ depends on $\mathbf{z}_k$ as $\bar{\mathbf{s}}_{k+1}$ is a predictive estimate while $\mathbf{s}_{k+1}$ represented a filtering estimate.

by $\bar{L}(\hat{\mathbf{z}}_{k|k-1}(\boldsymbol{\theta}_k), \mathbf{P}^{\mathbf{zz}}_{k|k-1}(\boldsymbol{\theta}_k), \mathbf{z}_k)$. However, for the sake of convenience, the notation $\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k)$ expressing the dependence of the cost on the information state, the measurement, and parameters will be used in the sequel.

In general, the cost $\bar{L}$ may value estimate *quality* given by a norm of measurement prediction error $\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k) = \|\tilde{\mathbf{z}}_{k|k-1}(\boldsymbol{\theta}_k)\|$ or *consistency* given by normalized innovation squared (NIS) $\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k) = \tilde{\mathbf{z}}_{k|k-1}(\boldsymbol{\theta}_k)^\intercal (\mathbf{P}^{\mathbf{zz}}_{k|k-1}(\boldsymbol{\theta}_k))^{-1} \tilde{\mathbf{z}}_{k|k-1}(\boldsymbol{\theta}_k)$ or *robustness* given by penalizing large estimate errors, computational costs, or by a combination thereof.

With the above considerations, the Bellman recursion for the parameter adaptation has the form

$$V(\bar{\mathbf{s}}_k, \mathbf{z}_k) = \min_{\boldsymbol{\theta}_k} \big[\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k) + \gamma\, \mathsf{E}\left[V\left(\boldsymbol{\Phi}_{\text{filter}}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k), \mathbf{z}_{k+1}\right)\right]\big]. \tag{18}$$

The expectation on the right-hand side of (18) is due to the presence of the future measurement $\mathbf{z}_{k+1}$ and also due to the cases when the filter information state dynamics (16) contains random effects, which is the case of PFs. The optimal adaptive parameter policy π providing $\boldsymbol{\theta}_k$ is then

$$\pi(\bar{\mathbf{s}}_k, \mathbf{z}_k) = \arg\min_{\boldsymbol{\theta}_k} \big[\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k). + \gamma\, \mathsf{E}\left[V\left(\boldsymbol{\Phi}_{\text{filter}}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k), \mathbf{z}_{k+1}\right)\right]\big]. \tag{19}$$

By this (PO)MDP formulation, the estimation problem becomes sequentially coupled, adaptive, and non-myopic. It accounts for the fact that approximation quality depends on earlier parameter choices.

In contrast, the classical approach to state estimation is inherently myopic. It focuses on obtaining the best possible estimate at each moment in time by calculating an approximate estimate that is closest to the optimal value, such as the conditional mean in the case of MSE. In the classical formulation, one cannot worsen the current estimate to achieve a substantial improvement in the future.

B. Reinforcement learning solution to MDP

Analytical solution to (19) is not tractable, and in this paper, an ML solution is chosen. In particular, the RL approach is adopted, which falls under unsupervised ML. We employ the on-policy actor-critic algorithm with temporal-difference learning TD(0) [30] with the following assumptions: The parameter space Θ is assumed discrete $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}^{(i)}, i = 1, \dots |\Theta|\}$.

The adaptive parameter policy $\pi(\bar{\mathbf{s}}_k, \mathbf{z}_k)$ (actor) is assumed stochastic, given by a conditional probability distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\bar{\mathbf{s}}_k, \mathbf{z}_k)$ and represented by a neural network (NN)

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\bar{\mathbf{s}}_k, \mathbf{z}_k; \phi) = \text{softmax}(f_\phi(\bar{\mathbf{s}}_k, \mathbf{z}_k)), \tag{20}$$

where $f_\phi(\cdot)$ is a feedforward multilayer perceptron (MLP) NN parametrized by ϕ. The Bellman function V (critic) is approximated by another MLP NN $V_\psi(\bar{\mathbf{s}}_k, \mathbf{z}_k)$ parametrized by ψ.

Both actor and critic are updated based on the temporal-difference (TD) error computed as (c.f. (18))

$$\delta_k = \bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k) + \gamma\, V_\psi(\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1}) - V(\bar{\mathbf{s}}_k, \mathbf{z}_k). \tag{21}$$

The TD error is the discrepancy between the current estimate $V(\bar{\mathbf{s}}_k, \mathbf{z}_k)$ and one-step bootstrap target $\bar{L}(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k) + \gamma\, V_\psi(\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1})$. The critic then minimizes the squared error

$$\mathcal{L}_{\text{critic}} = \delta_k^2 \tag{22}$$

to learn the Bellman function. The actor is updated to minimize the policy gradient objective

$$\mathcal{L}_{\text{actor}} = -\delta_k \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\bar{\mathbf{s}}_k, \mathbf{z}_k; \phi). \tag{23}$$

To stabilize the TD learning process, a target critic network is used. Instead of bootstrapping from the current critic $V_\psi(\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1})$, the TD target is computed using a lagged copy $V_{\bar{\psi}}(\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1})$. This prevents the critic from chasing a moving target, mitigates divergence due to function approximation, and reduces variance in the value update. The target parameters are updated via soft (Polyak) averaging [31].

V. Algorithm for Parameter Adaptation

The idea of parameter adaptation for nonlinear Bayesian filters is general and can be applied to any algorithm, such as the PF, for adaptation of the particle number or the proposal density, the Gaussian mixture filter [7] for the specification of the number of Gaussian components, or the Gaussian assumed filters such as the UKF for the specification of the scaling parameter, or for the SIF for specification of the number of iterations.

The algorithm for offline learning for the Gaussian assumed filter is described in Algorithm 2. To simplify the exposition, only parameters appearing in the update step, i.e., the calculation of the predictive measurement moments (5-7) are adapted. The steps labeled 1 and 2 correspond to the predict and update steps in Algorithm 1.

After learning the policy $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\bar{\mathbf{s}}_k, \mathbf{z}_k; \phi)$, the Gaussian assumed filter with parameter adaptation can be used. Its algorithm is described in Algorithm 3.

Notice that during offline learning, the parameter is sampled from the policy, while during online estimation, the greedy parameter value with respect to the policy is selected.

VI. Numerical illustration

The proposed learning of adaptive parameter policy is demonstrated for the UKF scaling parameter and the SIF number of iterations. For the demonstration, two estimation problems will be considered: the univariate nonstationary growth model (UNGM) used often for benchmarking due to its strong nonlinearity and multimodal posterior, and the coordinated turn model (CTM) with bearing measurements. Both models are specified first.

Algorithm 2: Offline Monte-Carlo Actor–Critic Training for Adaptive Parameter Policy

Input: Discrete parameter set $\Theta = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(|\Theta|)}\}$, discount factor γ, number of episodes N_{MC}, horizon T, hyperparameter τ, actor $\pi_\theta(\theta|\bar{\mathbf{s}}, \mathbf{z}; \phi)$, critic $V_\psi(\bar{\mathbf{s}}, \mathbf{z})$
Output: Trained actor parameters ϕ and critic parameters ψ

Initialize actor parameters ϕ, critic parameters ψ, and target critic parameters $\bar{\psi} \leftarrow \psi$
for $m = 1$ **to** N_{MC} **do**
  Reset simulator and filter state $(\hat{\mathbf{x}}_{0|0}, \mathbf{P}^{\mathbf{xx}}_{0|0})$
  for $k = 1$ **to** $T - 1$ **do**
1 $(\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}^{\mathbf{xx}}_{k|k-1}) \leftarrow \text{Predict}(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{P}^{\mathbf{xx}}_{k-1|k-1})$
    Construct $\bar{\mathbf{s}}_k$ from $\hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}^{\mathbf{xx}}_{k|k-1}$
    Sample $\boldsymbol{\theta}_k \sim \pi_{\boldsymbol{\theta}}(\cdot|\bar{\mathbf{s}}_k, \mathbf{z}_k; \phi)$
2 $(\hat{\mathbf{x}}_{k|k}, \mathbf{P}^{\mathbf{xx}}_{k|k}, \tilde{\mathbf{z}}_{k|k-1}, \mathbf{P}^{\mathbf{zz}}_{k|k-1}) \leftarrow$ Update$(\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}^{\mathbf{xx}}_{k|k-1}, \boldsymbol{\theta}_k)$
    Compute cost $L_k(\bar{\mathbf{s}}_k, \mathbf{z}_k, \boldsymbol{\theta}_k)$
    $(\hat{\mathbf{x}}_{k+1|k}, \mathbf{P}^{\mathbf{xx}}_{k+1|k}) \leftarrow \text{Predict}(\hat{\mathbf{x}}_{k|k}, \mathbf{P}^{\mathbf{xx}}_{k|k})$
    Construct $\bar{\mathbf{s}}_{k+1}$ from $\hat{\mathbf{x}}_{k+1|k}$ and $\mathbf{P}^{\mathbf{xx}}_{k+1|k}$
    $\delta_k \leftarrow L_k(\boldsymbol{\theta}_k, \bar{\mathbf{s}}_k, \mathbf{z}_k) + \gamma V_{\bar{\psi}}(\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1}) - V_\psi(\bar{\mathbf{s}}_k, \mathbf{z}_k)$
    $\mathcal{L}_{\text{critic}} \leftarrow \delta_k^2$
    $\mathcal{L}_{\text{actor}} \leftarrow -\delta_k \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_k|\bar{\mathbf{s}}_k, \mathbf{z}_k; \phi)$
    Update ψ and ϕ using Adam optimizer
    Update target critic parameters: $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$

Algorithm 3: Gaussian assumed filter with adaptive parameter policy

Input: Adaptive parameter policy $\pi_\theta(\theta|\bar{\mathbf{s}}, \mathbf{z}; \phi)$
Initialization:
Set $k \leftarrow 0$
Define the initial condition

$$p(\mathbf{x}_0|\mathbf{z}^0) = p(\mathbf{x}_0) = \mathcal{N}\{\mathbf{x}_0; \hat{\mathbf{x}}_{0|0}, \mathbf{P}^{\mathbf{xx}}_{0|0}\},$$

with $\hat{\mathbf{x}}_{0|0} = \bar{\mathbf{x}}_0$ and $\mathbf{P}^{\mathbf{xx}}_{0|0} = \mathbf{P}_0$
while *new measurement* $\mathbf{z}_{k+1}$ *available* **do**
  Predict:
  Compute the state prediction mean $\hat{\mathbf{x}}_{k+1|k}$ and covariance $\mathbf{P}^{\mathbf{xx}}_{k+1|k}$ as in Algorithm 1.
  Construct $\bar{\mathbf{s}}_{k+1}$ from $\hat{\mathbf{x}}_{k+1|k}$ and $\mathbf{P}^{\mathbf{xx}}_{k+1|k}$
  Select $\boldsymbol{\theta}^*_{k+1} = \arg\max_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\bar{\mathbf{s}}_{k+1}, \mathbf{z}_{k+1}; \phi)$
  Update:
  Compute the measurement prediction mean $\hat{\mathbf{z}}_{k+1|k}$ and covariances $\mathbf{P}^{\mathbf{zz}}_{k+1|k}$ and $\mathbf{P}^{\mathbf{xz}}_{k+1|k}$ using the optimum parameter $\boldsymbol{\theta}^*_{k+1}$.
  Subsequently, compute the gain $\mathbf{K}_{k+1}$ and update the mean $\hat{\mathbf{x}}_{k+1|k+1}$ and covariance $\mathbf{P}^{\mathbf{xx}}_{k+1|k+1}$ as in Algorithm 1.
  $k \leftarrow k + 1$

A. Univariate Nonstationary Growth Model

The UNGM is a scalar nonlinear state–space model

$$x_k = f(x_{k-1}, k-1) + w_{k-1}, \qquad w_{k-1} \sim \mathcal{N}\{0, Q\}, \quad (24)$$

with $f(x, k) = 0.5x + \frac{25x}{1+x^2} + 8\cos(0.05\,k)$ and measurement equation

$$y_k = \frac{x_k^2}{20} + v_k, \qquad v_k \sim \mathcal{N}\{0, R\}. \quad (25)$$

The model was simulated for $k = 1, \dots 500$ with Gaussian initial state $x_1 \sim \mathcal{N}\{0,\, 5)$ and variances $Q = 1$ and $R = 0.1$.

B. Coordinated Turn Model

The CTM considers a four-dimensional state $\mathbf{x}_k = [\mathsf{x}_k, \dot{\mathsf{x}}_k, \mathsf{y}_k, \dot{\mathsf{y}}_k]^\intercal \in \mathbb{R}^4$ consisting of positions $[\mathsf{x}, \mathsf{y}]$ and velocities $[\dot{\mathsf{x}}, \dot{\mathsf{y}}]$ in 2D space. It evolves according to a linear dynamics with additive Gaussian noise

$$\mathbf{x}_k = \mathbf{F}\,\mathbf{x}_{k-1} + \mathbf{w}_{k-1}, \qquad \mathbf{w}_{k-1} \sim \mathcal{N}\{\mathbf{0}, \mathbf{Q}\}, \quad (26)$$

where the transition matrix is

$$\mathbf{F} = \begin{bmatrix} 1 & \frac{\sin(\omega\Delta t)}{\omega} & 0 & -\frac{1-\cos(\omega\Delta t)}{\omega} \\ 0 & \cos(\omega\Delta t) & 0 & -\sin(\omega\Delta t) \\ 0 & \frac{1-\cos(\omega\Delta t)}{\omega} & 1 & \frac{\sin(\omega\Delta t)}{\omega} \\ 0 & \sin(\omega\Delta t) & 0 & \cos(\omega\Delta t) \end{bmatrix} \quad (27)$$

with $\Delta t = 1$ s denoting the sampling period and ω being a known constant turn rate. The process noise covariance is

$$\mathbf{Q} = q\left(\mathbf{I}_2 \otimes \begin{bmatrix} \frac{\Delta t^3}{3} & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & \Delta t \end{bmatrix}\right), \quad (28)$$

with $\mathbf{I}_2$ being the identity matrix and q being noise intensity. The position is measured via bearing and is given by

$$y_k = \operatorname{atan2}\big(\mathsf{y}_k, \mathsf{x}_k\big) + v_k, \qquad v_k \sim \mathcal{N}\{0, R\}. \quad (29)$$

Note that the innovation $\tilde{\mathbf{z}}_{k|k-1}$ in the filter is wrapped to the interval $(-\pi, \pi]$.

The model was simulated for $k = 1, \dots 150$ with $\Delta t = 1$, $\omega = 0.5$, intensity $q = 1$, variance $R = 0.04$, and Gaussian initial state

$$\mathbf{x}_1 \sim \mathcal{N}\{[80,\, 0,\, 0\,20]^\intercal, \operatorname{diag}([10^3,\, 10^2,\, 10^3,\, 10^2])\}, \quad (30)$$

where diag denotes a diagonal matrix.

C. Actor critic RL parameters

The actor was a feed-forward NN with two 64-unit ReLU hidden layers mapping $\bar{\mathbf{s}}_k$ and $\mathbf{z}_k$ to a softmax distribution over discrete parameter choices. The critic was a feed-forward NN with two 64-unit ReLU hidden layers. The actor and critic were optimized with Adam using learning rates of 10^{-4} and 5×10^{-4}, respectively, with gradient decay 0.9 and squared-gradient decay 0.999. A target critic with soft updates ($\tau = 0.01$) was employed for stabilization, and entropy regularization with coefficient 10^{-3} was used to promote exploration. The forgetting factor was set to $\gamma = 0.5$. The NNs were trained using $N_{\mathrm{MC}} = 10^3$ simulations.

D. Algorithms

The learning of adaptive parameter policies is demonstrated for the adaptation of *(i)* the scaling parameter κ of the UKF and *(ii)* the number of iterations N_{it} used in the SIF. In the case of the UKF, the computational cost was not included in the criterion since all choices of κ are equivalent in terms of computational complexity. On the other hand, the number of iterations used by the SIF to compute the predictive measurement moments $\hat{\mathbf{z}}_{k+1|k}$, $\mathbf{P}^{\mathbf{zz}}_{k+1|k}$, $\mathbf{P}^{\mathbf{xz}}_{k+1|k}$ has a significant impact on the computational complexity, which was, thus, included in the optimized cost.

The information state $\bar{\mathbf{s}}_k$ for both cases consisted of the predictive moments computed by the algorithm, the measurement[2], and the innovation $\tilde{\mathbf{z}}_{k|k-1}$ and was defined as

$$\bar{\mathbf{s}}_k = [\hat{\mathbf{x}}_{k|k-1}^\intercal,\, \mathrm{tr}(\mathbf{P}_{k|k-1}^{\mathbf{xx}}),\, \log\det(\mathbf{P}_{k|k-1}^{\mathbf{xx}}),\, \mathbf{z}_k^\intercal,\, \tilde{\mathbf{z}}_{k|k-1}^\intercal]^\intercal. \tag{31}$$

Several baseline settings were considered to analyze the performance of the adaptive parameter policy.

- First, a **default** baseline employing the parameter value typically specified in the target application, representing standard practice where the filter is deployed with a fixed, heuristically chosen tuning. For the UKF, it is $\kappa = \max(0, 3 - n_x)$, for the SIF, it is $N_{\text{it}} = 10$.
- Second, a set of **fixed** parameter baselines across various constant values providing a controlled overview of performance, highlighting the sensitivity of estimation accuracy and consistency to tuning choices.
- Third, a **myopic** baseline applying the same optimization criterion as the adaptive method but restricting the decision to a single-step horizon ($\gamma = 0$), thereby isolating the effect of non-myopic reasoning and long-term trade-offs.
- Fourth, an **optimal** baseline obtained by selecting the parameter that maximizes the likelihood [32]. Note that this baseline applies only to UKF since for the SIF, the optimal choice is always the maximum number of iterations that is possible.

Collectively, these baselines establish a range of reference operating points, from conventional heuristic tuning to optimal values, enabling a systematic assessment of the benefits of adaptive, non-myopic parameter adaptation. Note that the fixed and default baselines used fixed parameter values for all time instants, whereas the myopic and optimal baselines, and the proposed parameter adaptation, use parameters that vary over time according to their respective criteria.

The adaptation considered the following costs:

- NIS cost penalizing consistency

$$\bar{L}_{\text{NIS}} = (\text{NIS} - n_z)^2, \tag{32}$$

  where $\text{NIS} = \tilde{\mathbf{z}}_{k|k-1}^\intercal (\mathbf{P}_{k|k-1}^{\mathbf{zz}})^{-1} \tilde{\mathbf{z}}_{k|k-1}$ (for consistent measurement prediction estimates, $\mathsf{E}[\text{NIS}] = n_z$)
- a NIS-based cost penalizing only optimistic estimates

$$\bar{L}_{\text{logmaxNIS}} = \log\left(1 + \max\left(0, (\text{NIS} - n_z)\right)\right), \tag{33}$$

- a quality-oriented cost as an L_2 norm of the state innovation

$$\bar{L}_{\text{stateInnov}} = \|\mathbf{K}_k \tilde{\mathbf{z}}_{k|k-1}\|. \tag{34}$$

[2]Remind that the measurement $\mathbf{z}_k$ is an argument of the Bellman function V.

E. Performance criteria

The performance of the proposed parameter adaptation and baselines was measured using the root mean-squared error (RMSE) defined as

$$\text{RMSE}_k = \sqrt{\frac{1}{M} \sum_{\ell=1}^{M} \tilde{\mathbf{x}}_{k|k}(\ell)^\intercal \, \tilde{\mathbf{x}}_{k|k}(\ell)} \tag{35}$$

based on $M = 10^4$ Monte Carlo (MC) simulations with $\tilde{\mathbf{x}}_{k|k}(\ell) \coloneqq \mathbf{x}_k(\ell) - \hat{\mathbf{x}}_{k|k}(\ell)$ being the estimate error, $\mathbf{x}_k(\ell)$ being the true state at ℓ-th MC simulation and $\hat{\mathbf{x}}_{k|k}(\ell)$ being its filtering estimate. To assess higher-order information provided by the filters, the average normalized estimate error squared (ANEES) [33] defined by

$$\text{ANEES}_k = \frac{1}{M} \sum_{\ell=1}^{M} \tilde{\mathbf{x}}_{k|k}(\ell)^\intercal \, \mathrm{cov}[\mathbf{x}_k | \mathbf{z}^k]^{-1} \, \tilde{\mathbf{x}}_{k|k}(\ell) \tag{36}$$

is used. ANEES assesses the consistency of the estimator, i.e., alignment of the conditional covariance matrix $\mathrm{cov}[\mathbf{x}_k|\mathbf{z}^k]$ and the estimate error $\tilde{\mathbf{x}}_{k|k}$. This value should be close to n_x. Higher ANEES values mean that the estimator is too optimistic, while values smaller than one mean the estimator is too pessimistic.

Note that the values of the costs $\bar{L}$ described above are only loosely connected to ANEES. The costs are defined in measurement space and use the actual value of the measurement as a reference value and its prediction. The ANEES, on the other hand, is defined in the state space and uses the true value of the state and the filtering estimate.

F. Results

Time averages of RMSE and ANEES for the UNGM and UKF are given in Figures 1 and 2. In addition, the RMSE–ANEES plot 3 visualizes the trade-off between estimation accuracy and statistical consistency by mapping each filter configuration to a point whose horizontal position reflects time-averaged RMSE and vertical position reflects time-averaged ANEES. Analogously, the results for the CTM and UKF are given in Figures 4, 5, and 6.

In both UNGM and CTM problems, the proposed parameter adaptation achieves the best consistency in terms of ANEES compared to the myopic, optimal, and default baselines. In terms of RMSE, the proposed parameter adaptation achieves the lowest values compared to the myopic, optimal, and default baselines. The same estimate accuracy was achieved by the UKF with one of the fixed parameter values. This value, however, was not known in advance. Also, it must be noted that the parameter adaptation used the NIS cost that values consistency in the measurement space, which is rather closer to state consistency assessed by ANEES than to accuracy RMSE. Both UNGM and CTM problems also demonstrated better performance of parameter adaptation with longer horizon (despite a relatively low forgetting factor $\gamma = 0.5$) compared to the myopic baseline that corresponds to $\gamma = 0$.

The SIF was applied to CTM only since UNGM has one-dimensional state and SIF reduces to an MC KF in such a case.

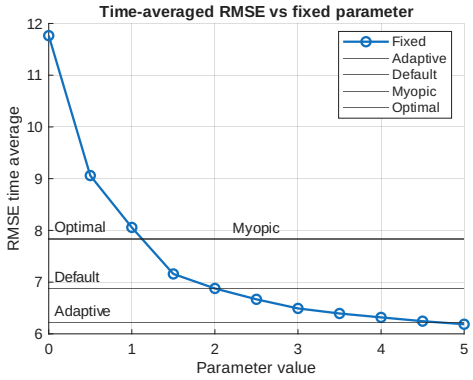

Figure 1. Time averaged RMSE for UKF and UNGM.

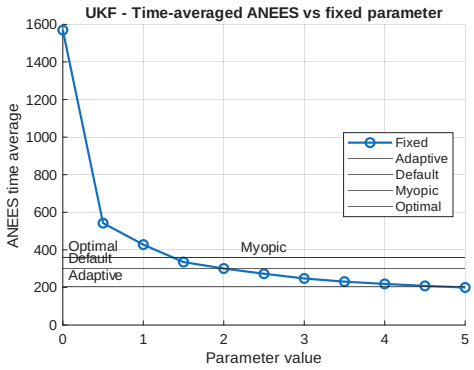

Figure 2. Time averaged ANEES for UKF and UNGM.

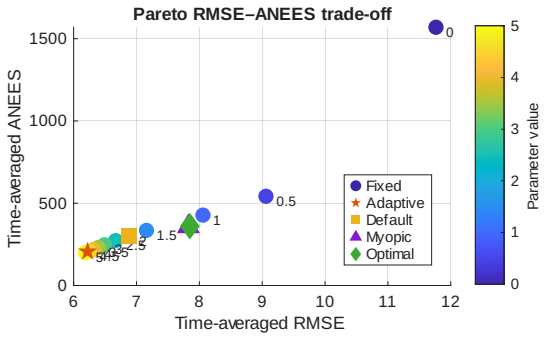

Figure 3. Time averaged ANEES vs. RMSE for UKF and UNGM.

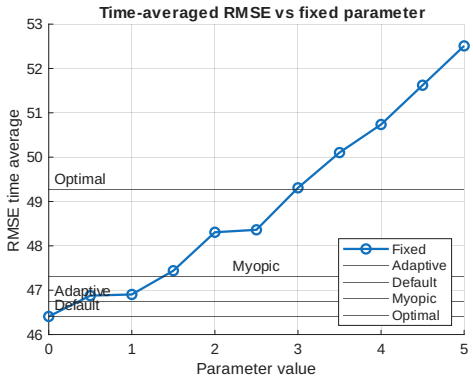

Figure 4. Time averaged RMSE for UKF and CTM.

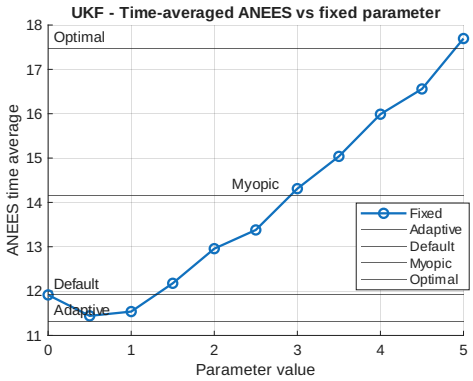

Figure 5. Time averaged ANEES for UKF and CTM.

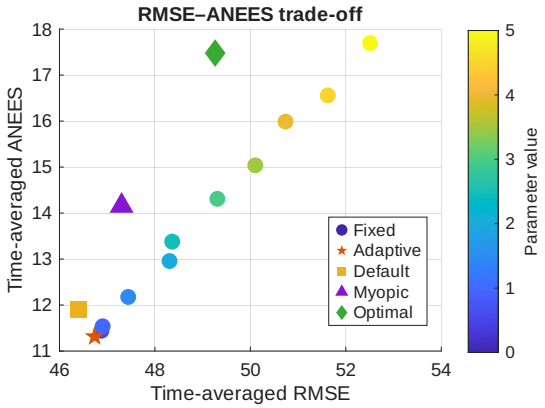

Figure 6. Time averaged RMSE vs. ANEES for UKF and CTM.

The results are given in Table I. In this case, only two baselines are used: the one with myopic optimization and the default parameter. The reason is that, in this case, the computational cost is included in the cost, and the computational cost weight can be adjusted to value either quality or cost. A low weight leads to an optimum N_{it} close to the maximum number of iterations, while a high weight results in the optimum N_{it} near the minimum number of iterations. The results again indicate that parameter adaptation leads to better accuracy and consistency than the myopic and fixed-parameter baselines. The last row of the table also contains the time-average value of the corresponding cost evaluated myopically (i.e., over a single time step). In contrast to RMSE and ANEES, this cost also accounts for computational costs, as shown in the column labels. Here, the parameter adaptation also yields values smaller than those from myopic adaptation. While this situation may seem counterintuitive, it is caused by the myopic setting selecting the best value of the parameter, taking into account a single time instant, without regard for the future. Such greedy optimization cannot achieve the minimum cost over a long horizon, which is 500 steps in this case.

VII. Conclusion

In this paper, we investigated parameter adaptation for non-linear state estimation from a decision-theoretic perspective,

Table I
SIF FOR CTM: TIME-AVERAGED PERFORMANCE RMSE, ANEES, AND INSTANTANEOUS COST $\bar{L}$ FOR THREE COST FUNCTIONS.

	$\bar{L}_{\text{stateInnov}} + \frac{1}{50} N_{\text{it}}$			$\bar{L}_{\text{logmaxNIS}} + \frac{1}{50} N_{\text{it}}$			$\bar{L}_{\text{NIS}} + \frac{1}{50} N_{\text{it}}$		
Metric	Adaptive	Fixed	Myopic	Adaptive	Fixed	Myopic	Adaptive	Fixed	Myopic
Time Avg. RMSE	5.630	6.013	5.778	6.866	6.462	6.771	6.359	6.537	6.646
Time Avg. ANEES	14.369	18.389	16.987	7.379	8.024	8.195	7.051	8.271	7.483
Time Avg. cost	49.872	62.251	51.770	1.809	1.881	1.906	6.140×10^7	6.356×10^7	$1.071 \times^8$

treating it as a sequential optimization over the filter parameter values. By integrating reinforcement learning with classical filtering, the proposed approach enables a non-myopic parameter adaptation that explicitly balances estimation accuracy, consistency, and computational costs. Empirical results from various cost forms show that adaptive parameter policies consistently outperform traditional fixed-parameter baselines and enhance myopic optimization. These findings indicate that treating filter parameter adaptation as a sequential decision problem offers a systematic and practical approach to robustly deploying nonlinear estimators.

REFERENCES

[1] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[2] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *IEEE Proceedings*, vol. 92, no. 3, pp. 401–421, 2004.

[3] J. Duník, O. Straka, and M. Šimandl, "Stochastic integration filter," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1561–1566, 2013.

[4] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," *Journal of Geophysical Research: Oceans*, vol. 99, no. C5, pp. 10 143–10 162, 1994.

[5] J. Anderson and S. Anderson, "A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts," *Monthly Weather Review*, vol. 127, no. 12, pp. 2741–2758, 1999.

[6] S. Yun, R. Zanetti, and B. A. Jones, "Kernel-based ensemble Gaussian mixture filtering for orbit determination with sparse data," *Advances in Space Research*, vol. 69, no. 12, pp. 4179–4197, June 2022, https://doi.org/10.1016/j.asr.2022.03.041.

[7] O. Straka and U. D. Hanebeck, "Efficient Gaussian mixture filters based on transition density approximation," in *2025 28th International Conference on Information Fusion (FUSION)*, 2025, pp. 1–8.

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

[9] M. Simandl and O. Straka, "Sampling density design for particle filters," in *Proceedings of the 13th IFAC Symposium on System Identification*, Rotterdam, 2003.

[10] J. Matoušek, J. Duník, and M. Brandner, "Design of efficient point-mass filter with illustration in terrain aided navigation," in *26th International Conference on Information Fusion (FUSION)*, Charleston, USA, 2023.

[11] O. Straka, J. Duník, and M. Šimandl, "Unscented Kalman filter with advanced adaptation of scaling parameter," *Automatica*, vol. 50, no. 10, pp. 2657–2664, 2014.

[12] J. Dunik, M. Simandl, and O. Straka, "Unscented Kalman filter: Aspects and adaptive setting of scaling parameter," *IEEE Transactions on Automatic Control*, vol. 57, no. 9, pp. 2411–2416, 2012.

[13] J. Havlík, O. Straka, J. Duník, and J. Ajgl, "On nonlinearity measuring aspects of stochastic integration filter," in *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics*, ser. ICINCO 2016. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda, 2016, p. 353–361.

[14] A. A. Popov and R. Zanetti, "An adaptive covariance parameterization technique for the ensemble Gaussian mixture filter," *SIAM Journal on Scientific Computing*, vol. 46, no. 3, 2024.

[15] Z. Fan, D. Shen, Y. Bao, K. Pham, E. Blasch, and G. Chen, "RNN-UKF: Enhancing hyperparameter auto-tuning in unscented Kalman filters through recurrent neural networks," in *2024 27th International Conference on Information Fusion (FUSION)*, 2024, pp. 1–8.

[16] L. A. Scardua and J. J. da Cruz, "Complete offline tuning of the unscented Kalman filter," *Automatica*, vol. 80, pp. 54–61, 2017.

[17] Z. Bekhtaoui, A. Meche, M. Dahmani, and K. A. Meraim, "Maneuvering target tracking using q-learning based kalman filter," in *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)*, 2017, pp. 1–5.

[18] G. Shaaban, H. Fourati, C. Prieur, and A. Kibangou, "Q-learning-based noise covariance matrices adaptation in Kalman filter for inertial navigation," *IFAC-PapersOnLine*, vol. 58, no. 21, pp. 96–101, 2024, 4th IFAC Conference on Modelling, Identification and Control of Nonlinear Systems MICNON 2024.

[19] L. Hu, C. Wu, and W. Pan, "Lyapunov-based reinforcement learning state estimator," 2021. [Online]. Available: https://arxiv.org/abs/2010.13529

[20] G. Revach, N. Shlezinger, R. J. G. van Sloun, and Y. C. Eldar, "Kalmannet: Data-driven kalman filtering," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3905–3909.

[21] A. Ghosh, A. Honoré, and S. Chatterjee, "Danse: Data-driven non-linear state estimation of model-free process in unsupervised learning setup," *IEEE Transactions on Signal Processing*, vol. 72, pp. 1824–1838, 2024.

[22] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Pri´ncipe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.

[23] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2007.

[24] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[25] O. Straka and I. Punčochář, "Distributed design for active fault diagnosis," *International Journal of Systems Science*, vol. 53, no. 3, pp. 562–574, 2022.

[26] M. J. A. Strens, "A bayesian framework for reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, p. 943–950.

[27] C. D. Karlgaard and H. Schaub, "Huber-based divided difference filtering," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 3, pp. 885–891, 2007.

[28] O. Straka and J. Duník, "Stochastic integration student's-t filter," in *2017 20th International Conference on Information Fusion (Fusion)*, 2017.

[29] K. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174–205, 1965.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[31] J. Adamczyk, V. Makarenko, S. Tiomkin, and R. V. Kulkarni, "Average-reward soft actor-critic," 2025. [Online]. Available: https://arxiv.org/abs/2501.09080

[32] J. Duník, M. Šimandl, and O. Straka, "Adaptive choice of scaling parameter in derivative-free local filters," in *2010 International Conference on Information Fusion*, Edinburgh, Great Britain, 2010.

[33] X. R. Li and Z. Zhao, "Measuring estimator's credibility: Noncredibility index," in *Proceedings of 2006 International Conference on Information Fusion*, Florence, Italy, 2006.