Nonlinear Bayesian Estimator for Parameter Learning: A Fixed-Point Characterization

Sasan Vakili[1], Daniël Woonings[1], Pradyumna Paruchuri[1], and Peyman Mohajerin Esfahani[1,2]

[1]Delft University of Technology, The Netherlands
[2]University of Toronto, Canada

Abstract. This paper presents a nonlinear parameter estimator for Wiener-type state-space models obtained as a fixed-point architecture that couples two affine minimum mean-squared error (MMSE) estimators: one for the unknown parameters and one for latent variables. The architecture retains the functional structure of the optimal affine MMSE parameter estimator while incorporating Dynamic Basis Statistics (DBS) estimates that summarize nonlinear basis-function evaluations. Two DBS construction strategies are developed, leading to two nonlinear estimator frameworks. The dual basis-parameter estimator combines an affine basis estimator with the affine parameter estimator, whereas the dual state-parameter estimator first computes affine state estimates and their covariances, then maps these state-estimate statistics through a Gaussian DBS operator to obtain DBS estimates. Both dual estimators admit fixed-point characterizations that alternate between estimating each component using the updated prior of the other, obtained from that component's plug-in estimate statistics from the previous iteration. The efficacy of the proposed methods is examined via extensive Monte Carlo experiments, showing that the dual basis-parameter estimator attains parameter mean-squared errors comparable to those of the purely affine parameter estimator, while the dual state-parameter estimator achieves the lowest parameter mean-squared error, outperforming both the dual basis-parameter and purely affine parameter estimators, as well as sequential Monte Carlo variants of classical Particle Gibbs and Expectation-Maximization schemes.

Keywords: Bayesian inference, **state estimation**, **parameter learning**, **system identification**, **state-space model**, **Wiener model**

1. Introduction

The identification of unknown parameters in mathematical models is a foundational problem that has been studied extensively using a wide range of techniques, including subspace methods in system identification [37] and statistical inference [7]. Among the two main paradigms of parameter inference, namely the frequentist and Bayesian approaches, the Bayesian approach has received increasing attention in recent years due to its ability to quantify parameter uncertainty and incorporate prior information within a coherent probabilistic framework [26, 38]. Identifying a parametrized unknown output function becomes particularly challenging when the inputs are corrupted by correlated noise arising from underlying state dynamics, as in state-space models, because these additional sources

Date: June 30, 2026.

This work was supported by the Marie Skłodowska-Curie Grant under Agreement 956200, the ERC Starting Grant TRUST-949796, and the NSERC Discovery grant RGPIN-2025-06544.

of randomness further complicate the inference process [47]. Consider a *known* discrete-time linear time-varying dynamical system

$$\mathrm{x}_{t+1} = \mathrm{A}_t\mathrm{x}_t + \mathrm{B}_t\mathrm{u}_t + \mathrm{w}_{t+1}, \qquad \mathrm{y}_t = h_\theta(\mathrm{x}_t) + \mathrm{v}_t, \tag{1}$$

where t is the time index, $\mathrm{x}_t \in \mathbb{R}^{n_\mathrm{x}}$ denotes the system state vector, $\mathrm{A}_t \in \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$ and $\mathrm{B}_t \in \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{u}}$ are known system matrices, and $\mathrm{u}_t \in \mathbb{R}^{n_\mathrm{u}}$ represents the sequence of known or controlled inputs. The process noise $\mathrm{w}_{t+1} \in \mathbb{R}^{n_\mathrm{x}}$ is modelled as a zero-mean random vector with known covariance $\Sigma_{\mathrm{w}_{t+1}}$, that is, $\mathrm{w}_{t+1} \sim \mathbb{P}(0, \Sigma_{\mathrm{w}_{t+1}})$. The observation $\mathrm{y}_t \in \mathbb{R}^{n_\mathrm{y}}$ is generated via an *unknown* parameterized function $h_\theta \colon \mathbb{R}^{n_\mathrm{x}} \to \mathbb{R}^{n_\mathrm{y}}$ corrupted by measurement noise $\mathrm{v}_t \in \mathbb{R}^{n_\mathrm{y}}$ and $\mathrm{v}_t \sim \mathbb{P}(0, \Sigma_{\mathrm{v}_t})$. Throughout this work, and without loss of generality, we focus on the single-output case $n_\mathrm{y} = 1$, since the extension to multi-output measurements ($\mathrm{y}_t \in \mathbb{R}^{n_\mathrm{y}}$) is straightforward and essentially reduces to a set of independent single-output problems [52, Rem. 2.1]. The objective is to estimate the unknown parameter vector θ characterizing the observation model $h_\theta(\cdot)$, given only the noisy observations y_t and known inputs u_t. We further parameterize the output function $h_\theta \colon \mathbb{R}^{n_\mathrm{x}} \to \mathbb{R}$ as a linear combination of known basis functions $\phi_n \colon \mathbb{R}^{n_\mathrm{x}} \to \mathbb{C}$, i.e.,

$$h_\theta(\mathrm{x}_t) = \sum_{n=0}^{\mathrm{N}} \theta_n \phi_n(\mathrm{x}_t) = \langle \phi(\mathrm{x}_t), \theta \rangle, \tag{2}$$

where the number of basis functions is $\mathrm{N}+1$, the coefficient vector $\theta = [\theta_0, \ldots, \theta_\mathrm{N}]^\mathsf{T}$ contains the unknown parameters, and $\phi(\mathrm{x}_t) = [\phi_0(\mathrm{x}_t), \ldots, \phi_\mathrm{N}(\mathrm{x}_t)]^\mathsf{T}$ is the vector of basis functions evaluated at x_t. Additionally, we incorporate prior information about the unknown parameters through a probability distribution characterized by mean μ_θ and covariance Σ_θ, i.e., $\theta \sim \mathbb{P}(\mu_\theta, \Sigma_\theta)$. This formulation encompasses many nonlinear system identification problems.

Such systems are important in many application domains that are modelled as block-oriented structures consisting of a linear dynamical process followed by a static nonlinear observation model, commonly known as Wiener models [49]. In this setting, parameter estimation is inherently challenging because the system states, x_t, are stochastic due to the process noise, w_{t+1}, which propagates uncertainty through the linear dynamics. Given this stochastic nature of the system states appearing as inputs to the nonlinear function $h_\theta(\cdot)$, the probability density function of the observations y_t is, in general, not available in closed-form [21, 47]. This intractability fundamentally complicates Bayesian parameter estimation and necessitates the use of approximate inference techniques [7]. In particular, the optimal Bayesian minimum mean squared error (MMSE) estimator requires integrating over the joint distribution of the unobserved states and unknown parameters, which is computationally prohibitive. As a result, research in this area has mainly focused on fully Bayesian inference techniques [20, 51] as well as more tractable alternatives such as maximum a posteriori (MAP) estimation [11].

Related literature. Fully Bayesian methods aim to characterize the entire posterior distribution, typically only approximately, using techniques such as Markov Chain Monte Carlo (MCMC) [45, Ch. 6], sequential Monte Carlo (SMC) [1, 14, 42], and variational inference [8, 16]. MCMC algorithms construct a Markov chain that converges in distribution to the posterior [46]. Classic examples include the Metropolis-Hastings (MH) algorithm [39, 29] and the Gibbs sampler [25]. However, in nonlinear, non-Gaussian state-space models, such conventional MCMC algorithms cannot be applied directly, since MH method requires evaluating the likelihood up to a normalizing constant, and Gibbs sampler relies on sampling from exact conditional distributions, both of which are generally infeasible in this setting.

To address these challenges, several approximate variants have been developed, such as MCMC methods that incorporate the unscented Kalman filter (UKF-MCMC) [24], and combinations of MCMC and SMC techniques [43], collectively referred to as particle MCMC (PMCMC) algorithms. By exploiting different theoretical properties of the SMC framework [14], the MH and Gibbs sampler algorithms can be extended to the particle marginal Metropolis-Hastings (PMMH) method [48] and the particle Gibbs (PG) algorithm [15], respectively. Although these methods provide reliable uncertainty quantification and improved performance, they are typically computationally demanding [17] and scale poorly to high-dimensional problems [3, 44]. In this work, the particle Gibbs with ancestor sampling (PGAS) algorithm [35], a prominent particle sampler for Bayesian learning of state-space models, is included as one of the benchmark methods for numerical evaluation in Section 4.

In contrast to fully Bayesian approaches, MAP estimation seeks point estimates by identifying the mode of the posterior. The expectation maximization (EM) algorithm is widely used for maximum likelihood (ML) and MAP estimation in the presence of latent variable models. EM alternates between estimating the distribution over latent variables (E-step) and updating model parameters (M-step). In dynamical systems with unobserved latent states [19, 48], the E-step entails a smoothing problem, where the latent states distribution given all observations is estimated, while the M-step maximizes the expected complete-data log-likelihood, typically incorporating prior information.

When both steps have closed-form solutions, as in linear Gaussian models [27], EM guarantees a monotonic increase in the likelihood [26, Ch. 13]. However, for nonlinear or non-Gaussian models, these quantities are generally intractable, so approximate methods are used in the E-step, such as extended Kalman smoothing [30, Ch. 6], stochastic methods [18], and variational approaches [5]. Among approximate E-step methods, particle filters and kernels such as PGAS sample the smoothing distribution to approximate the likelihood [34, 48]. These approximations, while often effective, do not guarantee monotonic likelihood improvement and can lead to suboptimal modes or slow convergence [2]. Nevertheless, we use the EM algorithm [48] as a benchmark in our experiments of Section 4.

A recent perspective restricts attention to the class of affine estimators rather than approximating full posterior distributions [52]. Specifically, this approach leverages the fact that the optimal Bayesian MMSE estimator is affine when the joint distribution of the observations and parameters is elliptical. The resulting closed-form optimal affine estimator for unknown parameters relies on computing the so-called Dynamic Basis Statistics (DBS), which capture the influence of stochastic system states on the estimation process. In other words, when $h_\theta(\cdot)$ is expressed as a linear combination of known basis functions as in (2), the DBS capture the statistics of these basis functions evaluated at different time steps along the stochastic state trajectory.

To illustrate how state uncertainty affects the estimation of the parameter vector θ, we consider the following simple example with a linear basis function $\phi(\mathrm{x}_t) = \mathrm{x}_t$ in (1), where θ is a scalar parameter to be inferred from a single measurement. Specifically, we focus on a one-dimensional dynamical system with $n_\mathrm{x} = 1$ and only an initial-state prior $\mathrm{x}_0 \sim \mathbb{P}(\mu_{\mathrm{x}_0}, \Sigma_{\mathrm{x}_0})$. This example demonstrates that the presence of a latent state variable can give rise to a nonlinear parameter estimator. In this simple setting, the empirical MMSE estimate is tractable and is taken to represent the optimal Bayesian MMSE estimator. The estimator obtained from the optimal Bayesian MMSE estimate of θ exhibits pronounced nonlinearity and deviates significantly from the affine estimator.

Example 1 (1-D linear state-space model). Let the output function h_θ in (1) be defined using a single linear basis function, $\phi_n(\mathrm{x}) = \mathrm{x}$, so that

$$\mathrm{y} = \theta \mathrm{x} + \mathrm{v}, \tag{3}$$

where $\mathrm{x} \in \mathbb{R}$ is the initial state with $\mathrm{x} \sim \mathcal{N}(\mu_\mathrm{x}, \sigma_\mathrm{x}^2)$, $\theta \in \mathbb{R}$ is an unknown scalar parameter, and $\mathrm{v} \sim \mathcal{N}(0, \sigma_\mathrm{v}^2)$ is measurement noise with $\sigma_\mathrm{v}^2 = 0.16$. In this simulation, the prior on θ is taken to be the uniform distribution on $[-1, 5]$ with mean $\mu_\theta = 2$ and variance $\sigma_\theta^2 = 3$. Figure 1 shows the optimal affine and empirical MMSE estimators for θ as functions of the measurement y, i.e., $\widehat{\theta}(\mathrm{y})$, for two choices of initial-state statistics: $(\mu_\mathrm{x}, \sigma_\mathrm{x}^2) = (0.5, 0.05)$ (left plot) and $(\mu_\mathrm{x}, \sigma_\mathrm{x}^2) = (0.5, 4)$ (right plot). The empirical MMSE estimator (red) is computed using MCMC methods, while the affine MMSE estimator (blue) is obtained from the analytical expression in [52]. This example illustrates that, depending on the prior on x and the measurement range, the true MMSE estimator may be close to the optimal affine MMSE estimator in some regions of y, while exhibiting nonlinear deviations in others.

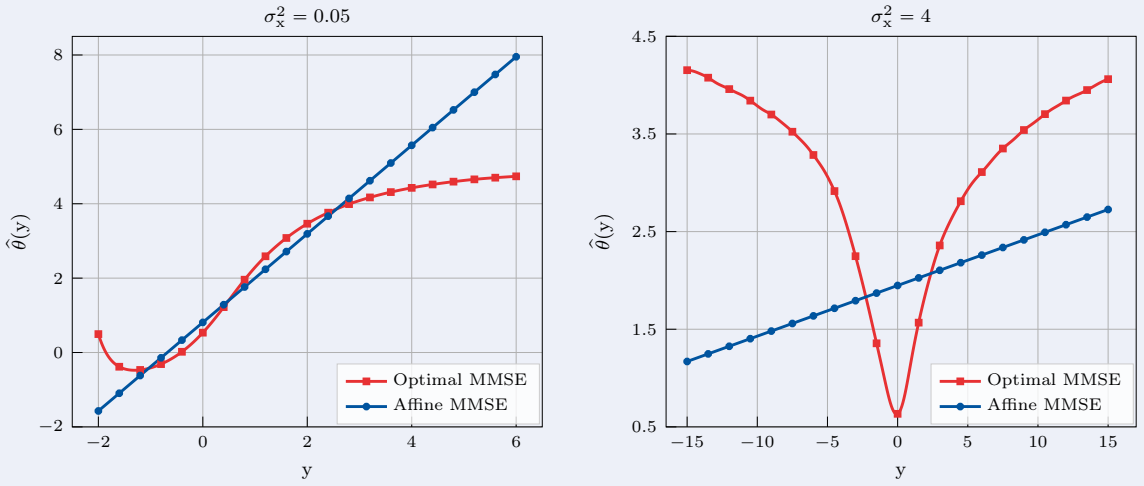


FIGURE 1. Optimal MMSE in (5) vs. affine MMSE in (10a) for estimating θ given y

In this work, we propose an analytical approach to construct a nonlinear estimator that improves upon the affine class. We revisit Example 1 throughout the paper to illustrate the behaviour of the proposed estimation algorithms.

Contributions. Building on this direction, we propose a nonlinear parameter estimator that retains the functional structure of the affine MMSE estimator for the unknown parameters while using new DBS estimates to reduce the mean-squared estimation error, as illustrated in Figure 2. As shown in the block diagram, the proposed estimator consists of two coupled components: block (I), described in Section 2, implements the affine MMSE parameter estimator, which takes as input the entire measurement sequence $\{\mathrm{y}_0, \ldots, \mathrm{y}_T\}$, the prior $(\mu_\theta, \Sigma_\theta)$ on the unknown parameters, and DBS estimates $(\widehat{\Phi}, \Sigma_{\widehat{\Phi}})$ summarizing the basis-function evaluations. This block further produces the parameter estimates and their estimation-error covariance $(\widehat{\theta}, \Sigma_{\widehat{\theta}})$, which are used as input to the DBS estimator in block (II). Block (II), discussed in Section 3, implements the DBS estimator, which uses the parameter estimates $(\widehat{\theta}, \Sigma_{\widehat{\theta}})$, the same measurement sequence, and information about the state-trajectory distribution $p(\mathrm{x}_0, \ldots, \mathrm{x}_T)$ to construct updated DBS statistics. This nonlinear estimator architecture

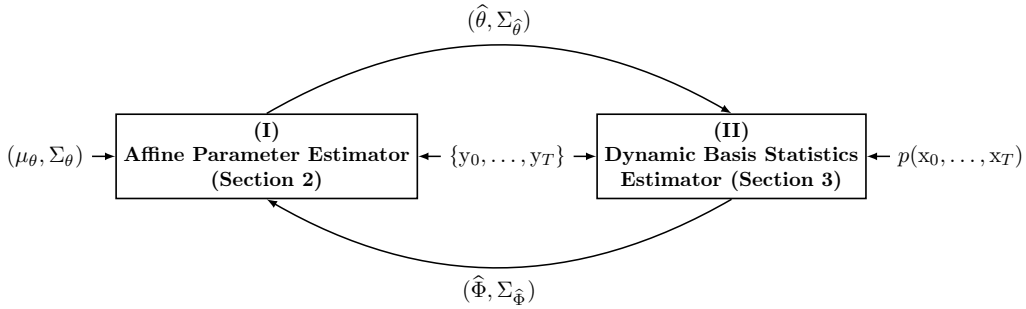
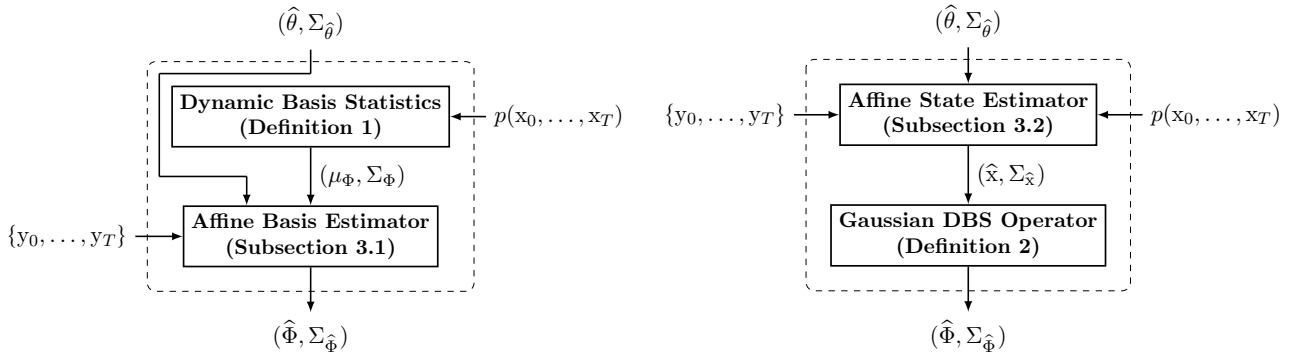


FIGURE 2. Architecture of the proposed nonlinear parameter estimator

induces an interdependence between the two blocks: the parameter estimator requires DBS statistics, and the DBS estimator, in turn, depends on the parameter estimates. We resolve this interdependence via a fixed-point characterization that iteratively refines both sets of statistics. Leveraging the closed-form structure of the affine MMSE estimator, we investigate two DBS estimation methods for the DBS estimator in block (II) of Figure 2, as detailed in Figure 3:

- **Basis-estimation approach (II-A):** This approach directly estimates basis-function evaluations over the entire trajectory. Within a Bayesian setting, we derive a closed-form expression for the optimal affine MMSE estimator of these unknown basis evaluations (Lemma 3.1), which depends on the prior mean and covariance of the unknown parameters. As shown in subfigure II-A, the DBS estimator uses information about the state-trajectory distribution $p(x_0, \ldots, x_T)$ to construct prior DBS statistics (μ_Φ, Σ_Φ) for the basis evaluations via Definition 1. This prior, together with the measurement sequence and the parameter-estimate statistics $(\widehat{\theta}, \Sigma_{\widehat{\theta}})$, is then processed by the affine basis estimator to obtain DBS estimates $(\widehat{\Phi}, \Sigma_{\widehat{\Phi}})$. Using this DBS estimator in block (II) and combining it with the affine parameter estimator in block (I) yields the dual basis-parameter (DB-P) estimator within the architecture of Figure 2. However, numerical experiments show that this approach can underperform the purely affine MMSE parameter estimator in certain regimes (cf. Figure 6), highlighting the limitations discussed in Remark 3.3 and the importance of explicitly estimating the latent state dynamics.
- **State-estimation approach (II-B):** In contrast to the basis-estimation approach, this method first estimates the latent state trajectory and then constructs the DBS from the resulting state estimates. Assuming Gaussian process noise, $w_{t+1} \sim \mathcal{N}(0, \Sigma_{w_{t+1}})$, we derive a closed-form expression for the optimal affine MMSE state estimator (Lemma 3.5). As shown in subfigure II-B, the affine state estimator uses information about the state-trajectory distribution $p(x_0, \ldots, x_T)$, together with the measurement sequence and the parameter-estimate statistics $(\widehat{\theta}, \Sigma_{\widehat{\theta}})$ in place of its prior, to compute state estimates and their estimation-error covariance. These state-estimate statistics $(\widehat{x}, \Sigma_{\widehat{x}})$ are then interpreted as the mean and covariance of a Gaussian prior over the system states and passed to the Gaussian DBS operator in Definition 2 to obtain DBS estimates $(\widehat{\Phi}, \Sigma_{\widehat{\Phi}})$. Using this DBS estimator in block (II) and pairing it with the affine parameter estimator in block (I) yields the dual state-parameter (DS-P) estimator within the architecture of Figure 2. Extensive numerical experiments indicate that this approach achieves a lower average parameter estimation error than both the dual basis-parameter (DB-P) estimator and a purely affine parameter estimator in two different configurations (cf. Figures 6, 7, and 8).



(II-A) Basis-estimation approach (II-B) State-estimation approach

FIGURE 3. Architectures of the Dynamic Basis Statistics estimator

Combining either of the proposed DBS estimation methods with the affine parameter estimator induces an interdependence between the parameters and the DBS statistics, resolved by the algebraic equations defining DB-P and DS-P estimators. These estimators naturally lead to fixed-point characterizations, implemented by Algorithm 1, that alternate between estimating each set of unknowns using the updated prior of the other set derived from that set's estimate statistics from the previous iteration (see the plug-in interpretation in Remark 2.5 for the prior-update concept). This mechanism is enabled by the structure of the affine MMSE estimators, whose coefficients depend only on the means and covariances of the unknowns rather than their full distributions. All affine estimators in DB-P and DS-P admit explicit expressions for Fourier basis functions, yielding fully explicit DBS formulas in this case [52]. To facilitate reproducibility, an open-source MATLAB library implementing all proposed methods for Fourier basis functions is provided at https://github.com/sasanvakili/nonlinearBayesian4Wiener.

Similar ideas of alternating between state and parameter estimates have been explored in the literature, most notably in the Dual Kalman Smoother (DKS) of [53], the iterative smoothing scheme of the dual estimation method [30, Ch. 5], and in SMC-based approaches such as PGAS [35] and SMC-EM (particle EM) [48]. Both DKS and PGAS target the joint smoothing distribution $p(\theta, \{\mathrm{x}_t\}_{t=0}^T \,|\, \{\mathrm{y}_t\}_{t=0}^T)$ and conceptually treat the two unknown sets as separate blocks, updating one while treating the other as fixed at its current estimate. DKS approximates the conditional posteriors of the latent trajectory and the parameters by Gaussian distributions and alternates between two conditional smoother blocks: it iteratively estimates the system states using the latest parameter estimates via an extended Kalman smoother, and then estimates the parameters from the currently smoothed states until convergence. By contrast, PGAS does not make this Gaussian approximation and instead alternates between sampling the latent trajectory and the parameters from their full conditional posteriors. EM-based methods, on the other hand, treat the parameters as fixed in the E-step to approximate the state-trajectory posterior $p(\{\mathrm{x}_t\}_{t=0}^T \,|\, \theta, \{\mathrm{y}_t\}_{t=0}^T)$, and then update θ in the M-step by maximizing an objective function obtained from integrating over this posterior. Our approach differs from these schemes in that it propagates both first- and second-order statistics between the coupled estimators. Extensive numerical experiments show that the proposed dual state-parameter (DS-P) estimator achieves a lower average parameter estimation error than both PGAS and the EM-based method (cf. Figure 9).

Organization. Section 2 formulates the MMSE estimation problem for the parameterized output function (2), reviews the affine parameter estimator of [52], and motivates the nonlinear parameter

estimator structure. Section 3 develops two approaches for constructing DBS estimates and introduces their fixed-point characterizations when combined with the parameter estimator, resulting in the dual basis-parameter and dual state-parameter estimators. Section 4 presents numerical experiments comparing the proposed dual estimators with the affine parameter estimator and SMC-based methods. Detailed proofs of the mathematical statements are provided in Appendix A, titled "Technical Proofs."

Notation. Throughout this paper, $\mathbb{R}$ and $\mathbb{R}^{n\times m}$ denote the set of real numbers and the set of $n \times m$ real matrices, respectively. The symbol $\mathbb{I}$ refers to the identity matrix. Subscripts denote elements of a vector (e.g., $\mathrm{x}_{t,i}$ represents the i^{th} element of the vector x_t), while superscripts represent the time index of vector or matrix elements (e.g., μ_ϕ^t for a vector and $\Sigma_\Phi^{tt'}$ for a matrix). The trace operator is denoted by $\mathrm{tr}(\cdot)$, the transpose of a matrix A is denoted by A^T, and $\mathrm{diag}(\mathrm{A}_1, \ldots, \mathrm{A}_k)$ represents a block-diagonal matrix with diagonal entries $\mathrm{A}_1, \ldots, \mathrm{A}_k$. Given $\mathrm{A} \in \mathbb{R}^{m\times n}$, a matrix with columns $\mathrm{a}_1, \ldots, \mathrm{a}_n \in \mathbb{R}^m$, $\mathrm{vec}(\mathrm{A})$ is the vector $[\mathrm{a}_1^\mathsf{T}, \ldots, \mathrm{a}_n^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{mn}$. The inner product of two vectors x and y is given by $\langle \mathrm{x}, \mathrm{y}\rangle = \mathrm{x}^\mathsf{T}\mathrm{y}$, and the corresponding 2-norm is $\|\mathrm{x}\| = \sqrt{\langle \mathrm{x}, \mathrm{x}\rangle}$. The inner product of two matrices $\mathrm{A}, \mathrm{B} \in \mathbb{R}^{m\times n}$ is defined as $\langle \mathrm{A}, \mathrm{B}\rangle = \mathrm{tr}(\mathrm{A}^\mathsf{T}\mathrm{B})$. The conic inequality $\mathrm{A} \preceq \mathrm{B}$ means that the matrix difference $\mathrm{B} - \mathrm{A}$ is positive semidefinite, i.e., $\mathrm{B} - \mathrm{A} \succeq 0$. The notation $\mathbb{P}(\mu, \Sigma)$ refers to an arbitrary distribution with mean μ and covariance matrix Σ, while a multivariate normal (Gaussian) distribution is denoted by $\mathcal{N}(\mu, \Sigma)$ and a uniform distribution over $[a, b]$ is denoted by $\mathcal{U}(a, b)$. The symbol $\sim$ stands for *"distributed according to"*. For a random vector $\theta \in \mathbb{R}^n$, an estimate and its corresponding estimation-error covariance are denoted by $\widehat{\theta}$ and $\Sigma_{\widehat{\theta}}$, respectively, and the indices of these quantities indicate the type of estimator: $(\widehat{\theta}_{\text{opt}}, \Sigma_{\widehat{\theta}_{\text{opt}}})$ are obtained from the optimal Bayesian MMSE estimator, $(\widehat{\theta}_{\text{af}}, \Sigma_{\widehat{\theta}_{\text{af}}})$ from the affine estimator, and $(\widehat{\theta}_{\text{nl}}, \Sigma_{\widehat{\theta}_{\text{nl}}})$ from a nonlinear estimator.

2. Bayesian estimation for the Wiener model

In this section, we formulate the Bayesian MMSE estimation problem for the Wiener model and review key results on the Bayesian affine MMSE estimator from [52]. We then discuss the relevance of these results and their connections to the proposed nonlinear estimator.

2.1. Problem description

Let the process noise w_t, the measurement noise v_t, the initial state x_0, and the vector of parameters θ be independent of one another at all times. Given the parametrized observation model (2) and state representation in (1), together with the sequence of measurements $\mathrm{y} = [\mathrm{y}_0, \ldots, \mathrm{y}_T]^\mathsf{T}$, our objective is to design an estimator $\widehat{\theta}(\cdot)$ that minimizes the following mean squared error (MSE):

$$\min_{\widehat{\theta}(\cdot)\in\mathscr{F}_\theta} \mathbb{E}\big[\big\|\theta - \widehat{\theta}(\mathrm{y})\big\|^2\big], \tag{4}$$

where $\mathscr{F}_\theta := \left\{\widehat{\theta}\colon \mathbb{R}^{T+1} \to \Theta\right\}$ denotes the class of all estimators, $T+1$ is the length of the data sequence, and Θ is the parameter space. By definition, among all $\widehat{\theta}$ in $\mathscr{F}_\theta$, the optimal Bayesian MMSE estimator $\widehat{\theta}_{\text{opt}}$ for (4) coincides with the posterior mean of θ given the measurements [32, Ch. 4]. Accordingly, the parameter estimate $\widehat{\theta}_{\text{opt}}$ and its error covariance $\Sigma_{\widehat{\theta}_{\text{opt}}}$ are

$$\widehat{\theta}_{\text{opt}} = \mathbb{E}[\theta|\mathrm{y}] = \int_\Theta \theta p(\theta|\mathrm{y})\mathrm{d}\theta, \quad \Sigma_{\widehat{\theta}_{\text{opt}}} = \mathbb{E}[(\theta - \widehat{\theta}_{\text{opt}})(\theta - \widehat{\theta}_{\text{opt}})^\mathsf{T}]. \tag{5}$$

However, this estimator is generally computationally intractable due to the complexity of the posterior distribution $p(\theta|\mathrm{y})$ and the lack of an explicit solution for the associated integral. In the following subsection, we briefly review the affine MMSE parameter estimator, which admits a closed-form solution [52] and serves as the foundation for the new contributions presented in this work.

2.2. From affine to nonlinear estimators

To obtain a tractable alternative to the Bayesian MMSE estimator in (5), we first restrict attention to the class of affine estimators, characterize the corresponding MMSE solution, and subsequently use this construction as a stepping stone towards more general nonlinear estimators. The following generic result characterizes the optimal affine MMSE estimator in terms of the mean and covariance of a pair of random vectors, serving as the basis for our subsequent specialization to the parameter estimation.

Lemma 2.1 (**Affine MMSE estimator**). *Consider a random $\xi = (\xi_1, \xi_2)$ with components $\xi_1 \in \mathbb{R}^{n_1}$ and $\xi_2 \in \mathbb{R}^{n_2}$, whose joint distribution is*

$$\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} \sim \mathbb{P}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),$$

where $\mu_i := \mathbb{E}[\xi_i]$ and $\Sigma_{ij} := \mathbb{E}[(\xi_i - \mu_i)(\xi_j - \mu_j)^{\mathsf{T}}]$ for $i, j \in \{1, 2\}$ denote the means and covariances, respectively. Suppose that Σ_{22} is invertible and restrict the estimation of ξ_1 from ξ_2 to affine estimators,

$$\mathscr{A}_\zeta := \left\{ \widehat{\xi}_1(\xi_2) = \Psi\xi_2 + \psi \,\middle|\, \Psi \in \mathbb{R}^{n_1 \times n_2},\, \psi \in \mathbb{R}^{n_1} \right\}. \tag{6}$$

For the Bayesian MMSE problem $\mathcal{J}^\star := \min_{\widehat{\xi}_1(\cdot)\in\mathscr{A}_\zeta} \mathbb{E}[\|\xi_1 - \widehat{\xi}_1(\xi_2)\|^2]$, the solution comprises the affine MMSE estimate, its error covariance, and optimal cost:

$$\begin{cases} \widehat{\xi}_1(\xi_2) = \Psi^\star\xi_2 + \psi^\star, \\ \Sigma_{\widehat{\xi}_1} := \mathbb{E}\big[\big(\xi_1 - \widehat{\xi}_1(\xi_2)\big)\big(\xi_1 - \widehat{\xi}_1(\xi_2)\big)^{\mathsf{T}}\big] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ \mathcal{J}^\star = \mathrm{tr}(\Sigma_{\widehat{\xi}_1}), \end{cases} \tag{7a}$$

with the optimal coefficients given by

$$\Psi^\star = \Sigma_{12}\Sigma_{22}^{-1}, \qquad \psi^\star = \mu_1 - \Psi^\star\mu_2. \tag{7b}$$

The proof of this lemma is provided in Appendix A.1. Finding the affine MMSE estimator for estimating θ from y using Lemma 2.1 requires the observation covariance Σ_{y} and the cross-covariance between the unknown parameters and the observation vector, $\Sigma_{\theta\mathrm{y}}$. This task is nontrivial because the observation model (2) depends on the inner product of two unknown quantities, namely the basis-function evaluations of the unknown system states $\phi(\mathrm{x})$ and the unknown parameters θ. A central component in characterizing the affine estimator for the MMSE problem (4), as detailed in [52], is the propagation of the state trajectory through the output basis functions. The resulting quantities, known as Dynamic Basis Statistics (DBS), are formalized as follows.

Definition 1 (**Dynamic Basis Statistics**). *Let the state sequence $\mathrm{x} = [\mathrm{x}_0^{\mathsf{T}}, \ldots, \mathrm{x}_T^{\mathsf{T}}]^{\mathsf{T}}$ be a trajectory of (1) for $t \in \{0, \ldots, T\}$, and let $\Phi(\mathrm{x}) = [\phi(\mathrm{x}_0), \ldots, \phi(\mathrm{x}_T)]$ be the matrix whose columns are the basis function vectors evaluated at each x_t as in (2). The Dynamic Basis Statistics (DBS) collection with*

respect to the joint distribution of the state trajectory is defined as

$$\mu_\Phi = [\mu_\phi^0, \dots, \mu_\phi^T], \qquad \Sigma_\Phi = \begin{bmatrix} \Sigma_\phi^{00} & \dots & \Sigma_\phi^{0T} \\ \vdots & \ddots & \vdots \\ \Sigma_\phi^{T0} & \dots & \Sigma_\phi^{TT} \end{bmatrix}, \tag{8}$$

where μ_ϕ^t *is the mean of* $\phi(\mathrm{x}_t)$ *and* $\Sigma_\phi^{tt'}$ *is the covariance between* $\phi(\mathrm{x}_t)$ *and* $\phi(\mathrm{x}_{t'})$*, given by*

$$\mu_\phi^t := \mathbb{E}[\phi(\mathrm{x}_t)], \qquad \Sigma_\phi^{tt'} := \mathbb{E}[\phi(\mathrm{x}_t)\phi^\mathsf{T}(\mathrm{x}_{t'})] - \mathbb{E}[\phi(\mathrm{x}_t)]\mathbb{E}[\phi(\mathrm{x}_{t'})]^\mathsf{T}. \tag{9}$$

This DBS collection enables marginalization over the latent system states, which in turn allows derivation of the closed-form solution for the Bayesian affine MMSE estimator.

Theorem 2.2 (**Affine MMSE parameter estimator** [52, Thm. 3.2])**.** *The Bayesian MMSE estimator of Problem 4 within the affine class, analogous to (6), and its corresponding estimation-error covariance with the optimal cost, as functions of* μ_Φ *and* Σ_Φ*, are given by*

$$\begin{cases} \widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi) = \Psi_\theta^\star(\mu_\Phi, \Sigma_\Phi)\mathrm{y} + \psi_\theta^\star(\mu_\Phi, \Sigma_\Phi), \\ \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\mu_\Phi, \Sigma_\Phi) = \Sigma_\theta - \Sigma_\theta\mu_\Phi\big(\mu_\Phi^\mathsf{T}\Sigma_\theta\mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_\mathrm{v}\big)^{-1}\mu_\Phi^\mathsf{T}\Sigma_\theta, \\ \mathcal{J}_\theta^\star(\mu_\Phi, \Sigma_\Phi) = \mathrm{tr}(\Sigma_{\widehat{\theta}_{\mathrm{af}}}(\mu_\Phi, \Sigma_\Phi)), \end{cases} \tag{10a}$$

where the coefficient matrices are defined as

$$\Psi_\theta^\star(\mu_\Phi, \Sigma_\Phi) = \Sigma_\theta\mu_\Phi\big(\mu_\Phi^\mathsf{T}\Sigma_\theta\mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_\mathrm{v}\big)^{-1}, \quad \psi_\theta^\star(\mu_\Phi, \Sigma_\Phi) = \mu_\theta - \Psi_\theta^\star(\mu_\Phi, \Sigma_\Phi)\mu_\Phi^\mathsf{T}\mu_\theta, \tag{10b}$$

$\mathcal{M}\colon \mathbb{R}^{((\mathrm{N}+1)(T+1))^2} \to \mathbb{R}^{(T+1)^2}$ *is a linear operator whose* $(t,t')^{th}$ *component is*

$$\mathcal{M}^{tt'}(\Sigma_\Phi) = \langle \Sigma_\phi^{tt'}, (\Sigma_\theta + \mu_\theta\mu_\theta^\mathsf{T}) \rangle, \tag{10c}$$

and Σ_v *is the covariance matrix of the measurement noise sequence* $\mathrm{v} = [\mathrm{v}_0, \dots, \mathrm{v}_T]^\mathsf{T}$*, which is the diagonal matrix* $\Sigma_\mathrm{v} = \mathrm{diag}(\sigma_{\mathrm{v}_0}^2, \dots, \sigma_{\mathrm{v}_T}^2)$*, when they are mutually independent.*

Remark 2.3 (**Functional dependence**)**.** *In principle, the estimator* $\widehat{\theta}_{\mathrm{af}}(\cdot)$ *is a function of three arguments. However, the notation used in (10a), i.e.,* $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)$*, emphasizes the dependence of the estimator on the DBS collection* (μ_Φ, Σ_Φ) *in the last two arguments. This estimator depends linearly on the measurements* y *and nonlinearly on* (μ_Φ, Σ_Φ)*. Note that the DBS collection itself is generated by a nonlinear transformation of the inputs and system dynamics and does not involve the measurements.*

Given Definition 1, the proof of Theorem 2.2 follows directly from the generic affine MMSE estimator of Lemma 2.1 by setting $\xi_1 = \theta$ and $\xi_2 = \mathrm{y}$ (see [52, Thm. 3.2] for a complete proof). While $\widehat{\theta}_{\mathrm{af}}(\cdot)$ is provably optimal within the affine class, the proximity of its MSE to that of the Bayesian MMSE estimator, $\widehat{\theta}_{\mathrm{opt}}$, remains an open question. When perfect state knowledge is available, the measurements contain uncertainty only through the measurement noise. In this case, the affine MMSE estimator returns $\widehat{\theta}_{\mathrm{opt}}$ if the prior on the unknown parameters, $\mathbb{P}(\mu_\theta, \Sigma_\theta)$, and the distribution of the measurement noise sequence, $\mathbb{P}(0, \Sigma_\mathrm{v})$, belong to an elliptical family [52, Rem. 3.1]. As observed in the simple case of Example 1, $\widehat{\theta}_{\mathrm{af}}(\cdot)$ can track $\widehat{\theta}_{\mathrm{opt}}$ well for certain priors on x and ranges of y, particularly when the state uncertainty is small. Let us revisit Example 1 to see how Theorem 2.2 specializes to an explicit expression for the affine estimator in that setting.

Example 1 (continued). In the setting of (3), the dynamical system is one-dimensional, and the observation model involves the linear basis function $\phi(\mathrm{x}) = \mathrm{x}$. In this case, the DBS from Definition 1 simplify to

$$\mu_\Phi = \mu_\mathrm{x}, \qquad \Sigma_\Phi = \sigma_\mathrm{x}^2,$$

and the affine MMSE estimator and its estimation-error covariance from Theorem 2.2 are

$$\begin{aligned} \widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\mathrm{x}, \sigma_\mathrm{x}^2) &= \mu_\theta + \frac{\sigma_\theta^2 \mu_\mathrm{x}}{\mu_\mathrm{x}^2\sigma_\theta^2 + \sigma_\mathrm{x}^2\sigma_\theta^2 + \sigma_\mathrm{x}^2\mu_\theta^2 + \sigma_\mathrm{v}^2}\,(\mathrm{y} - \mu_\mathrm{x}\mu_\theta), \\ \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\mu_\mathrm{x}, \sigma_\mathrm{x}^2) &= \sigma_\theta^2 - \frac{\sigma_\theta^4 \mu_\mathrm{x}^2}{\mu_\mathrm{x}^2\sigma_\theta^2 + \sigma_\mathrm{x}^2\sigma_\theta^2 + \sigma_\mathrm{x}^2\mu_\theta^2 + \sigma_\mathrm{v}^2}. \end{aligned} \tag{11}$$

However, as shown in Figure 1, the Bayesian MMSE estimator for Example 1 can become highly nonlinear as the uncertainty in the system state increases. Consequently, the MSE of the affine estimator in (11) may be far from that of $\widehat{\theta}_{\mathrm{opt}}$. More generally, for hidden Markov processes (HMPs), the MSE decomposes into two terms [22, 23]: the MSE achieved when the system state is perfectly known, and an additional cross-error term for which no explicit expression is generally available. An analogous decomposition holds for the class of affine estimators marginalized over the latent system states. The following lemma demonstrates that the expected error of the affine MMSE estimator is bounded below by the fundamental limit corresponding to complete knowledge of the system state.

Lemma 2.4 (**Lower bound on the MSE of the affine estimator**). *Consider the affine MMSE estimator* $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)$ *and its optimal error* $\mathcal{J}_\theta^\star(\mu_\Phi, \Sigma_\Phi)$ *as defined in* (10a). *, we have*

$$\mathcal{J}_\theta^\star(\mu_\Phi, \Sigma_\Phi) \geqslant \mathbb{E}_\mathrm{x}\Big[\mathcal{J}_\theta^\star(\Phi(\mathrm{x}), 0)\Big] = \mathbb{E}_\mathrm{x}\Big[\min_{\widehat{\theta}(\cdot)\in\mathscr{A}_\theta} \mathbb{E}\big[\|\theta - \widehat{\theta}(\mathrm{y})\|^2 \,|\, \mathrm{x}\big]\Big], \tag{12}$$

where x *is the state trajectory,* $\Phi(\mathrm{x})$ *is the DBS in Definition 1, the set* $\mathscr{A}_\theta$ *denotes the class of affine estimators of* θ *from* y*, analogous to* (6)*, and the inner expectation on the right-hand side is conditional on* x*, while the outer expectation is taken with respect to the marginal distribution of* x.

The proof can be found in Appendix A.2. Notably, the term $\mathcal{J}_\theta^\star\big(\Phi(\mathrm{x}), 0\big)$ in (12) is indeed the optimal estimation error achievable by an affine MMSE estimator when the state trajectory x is perfectly known. In this light, the lower bound in Lemma 2.4 suggests a fundamental performance limit attainable by affine estimators when the state trajectory is perfectly known. This gap motivates the development of nonlinear parameter estimators that more effectively exploit the DBS structure and the latent state information, with the aim of reducing the difference between the achievable MSE and the fundamental limit suggested by Lemma 2.4. Consequently, this study proposes a nonlinear parameter estimator by composing the affine MMSE mapping (10) with data-dependent DBS estimates. Let $\widehat{\Phi}(\mathrm{y})$ and $\Sigma_{\widehat{\Phi}}(\mathrm{y})$ denote estimates of the DBS components constructed from the measurements y. Substituting these estimates into affine MMSE expressions $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \cdot, \cdot)$ and $\Sigma_{\widehat{\theta}_{\mathrm{af}}}(\cdot, \cdot)$ in (10a) yields the nonlinear estimator

$$\begin{cases} \widehat{\theta}_{\mathrm{nl}}(\mathrm{y}) = \widehat{\theta}_{\mathrm{af}}\big(\mathrm{y}; \widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y})\big), \\ \Sigma_{\widehat{\theta}_{\mathrm{nl}}}(\mathrm{y}) = \Sigma_{\widehat{\theta}_{\mathrm{af}}}\big(\widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y})\big). \end{cases} \tag{13}$$

Thus, $\widehat{\theta}_{\mathrm{nl}}(\mathrm{y})$ has the same functional form as the affine MMSE estimator, but is nonlinear with respect to y because its DBS arguments are replaced by data-driven estimates.

Remark 2.5 (**Plug-in interpretation**). *The nonlinear estimator* (13) *can be interpreted as a plug-in construction with respect to the DBS statistics. Whereas the affine MMSE estimator* (10a) *treats* (μ_Φ, Σ_Φ) *as known hyperparameters, the nonlinear variant substitutes their data-dependent estimates* $\big(\widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y})\big)$ *into the same affine formulas. This procedure is analogous to empirical Bayes methods, where hyperparameters are first estimated from data and subsequently treated as fixed when evaluating the conditional estimator* [12]. *Consequently, the independence assumptions underlying the affine derivation no longer hold exactly, since the effective DBS now depend on* y. *The covariance* $\Sigma_{\widehat{\theta}_{\mathrm{nl}}}(\mathrm{y})$ *is thus computed under the approximation that the plugged-in DBS are fixed rather than random. Nevertheless, this construction retains the affine functional form while incorporating DBS estimates.*

The subsequent section presents two approaches for constructing such DBS estimates, each resulting in a fixed-point characterization of the nonlinear parameter estimator.

3. Fixed-Point Characterization of Nonlinear Estimation

The nonlinear parameter estimator (13) depends critically on accurate DBS estimates $(\widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y}))$ that improve parameter estimation performance beyond the affine counterpart. Since (μ_Φ, Σ_Φ) are determined by the latent state sequence, we investigate two distinct strategies: (i) direct estimation of the basis-function evaluations $\phi(\mathrm{x})$ in (2) over the entire trajectory, and (ii) estimation of the internal latent system states x followed by DBS construction from the resulting state estimates. Crucially, an interdependency arises in both strategies, as the DBS estimates require parameter estimates and vice versa, which are addressed by finding a fixed-point solution in the following subsections.

3.1. Dual basis-parameter estimation

The fundamental limit in Lemma 2.4 suggests that DBS estimates should satisfy $\widehat{\Phi}(\mathrm{y}) \to \Phi(\mathrm{x})$ and $\Sigma_{\widehat{\Phi}}(\mathrm{y}) \to 0$ to improve the parameter estimation performance of $\widehat{\theta}_{\mathrm{nl}}(\mathrm{y})$ compared to $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)$. This subsection implements the first strategy by estimating the basis-function evaluations $\phi(\mathrm{x})$ in (2) via an affine MMSE estimator, and then forming $\widehat{\Phi}(\mathrm{y})$ and $\Sigma_{\widehat{\Phi}}(\mathrm{y})$ from the resulting basis estimates and their estimation-error covariance. The following lemma is analogous to Theorem 2.2 and formalizes this construction as an immediate consequence of Lemma 2.1.

Lemma 3.1 (**DBS estimation via affine MMSE basis estimator**). *Let* $\Phi(\mathrm{x})$ *be the unknown basis-function matrix over the trajectory, with prior DBS* (μ_Φ, Σ_Φ) *as in Definition 1, and define the stacked basis-function vector* $\phi := \mathrm{vec}(\Phi(\mathrm{x})) = [\phi^{\mathsf{T}}(\mathrm{x}_0), \ldots, \phi^{\mathsf{T}}(\mathrm{x}_T)]^{\mathsf{T}}$. *Given the prior mean and covariance of the unknown parameters* $(\mu_\theta, \Sigma_\theta)$, *the Bayesian affine MMSE estimator of* ϕ *admits the closed form* $\widehat{\phi}_{\mathrm{af}}(\mathrm{y}; \mu_\theta, \Sigma_\theta) = \Psi^\star_\phi(\mu_\theta, \Sigma_\theta)\mathrm{y} + \psi^\star_\phi(\mu_\theta, \Sigma_\theta)$, *where*

$$\Psi^\star_\phi(\mu_\theta, \Sigma_\theta) = \Sigma_\Phi \mu_\Theta \big(\mu_\Theta^{\mathsf{T}} \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_{\mathrm{v}}\big)^{-1}, \quad \psi^\star_\phi(\mu_\theta, \Sigma_\theta) = \bar{\phi} - \Psi^\star_\phi(\mu_\theta, \Sigma_\theta)\mu_\Phi^{\mathsf{T}}\mu_\theta, \tag{14a}$$

with $\bar{\phi} = \mathrm{vec}(\mu_\Phi)$, $\mu_\Theta = \mathrm{diag}(\mu_\theta^0, \ldots, \mu_\theta^T)$ *where* $\mu_\theta^t = \mu_\theta$ *for each* $t \in \{0, \ldots, T\}$, Σ_{v} *identical to that in the affine MMSE parameter estimator* (10), *and the linear operator* $\mathcal{S}\colon \mathbb{R}^{(\mathrm{N}+1)^2} \to \mathbb{R}^{(T+1)^2}$ *having* $(t, t')^{th}$ *component*

$$\mathcal{S}^{tt'}(\Sigma_\theta) = \langle \Sigma_\theta, (\Sigma_\phi^{tt'} + \mu_\phi^t \mu_\phi^{t'^{\mathsf{T}}}) \rangle. \tag{14b}$$

Consequently, the resulting DBS estimates and the corresponding MMSE are

$$\begin{cases} \widehat{\Phi}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta) = [\widehat{\phi}^0_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta),\dots,\widehat{\phi}^T_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta)], \\ \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta) = \Sigma_\Phi - \Sigma_\Phi\mu_\Theta\big(\mu_\Theta^{\mathsf{T}}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_{\mathrm{v}}\big)^{-1}\mu_\Theta^{\mathsf{T}}\Sigma_\Phi, \\ \mathcal{J}^\star_\phi(\mu_\theta,\Sigma_\theta) = \mathrm{tr}(\Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta)), \end{cases} \tag{14c}$$

where $\widehat{\phi}^t_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta)$ *is the estimate of* $\phi(\mathrm{x}_t)$.

Remark 3.2 (**Observation covariance forms**). *The Bayesian affine MMSE estimators for the parameters in* (10) *and for the basis collection in* (14) *both rely on the observation covariance*

$$\Sigma_{\mathrm{y}} = \mu_\Phi^{\mathsf{T}}\Sigma_\theta\mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_{\mathrm{v}} = \mu_\Theta^{\mathsf{T}}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_{\mathrm{v}}, \tag{15}$$

where $\mathcal{M}(\Sigma_\Phi)$ *and* $\mathcal{S}(\Sigma_\theta)$ *are defined in* (10c) *and* (14b)*, respectively. The two expressions follow from the lifted observation model written either as* $\mathrm{y} = \Phi^{\mathsf{T}}(\mathrm{x})\theta + \mathrm{v}$ *in* [52] *or* $\mathrm{y} = \mathrm{G}_\theta^{\mathsf{T}}\phi + \mathrm{v}$ *in* (37)*, where* $\phi = \mathrm{vec}(\Phi(\mathrm{x}))$ *and* $\mathrm{G}_\theta = \mathrm{diag}(\theta^0,\dots,\theta^T)$ *with with* $\theta^t = \theta$ *for all* $t \in \{0,\dots,T\}$ *(see also Appendix A.3). The former highlights the dependence on DBS and is used for affine parameter estimation, whereas the latter highlights the dependence on the parameter prior and is used for affine estimation of DBS.*

For the proof of Lemma 3.1, see Appendix A.3. In the context of Example 1, the DBS estimation admits an explicit expression via Lemma 3.1. In this setting, because the basis function is linear, basis evaluations and states coincide. Hence, the DBS estimates in (14c) for a single measurement is essentially an optimal affine MMSE state estimator.

Example 1 (continued). Given the linear basis function $\phi(\mathrm{x}) = \mathrm{x}$ in (3), the DBS estimator described in (14c) is equal to the optimal affine MMSE state estimator

$$\begin{aligned} \widehat{\Phi}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\sigma^2_\theta) &= \mu_{\mathrm{x}} + \frac{\sigma^2_{\mathrm{x}}\mu_\theta}{\mu^2_\theta\sigma^2_{\mathrm{x}} + \sigma^2_\theta\sigma^2_{\mathrm{x}} + \sigma^2_\theta\mu^2_{\mathrm{x}} + \sigma^2_{\mathrm{v}}}(\mathrm{y} - \mu_{\mathrm{x}}\mu_\theta), \\ \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\mu_\theta,\sigma^2_\theta) &= \sigma^2_{\mathrm{x}} - \frac{\sigma^4_{\mathrm{x}}\mu^2_\theta}{\mu^2_\theta\sigma^2_{\mathrm{x}} + \sigma^2_\theta\sigma^2_{\mathrm{x}} + \sigma^2_\theta\mu^2_{\mathrm{x}} + \sigma^2_{\mathrm{v}}}. \end{aligned} \tag{16}$$

Observe that the operators in (10c) and (14b) evaluate to

$$\mathcal{M}(\sigma^2_{\mathrm{x}}) = \sigma^2_{\mathrm{x}}(\sigma^2_\theta + \mu^2_\theta), \qquad \mathcal{S}(\sigma^2_\theta) = \sigma^2_\theta(\sigma^2_{\mathrm{x}} + \mu^2_{\mathrm{x}}),$$

and that the denominators in (11) and (16) are equal, representing the observation covariance for a scalar measurement.

The DBS estimates in (14c) depend on the prior parameter statistics $(\mu_\theta,\Sigma_\theta)$. Leveraging the fact that these estimates would improve if we had perfect knowledge of θ, we apply a plug-in interpretation analogous to Remark 2.5 by treating the parameter estimates $\big(\widehat{\theta}(\mathrm{y}),\Sigma_{\widehat{\theta}}(\mathrm{y})\big)$ as fixed prior moments,

$$\begin{cases} \widehat{\Phi}_{\mathrm{nl}}(\mathrm{y}) = \widehat{\Phi}_{\mathrm{af}}(\mathrm{y};\widehat{\theta}(\mathrm{y}),\Sigma_{\widehat{\theta}}(\mathrm{y})), \\ \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}(\mathrm{y}) = \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\widehat{\theta}(\mathrm{y}),\Sigma_{\widehat{\theta}}(\mathrm{y})). \end{cases} \tag{17}$$

The mutual substitution of (17) and (13) creates a circular dependency: the DBS estimates require the parameter estimates, which in turn require the DBS estimates. This interdependency yields a

nonlinear estimator, the main message of this study, characterized as a solution (a.k.a. fixed-point) to the following algebraic equations.

Dual basis-parameter estimator: For any measurements vector y, there exists a tuple of basis-parameter estimators $\big(\widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}, \widehat{\Phi}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\Phi}_{\mathrm{nl}}}\big)$ such that

$$\begin{cases} \widehat{\theta}^{\star}_{\mathrm{nl}} = \widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \widehat{\Phi}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\Phi}_{\mathrm{nl}}}), \\ \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}} = \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\widehat{\Phi}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\Phi}_{\mathrm{nl}}}), \end{cases} \qquad \begin{cases} \widehat{\Phi}^{\star}_{\mathrm{nl}} = \widehat{\Phi}_{\mathrm{af}}(\mathrm{y}; \widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}), \\ \Sigma^{\star}_{\widehat{\Phi}_{\mathrm{nl}}} = \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}), \end{cases} \tag{DB-P}$$

where the functions $\big(\widehat{\theta}_{\mathrm{af}}(\cdot), \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\cdot)\big)$ and $\big(\widehat{\Phi}_{\mathrm{af}}(\cdot), \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\cdot)\big)$ are the closed form solutions of the affine Bayesian MMSE defined in (10a) and (14c), respectively.

The dual basis-parameter estimation framework thus comprises two affine MMSE estimators, each relying on the estimates of the other. Using the closed-form expressions of the affine MMSE parameter and DBS estimates in (11) and (16) for Example 1, we illustrate the dual basis-parameter (DB-P) estimator on the same data and compare it with the affine and empirical MMSE estimators.

Example 1 (continued). The dual basis-parameter (DB-P) estimator for θ given a single measurement, constructed using the affine MMSE parameter and basis estimators in (11) and (16), is shown in green in Figure 4. The algorithm used to construct the dual basis-parameter estimator (green curve) is described in Section 3.3. As can be seen, the resulting nonlinearity closely follows the one obtained from the empirical MMSE estimator (red curve).

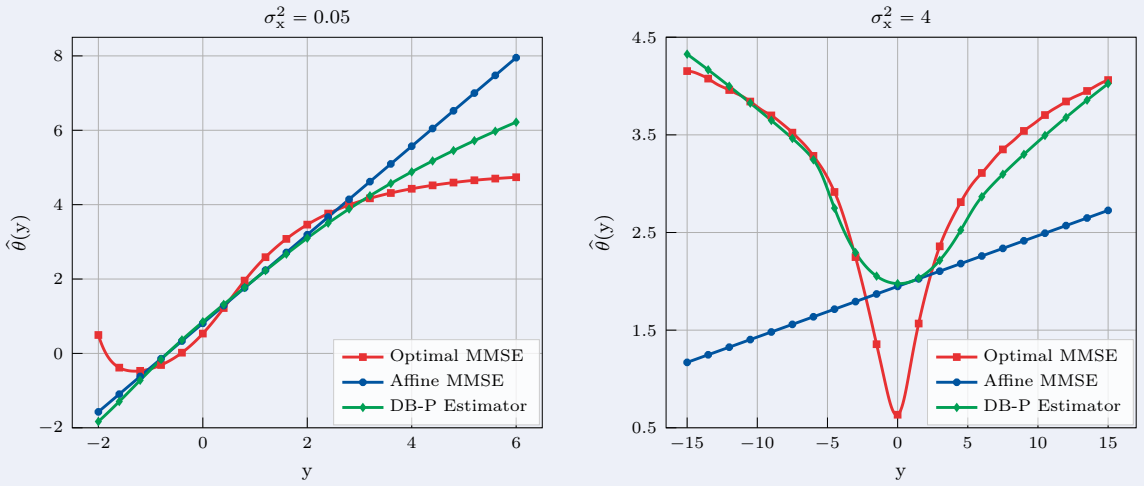


FIGURE 4. DB-P, optimal MMSE (5), and affine (10a) estimators.

In the above example, the basis function was linear, so estimating basis-function evaluations is equivalent to estimating the latent state. However, for nonlinear basis functions, this equivalence no longer holds, and the dual basis-parameter estimator can become both statistically and computationally inefficient. These limitations motivate an alternative, state-based DBS construction, in which the latent states are estimated directly, and the basis representation is treated as a derived quantity. In particular, the following remark provides a more detailed discussion of the main limitations of the dual basis-parameter estimation approach.

Remark 3.3 (**Limitations of the dual basis-parameter estimator**). *Two notable limitations of the dual basis-parameter estimator are*

(i) ***Elliptical-distribution approximation:*** *The affine MMSE estimators rely only on mean and covariance statistics and are exact when the joint distributions are Gaussian or, more generally, elliptical. In non-Gaussian settings, the performance of such estimators is limited to what one can achieve by approximating the true prior and posterior distributions of the unknown quantities by elliptical models, which may not hold in practice. In particular, the distribution of the stacked basis-function vector* $\phi = \mathrm{vec}(\Phi(\mathrm{x}))$ *is generally unknown (e.g., when using Fourier basis functions), making such elliptical approximations potentially poor.*

(ii) ***High-dimensional basis representation:*** *In practice, the number of unknown basis functions,* $(\mathrm{N}+1)(T+1)$*, typically exceeds the dimensionality of the state sequence* x*, which is* $(n_{\mathrm{x}})(T+1)$*. This high-dimensional representation increases both the statistical and computational complexity of the estimation problem.*

Given these limitations, an alternative approach exploits the system dynamics in (1) by first estimating the latent states and then computing DBS from these estimates statistics. This state-based DBS approach, developed in the next subsection, estimates the system states directly rather than the basis-function evaluations.

3.2. **Dual state-parameter estimation**

In contrast to the previous approach, which directly estimates the basis collection as in (14c), this approach constructs DBS estimates from the state estimate sequence $\widehat{\mathrm{x}}(\mathrm{y}) = [\widehat{\mathrm{x}}_0^{\mathsf{T}}(\mathrm{y}), \ldots, \widehat{\mathrm{x}}_T^{\mathsf{T}}(\mathrm{y})]^{\mathsf{T}}$ and its corresponding covariance $\Sigma_{\widehat{\mathrm{x}}}(\mathrm{y})$. This approach is particularly effective when the process noise is Gaussian, addressing a key limitation of the previous method discussed in Remark 3.3.

Assumption 1 (**Gaussian process noise**). *Throughout this subsection, the initial state and the process noise are drawn from a Gaussian distribution, i.e.,* $\mathrm{x}_0 \sim \mathcal{N}(\mu_{\mathrm{x}_0}, \Sigma_{\mathrm{x}_0})$ *and* $\mathrm{w}_{t+1} \sim \mathcal{N}(0, \Sigma_{\mathrm{w}_{t+1}})$.

Under this assumption, the state trajectory admits a Gaussian prior distribution. We adopt the *lifted* representation of the process model (1) as

$$\mathrm{x} = \mathrm{A}(\mathrm{B}\mathrm{u} + \mathrm{w}), \tag{18}$$

where the input sequence $\mathrm{u} = [\mu_{\mathrm{x}_0}^{\mathsf{T}}, \mathrm{u}_0^{\mathsf{T}}, \ldots, \mathrm{u}_{T-1}^{\mathsf{T}}]^{\mathsf{T}}$ starts with the initial state mean, and the process noise sequence $\mathrm{w} = [\mathrm{w}_0^{\mathsf{T}}, \mathrm{w}_1^{\mathsf{T}}, \ldots, \mathrm{w}_T^{\mathsf{T}}]^{\mathsf{T}}$ includes the initial state uncertainty, with $\mathrm{w}_0 \sim \mathbb{P}(0, \Sigma_{\mathrm{x}_0})$. The lifted matrices A and B have the following lower-triangular and block-diagonal structure, respectively:

$$\mathrm{A} = \begin{bmatrix} \mathbb{I} & 0 & 0 & \ldots & 0 \\ \mathrm{A}_0 & \mathbb{I} & 0 & \ldots & \vdots \\ \mathrm{A}_1\mathrm{A}_0 & \mathrm{A}_1 & \mathbb{I} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \prod_{i=0}^{T-1} \mathrm{A}_i & \prod_{i=1}^{T-1} \mathrm{A}_i & \ldots & \mathrm{A}_{T-1} & \mathbb{I} \end{bmatrix}, \qquad \mathrm{B} = \mathrm{diag}(\mathbb{I}, \mathrm{B}_0, \ldots, \mathrm{B}_{T-1}), \tag{19}$$

where $\prod_{i=t}^{t'} \mathrm{A}_i = \mathrm{A}_{t'} \dots \mathrm{A}_t$. The process noise vector $\mathrm{w} \sim \mathcal{N}(0, \Sigma_\mathrm{w})$ follows a block-diagonal covariance structure under temporal independence, $\Sigma_\mathrm{w} = \mathrm{diag}(\Sigma_{\mathrm{x}_0}, \Sigma_{\mathrm{w}_1}, \dots, \Sigma_{\mathrm{w}_T})$. Temporal correlations may also be incorporated if the corresponding covariance matrix is known. With the lifted representation (18), the state trajectory prior becomes $\mathrm{x} \sim \mathcal{N}(\mathrm{ABu}, \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T})$, which is Gaussian and enables a systematic computation of the DBS. Since a Gaussian distribution is fully characterized by its mean and covariance, the DBS computation in Definition 1 can be specialized to an operator requiring only these moments rather than the full distribution or higher moments. The following Gaussian DBS operator formalizes this specialization.

Definition 2 (**Gaussian DBS operator**). *For a Gaussian state trajectory $\nu \sim \mathcal{N}(\mu_\nu, \Sigma_\nu)$, the DBS can be computed from only the mean and covariance as*

$$(\mu_{\Phi(\nu)}, \Sigma_{\Phi(\nu)}) = \Phi_\#(\mu_\nu, \Sigma_\nu), \tag{20}$$

where $\Phi_\# \colon \mathbb{R}^{n_\mathrm{x}(T+1)} \times \mathbb{R}^{(n_\mathrm{x}(T+1))^2} \to \mathbb{R}^{(\mathrm{N}+1)(T+1)} \times \mathbb{R}^{((\mathrm{N}+1)(T+1))^2}$ maps the first two moments of the state trajectory to the basis statistics as in Definition 1.

Given the linearity of the basis function in Example 1, the Gaussian DBS operator reduces to the identity map. To highlight its nontrivial behaviour, we now consider another one-dimensional dynamical system in which the output function (2) contains a single sinusoidal basis function. In this setting, the Gaussian DBS operator transforms the mean and covariance of a single state into the exact statistics of the corresponding basis-function evaluation, illustrated in the following example.

Example 2 (1-D Wiener model with sinusoidal basis function). Consider the following sinusoidal nonlinearity in the measurement model:

$$\mathrm{y} = \theta \sin(f\mathrm{x}) + \mathrm{v}. \tag{21}$$

where $\mathrm{x} \in \mathbb{R}$ is a scalar random variable representing the initial state, with $\mathrm{x} \sim \mathbb{P}(\mu_\mathrm{x}, \sigma_\mathrm{x}^2)$, and f is a known frequency. The Gaussian DBS operator $\Phi_\#(\cdot, \cdot)$ in Definition 2 maps the statistics of x to the following DBS, evaluated according to [52, Lem. 4.1, Ex. 1]:

$$\begin{aligned} \mu_\Phi &= \exp(-\frac{1}{2} f^2 \sigma_\mathrm{x}^2) \sin(f\mu_\mathrm{x}), \\ \Sigma_\Phi &= \frac{1}{2} - \frac{1}{2}\exp(-2f^2\sigma_\mathrm{x}^2)\cos(2f\mu_\mathrm{x}) - \exp(-f^2\sigma_\mathrm{x}^2)\sin^2(f\mu_\mathrm{x}). \end{aligned} \tag{22}$$

For the lifted prior in (18), the Gaussian DBS operator provides an equivalent computation of the basis statistics from Definition 1 via $(\mu_\Phi, \Sigma_\Phi) = \Phi_\#(\mathrm{ABu}, \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T})$. Once state estimates $\hat{\mathrm{x}}(\mathrm{y})$ and estimation-error covariance $\Sigma_{\hat{\mathrm{x}}}(\mathrm{y})$ are obtained, we apply a plug-in interpretation analogous to Remark 2.5: the estimated mean and covariance are treated as fixed values of a Gaussian prior distribution for the subsequent DBS computation, i.e., $\mathrm{x} \sim \mathcal{N}(\hat{\mathrm{x}}(\mathrm{y}), \Sigma_{\hat{\mathrm{x}}}(\mathrm{y}))$, analogous to an empirical Bayes procedure applied to the state trajectory. This procedure naturally leads to data-driven DBS estimates via the Gaussian DBS operator,

$$\big(\widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y})\big) = \Phi_\#(\hat{\mathrm{x}}(\mathrm{y}), \Sigma_{\hat{\mathrm{x}}}(\mathrm{y})). \tag{23}$$

The affine MMSE state estimator is a natural choice for providing $\widehat{\mathrm{x}}(\mathrm{y})$ and $\Sigma_{\widehat{\mathrm{x}}}(\mathrm{y})$, as it computes the posterior mean and covariance under a Gaussian approximation of the state trajectory posterior. The derivation of the affine MMSE state estimator as a direct consequence of Lemma 2.1 requires the cross-covariance Σ_{xy} between the state trajectory and observation sequence. Under Assumption 1, this cross-covariance can be explicitly derived using Stein's identity.

Remark 3.4 (**Cross-covariance via Stein's identity**). *For a random vector $\nu \sim \mathcal{N}(\mu, \Sigma)$ and a differentiable function h with appropriate integrability conditions (sufficient for applying Fubini's theorem [40]), Stein's identity [50] states*

$$\mathbb{E}[(\nu - \mu)h(\nu)] = \Sigma\mathbb{E}[\nabla_\nu h(\nu)]. \tag{24}$$

While this study applies Stein's identity under Gaussian noise, the identity extends to elliptical distributions [31], enabling broader applicability to process noise models. Additionally, extensions to other distributions such as Poisson and t-distributions exist [36], suggesting potential avenues for non-elliptical process noise distributions.

The following lemma leverages this identity to derive a closed-form affine MMSE state estimator, which requires computing the mean of the Jacobian matrix of the basis-function evaluations with respect to the state distribution.

Lemma 3.5 (**Affine MMSE state estimator**). *Under Assumption 1, suppose that for each basis function $\phi_n(\cdot)$ defined in (2), the gradient vector $\nabla_{\mathrm{x}_t}\phi_n(\mathrm{x}_t)$, for $t \in \{0, \ldots, T\}$, has components that are continuous almost everywhere and satisfy the bounded expectation condition*

$$\nabla_{\mathrm{x}_t}\phi_n(\mathrm{x}_t) = \left[\frac{\partial\phi_n(\mathrm{x}_t)}{\partial \mathrm{x}_{t,1}}, \ldots, \frac{\partial\phi_n(\mathrm{x}_t)}{\partial \mathrm{x}_{t,n_\mathrm{x}}}\right]^\mathsf{T}, \quad \mathbb{E}\Big[\Big|\frac{\partial\phi_n(\mathrm{x}_t)}{\partial \mathrm{x}_{t,i}}\Big|\Big] < \infty, \quad n = 0, \ldots, \mathrm{N}, \quad i = 1, \ldots, n_\mathrm{x}, \tag{25}$$

where $\mathrm{x}_{t,i}$ denotes the i^{th} element of the state vector $\mathrm{x}_t \in \mathbb{R}^{n_\mathrm{x}}$. Then, the affine MMSE estimator for the state trajectory (18), its estimation-error covariance, and the optimal MSE, all as functions of $(\mu_\theta, \Sigma_\theta)$, are given by

$$\begin{cases} \widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y}; \mu_\theta, \Sigma_\theta) = \Psi^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta)\mathrm{y} + \psi^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta), \\ \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\mu_\theta, \Sigma_\theta) = \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T} - \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T}\mathrm{C}^\mathsf{T}\mu_\Theta(\mu_\Theta^\mathsf{T}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_\mathrm{v})^{-1}\mu_\Theta^\mathsf{T}\mathrm{C}\mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T}, \\ \mathcal{J}^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta) = \mathrm{tr}(\Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\mu_\theta, \Sigma_\theta)), \end{cases} \tag{26a}$$

with the optimal coefficient matrices

$$\Psi^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta) = \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T}\mathrm{C}^\mathsf{T}\mu_\Theta(\mu_\Theta^\mathsf{T}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_\mathrm{v})^{-1}, \quad \psi^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta) = \mathrm{A}\mathrm{B}\mathrm{u} - \Psi^\star_\mathrm{x}(\mu_\theta, \Sigma_\theta)\mu_\Phi^\mathsf{T}\mu_\Phi, \tag{26b}$$

where $(\mu_\Phi, \Sigma_\Phi) = \Phi_\#(\mathrm{A}\mathrm{B}\mathrm{u}, \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T})$ is the DBS for the Gaussian state trajectory prior from Definition 2, $\mu_\Theta = \mathrm{diag}(\mu_\theta^0, \ldots, \mu_\theta^T)$ with $\mu_\theta^t = \mu_\theta$ for $t \in \{0, \ldots, T\}$, the linear operator $\mathcal{S}(\cdot)$ is defined in (14b), and

$$\mathrm{C} = \mathrm{diag}(\mathrm{C}_0, \ldots, \mathrm{C}_T), \qquad \mathrm{C}_t = \mathbb{E}[\mathrm{J}_\phi(\mathrm{x}_t)], \tag{26c}$$

with $\mathrm{J}_\phi(\mathrm{x}_t) = \left[\nabla_{\mathrm{x}_t}\phi_0(\mathrm{x}_t), \ldots, \nabla_{\mathrm{x}_t}\phi_\mathrm{N}(\mathrm{x}_t)\right]^\mathsf{T}$ being the Jacobian matrix of basis functions evaluated at x_t.

A detailed proof is presented in Appendix A.4. The explicit computation and tractability of the affine MMSE state estimator depend on the choice of basis functions, since the closed-form expression (26a) requires both the DBS (μ_Φ, Σ_Φ) and the expected Jacobian of the basis functions for constructing

the matrix C in (26c). The sinusoidal basis of Example 2 satisfies the bounded expectation condition in (25), and the corresponding matrix C admits a closed-form expression, as discussed below.

Example 2 (continued). In the context of (21), the matrix C is given by

$$\mathrm{C} = \mathbb{E}[\frac{\partial \phi(\mathrm{x})}{\partial \mathrm{x}}] = \mathbb{E}[f\cos(f\mathrm{x})] = f\exp(-\frac{1}{2}f^2\sigma_{\mathrm{x}}^2)\cos(f\mu_{\mathrm{x}}), \tag{27}$$

where $\mathrm{J}_\phi(\mathrm{x})$ in (26c) reduces to a scalar derivative since x is a scalar variable. Given the explicit expressions for the DBS quantities in (22) and for C in (27), the affine MMSE state estimator (26a) yields

$$\begin{aligned}
\widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta) &= \mu_{\mathrm{x}} + \frac{\sigma_{\mathrm{x}}^2\mathrm{C}\mu_\theta}{\mu_\theta^2\Sigma_\Phi + \sigma_\theta^2\Sigma_\Phi + \sigma_\theta^2\mu_\Phi^2 + \sigma_{\mathrm{v}}^2}(\mathrm{y} - \mu_\Phi\mu_\theta),\\
\Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta) &= \sigma_{\mathrm{x}}^2 - \frac{\sigma_{\mathrm{x}}^4\mathrm{C}^2\mu_\theta^2}{\mu_\theta^2\Sigma_\Phi + \sigma_\theta^2\Sigma_\Phi + \sigma_\theta^2\mu_\Phi^2 + \sigma_{\mathrm{v}}^2}.
\end{aligned} \tag{28}$$

Furthermore, the affine MMSE parameter and basis estimators for the setting of (21) follow their descriptions in Example 1, given in (11) and (16), respectively, after replacing the pair $(\mu_{\mathrm{x}}, \sigma_{\mathrm{x}}^2)$ in those expressions with the DBS quantities obtained in (22).

In the more general Fourier-basis setting, both the DBS, (μ_Φ, Σ_Φ), and the matrix C admit explicit expressions via the characteristic function of the Gaussian distribution [52, Lem. 4.1, Ex. 1], relying on the lifted representation (18). The following remark shows that the affine MMSE state estimates preserve this lifted structure, enabling the same explicit formulas to be applied in the plug-in construction (23).

Remark 3.6 (Lifted plug-in prior for DBS estimates). *An equivalent representation of the affine MMSE state estimator in Lemma 3.5 is*

$$\widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta) = \mathrm{A}(\mathrm{Bu}+\zeta), \qquad \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta) = \mathrm{A}(\Sigma_{\mathrm{w}} - Z)\mathrm{A}^{\mathsf{T}},$$

with measurement-driven corrections

$$\begin{cases}
\zeta = \Sigma_{\mathrm{w}}\mathrm{A}^{\mathsf{T}}\mathrm{C}^{\mathsf{T}}\mu_\Theta(\mu_\Theta^{\mathsf{T}}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_{\mathrm{v}})^{-1}(\mathrm{y} - \mu_\Phi^{\mathsf{T}}\mu_\theta),\\
Z = \Sigma_{\mathrm{w}}\mathrm{A}^{\mathsf{T}}\mathrm{C}^{\mathsf{T}}\mu_\Theta(\mu_\Theta^{\mathsf{T}}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_{\mathrm{v}})^{-1}\mu_\Theta^{\mathsf{T}}\mathrm{CA}\Sigma_{\mathrm{w}}.
\end{cases}$$

Thus, the plug-in empirical Bayes Gaussian prior on x *induced by the affine estimator can be written in lifted form as* $\mathrm{x} \sim \mathcal{N}(\mathrm{A}(\mathrm{Bu}+\zeta), \mathrm{A}(\Sigma_{\mathrm{w}} - Z)\mathrm{A}^{\mathsf{T}})$. *For Fourier basis functions, this structure results in closed-form DBS estimates in (23) via the explicit characteristic-function formulas of* [52, Ex. 1].

The affine state estimator in (26a) is parameterized by the prior statistics $(\mu_\theta, \Sigma_\theta)$. Since knowledge of θ would improve these estimates, we adopt the empirical Bayes perspective of Remark 2.5 and construct a nonlinear state estimator by inserting the parameter estimates $\big(\widehat{\theta}(\mathrm{y}), \Sigma_{\widehat{\theta}}(\mathrm{y})\big)$ as if they were the true prior moments,

$$\begin{cases}
\widehat{\mathrm{x}}_{\mathrm{nl}}(\mathrm{y}) = \widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y};\widehat{\theta}(\mathrm{y}),\Sigma_{\widehat{\theta}}(\mathrm{y})),\\
\Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}}(\mathrm{y}) = \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\widehat{\theta}(\mathrm{y}),\Sigma_{\widehat{\theta}}(\mathrm{y})).
\end{cases} \tag{29}$$

Using (29) within the DBS estimates construction (23) for the nonlinear parameter estimator (13), and simultaneously feeding the resulting parameter estimates back into (29), induces a circular dependence: the state estimates require the parameter estimates, while the parameter estimates depend on DBS computed from those same state estimates. This coupling gives rise to a nonlinear dual state-parameter estimator, defined as a solution of the resulting system of algebraic equations.

Dual state-parameter estimator: For any measurements vector y, there exists a tuple of state-parameter estimators $\big(\widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}, \widehat{\mathrm{x}}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\mathrm{x}}_{\mathrm{nl}}}\big)$ such that

$$\begin{cases} \widehat{\theta}^{\star}_{\mathrm{nl}} = \widehat{\theta}_{\mathrm{af}}\big(\mathrm{y}; \Phi_{\#}(\widehat{\mathrm{x}}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\mathrm{x}}_{\mathrm{nl}}})\big), \\ \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}} = \Sigma_{\widehat{\theta}_{\mathrm{af}}}\big(\Phi_{\#}(\widehat{\mathrm{x}}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\mathrm{x}}_{\mathrm{nl}}})\big), \end{cases} \qquad \begin{cases} \widehat{\mathrm{x}}^{\star}_{\mathrm{nl}} = \widehat{\mathrm{x}}_{\mathrm{af}}\big(\mathrm{y}; \widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}\big) \\ \Sigma^{\star}_{\widehat{\mathrm{x}}_{\mathrm{nl}}} = \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}\big(\widehat{\theta}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\theta}_{\mathrm{nl}}}\big) \end{cases} \tag{DS-P}$$

where $\Phi_{\#}(\widehat{\mathrm{x}}^{\star}_{\mathrm{nl}}, \Sigma^{\star}_{\widehat{\mathrm{x}}_{\mathrm{nl}}})$ returns the DBS estimates $\big(\widehat{\Phi}(\mathrm{y}), \Sigma_{\widehat{\Phi}}(\mathrm{y})\big)$ according to (23) and Definition 2, and the functions $\big(\widehat{\theta}_{\mathrm{af}}(\cdot), \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\cdot)\big)$ and $\big(\widehat{\mathrm{x}}_{\mathrm{af}}(\cdot), \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\cdot)\big)$ are the closed form solutions of the affine Bayesian MMSE defined in (10a) and (26a), respectively.

The dual state-parameter (DS-P) estimation framework likewise couples two affine MMSE estimators, one for the parameters and one for the states, each depending on the estimates produced by the other. Although this approach addresses the limitations of DB-P estimator raised in Remark 3.3, the asymptotic performance of the parameter estimates with infinitely many samples depends critically on the properties of the underlying dynamical system [52, Prop. 4.2].

Remark 3.7 (**Asymptotic performance**). *While the state-estimation step in DS-P estimator reduces the effective process-noise uncertainty, consistency of the nonlinear parameter estimator requires that the state-estimation error remain suitably bounded over the trajectory, playing a role analogous to the multiple-trajectory condition in* [52, Prop. 4.3]. *Characterizing conditions under which the state-estimation error is bounded and the nonlinear parameter estimator is consistent is an interesting direction for future research. Nevertheless, for finite data, the numerical experiments in Section 4 demonstrate that the proposed dual state-parameter estimator can substantially outperform existing affine and nonlinear alternatives when Fourier bases are employed.*

The mutual dependence between the estimator characterizations in DB-P and DS-P naturally suggests viewing the nonlinear parameter estimator architecture (cf. Figure 2) as a fixed point of suitable mappings. This perspective motivates computing the estimator via an iterative fixed-point scheme, in which the coupled estimators are applied repeatedly until convergence. From this viewpoint, each iteration can be interpreted as propagating information between the DBS and parameter estimator blocks, gradually reconciling their respective posterior summaries. Once the iterative scheme has converged, the resulting parameter estimates generally exhibit a highly nonlinear dependence on the measurement value. Whether the proposed estimators produce estimates close to the optimal affine MMSE parameter estimator or deviate significantly from it depends sensitively on the measurement region and on the prior distribution of the latent system states. To illustrate this nonlinear behaviour in a concrete setting, we now revisit Example 2 and examine the DB-P and DS-P estimators for different measurement values, while the discussion of the algorithmic aspects of the fixed-point characterization is deferred to the next subsection.

Example 2 (continued). For the parameter-estimation simulations, consider the observation model (21) with sinusoidal frequency $f = \frac{\pi}{6}$, a scalar latent state $\mathrm{x} \sim \mathcal{N}(\mu_{\mathrm{x}}, \sigma_{\mathrm{x}}^2)$ (interpreted as the initial state of a dynamical system) with $\mu_{\mathrm{x}} = 0.5$ and two choices of variance, $\sigma_{\mathrm{x}}^2 = 0.05$ (left plot) and $\sigma_{\mathrm{x}}^2 = 4$ (right plot), an unknown scalar parameter $\theta \in \mathbb{R}$ with prior $\theta \sim \mathcal{U}(-1, 5)$, and measurement noise $\mathrm{v} \sim \mathcal{N}(0, \sigma_{\mathrm{v}}^2)$ with $\sigma_{\mathrm{v}}^2 = 0.16$. Figure 5 shows the estimates of the unknown parameter θ as a function of the measurement value, i.e., $\widehat{\theta}(\mathrm{y})$, for four estimators: the DS-P estimator (orange), the DB-P estimator (green), the empirical MMSE estimator computed via MCMC methods (red), and the affine MMSE estimator obtained from (10a) (blue). As anticipated from the fixed-point perspective, the dual estimators exhibit notable nonlinearities relative to the affine MMSE estimator. The algorithms used to construct these dual estimators are described in the next subsection.

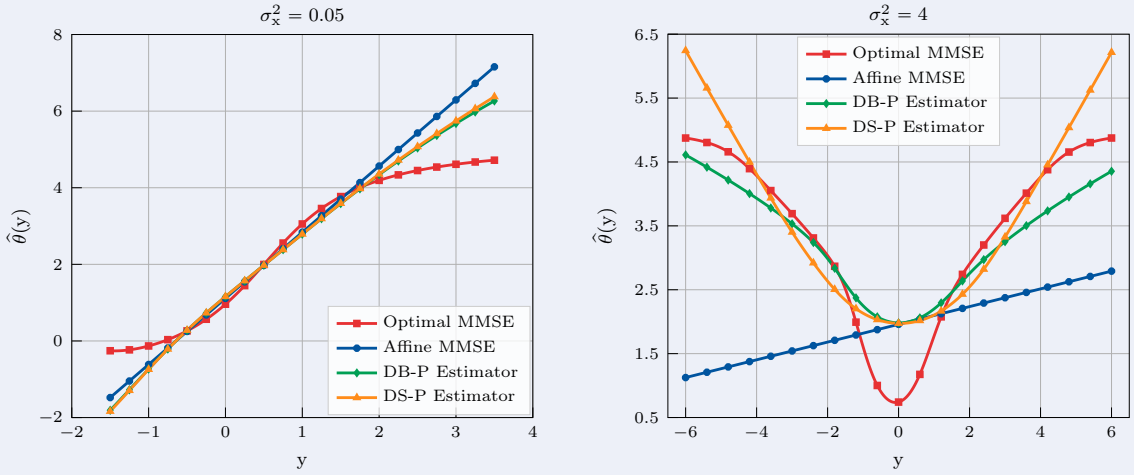


FIGURE 5. DS-P, DB-P, optimal MMSE (5), and affine MMSE (10a) estimators

3.3. Fixed-point algorithm

The dual basis-parameter (DB-P) and dual state-parameter (DS-P) estimators can be equivalently characterized by a fixed-point solution $\zeta^\star$ of an operator $\mathfrak{F}$, satisfying $\zeta^\star = \mathfrak{F}(\zeta^\star)$. We collect the estimates and their covariances into the tuple: $\zeta = \big(\widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}, \widehat{\Phi}_{\mathrm{nl}}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}\big)$ for DB-P and $\zeta = \big(\widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}, \widehat{\mathrm{x}}_{\mathrm{nl}}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}}\big)$ for DS-P. For a given measurement sequence y, the corresponding fixed-point operators are defined as

$$\begin{cases} \textbf{DB-P}: & \mathfrak{F}(\zeta) := \big(\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \widehat{\Phi}_{\mathrm{nl}}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}),\ \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\widehat{\Phi}_{\mathrm{nl}}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}),\ \widehat{\Phi}_{\mathrm{af}}(\mathrm{y}; \widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}),\ \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}})\big), \\ \textbf{DS-P}: & \mathfrak{F}(\zeta) := \big(\widehat{\theta}_{\mathrm{af}}\big(\mathrm{y}; \Phi_{\#}(\widehat{\mathrm{x}}_{\mathrm{nl}}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}})\big),\ \Sigma_{\widehat{\theta}_{\mathrm{af}}}\big(\Phi_{\#}(\widehat{\mathrm{x}}_{\mathrm{nl}}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}})\big),\ \widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y}; \widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}),\ \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}})\big), \end{cases}$$

respectively. A natural way to compute such a fixed point is via the simultaneous (Jacobi-style) iteration $\zeta^{(k+1)} = \mathfrak{F}(\zeta^{(k)})$, whose pseudo-code for both variants is summarized in Algorithm 1. To improve the stability of the iterations, it is sometimes recommended to use an averaged update with $0 < \alpha \leqslant 1$, at the cost of a slower convergence rate,

$$\zeta^{(k+1)} = \zeta^{(k)} + \alpha\big(\mathfrak{F}(\zeta^{(k)}) - \zeta^{(k)}\big), \tag{30}$$

where $\alpha = 1$ corresponds to the Jacobi-style iteration, used in Algorithm 1. Alternating updates (Gauss–Seidel-style), where the components of ζ are updated sequentially rather than simultaneously, are also common for coupled fixed-point problems and may improve convergence [10, Ch. 7].

Remark 3.8 (**Convergence condition**). *The convergence of the sequence of iterates* $\{\zeta^{(k)}\}_{k\geqslant 0}$ *in* (30) *depends on the properties of the operator* $\mathfrak{F}$. *A sufficient condition is that* $\mathfrak{F}$ *is nonexpansive* [4, Thm. 5.14]. *A rigorous analysis of the convergence of these iterations and the properties of* $\mathfrak{F}$ *for both dual estimators is left for future work; however, the empirical results provide encouraging evidence supporting this behaviour.*

Building on this fixed-point characterization, the iterative scheme in Algorithm 1 for computing the dual estimators initializes the means and covariances of the parameters and DBS-related quantities. At iteration $k = 1$, the estimates in the tuple $\zeta^{(1)}$ are obtained from a single pass of the Bayesian MMSE affine estimators for the unknown parameters together with the DBS or system states, depending on the chosen variant. In the DB-P variant, the associated DBS is computed once from the prior of the system states x according to Definition 1. In contrast, in the DS-P variant, the DBS is recomputed at every iteration by applying the Gaussian DBS operator $\Phi_{\#}(\cdot,\cdot)$ to the current state mean and covariance estimates. In this variant, however, the matrix C in (26c) is evaluated only once, since it depends solely on the prior of the state x, which does not change across iterations. The stopping criterion can be chosen based on the maximum number of iterations or on the deviation between subsequent iterates. In our implementation, we use a threshold ϵ on the improvement in the parameter MSE cost $\mathcal{J}_\theta(\cdot,\cdot)$ in (10a) between two iterations as the exit condition, i.e., $\left|\mathcal{J}_\theta^{(k)} - \mathcal{J}_\theta^{(k-1)}\right| < \epsilon$. Both algorithm variants, DB-P and DS-P, converged to a fixed-point solution in almost all extensive numerical experiments reported in Section 4, with the DS-P variant exhibiting rapid convergence. Moreover, both variants have a per-iteration computational complexity of $\mathcal{O}(T^3)$ when $\mathrm{N}, n_{\mathrm{x}} \ll T$, dominated by the covariance matrix inversions in the affine MMSE updates.

Algorithm 1 Dual estimator: basis-parameter (**DB-P**) or state-parameter (**DS-P**)

Variant: Choose either **DB-P** or **DS-P**
Require: Tolerance ϵ, maximum iterations K
Ensure: $(\widehat{\theta}_{\mathrm{nl}}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}})$ and **DB-P**: $(\widehat{\Phi}_{\mathrm{nl}}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}})$ or **DS-P**: $(\widehat{\mathrm{x}}_{\mathrm{nl}}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}})$

$(\widehat{\theta}_{\mathrm{nl}}^{(0)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(0)}) \leftarrow (\mu_\theta, \Sigma_\theta)$ ▷ θ prior information
$\mathcal{J}_\theta^{(0)} \leftarrow 0$ ▷ initial cost
$(\widehat{\Phi}_{\mathrm{nl}}^{(0)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(0)}) \leftarrow (\mu_\Phi, \Sigma_\Phi)$ ▷ DBS, Definition 1
for $k = 1, 2, \ldots, K$ **do**
  $(\widehat{\theta}_{\mathrm{nl}}^{(k)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(k)}) \leftarrow \big(\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \widehat{\Phi}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(k-1)}), \Sigma_{\widehat{\theta}_{\mathrm{af}}}(\widehat{\Phi}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(k-1)})\big)$ ▷ affine MMSE, (10a)
  $\mathcal{J}_\theta^{(k)} \leftarrow \mathcal{J}_\theta^{\star}(\widehat{\Phi}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(k-1)})$ ▷ associated cost
  DB-P: $(\widehat{\Phi}_{\mathrm{nl}}^{(k)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(k)}) \leftarrow \big(\widehat{\Phi}_{\mathrm{af}}(\mathrm{y}; \widehat{\theta}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(k-1)}), \Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\widehat{\theta}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(k-1)})\big)$ ▷ affine MMSE, (14c)
  DS-P: $(\widehat{\mathrm{x}}_{\mathrm{nl}}^{(k)}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}}^{(k)}) \leftarrow \big(\widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y}; \widehat{\theta}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(k-1)}), \Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}}(\widehat{\theta}_{\mathrm{nl}}^{(k-1)}, \Sigma_{\widehat{\theta}_{\mathrm{nl}}}^{(k-1)})\big)$ ▷ affine MMSE, (26a)
  DS-P: $(\widehat{\Phi}_{\mathrm{nl}}^{(k)}, \Sigma_{\widehat{\Phi}_{\mathrm{nl}}}^{(k)}) \leftarrow \Phi_{\#}(\widehat{\mathrm{x}}_{\mathrm{nl}}^{(k)}, \Sigma_{\widehat{\mathrm{x}}_{\mathrm{nl}}}^{(k)})$ ▷ Gaussian DBS, Definition 2
  if $\left|\mathcal{J}_\theta^{(k)} - \mathcal{J}_\theta^{(k-1)}\right| < \epsilon$ **then** ▷ stopping criterion
    break
  end if
end for

4. Numerical Experiments

In this section, we evaluate the proposed nonlinear estimators, dual basis-parameter (DB-P) and dual state-parameter (DS-P), against the affine MMSE parameter estimator (Theorem 2.2), hereafter called A-MMSE, across two setups: ($\mathrm{N} = 2, n_{\mathrm{x}} = 10$) with more states than parameters, and ($n_{\mathrm{x}} = 2, \mathrm{N} = 10$) with more parameters than states. We then benchmark the dual state-parameter estimator (DS-P) against two established sampling-based methods: Particle Gibbs with Ancestor Sampling (PGAS) [35] and SMC-EM [48]. Both methods perform SMC-based smoothing at each iteration to draw from the state posterior. SMC methods suffer from particle degeneracy, which necessitates an exponential growth in the number of particles with state dimension [6]. While modern methods such as Nested Sequential Monte Carlo (NSMC) [41] can improve SMC scalability in high dimensions, we restrict the benchmark comparisons to the second setup ($n_{\mathrm{x}} = 2, \mathrm{N} = 10$) for two reasons: (i) The first setup primarily illustrates a regime where the dual basis-parameter estimator underperforms the affine MMSE estimator, demonstrating the limitations of the elliptical-distribution approximation discussed in Remark 3.3; (ii) Adapting PGAS and SMC-EM to use NSMC-based smoothing would require substantial additional development and tuning, beyond the scope of a fair performance comparison. Focusing on the lower-dimensional state case ($n_{\mathrm{x}} = 2, \mathrm{N} = 10$) mitigates the impact of state dimensionality on SMC-based methods and enables a more meaningful and interpretable comparison. We use a marginally stable linear time-invariant (LTI) dynamical system ($\mathrm{A}_t = \mathbb{I}$) with a true function in the Fourier subspace

$$\begin{cases} \phi_0(\mathrm{x}) = 1 & n = 0 \\ \phi_n(\mathrm{x}) = \sum\limits_{\ell \in \{-1,1\}} \exp(j\langle \ell f_n, \mathrm{x} \rangle) & n \geqslant 1, \end{cases} \tag{31}$$

where $f_n \in \mathbb{R}^{n_{\mathrm{x}}}$ represents a *known* frequency. This Fourier basis choice enables explicit computation of the DBS and cross-covariance terms, and it allows us to examine how increasing state uncertainty affects estimator performance as the number of measurements grows. The experiments use time-invariant Gaussian noise: $\mathrm{v}_t \sim \mathcal{N}(0, \sigma_{\mathrm{v}}^2 \mathbb{I})$, $\mathrm{w}_{t+1} \sim \mathcal{N}(0, \sigma_{\mathrm{w}}^2 \mathbb{I})$, and $\mathrm{x}_0 \sim \mathcal{N}(\mu_{\mathrm{x}_0}, \sigma_{\mathrm{x}_0}^2 \mathbb{I})$, with $\sigma_{\mathrm{x}_0}^2 = \sigma_{\mathrm{w}}^2$. To assess the impact of process noise, we vary $\sigma_{\mathrm{w}}^2 \in \{0.001,\ 0.01\}$ while holding the measurement-noise variance fixed at $\sigma_{\mathrm{v}}^2 = 0.01$ across all experiments. The input trajectories are generated using the active-learning procedure from [52, Sec. 5], conducted prior to the experiments and optimized for the affine MMSE estimator. These same inputs are used for all methods in the study to ensure a fair comparison. For reproducibility, a MATLAB library implementing the affine MMSE estimator with optimal input design, as well as the dual basis-parameter and dual state-parameter estimators, is available at https://github.com/sasanvakili/nonlinearBayesian4Wiener.

For each configuration, we generate 100 independent samples of θ and 100 independent samples of (w, v) for each trajectory length T, resulting in 10,000 experiments in total. We evaluate estimators using the squared-error criterion $||\theta - \widehat{\theta}(\mathrm{y})||^2$ and ensure fair comparison by using identical realizations of w and v across all methods. In the provided plots, dashed lines show the empirically computed MSE, $\mathbb{E}[||\theta - \widehat{\theta}(\mathrm{y})||^2]$, for each method. Non-histogram plots include shaded regions indicating the 15^{th}-85^{th} percentile range of the squared error. Histograms display the probability density of $\log ||\theta - \widehat{\theta}(\mathrm{y})||^2$ on a logarithmic vertical scale to resolve differences in the squared-error distributions. All histograms within each plot use the same bin width, and each histogram integrates to unity.

Experiment setup 1 ($\mathbf{N = 2, n_x = 10}$). This setup considers an LTI system with $n_\mathrm{x} = 10$ states and $n_\mathrm{u} = 2$ control inputs. The system dynamics are $\mathrm{A}_t = \mathbb{I}$ and $\mathrm{B}_t = \Delta t[\mathbb{I}, \ldots, \mathbb{I}]^\intercal$, where $\Delta t = 0.1$ is the sampling interval and the identity matrices have dimensions conformable with n_x and n_u. The initial state is drawn from a standard normal distribution, $\mu_{\mathrm{x}_0} \sim \mathcal{N}(0, \mathbb{I})$ and the input sequence $\mathrm{u}_t = 4.5\sum_{\upsilon\in\Upsilon}[\cos(\upsilon t\Delta t),\ \sin(\upsilon t\Delta t)]^\intercal$, where $\Upsilon \in \{3, 5, 10, 20, 100\}$, is used to initialize the projected steepest descent algorithm in the active-learning procedure of [52, Sec. 5], with each component constrained to $[-200, 200]$. The optimal input $\mathrm{u}^\star$ is precomputed once and used for all methods evaluated under this configuration. The unknown parameters follow the Fourier basis representation in (31). We use three unknown parameters θ_n $(n = 0, 1, 2)$ with known frequency vectors $f_n \in \mathbb{R}^{10}$. Specifically, $f_0 = 0$ (the constant basis), and the nonzero frequencies are drawn independently as $f_n \sim \mathcal{N}(0, \mathbb{I})$ for $n \in \{1, 2\}$. The prior over the unknown parameters is uniform, $\theta_n \sim \mathcal{U}(2, 8)$ for each n, implying prior moments $\mu_{\theta_n} = 5$ and $\sigma^2_{\theta_n} = 3$. True parameter values are generated from this same prior.

Benchmark 1.1 (**Dual estimators: high-dimensional states**). This benchmark uses Experiment setup 1 with trajectory length $T = 100$ (i.e., 101 measurements) and two process-noise variances $\sigma^2_\mathrm{w} \in \{0.001, 0.01\}$, with 10,000 simulations per configuration. We compute parameter estimates $\widehat{\theta}(\mathrm{y})$ from A-MMSE (10a) and the dual estimators DB-P and DS-P. Figure 6 shows histograms of $\log \|\theta - \widehat{\theta}(\mathrm{y})\|^2$ over all simulations, with dashed vertical lines indicating the empirical MSE: orange for DB-P, blue for A-MMSE, and red for DS-P. Although one might expect DB-P to perform best, given that only two unknown basis variables and a scalar observation y_t are involved at each time, this setup provides a counterexample. In both noise regimes, DB-P leads to a higher empirical MSE than A-MMSE. This underperformance is consistent with the limitations of the elliptical prior approximation used in the basis estimation (cf. Remark 3.3). With a scalar measurement at each time and a high state dimension ($n_\mathrm{x} = 10$), the latent states remain substantially uncertain, limiting the benefit of dual state-parameter estimation relative to the affine approach. Nevertheless, DS-P consistently achieves the lowest empirical MSE (red), demonstrating that joint estimation of latent states and parameters improves parameter inference on average when the latent state process is Gaussian (cf. Assumption 1).

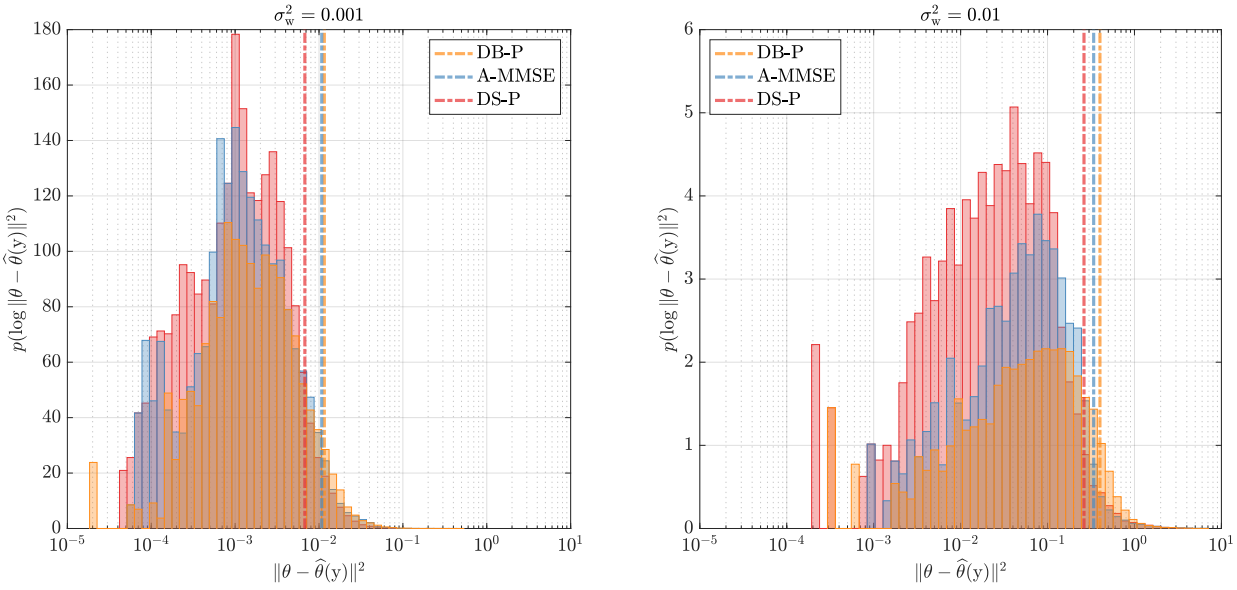


FIGURE 6. Parameter estimation squared-error distributions for Experiment setup 1

Experiment setup 2 ($\mathbf{N = 10, n_x = 2}$). This setup focuses on the complementary configuration of low-dimensional states and a larger number of parameters. The LTI system has two states, $\mathrm{x}_t \in \mathbb{R}^2$, and shares the same structure as Experiment setup 1: $\mathrm{u}_t \in \mathbb{R}^2$, $\mathrm{A}_t = \mathbb{I}$, and $\mathrm{B}_t = \Delta t \mathbb{I}$, with $\Delta t = 0.1$. The initial state is set to $\mu_{\mathrm{x}_0} = [3.2, 2.8]^\mathsf{T}$. The active-learning procedure uses the same input initialization structure $\mathrm{u}_t = 4.5 \sum_{\upsilon \in \Upsilon} [\cos(\upsilon t \Delta t),\ \sin(\upsilon t \Delta t)]^\mathsf{T}$, with $\Upsilon \in \{3, 5, 10, 20, 100\}$ and bounds $[-200, 200]$, and the resulting optimal trajectory $\mathrm{u}^\star$ is used for all methods in this setup. Under the Fourier basis representation in (31), we use eleven unknown parameters θ_n ($n = 0, \ldots, 10$) with known frequency vectors $f_n \in \mathbb{R}^2$ defined as: $f_0 = [0, 0]^\mathsf{T}$, $f_n = [n\frac{2\pi}{10}, 0]^\mathsf{T}$ for $n \in \{1, 2, 3\}$, and $f_n = [(n-7)\frac{2\pi}{10}, \frac{2\pi}{6}]^\mathsf{T}$ for $n \in \{4, \ldots, 10\}$. The unknown parameters follow the prior $\mathcal{U}(2, 8)$, corresponding to $\mu_{\theta_n} = 5$ and $\sigma^2_{\theta_n} = 3$, from which true parameter values are sampled.

Benchmark 2.1 (**Dual estimators: high-dimensional parameters**). This benchmark uses Experiment setup 2 with trajectory length $T = 100$ (i.e., 101 measurements) and two process-noise variances $\sigma^2_{\mathrm{w}} \in \{0.001, 0.01\}$, with 10,000 simulations per configuration. Using identical realizations of (w, v) across methods, we compute parameter estimates $\widehat{\theta}(\mathrm{y})$ from A-MMSE (10a) and the dual estimators DB-P and DS-P. Figure 7 shows histograms of $\log \|\theta - \widehat{\theta}(\mathrm{y})\|^2$ over all simulations, with dashed vertical lines indicating the empirical MSE for each estimator: orange for A-MMSE, blue for DB-P, and red for DS-P. In contrast to Experiment setup 1, this configuration has only two latent states but ten unknown parameters, leading to a practically relevant identification problem with high-dimensional parameters and low-dimensional dynamics. In both noise regimes, A-MMSE (orange) exhibits a higher empirical MSE than both dual estimators. The low state dimension ($n_{\mathrm{x}} = 2$) enables DS-P to exploit latent-state dynamics more effectively than DB-P, achieving the lowest empirical MSE (red) and demonstrating that explicit state estimation can substantially enhance parameter inference when Gaussian linear dynamics govern the latent states.

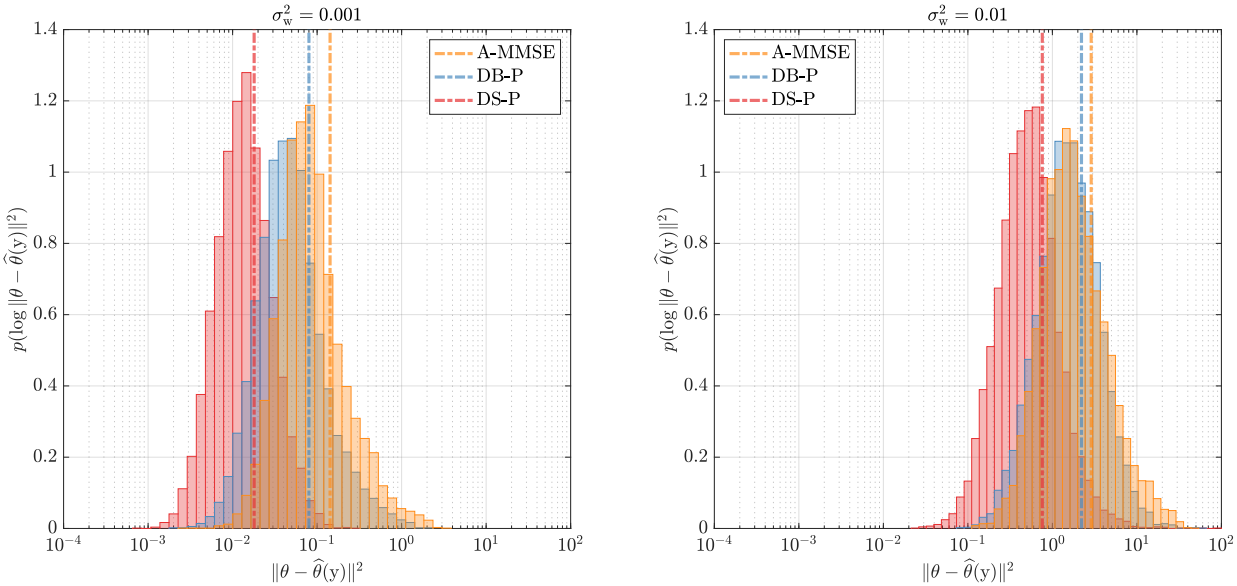


FIGURE 7. Parameter estimation squared-error distributions for Experiment setup 2

Benchmark 2.2 (**Trajectory horizon effects**). This experiment considers 10,000 simulations with trajectories of varying lengths, $T \in \{0, 4, 10, 13, 16, 20, 25, 32, 40, 50, 63, 79, 100\}$, to examine how parameter estimation error evolves as the number of measurements increases. Figure 8 compares the

squared errors for A-MMSE (orange), DB-P (blue), and DS-P (red) under two process-noise variances $\sigma_{\mathrm{w}}^2 \in \{0.001, 0.01\}$. Shaded regions show the 15^{th}-85^{th} percentile range of the squared error, while dashed lines indicate the empirical MSE. All estimators use an optimized input $\mathrm{u}^\star$ designed separately for each trajectory length. The results confirm that input optimization significantly reduces squared error for all three estimators when the number of measurements matches the number of unknown parameters. For both noise levels, the error reduction of A-MMSE and DB-P slows markedly as T increases and measurements become more strongly correlated. This behaviour aligns with the established limitations of A-MMSE for marginally stable systems [52, Prop. 4.2], and the similar trend for DB-P suggests it may suffer from related limitations. In contrast, DS-P exhibits a sustained downward trend in empirical MSE as T grows. In the low process-noise case ($\sigma_{\mathrm{w}}^2 = 0.001$, left plot), the red dashed line decreases steadily with increasing T. For higher process noise ($\sigma_{\mathrm{w}}^2 = 0.01$, right plot), the red dashed line shows slight local increases between $T = 32$ and 40 and between 50 and 63, before decreasing again for larger T. These small increases reflect heavier upper tails (above the 85^{th} percentile) of the squared-error distribution, as the central 15^{th}-85^{th} percentile range continues to decrease steadily. These results highlight the advantages of DS-P in the high-dimensional parameter regime and motivate a comparison with SMC-based methods in Benchmark 2.3.

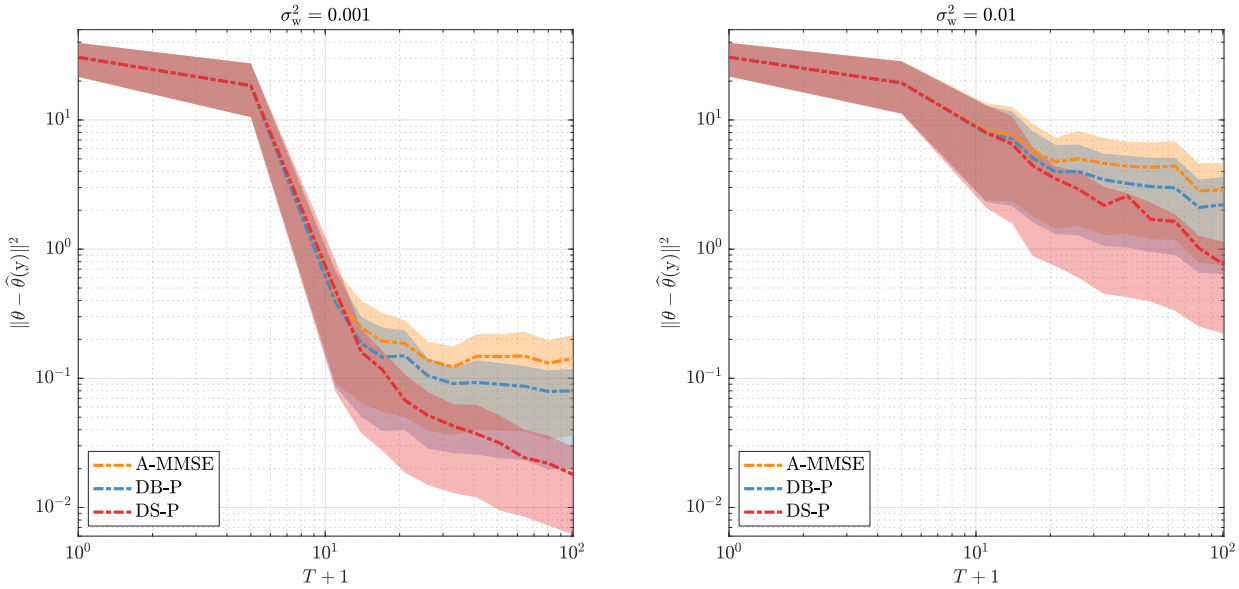


FIGURE 8. Parameter estimation squared-error vs. trajectory length T for Experiment setup 2

Benchmark 2.3 (**Dual state-parameter vs. SMC methods**). In this final benchmark, DS-P is compared with two well-known SMC-based approaches, PGAS and SMC-EM. Both methods alternate between updating the latent states given the current parameter estimates and updating the parameters given sampled state trajectories, whereas DS-P alternates between parameter and state updates using summary statistics of the other block rather than full conditional sampling. The comparison uses Experiment setup 2 with 10,000 simulations, trajectory length $T = 100$, and two process-noise levels $\sigma_{\mathrm{w}}^2 \in \{0.001, 0.01\}$. Identical realizations of (w, v) and the same optimized input $\mathrm{u}^\star$ from the active-learning procedure are used for all three methods. We briefly discuss each of the two SMC-based methods and refer interested readers to the cited references for further background and technical details. The following describes how each method is tuned for a fair benchmark comparison:

(i) **PGAS** [35, Alg. 4]: PGAS is a PMCMC method that targets the joint posterior over the latent state trajectory and the parameters. At each iteration k, it alternates between: (i) drawing a new state trajectory x^k using the PGAS Markov kernel [35, Alg. 3] given θ^{k-1}, and (ii) sampling θ^k from the conditional posterior $p(\theta \mid \mathrm{x}^k, \mathrm{y})$. In our configuration, we use $N_P = 6000$ particles and run the algorithm for $K = 5000$ iterations. The initial state trajectory is obtained from a bootstrap particle filter, and the initial parameters are drawn from the prior $\mathcal{U}(2, 8)$. At each iteration, the parameters are sampled from a multivariate truncated Gaussian distribution [33], corrected by an MH step to target the conditional posterior. The final estimate is the empirical mean of the retained samples, with the first 2500 samples discarded as burn-in.

(ii) **SMC-EM** [48, Alg. 4]: SMC-EM is a maximum-likelihood method that embeds particle smoothing within the EM framework. At each EM iteration k, it treats the full state trajectory as missing data and alternates between: (i) E-step: given θ^{k-1}, approximating the EM Q-function (the expected complete-data log-likelihood) by running a particle smoother to approximate the smoothing distribution of x; and (ii) M-step: maximizing this approximate Q-function with respect to θ to obtain θ^k. In our implementation, the original forward-backward smoother [48, Alg. 2-3] is replaced by the MH-FFBP joint smoother [9, Fig. 5], which is an MCMC-based refinement of forward filtering backward smoothing methods [21, 28]. We use a standard particle filter with $N_F = 6000$ particles in the forward pass [9, Fig. 1] and, in the backward pass, $N_S = 700$ smoothing particles with MH chain length $M = 300$. This choice follows the complexity trade-off analysis in [9] and aims to improve computational efficiency while maintaining particle diversity. With a uniform prior on θ, the M-step reduces to a constrained maximization of the approximate Q-function over the prior support, equivalent to a MAP update. The algorithm is initialized at $\theta^0 = \mu_\theta$ and run for a fixed $K = 1000$ iterations, since the particle-based approximation of the Q-function precludes reliable likelihood-based convergence diagnostics. The final estimate represents a local optimum of the Monte Carlo-based approximation of the posterior objective rather than a guaranteed stationary point of the true posterior.

From a practical standpoint, the three methods differ notably in tuning requirements and computational effort. PGAS and SMC-EM require careful selection of multiple hyperparameters (particle numbers N_P for PGAS and (N_F, N_S) for SMC-EM, plus the chain length M in the MH-FFBP smoother), which must be set conservatively large to mitigate Monte Carlo error, thereby increasing computational cost. In contrast, DS-P requires no such Monte Carlo tuning beyond specifying the prior mean and covariance. A further practical advantage of DS-P is its rapid convergence: we set the stopping tolerance $\epsilon = 10^{-6}$ in Algorithm 1 with a maximum of $K = 10{,}000$ iterations. For $\sigma_{\mathrm{w}}^2 = 0.001$, the state-parameter variant (DS-P) of Algorithm 1 requires at most 56 iterations to satisfy the stopping criterion across all experiments. For $\sigma_{\mathrm{w}}^2 = 0.01$, only 3 out of 10,000 experiments reach the maximum iteration limit, while the remaining runs converge in at most 229 iterations. In contrast, PGAS required 5000 iterations for sufficient posterior samples, and SMC-EM was run for 1000 iterations with no reliable convergence diagnostics.

Figure 9 compares the distributions of $\log \|\theta - \widehat{\theta}(\mathrm{y})\|^2$ over all 10,000 simulations for PGAS (green), SMC-EM (blue), and DS-P (red), with dashed vertical lines indicating the empirical MSEs. In the low process-noise case ($\sigma_{\mathrm{w}}^2 = 0.001$), SMC-EM produces an error distribution that overlaps partially

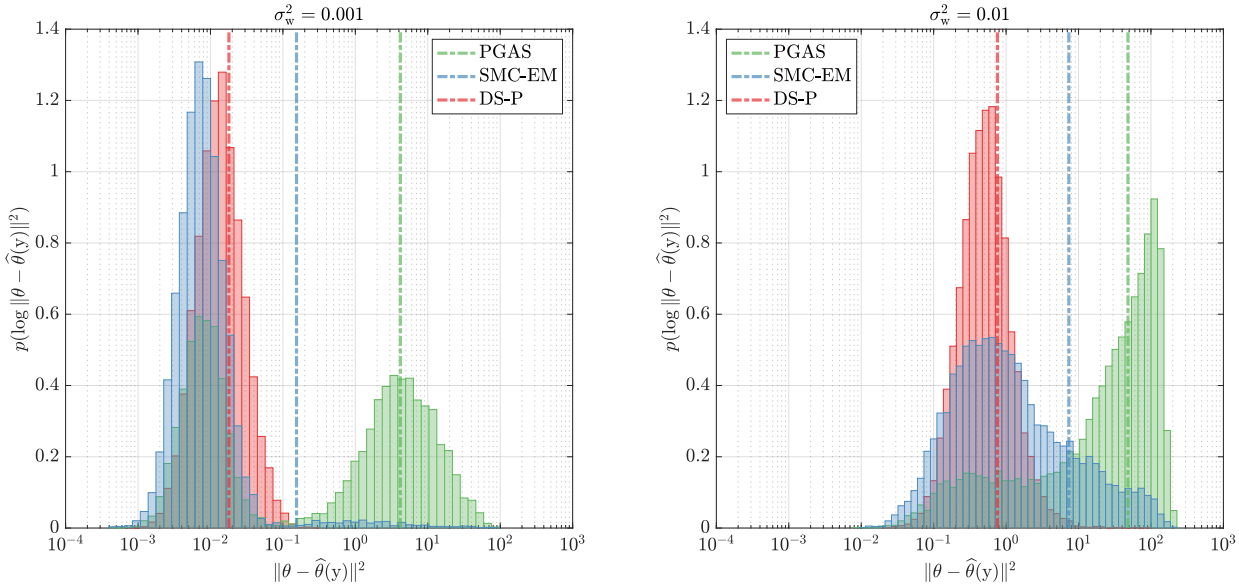


FIGURE 9. Squared-error distributions: dual state-parameter vs. SMC-based methods

with that of DS-P but also exhibits a substantial number of simulations with noticeably larger errors, raising its overall MSE. PGAS shows a similar pattern: one concentration of errors near DS-P and another at much larger values, again leading to a higher MSE. When the process noise increases to $\sigma_w^2 = 0.01$, the squared-error distributions of both SMC-EM and PGAS shift further away from the main concentration region of DS-P, and their empirical MSEs increase correspondingly. In this higher-noise regime, the SMC-based methods produce a small fraction of simulations with errors comparable to DS-P, while most realizations fall into higher-error regions. In both noise scenarios, DS-P attains the lowest empirical MSE and the most tightly concentrated error distribution, demonstrating more robust parameter-estimation performance as the process-noise level increases.

5. CONCLUSIONS

This work presents a nonlinear parameter estimator as a fixed-point characterization of two affine MMSE estimators, one for the unknown parameters and one for latent variables, which use the summary statistics of each other to iteratively refine their estimates. Two variants are developed, both requiring no tuning beyond specifying the prior: the dual basis-parameter (DB-P) estimator, which directly estimates the DBS via an affine MMSE basis estimator, and the dual state-parameter (DS-P) estimator, which uses affine MMSE state estimates as inputs for DBS estimates construction. Extensive numerical experiments validate that DS-P consistently achieves the lowest empirical MSE, outperforming DB-P, the affine MMSE parameter estimator, and two established SMC-based methods (PGAS and SMC-EM). The superior performance of DS-P stems from two mechanisms: the state-estimation step reduces the effective process-noise uncertainty, and the joint fixed-point iteration mutually refines both parameter and state estimates.

These results motivate several directions for future research. Although both algorithm variants converged to a fixed point in almost all experiments, a formal convergence analysis of the dual fixed-point iteration would provide valuable theoretical guarantees. Additionally, when the dual state-parameter

variant (DS-P) is used, the consistency of the nonlinear parameter estimator depends on the state estimation error remaining suitably bounded along the trajectory. Characterizing conditions that guarantee such boundedness would establish asymptotic performance guarantees for the proposed estimator. For finite data, however, extensive simulations indicate that the dual state-parameter estimator can substantially reduce parameter estimation error as the trajectory length increases (cf. Figure 8). Other promising extensions include generalizing the framework to nonparametric models, accommodating nonlinear state dynamics, and developing recursive implementations that enable scalability to large datasets by reducing computational complexity.

Appendix A. Technical proofs

A.1. **Lemma 2.1**

Proof. The proof follows the first-order optimality over the constants of the affine estimator (6). To this end, we apply the decompositions $\xi_1 = \mu_1 + \Delta\xi_1$ and $\xi_2 = \mu_2 + \Delta\xi_2$, where $\Delta\xi_1$ and $\Delta\xi_2$ are zero-mean random variables describing the uncertainty in ξ_1 and ξ_2, and then expand the objective as

$$\min_{\Psi,\,\psi} \mathbb{E}[(\Delta\xi_1 - \Psi\Delta\xi_2)^{\mathsf{T}}(\Delta\xi_1 - \Psi\Delta\xi_2)] - 2\mathbb{E}[(\Delta\xi_1 - \Psi\Delta\xi_2)^{\mathsf{T}}(\psi + \Psi\mu_2 - \mu_1)] + \mathbb{E}[(\psi + \Psi\mu_2 - \mu_1)^{\mathsf{T}}(\psi + \Psi\mu_2 - \mu_1)].$$

The second term of the first line vanishes because $\psi + \Psi\mu_2 - \mu_1$ is deterministic, and the random variables $(\Delta\xi_1, \Delta\xi_2)$ are zero mean, i.e., $\mathbb{E}[\Delta\xi_1] = 0$ and $\mathbb{E}[\Delta\xi_2] = 0$. The last term is minimized and its contribution set to zero by choosing $\psi^\star = \mu_1 - \Psi\mu_2$. With this choice, applying the trace operator, and defining the covariances $\Sigma_{11} \coloneqq \mathbb{E}[\Delta\xi_1\Delta\xi_1^{\mathsf{T}}]$, $\Sigma_{12} \coloneqq \mathbb{E}[\Delta\xi_1\Delta\xi_2^{\mathsf{T}}]$, and $\Sigma_{22} \coloneqq \mathbb{E}[\Delta\xi_2\Delta\xi_2^{\mathsf{T}}]$, where $\Delta\xi_1 = \xi_1 - \mu_1$ and $\Delta\xi_2 = \xi_2 - \mu_2$, the optimization reduces to minimizing

$$\mathcal{J}(\Psi) = \operatorname{tr}(\Sigma_{11} - 2\Psi\Sigma_{21} + \Psi\Sigma_{22}\Psi^{\mathsf{T}}),$$

with $\Sigma_{21} = \Sigma_{12}^{\mathsf{T}}$. The optimizer $\Psi^\star$ is found by setting the partial derivative of $\mathcal{J}(\Psi)$ with respect to Ψ to zero, i.e.,

$$\frac{\partial \mathcal{J}}{\partial \Psi}\Big|_{\Psi=\Psi^\star} = -2\Sigma_{12} + 2\Psi^\star\Sigma_{22} = 0,$$

which yields $\Psi^\star = \Sigma_{12}\Sigma_{22}^{-1}$. Since Σ_{22} is invertible, $\mathcal{J}(\Psi)$ is a strictly convex quadratic in Ψ, so this stationary point is the unique global minimizer. Substituting $\Psi^\star$ and $\psi^\star = \mu_1 - \Psi^\star\mu_2$ into the estimation-error covariance $\Sigma_{\hat{\xi}_1} \coloneqq \mathbb{E}[(\xi_1 - \Psi\xi_2 - \psi)(\xi_1 - \Psi\xi_2 - \psi)^{\mathsf{T}}]$, yields the expressions in (7a), which completes the proof. □

A.2. **Lemma 2.4**

Proof. By the law of total expectation [13, Ch. 4],

$$\mathcal{J}_\theta^\star(\mu_\Phi, \Sigma_\Phi) = \mathbb{E}\big[\|\theta - \widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)\|^2\big] = \mathbb{E}_{\mathrm{x}}\Big[\mathbb{E}\big[\|\theta - \widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)\|^2 \,\big|\, \mathrm{x}\big]\Big], \tag{32}$$

where $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi)$ is the affine MMSE estimator with its optimal MSE $\mathcal{J}_\theta^\star(\mu_\Phi, \Sigma_\Phi)$ as defined in (10a). Let $\mathscr{A}_\theta$ denote the class of affine estimators of θ from y, analogous to (6). Since $\widehat{\theta}_{\mathrm{af}}(\mathrm{y}; \mu_\Phi, \Sigma_\Phi) \in \mathscr{A}_\theta$, the conditional MSE of this estimator cannot be smaller than the minimum conditional MSE over all

affine estimators in $\mathscr{A}_\theta$ for each fixed x, i.e.,

$$\mathbb{E}_{\mathrm{x}}\Big[\mathbb{E}\big[\|\theta-\widehat{\theta}_{\mathrm{af}}(\mathrm{y};\mu_\Phi,\Sigma_\Phi)\|^2\,\big|\,\mathrm{x}\big]\Big]\geqslant\mathbb{E}_{\mathrm{x}}\Big[\min_{\widehat{\theta}(\cdot)\in\mathscr{A}_\theta}\mathbb{E}\big[\|\theta-\widehat{\theta}(\mathrm{y})\|^2\,\big|\,\mathrm{x}\big]\Big]. \tag{33}$$

Moreover, for each fixed state trajectory x, the optimizer of the minimum conditional MSE and its corresponding affine MMSE value are given by

$$\begin{cases}\widehat{\theta}_{\mathrm{af}}(\mathrm{y};\Phi(\mathrm{x}),0)=\operatorname{argmin}_{\widehat{\theta}(\cdot)\in\mathscr{A}_\theta}\mathbb{E}\big[\|\theta-\widehat{\theta}(\mathrm{y})\|^2\,\big|\,\mathrm{x}\big],\\ \mathcal{J}_\theta^\star(\Phi(\mathrm{x}),0)=\mathbb{E}\big[\|\theta-\widehat{\theta}_{\mathrm{af}}(\mathrm{y};\Phi(\mathrm{x}),0)\|^2\,\big|\,\mathrm{x}\big].\end{cases} \tag{34}$$

Combining (32), (33), and (34) arrives at the desired bound in (12). □

A.3. **Lemma 3.1**

Proof. Define the stacked basis-function vector

$$\phi\coloneqq\mathrm{vec}(\Phi(\mathrm{x}))=[\phi^{\mathsf{T}}(\mathrm{x}_0),\dots,\phi^{\mathsf{T}}(\mathrm{x}_T)]^{\mathsf{T}}, \tag{35}$$

where $\phi(\mathrm{x}_t)$ is defined in (2). By Definition 1, the prior DBS (μ_Φ,Σ_Φ) induce the prior mean and covariance of ϕ as

$$\bar{\phi}\coloneqq\mathbb{E}[\phi]=\mathbb{E}[\mathrm{vec}(\Phi(\mathrm{x}))]=\mathrm{vec}(\mu_\Phi),\qquad\mathbb{E}[(\phi-\bar{\phi})(\phi-\bar{\phi})^{\mathsf{T}}]=\Sigma_\Phi. \tag{36}$$

Next, introduce the lifted observation model over the trajectory. Stacking the sequence of measurements as $\mathrm{y}=[\mathrm{y}_0,\dots,\mathrm{y}_T]^{\mathsf{T}}$ and using (2) yields

$$\mathrm{y}=\mathrm{G}_\theta^{\mathsf{T}}\phi+\mathrm{v}, \tag{37}$$

where $\mathrm{G}_\theta=\mathrm{diag}(\theta^0,\dots,\theta^T)$ with $\theta^t=\theta$ for all $t\in\{0,\dots,T\}$, and $\mathrm{v}=[\mathrm{v}_0,\dots,\mathrm{v}_T]^{\mathsf{T}}$ has covariance matrix Σ_{v}. The affine MMSE estimator of ϕ given y follows from the generic affine MMSE estimator of Lemma 2.1 with $\xi_1=\phi$ and $\xi_2=\mathrm{y}$, resulting in the form $\widehat{\phi}_{\mathrm{af}}(\mathrm{y})=\Psi_\phi^\star\mathrm{y}+\psi_\phi^\star$, with coefficients

$$\Psi_\phi^\star=\Sigma_{\phi\mathrm{y}}\Sigma_{\mathrm{y}}^{-1},\qquad\psi_\phi^\star=\bar{\phi}-\Psi_\phi^\star\mu_{\mathrm{y}}, \tag{38}$$

where $\bar{\phi}$ is the mean of ϕ defined in (36), μ_{y} is the observation mean, $\Sigma_{\phi\mathrm{y}}$ is the cross-covariance between ϕ and y, and Σ_{y} is the observation covariance. To express these quantities, apply the decompositions: $\phi=\bar{\phi}+\Delta\phi$ and $\mathrm{G}_\theta=\mu_\Theta+\Delta\mathrm{G}_\theta$, where $\Delta\phi$ and $\Delta\mathrm{G}_\theta$ are zero-mean random quantities, and $\mu_\Theta\coloneqq\mathbb{E}[\mathrm{G}_\theta]=\mathrm{diag}(\mu_\theta^0,\dots,\mu_\theta^T)$ with $\mu_\theta^t=\mu_\theta$ for all $t\in\{0,\dots,T\}$. By construction, $\mathbb{E}[\Delta\phi\Delta\phi^{\mathsf{T}}]=\Sigma_\Phi$ and $\mathbb{E}[(\theta^t-\mu_\theta^t)(\theta^t-\mu_\theta^t)^{\mathsf{T}}]=\Sigma_\theta$. From the independence assumptions between ϕ, θ and v it follows that $\mathbb{E}[\Delta\mathrm{G}_\theta^{\mathsf{T}}\Delta\phi]=0$ and $\mathbb{E}[\Delta\mathrm{G}_\theta^{\mathsf{T}}\Delta\phi\mathrm{v}^{\mathsf{T}}]=0$. Using (37) together with these decompositions, and exploiting the zero-mean and independence properties, yields

$$\begin{aligned}\mu_{\mathrm{y}}&=\mathbb{E}[(\mu_\Theta+\Delta\mathrm{G}_\theta)^{\mathsf{T}}(\bar{\phi}+\Delta\phi)+\mathrm{v}]=\mu_\Theta^{\mathsf{T}}\bar{\phi}=\mu_\Phi^{\mathsf{T}}\mu_\theta,\\ \Sigma_{\phi\mathrm{y}}&=\mathbb{E}\big[\Delta\phi\big((\mu_\Theta+\Delta\mathrm{G}_\theta)^{\mathsf{T}}(\bar{\phi}+\Delta\phi)+\mathrm{v}-\mu_\Theta^{\mathsf{T}}\bar{\phi}\big)^{\mathsf{T}}\big]=\mathbb{E}[\Delta\phi\Delta\phi^{\mathsf{T}}]\mu_\Theta=\Sigma_\Phi\mu_\Theta,\\ \Sigma_{\mathrm{y}}&=\mathbb{E}[(\mathrm{y}-\mu_\Theta^{\mathsf{T}}\bar{\phi})(\mathrm{y}-\mu_\Theta^{\mathsf{T}}\bar{\phi})^{\mathsf{T}}]=\mu_\Theta^{\mathsf{T}}\Sigma_\Phi\mu_\Theta+\mathrm{P}+\mathrm{Q}+\mathrm{S}+\mathrm{S}^{\mathsf{T}}+\Sigma_{\mathrm{v}},\end{aligned} \tag{39}$$

where $\mathrm{P}=\mathbb{E}\big[\Delta\mathrm{G}_\theta^{\mathsf{T}}\Delta\phi\Delta\phi^{\mathsf{T}}\Delta\mathrm{G}_\theta\big]$, $\mathrm{Q}=\mathbb{E}\big[\Delta\mathrm{G}_\theta^{\mathsf{T}}\bar{\phi}\bar{\phi}^{\mathsf{T}}\Delta\mathrm{G}_\theta\big]$ and $\mathrm{S}=\mathbb{E}[\Delta\mathrm{G}_\theta^{\mathsf{T}}\bar{\phi}\Delta\phi^{\mathsf{T}}\Delta\mathrm{G}_\theta]$. Consider the $(t,t')^{\mathrm{th}}$ components of these matrices, denoted by $\mathrm{P}_{tt'}$, $\mathrm{Q}_{tt'}$, and $\mathrm{S}_{tt'}$. Using the block-diagonal structure of G_θ

and μ_Θ and denoting by $\Delta\theta$ the random parameter vector in each block, we have

$$\mathrm{P}_{tt'} = \mathbb{E}[\Delta\theta^\mathsf{T}\Delta\phi(\mathrm{x}_t)\Delta\phi^\mathsf{T}(\mathrm{x}_{t'})\Delta\theta] = \mathbb{E}[\mathrm{tr}(\Delta\theta^\mathsf{T}\Delta\phi(\mathrm{x}_t)\Delta\phi^\mathsf{T}(\mathrm{x}_{t'})\Delta\theta)] = \mathrm{tr}(\Sigma_\theta\Sigma_\phi^{tt'}),$$
$$\mathrm{Q}_{tt'} = \mathbb{E}[\Delta\theta^\mathsf{T}\mu_\phi^t\mu_\phi^{t'^\mathsf{T}}\Delta\theta] = \mathbb{E}[\mathrm{tr}(\Delta\theta^\mathsf{T}\mu_\phi^t\mu_\phi^{t'^\mathsf{T}}\Delta\theta)] = \mathrm{tr}(\Sigma_\theta\mu_\phi^t\mu_\phi^{t'^\mathsf{T}}),$$
$$\mathrm{S}_{tt'} = \mathbb{E}[\Delta\theta^\mathsf{T}\mu_\phi^t\Delta\phi^\mathsf{T}(\mathrm{x}_{t'})\Delta\theta] = \mathbb{E}[\mathrm{tr}(\Delta\theta^\mathsf{T}\mu_\phi^t\Delta\phi^\mathsf{T}(\mathrm{x}_{t'})\Delta\theta)] = \mathrm{tr}\big(\mathbb{E}[\Delta\theta\Delta\theta^\mathsf{T}]\mu_\phi^t\mathbb{E}[\Delta\phi^\mathsf{T}(\mathrm{x}_{t'})]\big) = 0,$$

which shows that $\mathrm{S} = 0$ since $\mathbb{E}[\Delta\phi(\mathrm{x}_{t'})] = 0$ by construction. Collecting these expressions, define the linear operator $\mathcal{S}\colon \mathbb{R}^{(\mathrm{N}+1)^2}\to\mathbb{R}^{(T+1)^2}$ by its $(t,t')^{\text{th}}$ element

$$\mathcal{S}^{tt'}(\Sigma_\theta) = \mathrm{tr}\big(\Sigma_\theta(\Sigma_\phi^{tt'} + \mu_\phi^t\mu_\phi^{t'^\mathsf{T}})\big).$$

Consequently, $\Sigma_\mathrm{y} = \mu_\Theta^\mathsf{T}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_\mathrm{v}$, which is invertible because $\Sigma_\mathrm{v} \succ 0$. Using this Σ_y together with μ_y and $\Sigma_{\phi\mathrm{y}}$ from (39) to substitute into (38) yields the closed-form expressions for $\Psi_\phi^\star(\mu_\theta,\Sigma_\theta)$ and $\psi_\phi^\star(\mu_\theta,\Sigma_\theta)$ in (14a), both as functions of $(\mu_\theta,\Sigma_\theta)$. Hence the affine MMSE basis estimator can be written as $\widehat{\phi}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta) = \Psi_\phi^\star(\mu_\theta,\Sigma_\theta)\mathrm{y} + \psi_\phi^\star(\mu_\theta,\Sigma_\theta)$, emphasizing the dependence of the estimator on the prior mean and covariance of the unknown parameters. Furthermore, the corresponding estimation-error covariance and MMSE of this estimator follow from (7a). Since $\widehat{\phi}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta)$ is stacked as in (35), its components $\widehat{\phi}_{\mathrm{af}}^t(\mathrm{y};\mu_\theta,\Sigma_\theta)$ directly define the column-wise DBS estimate $\widehat{\Phi}_{\mathrm{af}}(\mathrm{y};\mu_\theta,\Sigma_\theta)$, together with $\Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta)$ in (14c), with the corresponding optimal MMSE cost $\mathcal{J}_\phi^\star(\mu_\theta,\Sigma_\theta) = \mathrm{tr}(\Sigma_{\widehat{\Phi}_{\mathrm{af}}}(\mu_\theta,\Sigma_\theta))$. □

A.4. **Lemma 3.5**

Proof. The affine MMSE estimator of x given y follows from the generic affine MMSE estimator of Lemma 2.1 with $\xi_1 = \mathrm{x}$ and $\xi_2 = \mathrm{y}$, resulting in the form $\widehat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y}) = \Psi_\mathrm{x}^\star\mathrm{y} + \psi_\mathrm{x}^\star$, where the optimal coefficients based on (7b) are $\Psi_\mathrm{x}^\star = \Sigma_{\mathrm{xy}}\Sigma_\mathrm{y}^{-1}$ and $\psi_\mathrm{x}^\star = \mu_\mathrm{x} - \Psi_\mathrm{x}^\star\mu_\mathrm{y}$. Moreover, the estimation-error covariance and optimal MSE following (7a) are $\Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}} = \Sigma_\mathrm{x} - \Sigma_{\mathrm{xy}}\Sigma_\mathrm{y}^{-1}\Sigma_{\mathrm{xy}}^\mathsf{T}$ and $\mathcal{J}^\star = \mathrm{tr}(\Sigma_{\widehat{\mathrm{x}}_{\mathrm{af}}})$, respectively. From (18), we have $\mu_\mathrm{x} = \mathrm{ABu}$ and $\Sigma_\mathrm{x} = \mathrm{A}\Sigma_\mathrm{w}\mathrm{A}^\mathsf{T}$. Furthermore, from (39) in the proof of Lemma 3.1 (Appendix A.3), we have $\mu_\mathrm{y} = \mu_\Phi^\mathsf{T}\mu_\theta$ and $\Sigma_\mathrm{y} = \mu_\Theta^\mathsf{T}\Sigma_\Phi\mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_\mathrm{v}$. Thus, it remains to compute the cross-covariance matrix Σ_{xy} between the state trajectory x and the observation sequence y. In particular,

$$\Sigma_{\mathrm{xy}} = \mathbb{E}[(\mathrm{x} - \mathrm{ABu})(\Phi^\mathsf{T}(\mathrm{x})\theta + \mathrm{v} - \mu_\Phi^\mathsf{T}\mu_\theta)^\mathsf{T}] = \mathbb{E}[(\mathrm{x} - \mathrm{ABu})\mu_\theta^\mathsf{T}\Phi(\mathrm{x})], \tag{40}$$

where the observation model $\mathrm{y} = \Phi^\mathsf{T}(\mathrm{x})\theta + \mathrm{v}$ is used as in [52], and the equality follows from algebraic expansion using the independence of θ from x (and hence from $\Phi(\mathrm{x})$), which yields $\mathbb{E}[\mathrm{x}\theta^\mathsf{T}\Phi(\mathrm{x})] = \mathbb{E}[\mathrm{x}\mu_\theta^\mathsf{T}\Phi(\mathrm{x})]$, together with $\mathbb{E}[\mathrm{xv}^\mathsf{T}] = 0$ by independence of x and v. Consider the $(t,t')^{\text{th}}$ block of this cross-covariance matrix, $\Sigma_{\mathrm{xy}}^{tt'}$, which captures the cross-covariance between x_t and $\mathrm{y}_{t'}$. The state x_t can be written as $\mathrm{x}_t = \mathrm{A}_t(\mathrm{Bu} + \mathrm{w})$, where A_t is the $(t+1)^{\text{th}}$ block-row of A as in (19), i.e.,

$$\mathrm{A}_t = \Bigg[\prod_{i=0}^{t-1}\mathrm{A}_i \quad \prod_{i=1}^{t-1}\mathrm{A}_i \quad \dots \quad \mathbb{I} \quad 0 \quad \dots \quad 0\Bigg]. \tag{41}$$

The $(t,t')^{\text{th}}$ block of the cross-covariance matrix can then be expressed as

$$\Sigma_{\mathrm{xy}}^{tt'} = \mathbb{E}[(\mathrm{x}_t - \mathrm{A}_t\mathrm{Bu})\mu_\theta^\mathsf{T}\phi(\mathrm{x}_{t'})] = \sum_{n=0}^{\mathrm{N}}\mathbb{E}[(\mathrm{x}_t - \mathrm{A}_t\mathrm{Bu})\phi_n(\mathrm{x}_{t'})]\mu_{\theta_n}, \tag{42}$$

where $\phi(\mathrm{x}_{t'})$ is the $(t'+1)^{\text{th}}$ column of the basis collection matrix $\Phi(\mathrm{x})$, according to Definition 1, and μ_{θ_n} denotes the n^{th} component of the prior mean μ_θ for the unknown parameter vector θ. Under Assumption 1, $\mathrm{w} \sim \mathcal{N}(0, \Sigma_{\mathrm{w}})$, hence x_t and $\mathrm{x}_{t'}$ are both multivariate Gaussian random variables with $\mathrm{x}_t \sim \mathcal{N}(\mathrm{A}_t\mathrm{Bu}, \mathrm{A}_t\Sigma_{\mathrm{w}}\mathrm{A}_t^{\mathsf{T}})$ and $\mathrm{x}_{t'} \sim \mathcal{N}(\mathrm{A}_{t'}\mathrm{Bu}, \mathrm{A}_{t'}\Sigma_{\mathrm{w}}\mathrm{A}_{t'}^{\mathsf{T}})$, where $\mathrm{A}_{t'}$ is the $(t'+1)^{\text{th}}$ block-row of A analogous to (41). These variables can be represented as affine transformations of a standard Gaussian random vector $\nu \sim \mathcal{N}(0, \mathbb{I})$: $\mathrm{x}_t = \mathrm{A}_t\mathrm{Bu} + \mathrm{A}_t\Sigma_{\mathrm{w}}^{\frac{1}{2}}\nu$ and $\mathrm{x}_{t'} = \mathrm{A}_{t'}\mathrm{Bu} + \mathrm{A}_{t'}\Sigma_{\mathrm{w}}^{\frac{1}{2}}\nu$. Under the regularity conditions (25) on $\phi_n(\cdot)$ (continuity almost everywhere and bounded expectation of partial derivatives) for all $n \in \{0, \ldots, \mathrm{N}\}$, define $\mathrm{g}(\nu) \coloneqq \phi_n(\mathrm{A}_{t'}\mathrm{Bu} + \mathrm{A}_{t'}\Sigma_{\mathrm{w}}^{\frac{1}{2}}\nu)$. By Stein's identity [50],

$$\mathbb{E}[(\mathrm{x}_t - \mathrm{A}_t\mathrm{Bu})\phi_n(\mathrm{x}_{t'})] = \mathbb{E}[\mathrm{A}_t\Sigma_{\mathrm{w}}^{\frac{1}{2}}\nu\mathrm{g}(\nu)] = \mathrm{A}_t\Sigma_{\mathrm{w}}^{\frac{1}{2}}\mathbb{E}[\nabla_\nu \mathrm{g}(\nu)]. \tag{43}$$

Using the chain rule, $\mathbb{E}[\nabla_\nu \mathrm{g}(\nu)] = \Sigma_{\mathrm{w}}^{\frac{1}{2}}\mathrm{A}_{t'}^{\mathsf{T}}\mathbb{E}[\nabla_{\mathrm{x}}\phi_n(\mathrm{x}_{t'})]$, where $\nabla_{\mathrm{x}}\phi_n(\mathrm{x}_{t'})$ is the gradient of $\phi_n(\cdot)$ evaluated at $\mathrm{x}_{t'}$. Substituting this into (43) gives

$$\mathbb{E}[(\mathrm{x}_t - \mathrm{A}_t\mathrm{Bu})\phi_n(\mathrm{x}_{t'})] = \mathrm{A}_t\Sigma_{\mathrm{w}}\mathrm{A}_{t'}^{\mathsf{T}}\mathbb{E}[\nabla_{\mathrm{x}}\phi_n(\mathrm{x}_{t'})]. \tag{44}$$

Finally, inserting (44) into (42), we obtain

$$\Sigma_{\mathrm{xy}}^{tt'} = \sum_{n=0}^{\mathrm{N}} \mathrm{A}_t\Sigma_{\mathrm{w}}\mathrm{A}_{t'}^{\mathsf{T}}\mathbb{E}[\nabla_{\mathrm{x}}\phi_n(\mathrm{x}_{t'})]\mu_{\theta_n} = \mathrm{A}_t\Sigma_{\mathrm{w}}\mathrm{A}_{t'}^{\mathsf{T}}\mathrm{C}_{t'}^{\mathsf{T}}\mu_\theta, \tag{45}$$

where $\mathrm{C}_{t'} = \mathbb{E}[\mathrm{J}_\phi(\mathrm{x}_{t'})]$, and $\mathrm{J}_\phi(\mathrm{x}_{t'}) = \left[\nabla_{\mathrm{x}}\phi_0(\mathrm{x}_{t'}), \ldots, \nabla_{\mathrm{x}}\phi_{\mathrm{N}}(\mathrm{x}_{t'})\right]^{\mathsf{T}}$ is the Jacobian matrix evaluated at $\mathrm{x}_{t'}$. Assembling these blocks for all pairs (t, t') with $t, t' \in \{0, \ldots, T\}$ results in the full form of the cross-covariance matrix (40),

$$\Sigma_{\mathrm{xy}} = \mathrm{A}\Sigma_{\mathrm{w}}\mathrm{A}^{\mathsf{T}}\mathrm{C}^{\mathsf{T}}\mu_\Theta, \qquad \mathrm{C} = \mathrm{diag}(\mathrm{C}_0, \ldots, \mathrm{C}_T), \tag{46}$$

with $\mu_\Theta = \mathrm{diag}(\mu_\theta^0, \ldots, \mu_\theta^T)$ and $\mu_\theta^t = \mu_\theta$ for all $t \in \{0, \ldots, T\}$. Using Σ_{xy} from (46), along with the previously derived means and covariances, in the generic affine MMSE formulas provides the closed-form expressions for $\Psi_{\mathrm{x}}^\star(\mu_\theta, \Sigma_\theta)$ and $\psi_{\mathrm{x}}^\star(\mu_\theta, \Sigma_\theta)$ in (26b), both as functions of $(\mu_\theta, \Sigma_\theta)$. Hence, the affine MMSE state estimator $\hat{\mathrm{x}}_{\mathrm{af}}(\mathrm{y}; \mu_\theta, \Sigma_\theta) = \Psi_{\mathrm{x}}^\star(\mu_\theta, \Sigma_\theta)\mathrm{y} + \psi_{\mathrm{x}}^\star(\mu_\theta, \Sigma_\theta)$ explicitly emphasizes the dependence on the prior parameter statistics $(\mu_\theta, \Sigma_\theta)$. The corresponding estimation-error covariance $\Sigma_{\hat{\mathrm{x}}_{\mathrm{af}}}(\mu_\theta, \Sigma_\theta)$ and optimal MSE $\mathcal{J}_{\mathrm{x}}^\star(\mu_\theta, \Sigma_\theta)$ follow directly from (26a), completing the proof. □

References

[1] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72:269–342, 2010.

[2] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45:77–120, 2017.

[3] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43, 2017.

[4] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, Cham, 2nd edition, 2017.

[5] Matthew J. Beal and Zoubin Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:453–463, 2003.

[6] Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 316–335. Institute of Mathematical Statistics, 2008.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112:859–877, 2017.

[9] Pete Bunch and Simon Godsill. Improved particle approximations to the joint smoothing distribution using Markov chain Monte Carlo. *IEEE Transactions on Signal Processing*, 61:956–963, 2012.

[10] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Brooks/Cole, 9th edition, 2010.

[11] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[12] Bradley P. Carlin and Thomas A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2000.

[13] Matthew A. Carlton and Jay L. Devore. *Probability with Applications in Engineering, Science, and Technology*. Springer, 2017.

[14] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer International Publishing, 2020.

[15] Nicolas Chopin and Sumeetpal S. Singh. On particle Gibbs sampling. *Bernoulli*, 21:1855–1883, 2015.

[16] Jarrad Courts, Adrian G. Wills, Thomas B. Schön, and Brett Ninness. Variational system identification for nonlinear state-space models. *Automatica*, 147:110687, 2023.

[17] Hai-Dang Dau and Nicolas Chopin. On backward smoothing algorithms. *The Annals of Statistics*, 51:2145—-2169, 2023.

[18] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27:94–128, 1999.

[19] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38, 1977.

[20] Arnaud Doucet, Nando De Freitas, and Neil Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[21] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[22] Yariv Ephraim and Neri Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569, 2002.

[23] Yariv Ephraim and Neri Merhav. Lower and upper bounds on the minimum mean-square error in composite source signal estimation. *IEEE transactions on Information Theory*, 38:1709–1724, 2002.

[24] Nicholas Galioto and Alex Arkady Gorodetsky. Bayesian system ID: optimal management of parameter, model, and measurement uncertainty. *Nonlinear Dynamics*, 102:241–267, 2020.

[25] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[26] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013.

[27] Zoubin Ghahramani and Geoffrey E Hinton. Parameter estimation for linear dynamical systems. *Technical Report*, University of Toronto:CRG–TR–96–2, 1996.

[28] Simon J. Godsill, Arnaud Doucet, and Mike West. Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99:156–168, 2004.

[29] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[30] Simon Haykin. *Kalman Filtering and Neural Networks*. John Wiley & Sons, 2004.

[31] Zinoviy Landsman and Johanna Nešlehová. Stein's lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99:912–927, 2008.

[32] Bernard C. Levy. *Principles of Signal Detection and Parameter Estimation*. Springer, 2008.

[33] Yifang Li and Sujit K. Ghosh. Efficient sampling methods for truncated multivariate Normal and Student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, 9:712–732, 2015.

[34] Fredrik Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6274–6278, 2013.

[35] Fredrik Lindsten, Michael I. Jordan, and Thomas B. Schön. Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15:2145–2184, 2014.

[36] Jun S. Liu. Siegel's formula via Stein's identities. *Statistics & Probability Letters*, 21:247–251, 1994.

[37] Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, 2nd edition, 1999.

[38] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[39] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[40] Jan Mikusiński. The Fubini theorem. In *The Bochner Integral*, pages 91–105. Springer, 1978.

[41] Christian Naesseth, Fredrik Lindsten, and Thomas Schön. Nested sequential Monte Carlo methods. In *International Conference on Machine Learning*, pages 1292–1301, 2015.

[42] Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. Elements of sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12:307–392, 2019.

[43] Michael K. Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171:134–151, 2012.

[44] Bala Rajaratnam and Doug Sparks. MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *arXiv preprint arXiv:1508.00947*, 2015.

[45] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

[46] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[47] Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[48] Thomas B. Schön, Adrian Wills, and Brett Ninness. System identification of nonlinear state-space models. *Automatica*, 47:39–49, 2011.

[49] Maarten Schoukens and Koen Tiels. Identification of block-oriented nonlinear systems starting from linear approximations: A survey. *Automatica*, 85:272–292, 2017.

[50] Charles M. Stein. Estimation of the mean of a multivariate Normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

[51] Surya T. Tokdar and Robert E. Kass. Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:54–60, 2010.

[52] Sasan Vakili, Manuel Mazo Jr, and Peyman Mohajerin Esfahani. Optimal Bayesian affine estimator and active learning for the Wiener model. *arXiv preprint arXiv:2504.05490*, 2025.

[53] Eric Wan and Alex Nelson. Dual Kalman filtering methods for nonlinear prediction, smoothing and estimation. In *Advances in Neural Information Processing Systems*, volume 9, pages 793–799, 1996.