



Delft University of Technology

Document Version

Final published version

Citation (APA)

Vakili, S. (2026). *Nonlinear Parameter Estimators in Dynamic Environments: A Bayesian Approach*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:58ebf73e-c7a2-44d4-807a-83975e4e0b29>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

NONLINEAR PARAMETER ESTIMATORS IN DYNAMIC ENVIRONMENTS

A BAYESIAN APPROACH



SASAN VAKILI



NONLINEAR PARAMETER ESTIMATORS IN DYNAMIC ENVIRONMENTS

A BAYESIAN APPROACH

NONLINEAR PARAMETER ESTIMATORS IN DYNAMIC ENVIRONMENTS

A BAYESIAN APPROACH

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof. dr. ir. H. Bijl,
chair of the Board of Doctorates,
to be defended publicly on
Thursday 3 September 2026 at 15:00

by

Sasan VAKILI

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. B.H.K. De Schutter,	Delft University of Technology, <i>promotor</i>
Dr. P. Mohajerin Esfahani,	Delft University of Technology, <i>promotor</i> & University of Toronto, Canada

Independent members:

Prof. dr. ir. J.W. van Wingerden,	Delft University of Technology
Prof. dr. T.B. Schön,	Uppsala University, Sweden
Dr. ir. C.C. de Visser,	Delft University of Technology
Dr. M. Schoukens,	Eindhoven University of Technology
Prof. dr. R. Babuška,	Delft University of Technology, <i>reserve member</i>

This dissertation was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie Grant under Agreement 956200 and the ERC Starting Grant TRUST-949796.



Keywords: Bayesian inference, system identification, semidefinite programming, linear matrix inequalities, state-space models, Wiener models, active learning, nonlinear estimation

Cover: Sasan Vakili, concept and design with AI-assisted enhancement

Style: TU Delft House Style, with small modifications

The author has set this thesis in $\text{\LaTeX}$ using the Libertinus and Inconsolata fonts.

Copyright © 2026 by S. Vakili

ISBN 978-94-6563-015-1

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

Doubt is not a pleasant condition, but certainty is an absurd one.

Voltaire

To my parents, Sanaz and Solmaz

CONTENTS

Summary	xi
Samenvatting	xiii
Acknowledgments	xv
1 Introduction	1
1.1 Bayesian estimation in Wiener-type models	2
1.2 Existing approaches	4
1.2.1 Nonlinear system identification	5
1.2.2 Bayesian estimation methods.	6
1.3 Thesis contributions	10
1.4 Thesis outline	12
2 Time-Varying Parameter Estimators for Linear Systems	15
2.1 Introduction	16
2.2 Problem definition	17
2.3 Maximum a posteriori estimation	19
2.3.1 Lifted process and observation model	19
2.3.2 MAP loss function	20
2.4 Proposed solution	21
2.5 Numerical experiments.	23
2.6 Conclusion.	25
3 Affine Estimators and Active Learning for Wiener Models	27
3.1 Introduction	28
3.2 Problem description	30
3.3 Solution approaches	31
3.3.1 Classical regularized least-squares	32
3.3.2 Optimal Bayesian affine estimator	32
3.3.3 Technical proofs	34
3.4 Fourier basis	37
3.4.1 Lifted process model	38
3.4.2 Fourier basis statistics	38
3.4.3 Bayesian affine MMSE estimator consistency.	39
3.4.4 Technical proofs	40
3.5 Active learning.	44
3.6 Numerical experiments.	46
3.7 Conclusion.	50

4	Nonlinear Estimators via Fixed-Point Characterization	51
4.1	Introduction	52
4.2	Bayesian estimation for Wiener model	58
4.2.1	Problem description	58
4.2.2	From affine to nonlinear estimators	58
4.3	Fixed-point characterization	62
4.3.1	Dual basis-parameter estimation	62
4.3.2	Dual state-parameter estimation	65
4.3.3	Fixed-point algorithm	71
4.4	Numerical experiments.	73
4.5	Conclusion.	78
4.6	Technical proofs	79
5	Conclusions	85
5.1	Summary of contributions	85
5.2	Methodological insights	87
5.3	Future research directions	90
5.4	Concluding remarks	93
	Bibliography	95
	Glossary	108
	List of Publications	109

SUMMARY

This thesis develops a hierarchy of Bayesian estimation methods for Wiener-type state-space models, motivated by autonomous underwater vehicle (AUV) bathymetric mapping and related robotic perception problems. The focus is on a class of models in which a known linear dynamical process is observed through an unknown, possibly nonlinear output map whose parameters must be inferred from noisy input-output data. In this setting, the observation model is driven by latent, stochastic system states. The central objective is to design parameter estimators that are both statistically accurate and computationally tractable, enabling their embedding within navigation and mapping pipelines.

The work begins with a maximum a posteriori (MAP) estimator for identifying an unknown output map, formulated as a linear time-varying (LTV) observation-model identification problem. In this setting, the MAP estimation problem is posed over the entire state-parameter trajectory and shown to be non-convex. A semidefinite-programming (SDP) relaxation based on linear matrix inequalities (LMIs) is then derived to obtain a conservative but tractable approximation, whose solution serves as a warm start for quasi-Newton refinement. Numerical experiments validate the efficacy of the proposed method in terms of estimation accuracy and computational efficiency.

The thesis further develops a Bayesian affine minimum mean-squared error (MMSE) estimator for parametric Wiener models, in which the static nonlinear observation function is represented as a finite combination of known basis functions. A key ingredient is the dynamic basis statistics (DBS), which capture the mean and covariance of basis-function evaluations along the stochastic state trajectory, thereby summarizing how process noise propagates through the nonlinear observation model. Within this structure, a closed-form optimal affine MMSE estimator for the observation parameters is derived, expressed entirely in terms of DBS, prior statistics, and noise covariances. The resulting closed-form error expression is then used to design optimal inputs, also known as active learning, by selecting input signals that further reduce the mean-squared error of the optimal affine estimator. Specializing the framework to Fourier basis functions yields explicit DBS formulas that make the affine MMSE estimator computationally tractable in this setting, while also revealing an inherent inconsistency of the affine estimator for single-trajectory identification under stochastically unstable dynamics.

The framework is then extended beyond purely affine estimators by proposing a nonlinear estimator architecture based on fixed-point characterizations that couple two affine MMSE estimators: one for the unknown parameters and another for the latent variables. Two constructions are investigated. The dual basis-parameter (DB-P) estimator jointly uses an affine estimator for basis-function evaluations and the affine parameter estimator, whereas the dual state-parameter (DS-P) estimator first computes affine state estimates and covariances, then maps these statistics through a Gaussian DBS operator to obtain updated DBS. Both architectures give rise to coupled algebraic equations whose solutions can be implemented as fixed-point iterations that alternate between updating parameter estimates

and DBS or state statistics, using only means and covariances rather than full posterior distributions. Numerical experiments on representative Wiener-type models illustrate the effectiveness of these dual estimators and, in particular, the benefits of propagating both parameter and latent-variable statistics within a unified fixed-point framework.

In summary, the thesis develops a systematic perspective on Bayesian parameter estimation for Wiener-type models, starting with one-shot MAP formulations for LTV observation models and progressing to affine MMSE estimators and dual nonlinear fixed-point schemes built around DBS. Across these contributions, a recurring theme is the explicit exploitation of problem structure, namely, semidefinite relaxations in the LTV setting and DBS in the parametric nonlinear setting, to achieve estimators that are statistically principled without requiring full posterior inference. The affine MMSE estimator is further shown to be a natural building block for dual fixed-point architectures that propagate parameter and latent-variable statistics using only the mean and covariance, thereby extending the framework to more sophisticated nonlinear estimators. Collectively, these results offer a hierarchical set of Bayesian estimation tools for Wiener-type models and suggest a general approach to tractable, structure-exploiting estimators for nonlinear state-space problems.

SAMENVATTING

Dit proefschrift ontwikkelt een hiërarchie van Bayesiaanse schattingsmethoden voor Wiener-type toestandsruimtemodellen, gemotiveerd door bathymetrische kartering met AUV's en verwante problemen binnen de robotica. De focus ligt op een klasse modellen waarin een bekend lineair dynamisch proces wordt waargenomen via een onbekende, mogelijk niet-lineaire outputmap, waarvan de parameters moeten worden geschat uit met ruis verstoorde input-outputdata. In deze context wordt het observatiemodel aangedreven door latente, stochastische systeemtoestanden. Het centrale doel is het ontwikkelen van parameterschatters die zowel statistisch nauwkeurig als computationeel tractabel zijn, zodat ze kunnen worden ingebed in navigatie- en karteringspijplijnen.

Het onderzoek begint met een MAP-schatter voor het identificeren van een onbekende outputmap, geformuleerd als een LTV-identificatieprobleem voor observatiemodellen. In deze context wordt het MAP-schattingsprobleem geformuleerd over het volledige toestand-parametertraject en wordt aangetoond dat dit probleem niet-convex is. Vervolgens wordt een relaxatie op basis van semidefiniete programmering met lineaire matrixongelijkheden afgeleid om een conservatieve maar tractabele benadering te verkrijgen, waarvan de oplossing dient als warm start voor een quasi-Newton-verfijning. Numerieke experimenten bevestigen de effectiviteit van de voorgestelde methode in termen van schattingsnauwkeurigheid en computationele efficiëntie.

Daarna ontwikkelt het proefschrift een Bayesiaanse affine MMSE-schatter voor parametrische Wiener-modellen, waarin de statische niet-lineaire observatiefunctie wordt gerepresenteerd als een eindige combinatie van bekende basisfuncties. Een essentieel onderdeel hierbij is de DBS, die het gemiddelde en de covariantie van evaluaties van de basisfuncties langs het stochastische toestandstraject vastlegt en daarmee samenvat hoe procesruis zich door het niet-lineaire observatiemodel propageert. Binnen dit raamwerk wordt de optimale affine MMSE-schatter voor de observatieparameters in gesloten vorm afgeleid, volledig uitgedrukt in termen van de DBS, a-prioristatistieken en ruiscovarianties. De resulterende gesloten uitdrukking voor de schattingsfout wordt vervolgens gebruikt voor het ontwerpen van optimale inputs, ook bekend als active learning, door invoersignalen te kiezen die de gemiddelde kwadratische fout van de optimale affine schatter verder verlagen. Door het raamwerk toe te passen op Fourier-basisfuncties ontstaan expliciete DBS-formules die de affine MMSE-schatter in deze setting computationeel tractabel maken, terwijl tevens de inherente inconsistentie van de affine schatter bij identificatie op basis van een enkel traject onder stochastisch instabiele dynamica wordt aangetoond.

Vervolgens wordt het raamwerk uitgebreid voorbij puur affine schatters door een niet-lineaire schattingsarchitectuur voor te stellen op basis van vastepuntformuleringen, die twee affine MMSE-schatters koppelen: één voor de onbekende parameters en één voor de latente variabelen. Daarbij worden twee constructies onderzocht. De DB-P-schatter gebruikt gezamenlijk een affine schatter voor basisfunctiewaarden en de affine parameterschatter, terwijl de DS-P-schatter eerst affine toestandsschattingen en covarianties

berekent en deze statistieken vervolgens via een Gaussiaanse DBS-operator afbeeldt om de bijgewerkte DBS te verkrijgen. Beide architecturen leiden tot gekoppelde algebraïsche vergelijkingen waarvan de oplossingen kunnen worden geïmplementeerd als vastepunt-iteraties die afwisselend parameterschattingen en DBS- of toestandsstatistieken bijwerken, waarbij uitsluitend gemiddelden en covarianties worden gebruikt in plaats van volledige posterieure verdelingen. Numerieke experimenten op representatieve Wiener-type modellen illustreren de effectiviteit van deze duale schatters en, in het bijzonder, de voordelen van het gezamenlijk propageren van statistieken van parameters en latente variabelen binnen een verenigd vastepunt-raamwerk.

Samenvattend ontwikkelt het proefschrift een systematisch perspectief op Bayesiaanse parameterschatting voor Wiener-type modellen, van one-shot MAP-formuleringen voor LTV-observatiemodellen tot affine MMSE-schatters en duale niet-lineaire vastepunt-schema's gebaseerd op DBS. In al deze bijdragen staat het expliciet benutten van probleemstructuur centraal, namelijk semidefiniete relaxaties in de LTV-setting en DBS in de parametrisch niet-lineaire setting, om schatters te verkrijgen die statistisch gefundeerd zijn zonder volledige posterior-inferentie te vereisen. De affine MMSE-schatter blijkt daarbij een natuurlijk bouwblok voor duale vastepunt-architecturen die parameter- en latente-variabelenstatistieken propageren met uitsluitend het gemiddelde en de covariantie, waarmee het raamwerk wordt uitgebreid naar meer geavanceerde niet-lineaire schatters. Gezamenlijk bieden deze resultaten een hiërarchische verzameling Bayesiaanse schattings-instrumenten voor Wiener-type modellen en wijzen zij op een algemene benadering voor tractabele, structuurbenuttende schatters voor niet-lineaire toestandsruimtemodellen.

ACKNOWLEDGMENTS

Embarking on a PhD after several years in industry was not an easy decision. Moving to Europe and beginning this journey in a new environment, initially without knowledge of the local language, made it an even more significant step. Despite my initial hesitation about pursuing this path, the journey proved unique, with both joyful and difficult moments. Looking back, I am deeply grateful to the many people who supported, encouraged, and accompanied me throughout this time.

First and foremost, my deepest love and gratitude go to my family: my father, Ali, my mother, Monir, and my sisters, Sanaz and Solmaz. Your unconditional love, encouragement, and support have sustained me through every stage of this journey. I am also profoundly grateful to my uncle, Mohammad Vakili, and his family for their support during my time in Europe. Especially during the final months, your home, presence, and generous help gave me comfort and peace of mind when I needed them most.

I am deeply grateful to my supervisor, Peyman Mohajerin Esfahani, for his guidance, support, and trust throughout my doctoral studies. His willingness to support my work at a crucial stage of my PhD, as well as the perspective he provided along the way, have been invaluable. His distinctive way of thinking about scientific research and life has shaped my view, both professionally and personally.

I would like to thank Manuel Mazo Jr. for his role in facilitating my transition to TU Delft and for introducing me to Peyman. His emphasis on rigor in our collaborations strengthened my understanding of the fundamentals. I am grateful to Bart De Schutter for serving as my second promotor and for his valuable comments and feedback on this thesis. Despite his busy schedule and the demands of his role as department chair, he always found time to meet with me and offer guidance. I would also like to thank the members of my doctoral defence committee for their time, careful consideration of this dissertation, and valuable feedback. I am also very thankful to Mohammad Khosravi, with whom I worked on my first publication and had many enjoyable technical discussions.

Special thanks go to Cláudio Gomes, with whom I began collaborating before joining TU Delft. His words during difficult times meant the world to me, and this marathon, as he called it, is now approaching its finish line. I am also grateful to Arman Kolarijani and Amin Kolarijani for their advice throughout my time at TU Delft. Many thanks to Gabriel Gleizer for the enjoyable discussions we shared, which always renewed my energy.

I am indebted to Daniël Woonings for helping me with the Dutch translation of this dissertation's summary. I greatly appreciated the opportunity to supervise his master's thesis and the excellent work he carried out. Collaborating with him and Pradyumna Paruchuri on the final publication of my PhD improved the quality of the work presented in the final technical chapter of this thesis. My appreciation extends to the dedicated staff at DCSC, particularly Francy Ahaus-Smeets and Vanessa Krul, who always helped me find answers to administrative questions.

Warm thanks go to Amin Sheikhi, Mohammad Boveiri, and Haiwei Xie for the many conversations we shared over the years, especially the late-night oolong tea at Hanno. I would also like to acknowledge Amin Sheikhi and Gyula Max for the companionship and motivation they brought to our regular gym sessions. Special thanks go to Leila Gharavi for guiding me through setting up the cluster on my computer, and to Reza Rahimi for helping me navigate the administrative steps required for my defence. I am also grateful to Ilias Papagoudis for being such a great office mate. We often had long discussions at the whiteboard, solving mathematical problems together. I would also like to thank the other members of the research groups at DCSC, as well as the many colleagues and friends whose kindness made my time in Delft pleasant.

I gratefully acknowledge the Marie Skłodowska-Curie training network fellowship and the Reliable AI for Marine Robotics (REMARO) project, which made this doctoral journey possible. I thank the REMARO consortium, project team, and everyone who made it possible for me to join the project in the first place. Special thanks go to my fellow REMARO PhD candidates. Our summer schools, conversations, shared experiences, and mutual encouragement made this journey much more memorable and enjoyable. I will always cherish the unforgettable memories we created together.

To everyone who supported me in ways large or small during this journey, thank you.

*Sasan Vakili
Delft, August 2026*

1

INTRODUCTION

Bathymetric charts describe the seafloor topography by mapping water depth as a function of horizontal position. Accurate bathymetric maps are critical for maritime navigation and route planning, coastal-zone management, and hazard assessment [78, 99]. They also underpin the design, installation, and maintenance of subsea infrastructure, including telecommunication and power cables, pipelines, and offshore structures [108]. As demand for offshore energy, communication bandwidth, and marine resources continues to grow, so does the need for reliable, high-resolution seabed maps.

Conventional bathymetric surveys are typically performed by crewed surface vessels equipped with global navigation satellite systems (GNSS), such as the global positioning system (GPS), for georeferencing and single- or multibeam sonar systems for depth measurements [45]. Survey lines are often planned in “lawnmower” patterns to guarantee coverage and strip overlap [48]. These practices can achieve high accuracy but they require expensive ships, specialized crews, and significant fuel and time resources, which limit spatial and temporal coverage [78, 99]. After acquisition, individual soundings are interpolated onto regular grids to form digital terrain models, using methods such as kriging, splines, or other spatial interpolators [25, 78]. These methods primarily focus on denoising and gap-filling while making relatively weak assumptions about the underlying seabed structure. This combination can lead to conservative sampling densities and underuse of domain-specific statistical patterns.

The high operational costs and limited flexibility of traditional surveys have driven a shift towards more autonomous, scalable mapping strategies. Autonomous underwater vehicles (AUVs) have emerged as a key technology in this context [150]. Equipped with onboard navigation sensors, sonar payloads, and mission controllers, AUVs can operate close to the seafloor, follow complex trajectories, and conduct high-resolution surveys in regions that are difficult or hazardous for surface vessels. Multiple AUVs can operate concurrently, increasing coverage while reducing dependence on large survey ships [150]. AUV-based bathymetric simultaneous localization and mapping (SLAM) has become a vibrant research area, with recent reviews summarizing advances in navigation, mapping, and loop-closure techniques tailored to bathymetric data [83].

At the algorithmic level, recent work has begun to move from interpolation-centric terrain models towards probabilistic, model-driven representations of the seabed. Gaussian-process-based bathymetric SLAM, for example, models depth as a spatially correlated random field, enabling simultaneous estimation of vehicle pose and seafloor elevation with quantified uncertainty [74]. Other approaches exploit machine learning and statistical methods to retrieve bathymetry from multisensor data, including optical imagery and airborne laser bathymetry [72, 149]. Related bathymetric SLAM frameworks explicitly incorporate prior bathymetric maps into distributed particle filters to improve both navigation and mapping performance [8–10]. These methods leverage statistical patterns in the data and existing charts to improve map consistency and provide principled uncertainty estimates, which are valuable for risk-aware planning and follow-up survey design.

Despite these advances, underwater mapping remains challenging. Navigation is particularly difficult because absolute position fixes (e.g., GPS) are unavailable underwater, and acoustic communication is intermittent and bandwidth-limited [83, 99]. As a result, AUVs must rely on onboard estimation algorithms that fuse inertial, acoustic, and bathymetric cues to maintain an accurate belief over their state. At the same time, there is a growing awareness of the ecological impact of repeated surveys: extensive seafloor coverage can disturb sensitive benthic habitats, and regulatory frameworks increasingly constrain where and how long vehicles may operate [72, 99]. This tension between the need for detailed maps and the imperative to minimize environmental disturbance motivates methods that can extract as much information as possible from limited data.

In this context, there is a strong incentive to develop *probabilistic* mapping and estimation algorithms that (i) explicitly model the coupling between vehicle dynamics, sensor noise, and seafloor structure, (ii) provide both point estimates and uncertainty quantification for the seabed map, and (iii) can be integrated with planning and safety verification tools for autonomous operation. From a modeling perspective, this naturally leads to state-space formulations in which the vehicle dynamics are described by a (known or identified) linear or linearized model, and the seabed is represented as an unknown function of the vehicle state, observed through noisy measurements. Nonparametric models, such as Gaussian processes (GPs) and other kernel or data-driven methods, can flexibly capture unknown nonlinear seabed structures and have been successfully applied to bathymetric mapping and retrieval tasks [72, 74, 149]. However, when embedded in dynamical state-space models, the resulting probabilistic inference problems can become computationally demanding, and closed-form expressions are rarely available. As such, the present thesis focuses on Bayesian estimation of parametric observation models, with an emphasis on tractable, closed-form or fixed-point estimation architectures that are well suited for integration into larger navigation and mapping pipelines.

1.1 BAYESIAN ESTIMATION IN WIENER-TYPE MODELS

AUVs operating in deep-sea environments face challenges arising from the interplay between vehicle dynamics and an unknown nonlinear seabed observation model. The bathymetric mapping task, as illustrated in Figure 1.1, can be abstracted as a class of dynamical systems in which a known linear process evolves in time and is observed through an *unknown* static nonlinear output function. This structure is commonly referred to as a *Wiener* model: a linear dynamical block followed by a static nonlinearity [129]. We

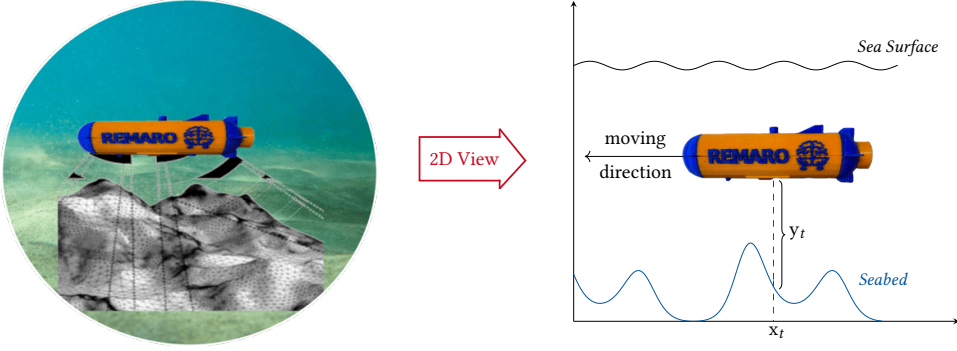


Figure 1.1: Autonomous underwater vehicle bathymetric mapping.

consider a discrete-time linear time-varying (LTV) state-space model

$$\mathbf{x}_{t+1} = \mathbf{A}_t \mathbf{x}_t + \mathbf{B}_t \mathbf{u}_t + \mathbf{w}_{t+1}, \quad t = 0, \dots, T-1, \quad (1.1)$$

where $\mathbf{x}_t \in \mathbb{R}^{n_x}$ is the system state vector, $\mathbf{A}_t \in \mathbb{R}^{n_x \times n_x}$ and $\mathbf{B}_t \in \mathbb{R}^{n_x \times n_u}$ are known system matrices, and $\mathbf{u}_t \in \mathbb{R}^{n_u}$ is a vector of known or controlled inputs. The process noise $\mathbf{w}_{t+1} \in \mathbb{R}^{n_x}$ is modeled as a zero-mean random vector with known covariance $\Sigma_{\mathbf{w}_{t+1}}$, and the initial state $\mathbf{x}_0$ is drawn from a known distribution. In the AUV bathymetric mapping example, $\mathbf{x}_t$ may encode the vehicle's position, orientation, and possibly additional internal states, while $\mathbf{u}_t$ represents commanded velocities or other control inputs. The system is observed through a static nonlinear output map

$$\mathbf{y}_t = \mathbf{h}_\theta(\mathbf{x}_t) + \mathbf{v}_t, \quad t = 0, \dots, T, \quad (1.2)$$

where $\mathbf{y}_t \in \mathbb{R}^{n_y}$ is the observation vector, $\mathbf{v}_t \in \mathbb{R}^{n_y}$ is measurement noise with known covariance $\Sigma_{\mathbf{v}_t}$, and $\mathbf{h}_\theta : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y}$ is an *unknown* static nonlinear function parameterized by the unknown variables $\theta \in \Theta$. In the bathymetric mapping setting, $\mathbf{y}_t$ can be interpreted as a depth-type measurement.

Given the static output map (1.2) and the state representation in (1.1), together with the sequence of measurements $\mathbf{y} = [\mathbf{y}_0, \dots, \mathbf{y}_T]^\top$, the Bayesian estimation framework seeks an estimator $\hat{\theta}(\cdot)$ that minimizes an expected loss function $\ell(\cdot, \cdot)$ associated with a chosen performance criterion [76, Ch. 4] as

$$\min_{\hat{\theta}(\cdot) \in \mathcal{F}_\theta} \mathbb{E}[\ell(\theta, \hat{\theta}(\mathbf{y}))] = \min_{\hat{\theta}(\cdot) \in \mathcal{F}_\theta} \int_{\mathbb{R}^{n_y(T+1)} \times \Theta} \ell(\theta, \hat{\theta}(\mathbf{y})) p(\theta, \mathbf{y}) d\theta d\mathbf{y}, \quad (1.3)$$

where $\mathcal{F}_\theta := \{\hat{\theta} : \mathbb{R}^{n_y(T+1)} \rightarrow \Theta\}$ denotes the class of all estimators, $T+1$ is the length of the data sequence, and Θ is the parameter space. It can be shown [76, Ch. 4] that solving (1.3) is equivalent to computing, for each realization of the observations $\mathbf{y}$, an estimator of the form

$$\hat{\theta}(\mathbf{y}) = \underset{\vartheta \in \Theta}{\operatorname{argmin}} \int_{\Theta} \ell(\theta, \vartheta) p(\theta | \mathbf{y}) d\theta, \quad (1.4)$$

which depends on the posterior distribution $p(\theta | y)$. Among the many loss functions that can be considered in the Bayesian framework, two common choices are norm-based losses, such as the squared Euclidean norm $\ell(\theta, \vartheta) = \|\theta - \vartheta\|^2$, and the indicator-type loss $\ell(\theta, \vartheta) = 1 - \mathbb{1}_{\|\theta - \vartheta\|_\infty \leq \epsilon}(\theta)$. The squared Euclidean norm loss yields the minimum mean-squared error (MMSE) estimator,

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}[\theta | y] = \int_{\Theta} \theta p(\theta | y) d\theta = \int_{\Theta} \theta p(y | \theta) p(\theta) d\theta, \quad (1.5)$$

whereas the indicator-type loss, also known as the Notch function loss [76, Ch. 4], penalizes all estimates outside an ϵ -neighborhood of the true parameter. In the limiting case $\epsilon \rightarrow 0$, the corresponding Bayesian estimator for this latter loss function converges to the maximum a posteriori (MAP) estimator, defined by

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta | y) = \underset{\theta}{\operatorname{argmax}} p(y | \theta) p(\theta), \quad (1.6)$$

where the last equalities in (1.5) and (1.6) follow from Bayes' rule. Both the MMSE and MAP estimators thus require two probabilistic ingredients:

- (i) **Prior:** a prior distribution for the unknown parameter vector, with density $p(\theta)$;
- (ii) **Likelihood:** a likelihood function that describes the relation between the data and the parameters, given by the conditional density $p(y | \theta)$.

In the bathymetric mapping context, the prior $p(\theta)$ could in principle be informed by coarse-resolution legacy charts, results from previous missions in the same region, or generic assumptions about seabed roughness and variability.

A central difficulty in computing the Bayesian estimator $\hat{\theta}$ in (1.4) arises from the posterior distribution $p(\theta | y)$. The state sequence $\{x_t\}_{t=0}^T$ is stochastic and *correlated in time* due to the process noise and the dynamics in (1.1). Consequently, the “inputs” to the output function (1.2), i.e., $\{x_t\}$, are neither directly observed nor independently distributed. This latent and correlated nature of the states complicates the specification and evaluation of the likelihood $p(y | \theta)$. Even when the likelihood is well defined, the resulting MAP estimator (1.6) generally leads to a non-convex optimization problem in θ , and the integral in the MMSE estimator (1.5) typically has no closed-form solution. For the Wiener-type models of interest, these challenges make exact Bayesian MMSE or MAP estimation analytically intractable and motivate the development of approximate inference techniques. The next section provides an overview of commonly used methods for addressing this intractability and discusses how the tractable estimators proposed in this thesis exploit the special structure of the Wiener setting.

1.2 EXISTING APPROACHES

The Bayesian estimation problems outlined in Section 1.1 sit at the intersection of nonlinear system identification and Bayesian state-space inference. This section briefly reviews common modeling choices for observation functions in Wiener-type systems and the main classes of inference algorithms used to tackle the resulting estimation problems.

1.2.1 NONLINEAR SYSTEM IDENTIFICATION

A standard way to model nonlinear dynamical systems is to start from a linear state-space description and introduce structured nonlinearities in the input or output channels [87, 103, 133]. Block-oriented models, such as Hammerstein, Wiener, and Hammerstein-Wiener structures, are particularly popular because they combine interpretability with moderate complexity [46, 129, 154]. In the Wiener case, which is central to this thesis, a linear dynamic block is followed by a static nonlinear function acting on the state or output signal. Several forms of parametrization of the static nonlinearity have been studied in the literature. Classical approaches use polynomials or piecewise-polynomial functions, often combined with separable least-squares or two-stage identification schemes [4, 129]. Other works employ basis-function expansions, such as radial basis functions (RBFs), so that the nonlinear map becomes linear in a finite set of parameters [3, 109, 129].

Beyond parametric models, nonparametric and kernel-based methods provide flexible alternatives for representing static nonlinearities. These include kernel-based regularization methods for linear and nonlinear systems [31, 110] and Gaussian process (GP) models for static maps embedded in dynamical systems [117]. In the bathymetric mapping setting, GP-based models have been used to represent seafloor elevation fields and to perform bathymetric SLAM with uncertainty quantification [74]. While highly expressive, such nonparametric models can be computationally demanding when combined with full state-space dynamics and joint state-parameter inference.

The present thesis focuses on *parametric* observation models in Wiener-type systems, in which the static nonlinearity is represented as follows:

- (i) **Chapter 2:** An LTV observation model, in which the nonlinear map is approximated by a piecewise linear time-varying model

$$y_t = C_t x_t + v_t, \quad t = 0, \dots, T, \quad (1.7)$$

where C_t represents the *unknown* slope (or gain) of each linear segment. Figure 1.2 illustrates such a piecewise affine approximation in the context of bathymetric mapping, where the solid red curve represents an underlying nonlinear function, the blue markers denote sampled data points, and the blue dashed segments depict the piecewise linear interpolation between samples.

- (ii) **Chapters 3 and 4:** A basis expansion with a finite-dimensional parameter vector,

$$y_t = \sum_{n=0}^N \theta_n \phi_n(x_t) + v_t, \quad t = 0, \dots, T, \quad (1.8)$$

where the unknown observation function is represented as a finite sum of known basis functions $\phi_n(x_t)$ with *unknown* coefficients θ_n .

While the piecewise linear model in Chapter 2 can perform well within observed input ranges, it cannot extrapolate to unseen states and lacks probabilistic error guarantees. The nonlinear basis-function expansion adopted in Chapters 3 and 4 helps overcome these limitations by providing a smoother functional representation and, in the Bayesian setting, explicit access to closed-form parameter estimates and their estimation-error covariance. In addition, these two chapters specialize the novel estimation methods to Fourier basis

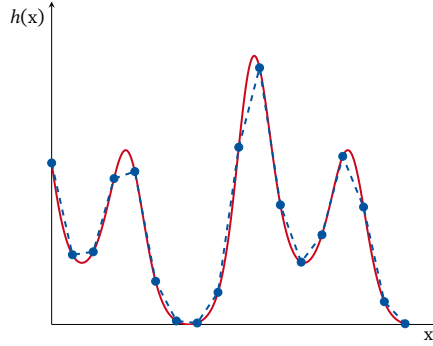


Figure 1.2: Piecewise affine approximation of a nonlinear function from sampled data.

functions and thereby connect to the framework of random Fourier features, which approximate shift-invariant kernel machines via randomized low-dimensional feature maps with probabilistic approximation guarantees [115]. In principle, employing such random feature constructions in conjunction with an extension of the estimation methods proposed in Chapters 3 and 4 can reduce computational cost while retaining kernel-style flexibility and theoretical approximation guarantees. This combination is attractive for bathymetric mapping in regimes where nonparametric function classes are more suitable.

1.2.2 BAYESIAN ESTIMATION METHODS

When both the process and observation models are linear and the noise is Gaussian, the Bayesian filtering and smoothing problems admit closed-form solutions via the Kalman filter and the Rauch–Tung–Striebel (RTS) smoother [70, 119]. These classical algorithms provide optimal MMSE estimates of the state trajectory and form the foundation of many state-estimation and navigation systems [7, 126]. Parameter estimation in dynamical systems, however, is more challenging because it involves two coupled sets of unknowns: the parameters θ and the latent state trajectory $\{x_0, \dots, x_T\}$. The presence of the unknown system states fundamentally complicates the design of the parameter estimator $\hat{\theta}(y)$ in (1.4) and has led to a wide range of solution methods.

In this thesis, we classify parameter-estimation methods according to their algorithmic structure for handling the latent states, distinguishing between *one-shot optimization* and *fixed-point characterization* techniques, as shown in Figure 1.3. In the one-shot optimization approach, a single loss function is minimized with respect to either θ alone or both θ and the latent state trajectory $\{x_0, \dots, x_T\}$. As such, the state trajectory is either marginalized out analytically or optimized jointly within the same objective, without introducing a separate loss function dedicated solely to the latent states. We refer to this paradigm as *cooperative*, since the unknown parameters and states are treated within one collective optimization problem whose ultimate purpose is to construct $\hat{\theta}(y)$. By contrast, fixed-point characterization techniques introduce two coupled estimators, one for θ and one for $\{x_0, \dots, x_T\}$. Each estimator is defined as the minimizer of its own loss function of the form (1.4), while using the output of the other estimator as part of its input. This

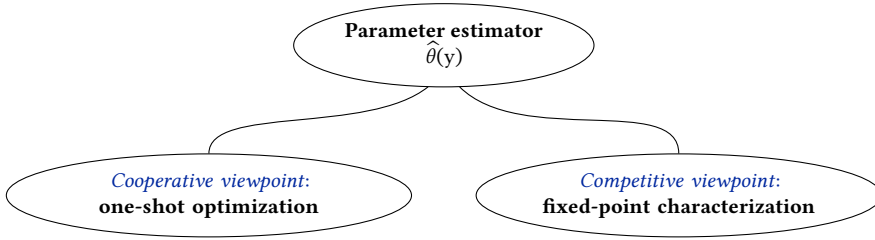


Figure 1.3: Taxonomy of parameter-estimation methods considered in this thesis, grouped according to the algorithmic structure of the unknown-parameter estimator.

competitive paradigm leads to iterative schemes in which the parameter and state estimators are updated in turn, and under suitable conditions the iterates converge to a fixed point that simultaneously satisfies the local optimality conditions of both estimators.

ONE-SHOT OPTIMIZATION: A COOPERATIVE VIEWPOINT

In this class, the estimation problem is formulated as the minimization of a single loss function, without alternating between separate state- and parameter-estimation steps. Within this one-shot optimization paradigm, the decision variables are either (i) associated only with parameter-only estimators, where the latent state process enters the objective function solely through its contribution to the marginal data likelihood [41, 61], or (ii) the states and parameters jointly, by embedding the parameters into an augmented state vector and applying a single filter or smoother to this enlarged model [30, 71, 87].

In the parameter-only setting, when the system dynamics are linear and Gaussian, a standard approach is to treat the latent state trajectory as a nuisance variable and express the likelihood of the measurements given the parameters via the prediction-error decomposition obtained from a single run of the Kalman filter [41, 61, 73]. The resulting likelihood is then optimized directly with respect to the parameters, while state trajectories are subsequently reconstructed, if needed, in a separate smoothing step under the fitted model [41]. For nonlinear state-space models, related ideas have been pursued by constructing tractable variational lower bounds on the marginal likelihood and optimizing these bounds with respect to the model parameters (and associated variational quantities), rather than designing an explicit alternating state-parameter estimator [34]. This non-iterative parameter-only viewpoint also underlies the MAP and MMSE estimators developed in Chapters 2 and 3, respectively, where state-related uncertainty is summarized through suitable statistics and the resulting optimization problem is solved with respect to the decision variables of the parameter-only estimators.

A structurally different subclass within one-shot optimization comprises augmented state-parameter methods. In these approaches, the unknown parameters are modeled as additional, often static or slowly varying, state components. Consequently, the prediction-update recursions propagate a joint estimate of the state-parameter vector without explicitly separating the two estimation tasks [30, 71]. These estimators thus perform joint state-parameter inference, but differ from parameter-only methods and alternating estimators in that they neither marginalize out the latent state process nor alternate between separate state- and parameter-inference steps. Instead, state and parameter information is updated

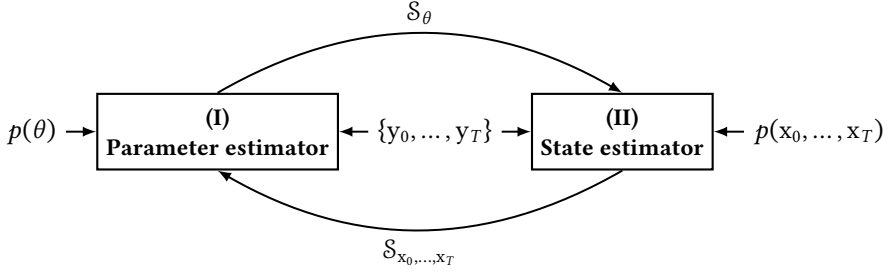


Figure 1.4: Block diagram of fixed-point characterization algorithms for joint state and parameter estimation. The parameter estimator block (I) produces parameter summaries $\mathcal{S}_\theta$ that are fed into the state estimator block (II), while the state estimator block returns state-related summaries $\mathcal{S}_{x_0, \dots, x_T}$ back to the parameter estimator block. Depending on the method, these summaries may correspond to point estimates, smoothing distributions, particle-based trajectory samples, or pairs of estimates and estimation-error covariances.

together within a single recursive inference mechanism, so that the augmented state estimation directly influences the parameter estimates, resulting in simultaneous state-parameter estimation [19].

FIXED-POINT CHARACTERIZATION: A COMPETITIVE VIEWPOINT

The fixed-point characterization techniques, on the other hand, comprise approaches that differ mainly in how they estimate the latent state trajectory or its posterior distribution and feed this information into a parameter estimator. In this class of methods, state and parameter estimator steps are iterated and exchange information, as illustrated in Figure 1.4. In this abstract architecture, $\{y_0, \dots, y_T\}$, $p(x_0, \dots, x_T)$, and $p(\theta)$ denote the measurement sequence, the prior density of the full state trajectory, and the prior density of the unknown parameters, respectively. The “Parameter estimator” block (I) and “State estimator” block (II) represent the parameter- and state-estimation components, and their outputs $\mathcal{S}_\theta$ and $\mathcal{S}_{x_0, \dots, x_T}$ denote statistical summaries that, depending on the specific method, may correspond to point estimates [104], smoothing distributions [38], particle-based trajectory samples [2, 71], or pairs of estimates and estimation-error covariance, as in the dual estimator proposed in Chapter 4.

A canonical method in this category is the expectation-maximization (EM) algorithm, which is a standard tool for maximum likelihood and MAP estimation in latent variable models [38]. In the dynamical system context, EM alternates between

- (i) **E-step:** computes (or approximates) the posterior distribution of the latent state trajectory given the current parameter estimates. This step corresponds to the “State estimator” block (II) in Figure 1.4, and the posterior distribution $p(x_0, \dots, x_T | y_0, \dots, y_T, \theta)$ plays the role of the state-related summary statistics $\mathcal{S}_{x_0, \dots, x_T}$;
- (ii) **M-step:** maximizes, with respect to the model parameters, the expected complete-data log-likelihood under this current smoothing distribution of the latent states [53, 132]. This step corresponds to the “Parameter estimator” block (I) in Figure 1.4, and the resulting point estimate $\hat{\theta}$ plays the role of the parameter summary statistics $\mathcal{S}_\theta$.

For linear Gaussian models, both steps admit closed-form expressions, and EM can be implemented using Kalman filtering and RTS smoothing in the E-step [53, 126, 132].

For nonlinear or non-Gaussian state-space models, exact smoothing is typically intractable, and EM must rely on approximations in the E-step. Approaches include extended or unscented Kalman smoothers, particle smoothers, and other approximate Bayesian smoothers, often combined with gradient-based or numerical optimization in the M-step [58, 127], except in special cases where the nonlinearities are modeled by Gaussian radial basis functions (RBFs), for which the M-step admits a closed-form solution [54]. While EM retains an appealing interpretation and can work well in practice, approximate E-steps can violate the monotonic-likelihood property of the exact EM algorithm, which in turn can lead to sensitivity to initialization and convergence to suboptimal modes [5]. Additionally, in these cases the resulting objective function in the M-step is generally non-convex and computationally demanding.

Beyond EM, a second family of this iterative methods explicitly separates the state and parameter updates into two coupled estimator components, where each solves a MAP problem given that the other unknown is fixed. Dual Kalman filters (and their smoother counterparts) are classical examples: one estimator focuses on the latent state trajectory given the parameter estimates, while a second estimator focuses on the parameters given the state estimates, and the two are iterated until convergence or until the estimates stabilize [63, 65, 104, 114]. This architecture can also be viewed in Figure 1.4, where the resulting point estimates $\{\hat{x}_0, \dots, \hat{x}_T\}$ and $\hat{\theta}$ play the role of the state and parameter summary statistics $\mathcal{S}_{x_0, \dots, x_T}$ and $\mathcal{S}_\theta$, respectively.

In these schemes, the updated state and parameter estimates produced by one component are fed into the other at the next iteration. Variants of this idea have been developed using extended or unscented Kalman filters for the state and parameter blocks, as well as ensemble-based and other approximate Gaussian filters [125, 147, 148]. These methods share the same alternating state-parameter structure at the algorithmic level as EM, but are conceptually distinct from EM in that they propagate point estimates and solve separate MAP problems rather than optimizing an explicit marginal likelihood. Consequently, these methods can be viewed as implementing a deterministic two-block fixed-point iteration for joint Bayesian state-parameter estimation [90].

Another instantiation of the fixed-point characterization paradigm arises in fully Bayesian methods for joint state-parameter inference in nonlinear state-space models. In these approaches, at each iteration, sequential Monte Carlo (SMC) smoothers draw entire state trajectories from an approximation of the conditional state posterior distribution $p(x_0, \dots, x_T | y_0, \dots, y_T, \theta)$ under the current parameter values, typically by combining a forward particle filter with a backward-smoothing pass to obtain weighted trajectory samples that represent the latent state posterior [21, 32, 40, 57]. Using these candidate trajectories from the smoothing distribution, the parameters are then updated from their conditional distribution given the sampled trajectory, as in particle Markov chain Monte Carlo (PMCMC) schemes such as particle marginal Metropolis–Hastings (MH) and particle Gibbs [2].

In this setting, the particle smoother naturally forms the “State estimator” block (II) of the alternating architecture in Figure 1.4, and parameter updates, i.e., “Parameter estimator” block (I), are performed on the basis of entire smoothed trajectories rather than solely

on filtering distributions [71]. As such, in Figure 1.4, trajectory samples $\{\hat{x}_0, \dots, \hat{x}_T\}$ and $\hat{\theta}$ play the role of the state and parameter summary statistics $\mathcal{S}_{x_0, \dots, x_T}$ and $\mathcal{S}_\theta$, respectively. These algorithms have been successfully employed in Bayesian identification of stochastic nonlinear state-space models, where accurate smoothing is crucial for learning dynamics and quantifying uncertainty [138]. Although these methods are usually presented as components of a single composite sampler, their roles mirror the deterministic alternating schemes discussed earlier: smoothed trajectory samples provide posterior information about the latent dynamics to the parameter update, and updated parameters in turn define the state-space model used in the next smoothing or filtering step. In this sense, particle smoothing with parameter updates provides a fully stochastic counterpart to the fixed-point characterization algorithms in the taxonomy, with the iterations implemented via coupled Monte Carlo updates rather than closed-form or deterministic estimators, but still grounded in the same joint state-parameter estimation objective [26].

A fourth fixed-point characterization architecture, developed in Chapter 4, is based on dual affine Bayesian MMSE estimators that operate on first- and second-moment statistics rather than full smoothing distributions or pure point estimates. In this framework, one affine MMSE estimator targets the nonlinear observation parameters, while a second affine MMSE estimator targets either basis-function statistics or the latent states themselves, and the two components are coupled via dynamic basis statistics (DBS), which summarize nonlinear basis-function evaluations in terms of means and covariances. Each dual estimator admits a fixed-point characterization in which the parameter- and state-related moments are updated in a Jacobi-style two-block iteration process, with the outputs of one affine MMSE block providing the prior statistics for the other. As such, $\mathcal{S}_{x_0, \dots, x_T}$ and $\mathcal{S}_\theta$ in Figure 1.4 correspond to pairs of estimates and estimation-error covariances rather than complete distributions or point estimates. This moment-based dual structure differs from the EM and dual Kalman schemes, which respectively alternate between smoothing distributions and point-estimate smoothers, and from PMCMC methods, which alternate between state trajectory and parameter samples. At the same time, our proposed dual structure retains a deterministic, closed-form update structure grounded in the same joint Bayesian state-parameter estimation scheme. We summarize the specific contributions of each chapter with respect to this taxonomy and outline the overall structure of the dissertation in the subsequent sections.

1.3 THESIS CONTRIBUTIONS

This thesis develops a hierarchy of Bayesian estimation methods for Wiener-type state-space models, organized according to the taxonomy of parameter estimators in Figure 1.3. Throughout the thesis, the numerical experiments are synthetic, with system dynamics, parameter variations, and noise levels chosen to be qualitatively representative of AUV bathymetric mapping and related robotic perception scenarios.

MAP estimator for LTV observation models. At one end of the spectrum, Chapter 2 develops a one-shot optimization method for Bayesian maximum a posteriori (MAP) estimation of linear time-varying (LTV) observation maps and shows how a non-convex smoothing problem can be approximated by a tractable SDP relaxation that warm-starts a quasi-Newton local optimizer with superlinear convergence rate. Numerical experiments

indicate that the resulting estimator achieves lower parameter-estimation error than EM and dual Kalman smoother (DKS) approaches from the fixed-point characterization category, particularly as the process-noise uncertainty increases. This contribution is based on [139]:

“Linear Time-Varying Parameter Estimation: Maximum A Posteriori Approach via Semidefinite Programming,” Sasan Vakili, Mohammad Khosravi, Peyman Mohajerin Esfahani, and Manuel Mazo Jr., *IEEE Control Systems Letters*, vol. 8, pp. 73–78, 2023.

Affine MMSE estimators for Wiener models. Chapter 3 focuses on a one-shot optimization method for parametric Wiener models with nonlinear basis-function observation maps. We derive a closed-form optimal affine MMSE estimator together with an active input design strategy that exploits the analytical MMSE expression. This work targets Bayesian MMSE estimation by marginalizing the latent states through the dynamic basis statistics (DBS), which summarize the effect of state uncertainty on nonlinear basis-function evaluations in terms of their means and covariances. We further analyze the consistency of the estimator under suitable assumptions. Extensive numerical experiments demonstrate that explicitly capturing this contribution of latent-state uncertainty through DBS leads to almost surely lower parameter estimation mean-squared error (MSE) than classical regularized least-squares estimators, which implement a prediction-error minimization approach within the one-shot optimization category of the parameter-estimator taxonomy in Figure 1.3. This contribution is based on [140]:

“Optimal Bayesian Affine Estimator and Active Learning for the Wiener Model,” Sasan Vakili, Manuel Mazo Jr., and Peyman Mohajerin Esfahani, *Transactions on Automatic Control*, under revision, 2025.

Dual affine MMSE estimators and fixed-point characterization. Chapter 4 develops nonlinear dual estimator architectures that sit at the fixed-point characterization end of the taxonomy in Figure 1.3. Building on the affine MMSE parameter estimator and the dynamic basis statistics (DBS) framework from Chapter 3, we construct two coupled affine MMSE estimators: one for the observation parameters and one for latent quantities. These two affine estimators are interconnected through DBS so that each block uses the other’s mean and covariance estimates as prior information. The dual basis-parameter (DB-P) estimator combines an affine estimator for basis-function evaluations with the affine parameter estimator, while the dual state-parameter (DS-P) estimator first computes affine state estimates and their covariances and then maps these state statistics through a Gaussian DBS operator to obtain basis statistics, thereby exploiting the system dynamics more directly. Both dual architectures admit fixed-point characterizations implemented by Jacobi-style two-block iterations, as in Figure 1.4, that exchange only first- and second-order moments between the parameter and state (or basis) estimator blocks, providing a deterministic moment-based alternative to EM, dual Kalman, and PMCMC schemes. Extensive Monte Carlo experiments show that the DS-P estimator systematically achieves the lowest empirical parameter MSE, outperforming the affine MMSE estimator from the one-shot optimization class, the DB-P estimator, and fully Bayesian particle-Gibbs and EM methods from the fixed-point characterization class in Wiener models with Fourier basis functions. This contribution is based on [141]:

“Nonlinear Bayesian Estimator for Parameter Learning: A Fixed-Point Characterization,” Sasan Vakili, Daniël Woonings, Pradyumna Paruchuri, and Peyman Mohajerin Esfahani, *Journal of Machine Learning Research*, under review, 2026.

1.4 THESIS OUTLINE

The bulk of the thesis is based on journal papers that are published or under review. Each technical chapter is written to be largely self-contained, with its own notation, assumptions, and detailed proofs, so that it can be read independently of the others. The dissertation is organized as follows:

- **Chapter 1: Introduction.**

The present chapter motivates the problem from the perspective of AUV-based bathymetric mapping, formulates the Bayesian estimation problems in Wiener-type state-space models, reviews existing approaches in nonlinear system identification and Bayesian estimation methods, and introduces the taxonomy of parameter estimators that underpins the subsequent chapters.

- **Chapter 2: Time-varying parameter estimators for linear systems.**

This chapter develops a MAP estimator formulation for the parameters of linear time-varying (LTV) observation models, derives a nonconvex optimization problem based on the full trajectory model, and presents a semidefinite-programming (SDP) relaxation that warm-starts a quasi-Newton refinement. We illustrate the proposed MAP estimator on a synthetic LTV system with randomly generated parameters that mimic plausible variation rates in observation models. The synthetic experimental study benchmarks the proposed estimator against the DKS and EM approaches, and compares estimation accuracy and computational effort.

- **Chapter 3: Affine estimators and active learning for Wiener models.**

This chapter introduces the Bayesian MMSE setting for Wiener models, defines the dynamic basis statistics (DBS), and derives the optimal affine MMSE parameter estimator together with its key properties, including estimation mean and covariance statistics, error monotonicity, and consistency. The framework is then specialized to Fourier bases, with explicit DBS formulas and the associated inconsistency phenomenon discussed for this choice of basis functions. Finally, the chapter presents an active input design procedure and evaluates the estimator’s performance, both with and without this procedure, against classical regularized least-squares methods. The performance of the affine MMSE estimator is evaluated on synthetic Wiener models designed to reflect both low and high process-noise regimes, with parameter magnitudes that stress-test the method in uncertain environments.

- **Chapter 4: Nonlinear estimators via fixed-point characterization.**

This chapter formulates the nonlinear dual estimator architecture that couples affine MMSE parameter and latent-variable estimators via dynamic basis statistics (DBS), leading to fixed-point characterizations and Jacobi-style iterations. The dual basis-parameter (DB-P) and dual state-parameter (DS-P) variants are developed, and their per-iteration complexity is analyzed. The convergence behavior of these dual estimators is examined empirically, and extensive Monte Carlo studies provide a com-

prehensive comparison of the two dual estimators with the affine MMSE baseline and SMC-based methods, including Particle Gibbs and EM. The dual estimators are tested on synthetic Wiener-type models similar to those in Chapter 3, again with parameters and noise levels chosen to stress-test the method.

- **Chapter 5: Conclusions.**

The final chapter synthesizes the main insights from the preceding chapters and discusses the implications for Bayesian identification of Wiener-type models. The thesis then outlines several directions for future research, including asymptotic analysis, extensions to richer model classes, recursive implementations, and real-time implementations for AUV-based bathymetry.

2

TIME-VARYING PARAMETER ESTIMATORS FOR LINEAR SYSTEMS

This chapter studies the problem of identifying a linear time-varying output map from measurements and linear time-varying system states, which are perturbed with Gaussian observation noise and process uncertainty, respectively. Employing a stochastic model as prior knowledge for the parameters of the unknown output map, we reconstruct their estimates from input-output pairs via a Bayesian approach to optimize the posterior probability density of the output map parameters. The resulting problem is a non-convex optimization, for which we propose a tractable linear matrix inequalities approximation to warm-start a first-order subsequent method. The efficacy of our algorithm is shown experimentally against classical expectation-maximization and dual Kalman smoother approaches.

2.1 INTRODUCTION

Bayesian approaches for estimating characteristics of dynamical systems have been a subject of studies for decades and have recently received extensive attention [7, 144]. In systems theory, the significance of the Bayesian approach is highlighted in state estimation of dynamical systems [69, 118], e.g., through the celebrated recursive Kalman filter. The Rauch–Tung–Striebel (RTS) smoother counterparts [118], on the other hand, are (offline) iterative non-causal algorithms incorporating future measurements into the current state estimation.

An alternative to Bayesian estimation, which requires a prior distribution of the parameters of interest, is the minimax estimation approach, assuming instead the knowledge of ambiguity sets. The *least favorable* uncertainty model from this ambiguity set is then used for estimation [106, 131, 152, 153]. In this chapter, we focus instead on designing a classical smoother for a different problem: system parameters estimation from input-output measurements via Bayesian estimation. This problem arises in, e.g., robot mapping in unknown environments, such as Autonomous Underwater Vehicles (AUVs) operating in the deep sea where global positioning is expensive due to low visibility and lack of radio communications.

Given unknown parameters with random states, applying a Bayesian estimation framework leads to severe non-convexities in the resulting optimization problem. As such, iterative schemes are typically employed to overcome these non-convexities. Assuming the parameters also follow a statistical formulation, two main types of smoother approaches, *dual Kalman smoother* (DKS) and *expectation-maximization* (EM), are available in the literature [63].

Dual Kalman smoothers (and filters) attempt to maximize the joint probability space of parameters and state (conditioned on input and output observations), iterating between estimating the system states using the last parameters' estimates followed by estimating the parameters from the currently estimated states. Although DKS is computationally efficient according to its recursive structure, its estimation performance can be significantly suboptimal due to the bilinearity between the parameters and states. Unlike DKS, *expectation-maximization* learns the parameters of statistical models by maximizing the posterior distribution of the parameters from the observed data and their prior density function when incomplete data or hidden variables exist [11, 86, 155].

Considering the states of a dynamical system as hidden variables [53, 124, 127], EM estimates the parameters of a dynamical system in two steps by integrating all possible values of the states in which the model could have generated the observations. The distribution over hidden variables is maximized in the *E-Step* using the parameters estimates from the previous iteration. Subsequently, the *M-Step* maximizes a lower bound of the original cost by fixing the hidden variables distribution to the one optimized in the *E-Step*. A closed-form solution of the *M-Step* is provided in [53] for estimating the parameters of linear time-invariant (LTI) dynamical systems and in [63, Chapter 6] for estimating the parameters of a Gaussian radial basis function (RBF) approximator. Both solutions consider the maximum likelihood case, where no prior exists for the parameters.

Finding a closed-form expression for the parameters update in the *M-Step* of a maximum a posteriori (MAP) smoothing problem when the parameters are time-varying and in the presence of *a priori* knowledge is non-trivial. This challenge leads to a slow conver-

gence of the EM algorithm utilizing computationally demanding approaches to solve the optimization in *M-step*, e.g., first-order methods. The slow convergence of EM is shown experimentally in [102], and further analyses in [5, 151] demonstrate the slow convergence rate of the gradient variant of the EM algorithm for Gaussian mixture models (GMMs).

Alternatives to the iterative schemes can be found in the parameter estimation problem of an elliptically contoured distribution [68, Page 107], employing recent conic geometric optimization methods [134]. These methods, however, require reformulating the MAP problem via techniques, such as those proposed in [66, Section 3], which result in losing the output map's original structure. Retaining such a structure to leverage the available *a priori* knowledge is essential to our problem.

In this work, we consider systems with known linear time-varying (LTV) dynamics affected by process and measurement Gaussian noise but with unknown time-varying output maps. We propose a method to estimate the unknown parameters of the output map having as *a priori* information, i.e., a linear stochastic system encoding the evolution of the parameters. We derive an optimization problem applying a fully Bayesian approach, maximizing the exact posterior distribution of the parameters when unfolded over the whole time horizon. A tractable conservative approximation to the resulting optimization problem is derived via semidefinite-programming (SDP) using linear matrix inequalities (LMIs) techniques. The solution from this approximation then provides a warm-start for a first-order quasi-Newton algorithm that enjoys a locally superlinear convergence rate. This combination allows us to enjoy both the computational advantage of DKS and outperform the statistical performance of EM. We illustrate the efficacy and performance of our proposed method in comparison with DKS and EM through a Monte Carlo experiment with different signal-to-noise ratios (SNRs) in Section 2.5.

Notation. Throughout this chapter, $\mathbb{Z}_+$, $\mathbb{R}$ and $\mathbb{R}^{n \times m}$ denote the set of positive integers, real numbers, and n by m real matrices, respectively. We indicate $\text{diag}(A_1, \dots, A_k)$ as a block diagonal matrix with diagonal entries of given matrices $A_1, \dots, A_k$. The symbol $\mathbf{I}$ denotes the identity matrix, the transpose of a matrix A is denoted by $A^\top$, and $\text{tr}(\cdot)$ is the trace operator. Given $A \in \mathbb{R}^{m \times n}$, a matrix with columns $a_1, \dots, a_n \in \mathbb{R}^m$, we define $\text{vec}(A)$ as the vector $[a_1^\top, \dots, a_n^\top]^\top \in \mathbb{R}^{m \times n}$. For a positive symmetric matrix $A \in \mathbb{R}^{n \times n}$, $\Lambda(A) := (\lambda_i(A))_{i=1}^n$ indicates the vector of eigenvalues of A in descending order, i.e., $\lambda_i(A)$ is the i^{th} largest eigenvalue of A . A multivariate normal (Gaussian) distribution with mean μ and covariance matrix Σ is denoted by $\mathcal{N}(\mu, \Sigma)$, and the symbol $\sim$ stands for “distributed according to.”

2.2 PROBLEM DEFINITION

Consider a discrete-time linear time-varying dynamical system described by the process model

$$x_{t+1} = A_t x_t + B_t u_t + w_t, \quad t \in \mathbb{Z}_+, \quad (2.1)$$

where t denotes the time index, $x_t \in \mathbb{R}^{n_x}$ is the vector of state variables, $A_t \in \mathbb{R}^{n_x \times n_x}$ is the state transition matrix, $u_t \in \mathbb{R}^{n_u}$ is the vector of inputs, $B_t \in \mathbb{R}^{n_x \times n_u}$ is the input matrix, and $w_t \in \mathbb{R}^{n_x}$ is an independent realization at time t of the process noise with Gaussian distribution $\mathcal{N}(0, \Sigma_{w_t})$. The initial state of system (2.1), denoted by x_0 , is also assumed to be drawn from a Gaussian distribution $\mathcal{N}(\mu_{x_0}, \Sigma_{x_0})$. For $t \in \mathbb{Z}_+$, the state of the system is

observed at time instant t through a perturbed linear time-varying map

$$y_t = C_t x_t + v_t, \quad t \in \mathbb{Z}_+, \quad (2.2)$$

where $y_t \in \mathbb{R}^{n_y}$ denotes the output measurements, $C_t \in \mathbb{R}^{n_y \times n_x}$ is an *unknown time-varying* observation matrix, and $v_t \in \mathbb{R}^{n_y}$ is the vector of measurement noise signals with Gaussian distribution $\mathcal{N}(0, \Sigma_{v_t})$. Let θ_t be the vector of all parameters at each time index t ,

$$\theta_t := \text{vec}(C_t^\top), \quad (2.3)$$

which implies that C_t and θ_t uniquely characterize each other. We introduce the following assumption, providing a form of *a priori* information. This plays a role akin to that of a regularizer in non-Bayesian techniques, such as in *Supervised Learning*, where algorithms without such regularizers are prone to overfitting.

Assumption 2.1 (Output map dynamics). *The dynamics of the output map are governed by the difference equation*

$$\theta_{t+1} = \theta_t + \eta_t, \quad t \in \mathbb{Z}_+, \quad (2.4)$$

where t denotes the time index, $\theta_t \in \mathbb{R}^{n_y \times n_x}$ is the vector of parameters driven by the vector of process noise $\eta_t \in \mathbb{R}^{n_y \times n_x}$ with Gaussian distribution $\mathcal{N}(\mu_{\eta_t}, \Sigma_{\eta_t})$. Further, assume that the initial parameter in (2.4), denoted by θ_0 , is drawn from the normal distribution $\mathcal{N}(\mu_{\theta_0}, \Sigma_{\theta_0})$.

Assumption 2.1 imposes a Gaussian random walk dynamics on the evolution of the parameters, which is the minimal structure and assumption on the variations of the parameters because of the maximum entropy feature of the Gaussian distributions. This assumption allows us to employ a stochastic belief of a deterministic reality in the *Bayesian* viewpoint. Let the inputs and outputs of system (2.1)-(2.2) be measured for $t = 0, \dots, T$, where $(T + 1) \in \mathbb{Z}_+$ denotes the length of the measurement data. More precisely, the input-output trajectory data is given by $\mathcal{D} = \{(u_t, y_t) | t = 0, \dots, T\}$. In addition, we make the following assumption:

Assumption 2.2 (Noise). *The process, measurement, and output map noise realizations, i.e., w_t , v_t and η_t , respectively, for all $t \in \{0, \dots, T\}$, are independent. Furthermore, the means μ_{x_0} , μ_{θ_0} , μ_{η_t} , and covariance matrices Σ_{x_0} , Σ_{w_t} , Σ_{v_t} , Σ_{θ_0} and Σ_{η_t} , for $t \in \{0, \dots, T\}$, are known.*

Remark 2.2.1 (A priori knowledge). *While we assume μ_{θ_0} , μ_{η_t} , Σ_{θ_0} , and Σ_{η_t} to be readily known, in practical applications, these parameters can be obtained through various means depending on the context, e.g., employing prior knowledge of the nominal model, empirically from previous experiments' data, or employing a suitable hyperparameter estimation method when $\mu_{\eta_t} = \mu_{\theta_0}$ and $\Sigma_{\eta_t} = \Sigma_{\theta_0}$, for $t \in \mathbb{Z}_+$.*

Ultimately, the question is whether the observation model (2.2) could be estimated. More precisely, we would like to address the following problem:

Problem 2.1 (Time-Varying parameter estimator). *Given the process and observation models (2.1) and (2.2), input-output measurement data $\mathcal{D}$, and under Assumptions 2.1-2.2, estimate the unknown time-varying observation matrices C_t in an efficient and tractable way.*

To address the problem 2.1, we develop a MAP approach in the next section, followed by a tractable reformulation using LMIs techniques in Section 2.4.

2.3 MAXIMUM A POSTERIORI ESTIMATION

In this section, we propose a Bayesian method for estimating the unknown observation matrices $C_0, \dots, C_T$. The three main elements in Bayesian estimation methods are a prior density function, an observation model, and a loss function, which we briefly explain for solving our problem with the MAP approach.

2.3.1 LIFTED PROCESS AND OBSERVATION MODEL

Let us represent the process model (2.1) in the following *lifted matrix form*:

$$\mathbf{x} = \mathbf{A}(\mathbf{u} + \mathbf{w}_x), \quad (2.5)$$

where $\mathbf{x} = [\mathbf{x}_0^\top, \dots, \mathbf{x}_T^\top]^\top$ is the system states vector for the entire horizon up to time T , while the input vector is modified to include the initial state $\mathbf{u} = [\mu_{x_0}^\top, (\mathbf{B}_0 \mathbf{u}_0)^\top, \dots, (\mathbf{B}_{T-1} \mathbf{u}_{T-1})^\top]^\top$, and the noise vector comprises the initial-state uncertainty and the process-noise terms, $\mathbf{w}_x = [\mathbf{w}_{x_0}^\top, \mathbf{w}_0^\top, \dots, \mathbf{w}_{T-1}^\top]^\top$, where $\mathbf{w}_{x_0} \sim \mathcal{N}(0, \Sigma_{x_0})$. Given that the process noise and initial state uncertainty are uncorrelated from Assumption 2.2, we can specify $\mathbf{w}_x$ in terms of a multivariate normal distribution $\mathcal{N}(0, \Sigma_{\mathbf{w}_x})$, in which $\Sigma_{\mathbf{w}_x} = \text{diag}(\Sigma_{x_0}, \Sigma_{w_0}, \dots, \Sigma_{w_{T-1}})$. The *lifted* transition matrix $\mathbf{A}$ has the following lower triangular form:

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & 0 & 0 & \dots & 0 \\ \mathbf{A}_0 & \mathbf{I} & 0 & \dots & \vdots \\ \mathbf{A}_1 \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{I} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{A}_{T-1} \dots \mathbf{A}_0 & \mathbf{A}_{T-1} \dots \mathbf{A}_1 & \dots & \mathbf{A}_{T-1} & \mathbf{I} \end{bmatrix}.$$

Similarly, the observation model (2.2) for the entire trajectory can be expressed as

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v}, \quad (2.6)$$

where $\mathbf{y} = [\mathbf{y}_0^\top, \dots, \mathbf{y}_T^\top]^\top$ is the vector of all measurements, $\mathbf{v} \sim \mathcal{N}(0, \Sigma_v)$ is the vector of all measurement noise realizations with $\Sigma_v = \text{diag}(\Sigma_{v_0}, \Sigma_{v_1}, \dots, \Sigma_{v_T})$, and $\mathbf{C}$ is the lifted observation matrix

$$\mathbf{C} = \text{diag}(\mathbf{C}_0, \mathbf{C}_1, \dots, \mathbf{C}_T).$$

Finally, we describe the dynamics of the output map parameters θ for the entire trajectory

$$\theta = \mu_\theta + \mathbf{w}_\theta, \quad (2.7)$$

where $\theta = [\theta_0^\top, \dots, \theta_T^\top]^\top$, and μ_θ results from the summation of the biases of the initial parameter and the noise $\mu_\theta = [\mu_{\theta_0}^\top, \mu_{\theta_0}^\top + \mu_{\eta_0}^\top, \dots, \mu_{\theta_0}^\top + \sum_{i=0}^{T-1} \mu_{\eta_i}^\top]^\top$. Similarly, the noise signal $\mathbf{w}_\theta$ results from the integration over the entire horizon including the uncertainty of the initial parameter as $\mathbf{w}_\theta = \mathbf{D}\eta$, where $\eta = [\eta_{\theta_0}^\top, \eta_0^\top, \dots, \eta_{T-1}^\top]^\top$ and $\eta_{\theta_0} \sim \mathcal{N}(0, \Sigma_{\theta_0})$, with

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & 0 & \dots & 0 \\ \mathbf{I} & \mathbf{I} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{I} & \mathbf{I} & \dots & \mathbf{I} \end{bmatrix}.$$

Since the parameters are assumed independent (Assumption 2.2), we have $\mathbf{w}_\theta \sim \mathcal{N}(0, \Sigma_{\mathbf{w}_\theta})$, $\Sigma_{\mathbf{w}_\theta} = D\Sigma_\eta D^\top$, with $\Sigma_\eta = \text{diag}(\Sigma_{\theta_0}, \Sigma_{\eta_0}, \dots, \Sigma_{\eta_{T-1}})$. Ultimately, the model (2.7) is used to specify the prior density function. In what follows, we represent C as $C(\theta)$ to emphasize the dependence of C on θ according to (2.3). Consequently, substituting $\mathbf{x}$ in (2.6) with the expression from (2.5) results in the observation model, with unknown $C(\theta)$, describing the measurements $\mathbf{y}$ as a function of the applied inputs $\mathbf{u}$:

$$\mathbf{y} = C(\theta)\mathbf{A}\mathbf{u} + \mathbf{w}_y(\theta), \quad (2.8)$$

where $\mathbf{w}_y(\theta) = C(\theta)\mathbf{A}\mathbf{w}_x + \mathbf{v}$. Also, from Assumption 2.2:

$$\mathbf{w}_y(\theta) | \theta \sim \mathcal{N}(0, \Sigma_{\mathbf{w}_y}(\theta)),$$

where $\Sigma_{\mathbf{w}_y}(\theta) = C(\theta)\mathbf{A}\Sigma_{\mathbf{w}_x}\mathbf{A}^\top C^\top(\theta) + \Sigma_v$. This model is used later to specify the conditional probability density function of the measurements. Note that $\mathbf{w}_y(\theta) | \theta$ remains Gaussian with the derived covariance since both noise sources, $\mathbf{w}_x$ and $\mathbf{v}$, are Gaussian and independent.

2.3.2 MAP LOSS FUNCTION

In MAP estimation, one aims to find an estimate $\hat{\theta}$ for the parameters by minimizing the cost function $\mathbb{E}[1 - \mathbb{1}_{|\theta - \hat{\theta}|_\infty \leq \epsilon}(\theta)]$, where θ is the vector of random variables, $\mathbb{1}(\cdot)$ is an indicator function, and ϵ is a small scalar [30]. Minimizing the expectation of such a loss function implies maximizing the conditional probability density of θ given the vector of observations and inputs [76, Chapter 4],

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} p(\theta | \mathbf{y}, \mathbf{u}), \quad (2.9)$$

where $\hat{\theta}$ is the estimate of the true parameter θ . The following lemma formalizes this first step to compute the MAP estimation.

Lemma 2.3.1 (MAP optimization problem). Define $\mathcal{J} : \mathbb{R}^{n_y \times n_x \times (T+1)} \rightarrow \mathbb{R}$ as

$$\mathcal{J}(\theta) := \log \det(\Sigma_{\mathbf{w}_y}(\theta)) + \|\mathbf{y} - C(\theta)\mathbf{A}\mathbf{u}\|_{\Sigma_{\mathbf{w}_y}^{-1}(\theta)}^2 + \|\theta - \mu_\theta\|_{\Sigma_\theta^{-1}}^2. \quad (2.10)$$

The MAP estimation is then equivalent to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathcal{J}(\theta). \quad (2.11)$$

Proof. Using Bayes' rule, the MAP estimation can be reformulated as

$$\max_{\theta} p(\theta | \mathbf{y}, \mathbf{u}) = \max_{\theta} \frac{p(\mathbf{y} | \theta, \mathbf{u}) p(\theta | \mathbf{u})}{p(\mathbf{y} | \mathbf{u})}. \quad (2.12)$$

We first note that the denominator of (2.12) does not depend on θ and, hence, can be neglected without changing the optimizer. Moreover, we note that the dynamics of θ in (2.4) (or equivalently in the lifted form in (2.7)) do not depend on the input sequence of $\mathbf{u}$ (i.e., $p(\theta | \mathbf{u}) = p(\theta)$). Next, using a straightforward computation, one can derive the

probability density functions $p(\theta)$ and $p(y|\theta, u)$. Specifically, from (2.7), we know that the variable θ is Gaussian with the probability density function

$$p(\theta) = \frac{\exp\left(-\frac{1}{2}(\theta - \mu_\theta)^\top \Sigma_{w_\theta}^{-1}(\theta - \mu_\theta)\right)}{\sqrt{(2\pi)^{(T+1)n_y n_x} \det(\Sigma_{w_\theta})}}.$$

Similarly, we know from (2.8) that given θ and the input sequence u , the output sequence y is also Gaussian with the conditional probability density function

$$p(y|\theta, u) = \frac{\exp\left(-\frac{1}{2}(y - C(\theta)Au)^\top \Sigma_{w_y}^{-1}(\theta)(y - C(\theta)Au)\right)}{\sqrt{(2\pi)^{(T+1)n_y} \det(\Sigma_{w_y}(\theta))}}.$$

Finally, applying the monotonically increasing function $\log$ and observing that all terms in the denominators except $\det(\Sigma_{w_y}(\theta))$ are constant, we arrive at the minimization problem of the function $\mathcal{J}$ defined in (2.11). $\square$

In the next section, we propose a tractable conservative approximation using LMIs techniques to tackle the non-convex objective function $\mathcal{J}(\theta)$ defined in (2.10).

Remark 2.3.2 (Robust estimation). *Alternatively, a robust minimax estimation formulation similar to [153] could be employed. This approach, however, requires finding an ambiguity set to approximate non-Gaussian observation uncertainties due to the multiplication of Gaussian variables in $w_y(\theta)$.*

2.4 PROPOSED SOLUTION

The optimization problem (2.11) is non-convex not only because of the weight $\Sigma_{w_y}^{-1}(\theta)$, being quadratic in the parameters θ , in the second term but also because of the log-determinant operator in the first term. A typical approach is to use first-order algorithms to find a solution due to the mentioned non-convexities. These algorithms, however, only guarantee convergence to a local optimum. Therefore, selecting an appropriate initial starting point is crucial to the obtained quality of the solution. We propose to solve the problem in two steps: first, we perform a convex relaxation of (2.11) into a set of LMIs, which we use to compute an initial approximate minimizer; next, we employ this approximate minimizer to initialize (warm-start) a first-order optimization method, e.g., steepest descent [89] or quasi-Newton algorithms [107], to solve (2.11) thus refining our initial minimizer estimate.

Theorem 2.4.1 (LMI conservative approximation). *Consider the following LMIs:*

$$\begin{aligned} \min_{S, \theta, \gamma, \beta} \quad & \text{tr}(S - \mathbb{I}) + \gamma + \beta \\ \text{s.t.} \quad & \begin{bmatrix} -\Sigma_{w_x}^{-1} & A^\top C^\top(\theta) \\ C(\theta)A & \Sigma_v - S \end{bmatrix} \preceq 0, \\ & \begin{bmatrix} -S & (y - C(\theta)Au) \\ (y - C(\theta)Au)^\top & -\gamma \end{bmatrix} \preceq 0, \\ & \begin{bmatrix} -\Sigma_{w_\theta} & (\theta - \mu_\theta) \\ (\theta - \mu_\theta)^\top & -\beta \end{bmatrix} \preceq 0. \end{aligned} \tag{2.13}$$

Then, the optimal value of the nonlinear program (2.11) is upper bounded by

$$\mathcal{J}^* + \|\mathbf{y} - \mathbf{C}(\theta^*)\mathbf{A}\mathbf{u}\|_{(\Sigma_{\mathbf{w}_y}^{-1}(\theta^*) - \mathbf{S}^{-1})}^2,$$

where $\mathcal{J}^*$ and $(\mathbf{S}^*, \theta^*)$ are the optimal value and the optimizer of (2.13), respectively.

Proof. Consider a matrix $\mathbf{S} \succ 0$ upper bounding the covariance matrix $\Sigma_{\mathbf{w}_y}(\theta) \succ 0$ as

$$\Sigma_{\mathbf{w}_y}(\theta) = \mathbf{C}(\theta)\mathbf{A}\Sigma_{\mathbf{w}_x}\mathbf{A}^\top\mathbf{C}^\top(\theta) + \Sigma_v \leq \mathbf{S}. \quad (2.14)$$

Thus, $\lambda_i(\Sigma_{\mathbf{w}_y}(\theta)) \leq \lambda_i(\mathbf{S})$, for $i = 1, \dots, n_y(T+1)$, which implies that

$$\log \det(\Sigma_{\mathbf{w}_y}(\theta)) \leq \log \det(\mathbf{S}).$$

Since $\log \det(\mathbf{S}) = \sum_{i=1}^{n_y(T+1)} \log \lambda_i(\mathbf{S})$ and $\text{tr}(\mathbf{S}) = \sum_{i=1}^{n_y(T+1)} \lambda_i(\mathbf{S})$, we also have

$$\log \det(\Sigma_{\mathbf{w}_y}(\theta)) \leq \log \det(\mathbf{S}) \leq \text{tr}(\mathbf{S} - \mathbf{I}).$$

Using the *Schur complement*, one can see that (2.14) is equivalent to the following linear matrix inequality (LMI)

$$\begin{bmatrix} -\Sigma_{\mathbf{w}_x}^{-1} & \mathbf{A}^\top\mathbf{C}^\top(\theta) \\ \mathbf{C}(\theta)\mathbf{A} & \Sigma_v - \mathbf{S} \end{bmatrix} \preceq 0.$$

Similarly, considering $\gamma \geq 0$ and $\beta \geq 0$ such that

$$(\mathbf{y} - \mathbf{C}(\theta)\mathbf{A}\mathbf{u})^\top \mathbf{S}^{-1}(\mathbf{y} - \mathbf{C}(\theta)\mathbf{A}\mathbf{u}) \leq (\mathbf{y} - \mathbf{C}(\theta)\mathbf{A}\mathbf{u})^\top \Sigma_{\mathbf{w}_y}^{-1}(\theta)(\mathbf{y} - \mathbf{C}(\theta)\mathbf{A}\mathbf{u}) \leq \gamma, \quad (2.15)$$

and

$$(\theta - \mu_\theta)^\top \Sigma_{\mathbf{w}_\theta}^{-1}(\theta - \mu_\theta) \leq \beta, \quad (2.16)$$

we can apply again the *Schur complement* to the inequalities in (2.15) and (2.16) to obtain the last two LMIs in (2.13). Finally, replacing the terms in the cost function $\mathcal{J}(\theta)$ in (2.10) with their bounds and including the corresponding LMIs as constraints arrives at the LMIs (2.13). Note further that by definition, we have

$$\mathcal{J}^* + \|\mathbf{y} - \mathbf{C}(\theta^*)\mathbf{A}\mathbf{u}\|_{(\Sigma_{\mathbf{w}_y}^{-1}(\theta^*) - \mathbf{S}^{-1})}^2 = \mathcal{J}(\theta^*) + \text{tr}(\mathbf{S}^* - \mathbf{I}) - \log \det(\Sigma_{\mathbf{w}_y}(\theta^*)) \geq \mathcal{J}(\theta^*), \quad (2.17)$$

where the function $\mathcal{J}(\theta^*)$ is defined in (2.10), and the last inequality follows from (2.14). $\square$

The tightness of the inequality in (2.17) mainly depends on the gap between $\log \det(\mathbf{S})$ and $\text{tr}(\mathbf{S} - \mathbf{I})$ since $\log \det(\mathbf{S})$ is bounded from above by $\text{tr}(\mathbf{S} - \mathbf{I})$, which is negligible when $\lambda_i(\mathbf{S}) \approx 1$, for $i = 1, \dots, n_y(T+1)$. One may employ a suitable matrix $\mathbf{W}$ to scale the eigenvalues of $\mathbf{S}$, replace $\log \det(\mathbf{S})$ with $\log \det(\mathbf{WSW}) - 2 \log \det(\mathbf{W})$ and approximate $\log \det(\mathbf{WSW})$ with $\text{tr}(\mathbf{WSW} - \mathbf{I})$. Furthermore, the closeness of $\mathcal{J}^*$ and $\mathcal{J}(\theta^*)$ in (2.17) is proportional to the fitness quality of the measurements and whether $\mathbf{S}^*$ is close to the covariance matrix accordingly.

In addition, Theorem 2.4.1 provides an approximation of (2.11), producing an initial near-optimal solution. As indicated earlier, we propose to employ this solution to warm-start a local (non-convex) optimizer. Due to its fast convergence, we propose to employ, as a refining optimizer, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [107, Chapter 6], a variant of quasi-Newton methods. The BFGS algorithm approximates the Hessian matrix for its search directions relying on an analytical expression of the gradient $\nabla_{\theta} \mathcal{J}(\theta)$. The gradient of the cost function (2.10) with respect to the parameters θ

$$\nabla_{\theta} \mathcal{J}(\theta) = \left[\frac{\partial \mathcal{J}}{\partial \theta_1}, \dots, \frac{\partial \mathcal{J}}{\partial \theta_{n_y n_x (T+1)}} \right]^T, \quad (2.18)$$

can be easily derived applying the following chain rule:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \theta_{ijk}} = & 2\text{tr} \left[\left(A \Sigma_{w_x} A^T C^T(\theta) \Sigma_{w_y}^{-1}(\theta) \right) C_k^{ij}(\theta) \right. \\ & - \left(A \Sigma_{w_x} A^T C^T(\theta) \Sigma_{w_y}^{-1}(\theta) (y - C(\theta) A u) (y - C(\theta) A u)^T \Sigma_{w_y}^{-1}(\theta) \right) C_k^{ij}(\theta) \\ & \left. - \left(A u (y - C(\theta) A u)^T \Sigma_{w_y}^{-1}(\theta) \right) C_k^{ij}(\theta) \right] + ((\theta - \mu_{\theta})^T \Sigma_{w_{\theta}}^{-1}) \theta^{ijk}, \end{aligned}$$

where $C_k^{ij}(\theta)$ is the single-entry matrix of $C(\theta)$ with the block matrix of $C_k(\theta)$ having 1 at index (i, j) and zero elsewhere, and θ^{ijk} is the single-entry vector of θ with 1 at index ijk and zero elsewhere. The LMIs (2.13) initializes the original non-convex problem with a locally optimal solution. Thus, the computational complexity of the proposed method consists of the well-known computational complexity of solving the SDP problems [105], i.e., a one-time solution of (2.13), and the computation of the gradient (2.18) per iteration of the first-order method, i.e.,

$$\mathcal{O} \left(\frac{n_x^3(T+1)^3 + n_x^2(T+1)^2}{2} + n_y n_x^2(T+1)^2 + n_x n_y^2(T+1)^3 + (n_y)^3(T+1)^3 \right),$$

which is $\mathcal{O}(T^3)$ when $n_x, n_y \ll T$.

2.5 NUMERICAL EXPERIMENTS

In this section, we provide a numerical example to verify the efficacy and performance of the proposed method: employing the LMIs (2.13) to warm-start the solution of (2.11) via the BFGS optimizer. Additionally, we compare the resulting solution with the estimates obtained from EM and DKS algorithms. To have a fair comparison, we also employ the same BFGS optimizer for the *M-Step* of the EM estimation. We demonstrate our solution on the following system:

$$x_{t+1} = \begin{bmatrix} 0.7 & 0.25 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0.25 & 0.7 \end{bmatrix} x_t + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} (3.5 + \cos(2t)) + w_t,$$

with $\mu_{x_0} = [1, 0.5, 2]^T$. The observation is a two-dimensional model, i.e., the number of measurements per time instant is $n_y = 2$. The system has sampled input and measurement

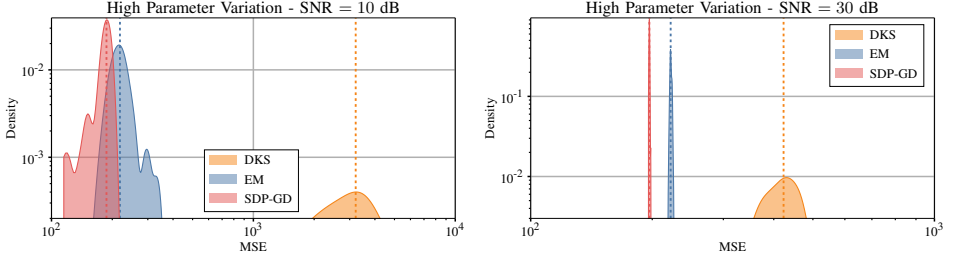
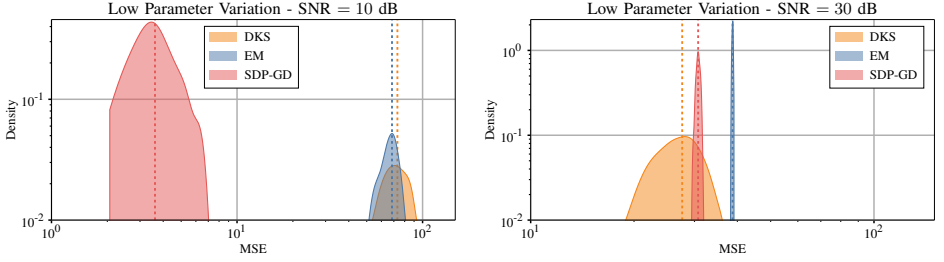
(a) High noise of Σ_η and two different SNRs(b) Low noise of Σ_η and two different SNRs

Figure 2.1: The MSE of the three methods across four scenarios for 100 experiments.

pairs in $\mathcal{D}$ every 100 milliseconds for 10 seconds, i.e., $T = 100$. Thus, the number of parameters to be estimated is $n_y \times n_x \times (T + 1) = 606$. The noise covariance of process, observation and output map dynamics, Σ_{w_t} , Σ_{v_t} and Σ_{η_t} , are assumed to remain constant across the entire horizon. Moreover, the initial state and parameter noise covariance are similar to the noise covariance of process and output map dynamics, respectively (i.e., $\Sigma_{x_0} = \Sigma_{w_t}$ and $\Sigma_{\theta_0} = \Sigma_{\eta_t}$). The output map noise biases μ_{η_t} are generated such that $\mu_{1,\eta_t} = 5 + \cos(0.4t)e^{-0.6t}$, $\mu_{2,\eta_t} = 1.5 + \sin(0.025t)e^{-0.6t}$, $\mu_{3,\eta_t} = 2$, $\mu_{4,\eta_t} = 5 + \cos(0.4t)e^{-0.6t}$, $\mu_{5,\eta_t} = 1.5 + \sin(0.025t)e^{-0.6t}$, and $\mu_{6,\eta_t} = 2$. The initial parameter bias, μ_{θ_0} , is derived from μ_{η_0} by setting $t = 0$, and the DKS and EM algorithms' initialization is according to these noise bias values. We examine the performance of our algorithm, hereafter called SDP-GD, compared with EM and DKS on four scenarios generated by employing High/Low SNRs for the process and observation noise, particularly 30 and 10 decibels (dBs), and High/Low parameter variation of:

$$\begin{cases} \text{High:} & \Sigma_{\eta_t} = \text{diag}(2.17, 0.076, 1.19, 1.38, 0.87, 1.27), \\ \text{Low:} & \Sigma_{\eta_t} = \text{diag}(6.9, 0.2, 3.8, 4.4, 2.8, 4) \times 10^{-2}. \end{cases}$$

Combined results from 100 experiments for each of the four scenarios, keeping the same ground-truth realization in each scenario, are illustrated in Figure 2.1. The plots demonstrate the median (vertical dotted lines) and distribution across experiments based on the mean-squared error (MSE), i.e. $\text{MSE} = \frac{1}{T+1} \sum_{t=0}^T \|\theta_t - \hat{\theta}_t\|_2^2$. One can observe how the DKS underperforms compared to the EM and our SDP-GD solutions in more challenging

Elapsed time per seconds (mean $\pm$ std)				
	Experiment Scenario	DKS	EM	SDP-GD
10 dB	Low Parameter Variation	18 $\pm$ 4	14265 $\pm$ 3112	1265 $\pm$ 109
	High Parameter Variation	27 $\pm$ 13	22547 $\pm$ 6768	1542 $\pm$ 178
30 dB	Low Parameter Variation	9 $\pm$ 3	7999 $\pm$ 802	1543 $\pm$ 132
	High Parameter Variation	21 $\pm$ 13	3796 $\pm$ 412	1605 $\pm$ 128

Table 2.1: The Average computation performance across all scenarios for 100 experiments.

scenarios where the process-observation model noise is high or in the presence of *High* parameter variation.

The average and standard deviation of the computation time of each method across 100 experiments are reported in Table 2.1. We performed all the experiments on a cluster node with 384G memory and 40 CPU cores (2 Intel Xeon Gold 6148 @ 2.40GHz). The elapsed execution times confirm our hypothesis that EM is computationally more expensive than the other alternatives. The performance of EM and DKS algorithms highly depends on the initialization, while in contrast, our proposed solution takes advantage of a warm-start initializer obtained from solving a convexified approximation of the original optimization problem. This initialization helps converge to a better local optimum faster than the EM algorithm. In addition, the *M-step* of the EM algorithm, in this problem, does not hold a closed-form solution, which results in utilizing a first-order method. This gradient *M-Step* also plays a part in the general slowness of the EM algorithm. Our algorithm, however, requires a one-time execution of the set of LMIs followed by an iterative quasi-Newton method with a superlinear convergence rate. Hence, it provides the best of both worlds, i.e., better estimations than EM and DKS with less computation time than EM.

2.6 CONCLUSION

We have introduced a method for estimating an unknown output map of a linear time-varying system by employing a stochastic characterization of the evolution of the output map parameters, which serves as *a priori* information for MAP estimation. The derived MAP optimization problem is solved by relaxing the optimization as a set of LMIs, whose solution provides a warm-start for a gradient descent algorithm. Compared with standard approaches to solve this problem, namely EM and DKS, we showed experimentally the superiority of our method in estimation performance and lower computational demands compared to EM. Future work will explore the incorporation of other types of *a priori* knowledge on the output map, the development of efficient causal filters following similar approaches, the minimax formulation for robust estimation, considering noise models with more general structures, and introducing methods for efficient design of the control input.

3

3

AFFINE ESTIMATORS AND ACTIVE LEARNING FOR WIENER MODELS

This chapter presents a Bayesian estimation framework for Wiener models, focusing on learning nonlinear output functions under known linear state dynamics. We derive a closed-form optimal affine estimator for the unknown parameters, characterized by the so-called “dynamic basis statistics” (DBS). Several properties of the resulting estimator are established, including Bayesian unbiasedness, closed-form approximate posterior statistics, error monotonicity with respect to trajectory length, and consistency under a persistent-excitation condition. In the special case of Fourier basis functions, we demonstrate that the closed-form description is computationally available, as the Fourier DBS enjoys explicit expressions. Furthermore, we identify an inherent inconsistency in the Fourier bases for single-trajectory measurements, regardless of the input excitation. Leveraging the closed-form estimation error, we develop an active learning algorithm synthesizing input signals to minimize estimation error. Numerical experiments validate the efficacy of our approach, showing significant improvements over regularized least-squares methods.

3.1 INTRODUCTION

System modeling and identification quality are essential for practical analysis, prediction, and control of complex phenomena across disciplines. Classical system identification techniques combine theoretical frameworks with empirical data to select appropriate model structures and estimate parameters for constructing accurate models of dynamic systems [36, 87, 133, 145]. While linear system identification has a well-established foundation, identifying nonlinear systems remains an evolving area of research [103].

Parametric approaches for nonlinear system identification, such as extending linear state-space models to incorporate polynomial nonlinear terms [109], have shown significant achievements [88]. However, these methods require a trade-off between bias and variance in model order selection [112, 128]. Recent regularization-based approaches address this challenge by exploring high-dimensional search spaces through kernel-based methods [31, 110]. These techniques enhance robustness in model selection by using continuous regularization parameters instead of discrete model orders. Furthermore, leveraging infinite-dimensional reproducing kernel Hilbert spaces (RKHS) via Gaussian process (GP) models [117] offers a probabilistic framework for nonlinear system identification [111], allowing prior knowledge about system properties, such as stability and smoothness, in the identification process.

While robust for system identification, kernel-based methods are computationally expensive for large-scale problems due to operations like matrix inversion in high dimensions. To address this challenge, randomized low-dimensional feature spaces, particularly Fourier basis functions, were introduced to accelerate kernel machine training [115]. Various techniques, including probabilistic variational methods [64], have since been proposed to approximate radial basis kernels using random Fourier features [84]. Additionally, approximations have been developed to handle input noise in GP regression [55, 95], enhancing applicability to real-world scenarios. This work addresses similar challenges for identifying an unknown nonlinear function influenced by time-varying correlated noise in its inputs. We propose an optimal Bayesian estimator for nonlinear functions represented as finite combinations of basis functions, focusing on Fourier bases.

Our problem can also be classified within the block-oriented nonlinear models, specifically Wiener systems characterized by a linear process followed by a static nonlinear observation model [129]. Existing Wiener system identification techniques span a range of methodologies. Some approaches use GP models for the static nonlinear block and approximate posterior densities using Markov chain Monte Carlo (MCMC) methods [82, 120]. Others model the static nonlinearity as a polynomial of known order or approximate it as a linear combination of predefined basis functions, employing the prediction error method (PEM) or Gaussian sum filtering with expectation-maximization (EM) algorithms for inference [20, 29, 137]. Additionally, studies on parameter estimation consistency using PEM and maximum likelihood under various noise assumptions [59, 60] demonstrate that traditional least-squares methods can lead to biased estimates when process noise is present.

Unlike these existing techniques, our work focuses solely on the Bayesian estimation of static nonlinear observation parameters in Wiener systems with known linear time-varying dynamics affected by process and measurement noise. This problem arises in many applications, such as robot mapping in unknown environments. For instance, Autonomous

Underwater Vehicles (AUVs) mapping the seabed in deep-sea environments face challenges due to the interplay between vehicle dynamics and unknown nonlinear seabed observation models. Our proposed Bayesian affine minimum mean-squared error (MMSE) estimator analytically computes estimates and estimation errors. By accounting for process noise correlations over time through information gained from the covariance of the observation model, our method avoids divergence issues observed in *approximate* PEM.

While our Bayesian affine MMSE estimator addresses parameter estimation robustness against noise correlations, the quality of identification depends heavily on the choice of input signals used to excite the system. Active learning and optimal input design maximize information gain by selecting informative samples [130] or constructing input signals that optimize experimental criteria like Fisher or mutual information [44, 96]. Numerous studies have explored strategies for optimizing input selection for both linear [79, 146] and nonlinear systems [52, 93, 142, 143], where different experimental design criteria have led to varied approaches. In this work, we derive an optimal input design by directly minimizing the analytical estimation error of our parameter estimates, specifically by minimizing the trace of the estimation-error covariance matrix. Integrating this optimal input design with our Bayesian affine MMSE estimator prior to conducting experiments enhances the efficacy of identifying static nonlinear observation parameters under correlated noise conditions.

Contributions. The main contributions are as follows:

- **Optimal Bayesian affine estimator:** Introducing a Bayesian setting, we derive the closed-form solution of the optimal affine estimator for the unknown parameters of the output function (Theorem 3.3.2), which is characterized in terms of the so-called “dynamic basis statistics” (DBS).
- **Optimal estimator features:** The proposed optimal estimator enjoys the following properties: (i) Bayesian unbiasedness (Proposition 3.3.3), (ii) Closed-form updates for approximate posterior statistics (Remark 3.3.4), (iii) Monotonic error reduction (Corollary 3.3.5), and (iv) Consistency under specific conditions (Proposition 3.4.3).
- **Fourier basis explicit solution:** The generic closed-form solution of the Bayesian estimator requires the respective DBS of the basis functions, which can be computationally demanding. We show that these statistics admit explicit expressions in the special case of Fourier basis (Lemma 3.4.1). We further identify an inherent inconsistency of single-trajectory measurements for the Fourier bases, irrespective of the input trajectory (even if unbounded and persistently excited), when the underlying dynamics are stochastically unstable (Proposition 3.4.2).
- **Active learning:** Leveraging the closed-form description of the estimation error, we propose a first-order algorithm to actively design an input signal that locally minimizes it (Section 3.5).

The theoretical results are validated through extensive numerical experiments, demonstrating the superiority of our Bayesian estimator with active learning over classical regularized least-squares methods (cf. Figure 3.2). To facilitate reproducibility, we provide an open-source MATLAB library available at “<https://github.com/sasanvakili/Bayesian4Wiener>.”

Organization. Section 3.2 introduces the modeling setting and problem formulation. Section 3.3 presents the solution approaches: classical regularized least-squares and the Bayesian affine MMSE estimator with its properties. Section 3.4 provides explicit expressions for the special case of Fourier basis functions and discusses the consistency condition. Section 3.5 describes a first-order algorithm for actively learning input signals. Section 3.6 outlines an experimental setup to examine the consistency condition and compares the proposed Bayesian affine estimator across four benchmarks, and Section 3.7 concludes the chapter. Detailed proofs of mathematical statements are provided in the “Technical Proofs” subsection of each corresponding section.

Notation. Throughout this chapter, $\mathbb{Z}_+$, $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^{n \times m}$, and $\mathbb{S}_+^n$ denote the set of positive integers, the real numbers, nonnegative real numbers, $n \times m$ real matrices, and the space of all symmetric positive semidefinite matrices in $\mathbb{R}^{n \times n}$, respectively. The symbol $\mathbb{I}$ refers to the identity matrix, subscripts denote elements of a vector or matrix, while superscripts represent the time index of vector or matrix elements. The trace operator is denoted by $\text{tr}(\cdot)$, the transpose of a matrix A is denoted by $A^\top$, and $\text{diag}(A_1, \dots, A_k)$ represents a block-diagonal matrix with diagonal entries $A_1, \dots, A_k$. The inner product of two vectors x and y is given by $\langle x, y \rangle = x^\top y$, and the respective 2-norm is $\|x\| = \sqrt{\langle x, x \rangle}$. For a matrix $A \in \mathbb{R}^{n \times n}$, the largest (smallest) absolute value of eigenvalues is denoted by $\lambda_{\max}(A)$ ($\lambda_{\min}(A)$). The conic inequality $A \preceq B$ means that the matrix difference $B - A$ is positive semidefinite, i.e., $B - A \succeq 0$. The notation $P(\mu, \Sigma)$ 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.”

3.2 PROBLEM DESCRIPTION

Consider a *known* discrete-time linear time-varying dynamical system where the states at time t are observed through an *unknown* observation model

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t u_t + w_{t+1}, \\ y_t &= h_\theta(x_t) + v_t. \end{aligned} \quad (3.1)$$

Here, $t = \{0, \dots, T\}$ represents the time index starting from 0 and ending at time T , $x_t \in \mathbb{R}^{n_x}$ is the vector of state variables, $A_t \in \mathbb{R}^{n_x \times n_x}$ is the state transition matrix, $u_t \in \mathbb{R}^{n_u}$ is the vector of inputs, $B_t \in \mathbb{R}^{n_x \times n_u}$ is the input matrix, and $w_{t+1} \in \mathbb{R}^{n_x}$ is the process noise, which has a distribution given by $P(0, \Sigma_{w_{t+1}})$. Additionally, the initial state x_0 is characterized by a mean vector μ_{x_0} and covariance matrix Σ_{x_0} . Observations are made through the scalar output measurements $y_t \in \mathbb{R}$, while $v_t \in \mathbb{R}$ represents the measurement noise with a distribution $P(0, \sigma_{v_t}^2)$. The output function $h_\theta : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ is defined as a finite linear combination of *known* basis functions $\phi_n : \mathbb{R}^{n_x} \rightarrow \mathbb{C}$:

$$h_\theta(x) = \sum_{n=0}^N \theta_n \phi_n(x) = \langle \phi(x), \theta \rangle, \quad (3.2)$$

where $\theta = [\theta_0, \dots, \theta_N]^\top$ is the vector of *unknown* parameters, $\phi(x) = [\phi_0(x), \dots, \phi_N(x)]^\top$ is the vector of basis functions evaluated at x , and $N + 1$ is the number of basis functions. We

note that the setting (3.1), i.e., linear dynamics followed by nonlinear output function, is referred to as the *Wiener model*.

Remark 3.2.1 (Modeling setting). Two important points are worth noting regarding the modeling setting of this study:

- (i) **Multivariate output measurements:** For measurements in higher dimensions where $y_t \in \mathbb{R}^{n_y}$, the output function (3.2) extends to a vector form $h_\theta(x) = [h_\theta^1(x), \dots, h_\theta^{n_y}(x)]^\top$, where the goal is to learn each function $h_\theta^i(x)$ separately, parameterized as in (3.2). A common assumption in many applications is that the parameters of each $h_\theta^i(x)$ are independent of those of other components, effectively reducing the learning of a multivariate output function to multiple single-variate outputs. Therefore, for the simplicity of the exposition, we focus on single-output measurements ($n_y = 1$) in this study.
- (ii) **Bayesian prior interpretation:** A fundamental aspect of the Bayesian framework is incorporating prior information about unknown parameters. This information is formalized mathematically through a probability distribution, which in this study is characterized solely by its mean and covariance, i.e., the prior distribution belongs to $\mathbb{P}(\mu_\theta, \Sigma_\theta)$, where μ_θ and Σ_θ are given modeling parameters.

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. Our objective is to identify the model parameters θ from the measurement data y_t at all time steps, represented as $y = [y_0, \dots, y_T]^\top$. This task can also be interpreted as estimating the parameters of a linear or nonlinear function influenced by time-varying correlated noise in its inputs. The problem can be formally described as follows:

Problem 3.1 (Bayesian estimator for Wiener model). Given measurements $y \in \mathbb{R}^{T+1}$ and the parameters prior information $\mathbb{P}(\mu_\theta, \Sigma_\theta)$, design the optimal estimator $\hat{\theta} : \mathbb{R}^{T+1} \rightarrow \Theta$ that minimizes the expected loss, $\min_{\hat{\theta}(\cdot)} \mathbb{E}[\ell(\theta, \hat{\theta}(y))]$, where Θ is the parameter space and $\ell(\cdot, \cdot)$ is a predefined loss function quantifying the discrepancy between the true parameters θ and their estimate $\hat{\theta}(y)$.

3.3 SOLUTION APPROACHES

The unknown function $h_\theta(x_t)$ in the observation model of (3.1) is assumed to belong exactly to the hypothesis class defined in (3.2). Consequently, the output function can be reformulated and expressed in *lifted matrix* form as

$$y = \Phi^\top \theta + v, \quad (3.3)$$

where $y = [y_0, \dots, y_T]^\top$, $\Phi = [\phi(x_0), \dots, \phi(x_T)]$ is the basis aggregation matrix, $\theta = [\theta_0, \dots, \theta_N]^\top$, and $v = [v_0, \dots, v_T]^\top$ is the measurement noise vector. The measurement noise follows $v \sim \mathbb{P}(0, \Sigma_v)$, where $\Sigma_v = \text{diag}(\sigma_{v_0}^2, \dots, \sigma_{v_T}^2)$ assuming v_t are independent from each other at all times. Let us recall that the prior information about θ is characterized by a probability distribution defined by its mean and covariance, $\mathbb{P}(\mu_\theta, \Sigma_\theta)$ (cf. Remark 3.2.1-ii). The propagation of the state trajectory of the dynamical system through the output basis function is a key object in characterizing our proposed Bayesian estimator.

Definition 3.1 (Dynamic basis statistics). Let $\mathbf{x}_t$ be the dynamics trajectory of the system (3.1), and $\phi(\mathbf{x})$ be the set of basis functions of the output function (3.2). The dynamic basis statistics (DBS), denoted by $(\mu_\phi^t, \Sigma_\phi^{tt'})$, is the mean and covariance of $\phi(\mathbf{x}_t)$ at two time instants (t, t') , i.e.,

$$\mu_\phi^t = \mathbb{E}[\phi(\mathbf{x}_t)], \quad \Sigma_\phi^{tt'} = \mathbb{E}[\phi(\mathbf{x}_t)\phi(\mathbf{x}_{t'})^\top] - \mathbb{E}[\phi(\mathbf{x}_t)]\mathbb{E}[\phi(\mathbf{x}_{t'})]^\top. \quad (3.4)$$

Given the randomness of the elements of Φ as defined in Definition 3.1, the observation model can be rewritten as $\mathbf{y} = (\mathbb{E}[\Phi] + \Delta\Phi)^\top \boldsymbol{\theta} + \mathbf{v}$, where $\Delta\Phi$ is a *zero-mean* random matrix. If $\Delta\Phi$ were deterministic, however, a solution to Problem 3.1 could be obtained via the classical least-squares methods of supervised learning, as discussed in the next subsection.

3.3.1 CLASSICAL REGULARIZED LEAST-SQUARES

Regularized least squares (RLS), also called ridge regression, identifies unknown parameters by extending the ordinary least-squares method with a penalty on the parameters, known as the $\mathcal{L}_2$ regularization [17, Ch. 3]. This regularization reduces model complexity, helping to avoid overfitting and improving generalization to new, unseen data. Since Φ in observation model (3.3) is a random matrix, as noted in (3.4), one can *approximate* its columns by $\tilde{\Phi} = [\tilde{\phi}(\mathbf{x}_0), \dots, \tilde{\phi}(\mathbf{x}_T)]$ in two ways:

- (i) Dead reckoning least squares (DLS): $\tilde{\phi}(\mathbf{x}_t) = [\phi_0(\mathbb{E}[\mathbf{x}_t]), \dots, \phi_N(\mathbb{E}[\mathbf{x}_t])]^\top$.
- (ii) Mean least squares (MLS): $\tilde{\phi}(\mathbf{x}_t) = \mu_\phi^t = [\mathbb{E}[\phi_0(\mathbf{x}_t)], \dots, \mathbb{E}[\phi_N(\mathbf{x}_t)]]^\top$.

Using either of these approximations, the regularized least-squares method provides an estimator by solving the optimization problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \tilde{\Phi}^\top \boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2, \quad (3.5a)$$

where λ is a hyperparameter. The closed-form optimal solution yields a linear estimator with respect to the measurements $\mathbf{y}$ [17, Ch. 3] as

$$\hat{\boldsymbol{\theta}}_{\text{LS}}(\mathbf{y}) = (\tilde{\Phi}\tilde{\Phi}^\top + \lambda\mathbf{I})^{-1}\tilde{\Phi}\mathbf{y}. \quad (3.5b)$$

This estimator requires inverting a square matrix with dimensions equal to the number of basis functions. Consequently, the computational complexity of (3.5b) is $\mathcal{O}_{\text{LS}}(N^2T + N^3)$, depending on the number of measurements and unknown parameters. This complexity further simplifies to $\mathcal{O}_{\text{LS}}(T)$ when $N \ll T$. These *approximate* PEM approaches are used to evaluate the performance of this method through numerical experiments in Section 3.6.

3.3.2 OPTIMAL BAYESIAN AFFINE ESTIMATOR

Given the input-output trajectory data, the optimal Bayesian estimator $\hat{\boldsymbol{\theta}}: \mathbb{R}^{T+1} \rightarrow \Theta$, which addresses Problem 3.1, is obtained by solving

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \mathbb{E}[\ell(\boldsymbol{\theta}, \boldsymbol{\vartheta}) | \mathbf{y}] = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \int_{\Theta} \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where the second equality follows from Bayes' rule. Here, $\ell(\theta, \hat{\theta}(y))$ is a loss function that defines the performance criterion for estimating the unknown parameters using $\hat{\theta}(y)$. The term $p(y|\theta)$ represents the likelihood, i.e., the probability density function of the data given the unknown parameters, which is derived from the relationship between the data and the parameters. Meanwhile, $p(\theta)$ is the prior probability density function of the unknown parameters. Consequently, the Bayesian estimation approach requires both a performance criterion and a prior distribution as its starting point.

Among various performance criteria, the mean squared loss function, defined as $\ell(\theta, \vartheta) = \|\theta - \vartheta\|^2$, leads to the minimum mean-squared error (MMSE) estimate. This estimate corresponds to the mean of the posterior distribution of the unknown parameters θ given the measurements [76, Ch. 4], i.e.,

$$\hat{\theta}(y) = \mathbb{E}[\theta|y] = \int_{\Theta} \theta p(y|\theta) p(d\theta). \quad (3.6)$$

However, computing the posterior solution (3.6) is often challenging due to either an incomplete specification of the likelihood distribution $p(y|\theta)$ or because the integral does not have a closed-form solution. For certain special classes of joint distributions, such as *elliptical* distributions, the posterior distribution admits an analytical solution. The distribution $\mathbb{P}$ of a random vector $v \in \mathbb{R}^{n_f}$ is called *elliptical*, denoted as $\mathbb{P} = \mathcal{E}_\rho^{n_f}(\mu, \Sigma)$, if its characteristic function is given by $\varphi(f) = \exp(j\langle f, \mu \rangle) \rho(f^\top \Sigma f)$, where $\mu \in \mathbb{R}^{n_f}$ is a location parameter, $\Sigma \in \mathbb{S}_+^{n_f}$ is the dispersion matrix, and $\rho : \mathbb{R}_+ \rightarrow \mathbb{R}$ is the characteristic generator [68, p. 107].

Remark 3.3.1 (MMSE estimate of elliptical distributions). *If the joint distribution of the unknown parameters θ and the measurements y is elliptical, then the conditional distribution $p(\theta|y)$ is elliptical, and its first moment, forming the optimal solution (3.6), is affine in the variable y [24, Thm. 5].*

A fundamental classical result in these cases is that the mean of the posterior belongs to the class of affine estimators when the joint distribution is Gaussian [7, Ch. 2]. Inspired by Remark 3.3.1 and to enhance computational efficiency, we restrict the family of estimators in Problem 3.1 to affine functions for the remainder of this study. The following theorem presents the optimal closed-form solution for this class of estimators.

Theorem 3.3.2 (Optimal Bayesian affine MMSE estimator). *The optimal Bayesian affine MMSE estimator of Problem 3.1 is of the form $\hat{\theta}_B(y) = \Psi^* y + \psi^*$, where*

$$\Psi^* = \Sigma_\theta \mu_\Phi (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + M + \Sigma_v)^{-1}, \quad \psi^* = \mu_\theta - \Psi^* \mu_\Phi^\top \mu_\theta, \quad (3.7a)$$

$\mu_\Phi = [\mu_\phi^0, \dots, \mu_\phi^T]$, the $(t+1, t'+1)^{th}$ element of matrix M is $M_{tt'} = \text{tr}(\Sigma_\Phi^{tt'} (\Sigma_\theta + \mu_\theta \mu_\theta^\top))$, in which $(\mu_\phi^t, \Sigma_\Phi^{tt'})$ is the respective DBS of $\phi(x_t)$ in the sense of Definition 3.1. Consequently, the optimal mean-squared estimation error is

$$\mathcal{J}_B^* = \text{tr}(\Sigma_\theta - \Sigma_\theta \mu_\Phi (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + M + \Sigma_v)^{-1} \mu_\Phi^\top \Sigma_\theta). \quad (3.7b)$$

We note that the technical proofs of the theoretical results are provided in Subsection 3.3.3. The computational complexity of the derived optimal Bayesian affine MMSE estimator is $\mathcal{O}_B(N^2 T^2 + T^3)$, depending on the number of measurements and unknown parameters. When $N \ll T$, the computational complexity simplifies to $\mathcal{O}_B(T^3)$, which is significantly higher than that of the DLS and MLS linear estimators discussed in Section 3.3.1. While computationally more expensive, the Bayesian MMSE estimator accounts for process noise correlations over time and provides unbiased estimates relative to the prior.

3

Proposition 3.3.3 (Bayesian unbiasedness). *The affine MMSE estimator (3.7a) is Bayesian unbiased, i.e., $E[\theta - \hat{\theta}_B(y)] = 0$, where θ represents the true unknown parameters and the expectation is taken with respect to the joint distribution of (θ, y) .*

Given that the estimator is unbiased with respect to the prior, one can update the prior with the result obtained from this MMSE estimation.

Remark 3.3.4 (Bayesian update via posterior distribution). *Using the measurements y and the proposed affine MMSE estimator, the unknown parameters follow an approximate posterior distribution, i.e., $\theta|y \sim \mathbb{P}(\mu_\theta^{\text{pos}}, \Sigma_\theta^{\text{pos}})$, with known mean and covariance, given by*

$$\begin{cases} \mu_\theta^{\text{pos}} = \mu_\theta + \Psi^*(y - \mu_\Phi^\top \mu_\theta), \\ \Sigma_\theta^{\text{pos}} = \Sigma_\theta - \Sigma_\theta \mu_\Phi (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + M + \Sigma_v)^{-1} \mu_\Phi^\top \Sigma_\theta. \end{cases} \quad (3.8)$$

The proposed optimal affine estimator achieves the *minimum variance* among all affine estimators and reduces the variance relative to the prior distribution, i.e., $\Sigma_\theta \succeq \Sigma_\theta^{\text{pos}}$, as evident from (3.8).

Corollary 3.3.5 (Bayesian error monotonicity). *Let $\mathcal{J}_B^*(t)$ be the MMSE of the affine estimator defined in (3.7) at time t corresponding the measurement vector $y = [y_0, \dots, y_t]^\top$. Then, with one extra measurement at time $(t+1)$, the estimation error decreases monotonically, i.e., $\mathcal{J}_B^*(t+1) \leq \mathcal{J}_B^*(t)$.*

In the following section, we present Fourier basis functions as a particular case of our generic solution and further illustrate the efficacy and performance of that through numerical analyses in Section 3.6.

3.3.3 TECHNICAL PROOFS

This subsection contains detailed proofs of the theoretical results introduced earlier.

Proof of Theorem 3.3.2. Using the affine estimator structure $\hat{\theta}(y) = \Psi y + \psi$, the proof consists of solving the following MMSE problem:

$$\min_{\Psi, \psi} E[\|\theta - \Psi y - \psi\|^2]. \quad (3.9)$$

Decompose $\Phi = \mu_\Phi + \Delta\Phi$, where $\mu_\Phi = E[\Phi]$ and $\Delta\Phi$ is a *zero-mean* random matrix. Specifically, $\mu_\Phi = [\mu_\phi^0, \dots, \mu_\phi^T]$, with $\mu_\phi^t = E[\phi(x_t)]$ as defined by the DBS in (3.4). Expanding the MMSE objective in (3.9) by substituting the measurement vector y with its model

from (3.3) and using the decompositions $\Phi = \mu_\Phi + \Delta\Phi$ and $\theta = \mu_\theta + \Delta\theta$, where $\Delta\theta \sim \mathbb{P}(0, \Sigma_\theta)$, results in the following optimization problem:

$$\begin{aligned} \min_{\Psi, \psi} & \mathbb{E} \left[(\Delta\theta - \Psi(\mu_\Phi + \Delta\Phi)^\top \Delta\theta - \Psi v - \Psi \Delta\Phi^\top \mu_\theta)^\top (\Delta\theta - \Psi(\mu_\Phi + \Delta\Phi)^\top \Delta\theta - \Psi v - \Psi \Delta\Phi^\top \mu_\theta) \right] \\ & - 2\mathbb{E} \left[(\Delta\theta - \Psi(\mu_\Phi + \Delta\Phi)^\top \Delta\theta - \Psi v - \Psi \Delta\Phi^\top \mu_\theta)^\top (\psi - \mu_\theta + \Psi \mu_\Phi^\top \mu_\theta) \right] \\ & + \mathbb{E} \left[(\psi - \mu_\theta + \Psi \mu_\Phi^\top \mu_\theta)^\top (\psi - \mu_\theta + \Psi \mu_\Phi^\top \mu_\theta) \right]. \end{aligned} \quad (3.10)$$

The second term in the above optimization is zero because all random variables are zero-mean, i.e., $\mathbb{E}[\Delta\theta] = 0$, $\mathbb{E}[\Delta\Phi] = 0$, $\mathbb{E}[v] = 0$, $\mathbb{E}[\Delta\theta^\top \Delta\Phi] = 0$ from the stochastic independence of θ and w_t at all times, and $\psi - \mu_\theta + \Psi \mu_\Phi^\top \mu_\theta$ is a constant. In addition, the last term in (3.10) is zero if $\psi = \mu_\theta - \Psi \mu_\Phi^\top \mu_\theta$. Therefore, minimizing (3.10) over the variable Ψ yields

$$\psi^* = \mu_\theta - \Psi^* \mu_\Phi^\top \mu_\theta.$$

Expanding the first term of (3.10), applying the trace operator, noting that all the random variables are independent, and after some algebraic manipulation, the problem (3.10) reduces to minimizing the following cost function over Ψ :

$$\mathcal{J}(\Psi) = \text{tr}(\Sigma_\theta - 2\Psi^\top \Sigma_\theta \mu_\Phi + \Psi^\top \Psi (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + H + V + \Sigma_v)),$$

where $H = \mathbb{E}[\Delta\Phi^\top \Delta\theta \Delta\theta^\top \Delta\Phi]$ and $V = \mathbb{E}[\Delta\Phi^\top \mu_\theta \mu_\theta^\top \Delta\Phi]$. The minimum of $\mathcal{J}(\Psi)$ is obtained from setting its partial derivative to zero, hence,

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \Psi} \Big|_{\Psi=\Psi^*} &= -2\Sigma_\theta \mu_\Phi + 2\Psi^* (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + H + V + \Sigma_v) = 0 \\ \iff \Psi^* &= \Sigma_\theta \mu_\Phi (\mu_\Phi^\top \Sigma_\theta \mu_\Phi + H + V + \Sigma_v)^{-1}. \end{aligned}$$

Since $\Sigma_v = \text{diag}(\sigma_{v_0}^2, \dots, \sigma_{v_T}^2) \succ 0$, we have $\mu_\Phi^\top \Sigma_\theta \mu_\Phi + H + V + \Sigma_v \succ 0$. As such, this matrix is invertible and Ψ^* has a unique solution. It remains to find the elements of matrices H and V . The $(t+1, t'+1)$ th elements of matrices H and V is calculated based on the $(t+1)$ th and $(t'+1)$ th columns of matrix $\Delta\Phi$ defined as $\Delta\phi(x_t)$ and $\Delta\phi(x_{t'})$, respectively. Applying the trace operator to each of their elements,

$$\begin{aligned} H_{tt'} &= \mathbb{E}[\Delta\phi(x_t)^\top \Delta\theta \Delta\theta^\top \Delta\phi(x_{t'})] = \mathbb{E}[\text{tr}(\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top \Delta\theta \Delta\theta^\top)], \\ V_{tt'} &= \mathbb{E}[\Delta\phi(x_t)^\top \mu_\theta \mu_\theta^\top \Delta\phi(x_{t'})] = \mathbb{E}[\text{tr}(\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top \mu_\theta \mu_\theta^\top)]. \end{aligned}$$

Noting that $\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top$ and $\Delta\theta$ are independent, one could observe that

$$\begin{aligned} H_{tt'} &= \text{tr}(\mathbb{E}[\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top] \Sigma_\theta), \\ V_{tt'} &= \text{tr}(\mathbb{E}[\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top] \mu_\theta \mu_\theta^\top). \end{aligned}$$

Finally, rewriting $\Delta\phi(x_t) = \phi(x_t) - \mathbb{E}[\phi(x_t)]$, it is straightforward to obtain

$$\mathbb{E}[\Delta\phi(x_{t'}) \Delta\phi(x_t)^\top] = \mathbb{E}[\phi(x_{t'}) \phi(x_t)^\top] - \mathbb{E}[\phi(x_{t'})] \mathbb{E}[\phi(x_t)]^\top,$$

hence, we have

$$\begin{aligned} H_{tt'} &= \text{tr}((\mathbb{E}[\phi(x_{t'})\phi^\top(x_t)] - \mathbb{E}[\phi(x_{t'})]\mathbb{E}[\phi(x_t)]^\top)\Sigma_\theta), \\ V_{tt'} &= \text{tr}((\mathbb{E}[\phi(x_{t'})\phi^\top(x_t)] - \mathbb{E}[\phi(x_{t'})]\mathbb{E}[\phi(x_t)]^\top)\mu_\theta\mu_\theta^\top). \end{aligned}$$

Introducing matrix $M := H + V$, its $(t + 1, t' + 1)^{\text{th}}$ element is

$$M_{tt'} = H_{tt'} + V_{tt'} = \text{tr}(\Sigma_\Phi^{t't}(\Sigma_\theta + \mu_\theta\mu_\theta^\top)),$$

where $\Sigma_\Phi^{t't} = \mathbb{E}[\phi(x_{t'})\phi^\top(x_t)] - \mathbb{E}[\phi(x_{t'})]\mathbb{E}[\phi(x_t)]^\top$ according to the DBS definition 3.1. Since

$$\text{tr}(\Sigma_\Phi^{t't}(\Sigma_\theta + \mu_\theta\mu_\theta^\top)) = \text{tr}(\Sigma_\Phi^{t't}(\Sigma_\theta + \mu_\theta\mu_\theta^\top)),$$

we write $M_{tt'} = \text{tr}(\Sigma_\Phi^{t't}(\Sigma_\theta + \mu_\theta\mu_\theta^\top))$. Ultimately, substituting $\Psi^* = \Sigma_\theta\mu_\Phi(\mu_\Phi^\top\Sigma_\theta\mu_\Phi + M + \Sigma_v)^{-1}$ and $\psi^* = \mu_\theta - \Psi^*\mu_\Phi^\top\mu_\theta$ in (3.10) results in the last two terms to be zero and the optimal MMSE error $\mathcal{J}_B^*$ after a simple algebraic manipulation arrives at (3.7b), which concludes the proof. $\square$

Proof of Proposition 3.3.3. Showing that $\mathbb{E}[\theta - \hat{\theta}_B(y)] = 0$ indicates that the estimator is Bayesian unbiased. We substitute $\hat{\theta}_B(y)$ with its optimal estimator (3.7a), which leads to

$$\mathbb{E}[\theta - \hat{\theta}_B(y)] = \mathbb{E}[\theta] - \mathbb{E}[\Psi^*y + \psi^*] = \mu_\theta - \Psi^*\mathbb{E}[y] - \mu_\theta + \Psi^*\mu_\Phi^\top\mu_\theta = \Psi^*(\mu_\Phi^\top\mu_\theta - \mathbb{E}[y]).$$

Noting that $\mathbb{E}[y] = \mathbb{E}[\Phi]\mu_\theta$ from the stochastic independence of θ and w_t at all times, and that $\mathbb{E}[\Phi] = \mu_\Phi$, it follows that $\mathbb{E}[\theta - \hat{\theta}_B(y)] = 0$. $\square$

Proof of Corollary 3.3.5. Let us define $R := M + \Sigma_v$, then the optimal MMSE in (3.7b) is

$$\mathcal{J}_B^* = \text{tr}(\Sigma_\theta - \Sigma_\theta\mu_\Phi(\mu_\Phi^\top\Sigma_\theta\mu_\Phi + R)^{-1}\mu_\Phi^\top\Sigma_\theta).$$

Applying the matrix inversion lemma [15], we derive the equivalent expression

$$\Sigma_\theta - \Sigma_\theta\mu_\Phi(\mu_\Phi^\top\Sigma_\theta\mu_\Phi + R)^{-1}\mu_\Phi^\top\Sigma_\theta = (\Sigma_\theta^{-1} + \mu_\Phi R^{-1}\mu_\Phi^\top)^{-1}, \quad (3.11)$$

which allows us to rewrite (3.7b) as $\mathcal{J}_B^* = \text{tr}((\Sigma_\theta^{-1} + \mu_\Phi R^{-1}\mu_\Phi^\top)^{-1})$. Let $\mathcal{J}_B^*(t + 1)$ denote the estimation error at time $(t + 1)$ using measurements $y = [y_0, \dots, y_{(t+1)}]^\top$. Define matrices

$$\mu_\Phi(t + 1) = [\mu_\Phi(t), \mu_\phi^{(t+1)}], \quad R(t + 1) = \begin{bmatrix} R(t) & c(t + 1) \\ c^\top(t + 1) & r(t + 1) \end{bmatrix},$$

with $\mu_\Phi(t) = [\mu_\phi^0, \dots, \mu_\phi^t]$, $c(t + 1) = [M_{0(t+1)}, \dots, M_{t(t+1)}]^\top$, and $r(t + 1) = M_{(t+1)(t+1)} + \sigma_{v(t+1)}^2$. Using the *Schur complement* [15],

$$\begin{aligned} R^{-1}(t + 1) &= \begin{bmatrix} \mathbb{I} & -R^{-1}(t)c(t + 1) \\ 0 & \mathbb{I} \end{bmatrix} \begin{bmatrix} R^{-1}(t) & 0 \\ 0 & (r(t + 1) - c^\top(t + 1)R^{-1}(t)c(t + 1))^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \mathbb{I} & 0 \\ -c^\top(t + 1)R^{-1}(t) & \mathbb{I} \end{bmatrix}. \quad (3.12) \end{aligned}$$

We define $S(t+1) := \mu_\Phi(t+1)R^{-1}(t+1)\mu_\Phi^\top(t+1)$ and $S(t) := \mu_\Phi(t)R^{-1}(t)\mu_\Phi^\top(t)$, which decomposes as $\Delta S(t+1) := S(t+1) - S(t)$, where

$$\begin{cases} \Delta S(t+1) = \gamma(t+1)s(t+1)s^\top(t+1), \\ s(t+1) = \mu_\Phi(t)R^{-1}(t)c(t+1) - \mu_\Phi^{(t+1)}, \\ \gamma(t+1) = r(t+1) - c^\top(t+1)R^{-1}(t)c(t+1). \end{cases}$$

Given $R(t+1) = M(t+1) + \Sigma_v(t+1) \succ 0$, one can observe from (3.12) that

$$(r(t+1) - c^\top(t+1)R^{-1}(t)c(t+1))^{-1} > 0,$$

implying that $\Delta S(t+1) \succeq 0$. Since, $\Sigma_\theta^{-1} \succeq 0$ and $S(t) \succeq 0$, we obtain

$$(\Sigma_\theta^{-1} + S(t) + \Delta S(t+1))^{-1} \preceq (\Sigma_\theta^{-1} + S(t))^{-1}.$$

Thus, the estimation error decreases monotonically, i.e.,

$$\mathcal{J}_B^*(t+1) = \text{tr}((\Sigma_\theta^{-1} + S(t) + \Delta S(t+1))^{-1}) \leq \text{tr}((\Sigma_\theta^{-1} + S(t))^{-1}) = \mathcal{J}_B^*(t).$$

□

3.4 FOURIER BASIS

The question that arises concerns the extent to which the matrices μ_Φ and M of Theorem 3.3.2 in (3.7) depend on the respective DBS $(\mu_\Phi^t, \Sigma_\Phi^{tt'})$ introduced in Definition 3.1. By employing Fourier basis functions, one can leverage their unique properties to efficiently compute these expectations via their characteristic function. Furthermore, any sufficiently well-behaved function can be approximated as a sum of Fourier series [123], with the discrete Fourier transform (DFT) providing an efficient method for calculating the coefficients of these series. The Fourier basis function is defined as

$$\begin{cases} \phi_0(\mathbf{x}) = 1 & n = 0 \\ \phi_n(\mathbf{x}) = \sum_{\ell \in \{-1,1\}} \exp(j\ell f_n \mathbf{x}) & n \geq 1, \end{cases} \quad (3.13)$$

where $f_n \in \mathbb{R}^{n_x}$ represents a *known* frequency, and n denotes the frequency index. To ensure that the codomain of $h_\theta(\cdot)$ in (3.2) remains in $\mathbb{R}$, symmetry is imposed on the frequencies and their corresponding parameters to eliminate imaginary components. Specifically, the basis function is constructed such that identical parameters are assigned to terms with positive and negative frequencies, f_n and $-f_n$. Additionally, $\phi_0(\mathbf{x}) = 1$ corresponds to a Fourier basis with $f_0 = 0$. Using the Fourier basis functions defined in (3.13), we derive explicit expressions for the elements of the mean vectors μ_Φ^t and covariance matrices $\Sigma_\Phi^{tt'}$, which are required for the analytical formulation of the affine MMSE estimator outlined in Theorem 3.3.2. These expressions are presented in a compact form, which relies on the *lifted matrix* representation introduced in the following section.

3.4.1 LIFTED PROCESS MODEL

We represent the process model in (3.1) for the entire trajectory in the *lifted matrix* form

$$\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{u} + \mathbf{w}), \quad (3.14)$$

where $\mathbf{x} = [\mathbf{x}_0^\top, \dots, \mathbf{x}_T^\top]^\top$ consists of the system states vector, $\mathbf{u} = [\mu_{x_0}^\top, \mathbf{u}_0^\top, \dots, \mathbf{u}_{T-1}^\top]^\top$ is the input vector with the initial state mean as its first element, and $\mathbf{w} = [\mathbf{w}_0^\top, \mathbf{w}_1^\top, \dots, \mathbf{w}_T^\top]^\top$ denotes the noise vector in which the first element corresponds to the uncertainty of the initial state. As such, $\mathbf{w}_0 \sim \mathcal{P}(0, \Sigma_{x_0})$ and $\mathbf{w}$ is a *zero-mean* uncertainty, i.e., $\mathbf{w} \sim \mathcal{P}(0, \Sigma_w)$, $\Sigma_w \geq 0$. The covariance matrix is diagonal only if $\mathbf{w}_t$ are independent, i.e., $\Sigma_w = \text{diag}(\Sigma_{x_0}, \Sigma_{w_1}, \dots, \Sigma_{w_T})$. Furthermore, the matrices $\mathbf{A}$ and $\mathbf{B}$ have the following lower triangular and block-diagonal structures, respectively:

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & 0 & 0 & \dots & 0 \\ \mathbf{A}_0 & \mathbf{I} & 0 & \dots & \vdots \\ \mathbf{A}_1\mathbf{A}_0 & \mathbf{A}_1 & \mathbf{I} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \prod_{i=0}^{T-1} \mathbf{A}_i & \prod_{i=1}^{T-1} \mathbf{A}_i & \dots & \mathbf{A}_{T-1} & \mathbf{I} \end{bmatrix}, \quad \mathbf{B} = \text{diag}(\mathbf{I}, \mathbf{B}_0, \dots, \mathbf{B}_{T-1}), \quad (3.15)$$

where $\prod_{i=t}^{t'} \mathbf{A}_i = \mathbf{A}_{t'} \dots \mathbf{A}_t$. The lifted matrix representation in (3.14) provides a compact expression for the mean and covariance of the random variable $\mathbf{x}_t$ as a function of $\mathbf{u}$ and Σ_w , given by $\mathbf{x}_t = \mathbf{A}_t(\mathbf{B}\mathbf{u} + \mathbf{w})$ and $\mathbf{x}_t \sim \mathcal{P}(\mathbf{A}_t\mathbf{B}\mathbf{u}, \mathbf{A}_t\Sigma_w\mathbf{A}_t^\top)$, where $\mathbf{A}_t$ corresponds to the $(t+1)^{\text{th}}$ block-row of $\mathbf{A}$ in (3.15), i.e.,

$$\mathbf{A}_t = \begin{bmatrix} \prod_{i=0}^{t-1} \mathbf{A}_i & \prod_{i=1}^{t-1} \mathbf{A}_i & \dots & \mathbf{I} & 0 & \dots & 0 \end{bmatrix}. \quad (3.16)$$

This representation serves as the foundation for deriving the DBS of Fourier bases presented in the next subsection.

3.4.2 FOURIER BASIS STATISTICS

To compute the expectation of a Fourier basis, the concept of the characteristic function is directly relevant, as it provides a mechanism for deriving such expectations.

Definition 3.2 (Characteristic function). *The characteristic function of a random vector is defined as the sign-reversed Fourier transform of its probability density function [56], i.e., for a random vector $\mathbf{v} \sim \mathcal{P}(0, \mathbf{I})$, the characteristic function at frequency $\mathbf{f}$ is given by $\varphi(\mathbf{f}) := \mathbb{E}[\exp(j\langle \mathbf{f}, \mathbf{v} \rangle)]$.*

Using characteristic functions, the following lemma derives explicit expressions for the elements of μ_ϕ^t and $\Sigma_\phi^{tt'}$, which are essential for the optimal Bayesian affine MMSE estimator presented in Theorem 3.3.2.

Lemma 3.4.1 (Fourier DBS explicit expression). For Fourier basis functions (3.13) and the dynamics of (3.1), let the respective DBS introduced in Definition 3.1, denoted by $(\mu_\phi^t, \Sigma_\Phi^{tt'})$. Additionally, let A_t and $A_{t'}$ represent the matrices defined analogously to (3.16). Then, the following expressions hold:

(i) **DBS mean:** The first element of the mean vector is $\mu_{\phi_0}^t = 1$, and its $(n+1)^{\text{th}}$ element is

$$\mu_{\phi_n}^t = \sum_{\ell \in \{-1, 1\}} \exp(j \langle A_t^\top \ell f_n, \text{Bu} \rangle) \varphi(\Sigma_w^{\frac{1}{2}} A_t^\top \ell f_n), \quad n \geq 1. \quad (3.17a)$$

(ii) **DBS covariance:** The $(m+1, n+1)^{\text{th}}$ element of the covariance matrix is $\Sigma_{\phi_{mn}}^{tt'} = 0$ if $m = 0$ or $n = 0$. For $m, n \geq 1$, this element is

$$\begin{aligned} \Sigma_{\phi_{mn}}^{tt'} = \sum_{\ell, \ell' \in \{-1, 1\}} \exp(j \langle A_t^\top \ell f_m + A_{t'}^\top \ell' f_n, \text{Bu} \rangle) \\ \times \left(\varphi\left(\Sigma_w^{\frac{1}{2}} (A_t^\top \ell f_m + A_{t'}^\top \ell' f_n)\right) - \varphi(\Sigma_w^{\frac{1}{2}} A_t^\top \ell f_m) \varphi(\Sigma_w^{\frac{1}{2}} A_{t'}^\top \ell' f_n) \right). \end{aligned} \quad (3.17b)$$

The technical proofs of the theoretical statements presented in this section are deferred to Subsection 3.4.4. Using the explicit expressions provided in Lemma 3.4.1, one can analytically compute μ_ϕ^t and $\Sigma_\Phi^{tt'}$ when the process noise is drawn from known distributions.

Example 3.1 (Gaussian explicit characteristic). When the process noise is Gaussian, $w \sim \mathcal{N}(0, \Sigma_w)$, the system states x_t follow a Gaussian distribution as $x_t \sim \mathcal{N}(A_t \text{Bu}, A_t \Sigma_w A_t^\top)$, with A_t defined in (3.16). In this case, the explicit expressions for (3.17) are obtained from the analytical characteristic function of a multivariate normal distribution [47, Ch. 10],

$$\begin{aligned} \varphi(\Sigma_w^{\frac{1}{2}} A_t^\top \ell f_n) &= \exp\left(-\frac{1}{2} \ell f_n^\top A_t \Sigma_w A_t^\top \ell f_n\right), \\ \varphi(\Sigma_w^{\frac{1}{2}} (A_t^\top \ell f_m + A_{t'}^\top \ell' f_n)) &= \exp\left(-\frac{1}{2} (\ell f_m^\top A_t + \ell' f_n^\top A_{t'}) \Sigma_w (A_t^\top \ell f_m + A_{t'}^\top \ell' f_n)\right). \end{aligned} \quad (3.18)$$

From the analytical expressions of the optimal estimator in Theorem 3.3.2, under the assumption of Gaussian process noise, we examine the consistency property of the affine estimator in the following subsection.

3.4.3 BAYESIAN AFFINE MMSE ESTIMATOR CONSISTENCY

As noted in previous sections, the proposed Bayesian affine estimator (3.7) requires the computation of the inverse of the observation covariance matrix, which depends on the matrices μ_Φ and M . These matrices, and consequently the optimal MSE (3.7b), are significantly influenced by the underlying dynamical system. The following proposition identifies the conditions under which the estimation error fails to converge to zero as the number of dependent samples from a trajectory length T approaches infinity, highlighting cases where the estimator is inconsistent.

Proposition 3.4.2 (Inherent inconsistency). Consider a stochastically unstable linear time-invariant (LTI) system (3.1), where $A_t = A$ and $\lambda_{\max}(A) \geq 1$. Let the process noise be Gaussian, $w \sim \mathcal{N}(0, \Sigma_w)$, and have an observation model (3.2) with $N \geq 1$ spanned by the Fourier basis (3.13). Then, the optimal estimation error (3.7b) is uniformly away from zero, i.e., $\lim_{t \rightarrow \infty} \mathcal{J}_B^*(t) > 0$, for any unbounded and persistently excited input trajectory $\{u_t\}_{t \in \mathbb{Z}_+}$.

In cases where the underlying dynamical system is marginally stable or unstable, multiple *statistically independent* trajectories of data help reduce the estimation error and ensure that the estimator remains consistent, provided trajectories are persistently excited and the number of independent trajectories approaches infinity. We extend our formulation to accommodate τ multiple *statistically independent* trajectories of data. For each trajectory i , the time index t independently starts from 0 and continues for a horizon T_i , resulting in independent trajectories $\{0, \dots, T_1\}, \dots, \{0, \dots, T_\tau\}$. In this extended formulation, the *lifted matrix* forms of A and B in the process model (3.14) change to block-diagonal matrices, i.e., $A = \text{diag}(A_1, \dots, A_\tau)$ and $B = \text{diag}(B_1, \dots, B_\tau)$, where A_i and B_i are defined as in (3.15) for each trajectory i . All other aspects of the formulation remain identical to the single-trajectory case, with their sizes extended according to the total number of data points $T^\tau = (T_1 + 1) + \dots + (T_\tau + 1)$. The following proposition formally specifies the condition to provide consistency for the proposed optimal Bayesian affine MMSE estimator.

Proposition 3.4.3 (Consistency via independent trajectories). *Let $\mathcal{J}_B^*(\tau, T^\tau)$ denote the optimal estimation error (3.7b) using τ statistically independent trajectories (or “batches”) with the total length of $T^\tau = (T_1 + 1) + \dots + (T_\tau + 1)$ data points, where each trajectory i contains $(T_i + 1)$ measurements, i.e., $y_i = [y_0, \dots, y_{T_i}]^\top$. Then, for any prior distribution and any set of basis functions $\phi(x)$ in (3.2), the optimal Bayesian MSE (3.7b) converges to zero as τ tends to ∞ if the minimum eigenvalue of the so-called “information matrix” of the basis functions diverges to infinity, i.e.,*

$$\lim_{\tau \rightarrow \infty} \lambda_{\min} \left(\sum_{i=1}^{\tau} \frac{\mu_\phi^0(i) \mu_\phi^0(i)}{M_{00}(i) + \sigma_{v_0}^2(i)} \right) = \infty \implies \lim_{\tau \rightarrow \infty} \mathcal{J}_B^*(\tau, T^\tau) = 0, \quad (3.19)$$

where $M_{00}(i) = \text{tr}(\Sigma_\Phi^{00}(i)(\Sigma_\theta + \mu_\theta \mu_\theta^\top))$, $\sigma_{v_0}^2(i)$ is the measurement noise variance, and $\mu_\phi^0(i)$ and $\Sigma_\Phi^{00}(i)$ are the respective DBS as described in Definition 3.1,

$$\mu_\phi^0(i) = \mathbb{E}[\phi(x_0(i))], \quad \Sigma_\Phi^{00}(i) = \mathbb{E}[\phi(x_0(i))\phi(x_0(i))^\top] - \mathbb{E}[\phi(x_0(i))]\mathbb{E}[\phi(x_0(i))]^\top,$$

all evaluated at the initial state of the i^{th} trajectory.

Condition (3.19) shows that estimates of θ from the optimal Bayesian affine MMSE estimator (3.7) converge to their true values if the initial states in *statistically independent* trajectories are indeed persistently excited. However, the rate at which the estimation error decays also depends on the persistent excitation of dependent data within individual trajectories. Persistent excitation can be achieved through the strategic selection of inputs u by actively minimizing the optimal Bayesian MSE (3.7b). In the subsequent section, we address this input optimization framework, commonly referred to as active learning.

3.4.4 TECHNICAL PROOFS

We provide detailed proofs supporting the theoretical statements of this section.

Proof of Lemma 3.4.1. The random vector x_t follows the distribution $x_t \sim \mathcal{P}(A_t B u, A_t \Sigma_w A_t^\top)$, where A_t is defined in (3.16) and the covariance matrix $A_t \Sigma_w A_t^\top$ is symmetric and positive

definite. Therefore, $\mathbf{x}_t$ can be expressed as an affine transformation, $\mathbf{x}_t = \mathbf{A}_t \mathbf{B} \mathbf{u} + \mathbf{A}_t \Sigma_w^{\frac{1}{2}} \mathbf{v}$, of the standard random vector $\mathbf{v} \sim \mathcal{P}(0, \mathbf{I})$ under the condition that this mapping transforms the distribution of $\mathbf{v}$ to that of $\mathbf{w}$. Consequently, the characteristic function of the random vector $\mathbf{x}_t$ from that of $\mathbf{v}$, according to Definition 3.2, is

$$\mathbb{E}[\exp(j\langle f, \mathbf{x}_t \rangle)] = \exp(j\langle f, \mathbf{A}_t \mathbf{B} \mathbf{u} \rangle) \varphi(\Sigma_w^{\frac{1}{2}} \mathbf{A}_t^\top f). \quad (3.20)$$

Using the Fourier basis (3.13), the $(n+1)^{\text{th}}$ element of the mean vector μ_ϕ^t is given by

$$\mathbb{E}[\phi_n(\mathbf{x}_t)] = \begin{cases} 1 & n = 0 \\ \sum_{\ell \in \{-1, 1\}} \exp(j\langle \mathbf{A}_t^\top \ell f_n, \mathbf{B} \mathbf{u} \rangle) \varphi(\Sigma_w^{\frac{1}{2}} \mathbf{A}_t^\top \ell f_n) & n \geq 1. \end{cases}$$

In addition, the $(m+1, n+1)^{\text{th}}$ element of the covariance matrix $\Sigma_\phi^{tt'}$ is expressed as

$$\Sigma_{\phi_{mn}}^{tt'} = \mathbb{E}[\phi_m(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'})] - \mathbb{E}[\phi_m(\mathbf{x}_t)] \mathbb{E}[\phi_n(\mathbf{x}_{t'})],$$

which can also be derived from (3.13) and (3.20). When at least one index is zero ($m = 0$ or $n = 0$), the element simplifies to

$$\begin{aligned} \Sigma_{\phi_{00}}^{tt'} &= \mathbb{E}[\phi_0(\mathbf{x}_t) \phi_0(\mathbf{x}_{t'})] - \mathbb{E}[\phi_0(\mathbf{x}_t)] \mathbb{E}[\phi_0(\mathbf{x}_{t'})] = 0, \\ \Sigma_{\phi_{m0}}^{tt'} &= \mathbb{E}[\phi_m(\mathbf{x}_t) \phi_0(\mathbf{x}_{t'})] - \mathbb{E}[\phi_m(\mathbf{x}_t)] \mathbb{E}[\phi_0(\mathbf{x}_{t'})] = 0, \\ \Sigma_{\phi_{0n}}^{tt'} &= \mathbb{E}[\phi_0(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'})] - \mathbb{E}[\phi_0(\mathbf{x}_t)] \mathbb{E}[\phi_n(\mathbf{x}_{t'})] = 0. \end{aligned}$$

Finally, for the case where both $m \geq 1$ and $n \geq 1$, the term $\phi_m(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'})$ is a summation of the following four terms:

$$\phi_m(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'}) = \sum_{\ell, \ell' \in \{-1, 1\}} \exp(j\langle \ell f_m, \mathbf{x}_t \rangle) \exp(j\langle \ell' f_n, \mathbf{x}_{t'} \rangle).$$

Given $\mathbf{x}_t = \mathbf{A}_t \mathbf{B} \mathbf{u} + \mathbf{A}_t \Sigma_w^{\frac{1}{2}} \mathbf{v}$ and $\mathbf{x}_{t'} = \mathbf{A}_{t'} \mathbf{B} \mathbf{u} + \mathbf{A}_{t'} \Sigma_w^{\frac{1}{2}} \mathbf{v}$, $\phi_m(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'})$ can be simplified to

$$\phi_m(\mathbf{x}_t) \phi_n(\mathbf{x}_{t'}) = \sum_{\ell, \ell' \in \{-1, 1\}} \exp(j\langle \mathbf{A}_t^\top \ell f_m + \mathbf{A}_{t'}^\top \ell' f_n, \mathbf{B} \mathbf{u} \rangle) \exp(j\langle \Sigma_w^{\frac{1}{2}} (\mathbf{A}_t^\top \ell f_m + \mathbf{A}_{t'}^\top \ell' f_n), \mathbf{v} \rangle).$$

Thus, when $m, n \geq 1$, the $(m+1, n+1)^{\text{th}}$ element of the covariance matrix $\Sigma_\phi^{tt'}$ is

$$\Sigma_{\phi_{mn}}^{tt'} = -\mu_{\phi_m}^t \mu_{\phi_n}^{t'} + \sum_{\ell, \ell' \in \{-1, 1\}} \exp(j\langle \mathbf{A}_t^\top \ell f_m + \mathbf{A}_{t'}^\top \ell' f_n, \mathbf{B} \mathbf{u} \rangle) \varphi(\Sigma_w^{\frac{1}{2}} (\mathbf{A}_t^\top \ell f_m + \mathbf{A}_{t'}^\top \ell' f_n)),$$

which can be rewritten as (3.17b), completing the proof. $\square$

Proof of Proposition 3.4.2. Let $\mathcal{J}_B^*(t)$ be the optimal estimation error (3.7b) at time t given an output measurement trajectory with length $t+1$, and consider its equivalent representation in (3.11). From Corollary 3.3.5, we derive

$$\mathcal{J}_B^*(t) = \text{tr}((\bar{\Sigma}_\theta^{-1} + \mathbf{S}(t-1) + \Delta \mathbf{S}(t))^{-1}), \quad (3.21)$$

where $S(t-1) = \mu_\Phi(t-1)R^{-1}(t-1)\mu_\Phi^\top(t-1)$, $\Delta S(t) = \gamma^{-1}(t)s(t)s^\top(t)$,

$$\begin{cases} s(t) = \mu_\Phi(t-1)R^{-1}(t-1)c(t) - \mu_\phi^t, \\ \gamma(t) = r(t) - c^\top(t)R^{-1}(t-1)c(t) > 0, \\ r(t) = M_{tt} + \sigma_{v_t}^2, \\ c(t) = [M_{0t}, \dots, M_{(t-1)t}]^\top. \end{cases} \quad (3.22)$$

Recursively applying the decomposition (3.21) from $t-1$ to 1, we obtain

$$\mathcal{J}_B^*(t) = \text{tr}((\bar{\Sigma}_\theta^{-1} + S(0) + \Delta S(1) + \dots + \Delta S(t))^{-1}) = \text{tr}((\bar{\Sigma}_\theta^{-1} + S(0) + \sum_{i=1}^t \frac{1}{\gamma(i)} s(i)s^\top(i))^{-1}).$$

Since $\bar{\Sigma}_\theta^{-1} \geq 0$ and $S(0) \geq 0$, $\mathcal{J}_B^*(t)$ tends to 0 as t goes to ∞ , if and only if the smallest eigenvalue of the information matrix diverges to infinity, i.e.,

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathcal{J}_B^*(t) = 0 &\iff \lim_{t \rightarrow \infty} \lambda_{\max}((\bar{\Sigma}_\theta^{-1} + S(0) + \sum_{i=1}^t \frac{1}{\gamma(i)} s(i)s^\top(i))^{-1}) = 0, \\ &\iff \lim_{t \rightarrow \infty} \lambda_{\min}(\bar{\Sigma}_\theta^{-1} + S(0) + \sum_{i=1}^t \frac{1}{\gamma(i)} s(i)s^\top(i)) = \infty. \end{aligned}$$

Equivalently, the eigenvalues diverge if and only if for every unit vector $\hat{v} \in \mathbb{R}^{(N+1)}$, i.e., $\|\hat{v}\| = 1$, the above series diverges to infinity, i.e.,

$$\lim_{t \rightarrow \infty} \lambda_{\min}(\bar{\Sigma}_\theta^{-1} + S(0) + \sum_{i=1}^t \frac{1}{\gamma(i)} s(i)s^\top(i)) = \infty \iff \forall \hat{v} \in \mathbb{R}^{(N+1)}, \|\hat{v}\| = 1, \lim_{t \rightarrow \infty} \sum_{i=1}^t \frac{1}{\gamma(i)} (\hat{v}^\top s(i))^2 = \infty.$$

Hence, the necessary and sufficient condition for the convergence of the estimator to the true value is

$$\lim_{t \rightarrow \infty} \mathcal{J}_B^*(t) = 0 \iff \lim_{t \rightarrow \infty} \inf_{\|v\|=1} \left(\sum_{i=1}^t \frac{1}{\gamma(i)} (v^\top s(i))^2 \right) = \infty. \quad (3.23)$$

For the sake of contradiction, let us choose the vector $\mathbf{1}_{\{n\}}$ as the n^{th} standard basis vector,

$$\mathbf{1}_{\{n\}} = [0, \dots, 0, \underset{\substack{\uparrow \\ n^{\text{th}}}}{1}, 0, \dots, 0]^\top, \quad n \geq 1.$$

Substituting this unit vector into the convergence condition (3.23), we have

$$\frac{1}{\gamma(i)} (\mathbf{1}_{\{n\}}^\top s(i))^2 = \frac{1}{\gamma(i)} (\mathbf{1}_{\{n\}}^\top \mu_\Phi(i-1)R^{-1}(i-1)c(i) - \mu_{\phi_n}^i)^2 = \frac{1}{\gamma(i)} \left(\sum_{j=0}^{i-1} \frac{1}{\gamma(j)} \mu_{\phi_n}^j M_{ji} - \mu_{\phi_n}^i \right)^2,$$

and

$$\lim_{t \rightarrow \infty} \sum_{i=1}^t \frac{1}{\gamma(i)} (\mathbf{1}_{\{n\}}^\top s(i))^2 = \sum_{i=1}^{\infty} \frac{1}{\gamma(i)} \left(\sum_{j=0}^{i-1} \frac{1}{\gamma(j)} \mu_{\phi_n}^j M_{ji} - \mu_{\phi_n}^i \right)^2, \quad (3.24)$$

where $\gamma(0) = M_{00} + \sigma_{v_0}^2$. From the structure of $\mu_{\phi_n}^i$ and $\Sigma_{\phi_{nm}}^{ji}$ for the Fourier basis with Gaussian process noise in (3.18), i.e.,

$$\left\{ \begin{array}{ll} \mu_{\phi_n}^i = \chi_n^i \exp\left(-\frac{1}{2} f_n^\top A_i \Sigma_w A_i^\top f_n\right), & -2 \leq \chi_n^i \leq 2, \\ \Sigma_{\phi_{mn}}^{ji} = \xi_{mn}^{ji} \exp\left(-\frac{1}{2} f_n^\top A_i \Sigma_w A_i^\top f_n\right), & -2 \leq \xi_{mn}^{ji} \leq 2, \\ \chi_n^i = \sum_{\ell \in \{-1,1\}} \exp(j \langle A_i^\top \ell f_n, \text{Bu} \rangle), \\ \xi_{mn}^{ji} = \sum_{\ell, \ell' \in \{-1,1\}} \exp(j \langle A_j^\top \ell f_m + A_i^\top \ell' f_n, \text{Bu} \rangle) (\exp(-\ell f_m^\top A_j \Sigma_w A_j^\top \ell' f_n) - 1) \\ \quad \times \exp\left(-\frac{1}{2} f_m^\top A_j \Sigma_w A_j^\top f_m\right), \end{array} \right.$$

3

we have the following properties:

- (i) Bounded coefficients: $\mu_{\phi_n}^i < \infty$, $\mu_{\phi_n}^j < \infty$, $\Sigma_{\phi}^{ii} < \infty$ for all values of i and j . Also, from (3.22),

$$M_{ji} = \text{tr}(\Sigma_{\phi}^{ji}(\Sigma_{\theta} + \mu_{\theta} \mu_{\theta}^\top)) < \infty, \quad 0 < \gamma(i) \leq M_{ii} + \sigma_{v_i}^2 < \infty, \quad 0 < \gamma(j) \leq M_{jj} + \sigma_{v_j}^2 < \infty.$$

- (ii) Impact of stochastic instability: For the special case of marginally stable or unstable LTI system, i.e., $\lambda_{\max}(A) \geq 1$,

$$\tilde{v}^\top (A_i \Sigma_w A_i^\top) \tilde{v} = \sum_{k=0}^i \tilde{v}^\top A^k \Sigma_{w(i-k)} A^{k\top} \tilde{v} = \sum_{k=0}^i |\lambda_q(A)|^{2k} \tilde{v}^\top \Sigma_{w(i-k)} \tilde{v},$$

where $\tilde{v}$ is the eigenvector of matrix A corresponding to $\lambda_{\max}(A)$ and $\Sigma_{w_0} = \Sigma_{x_0}$. It is evident from the equality that $A_i \Sigma_w A_i^\top$ grows unbounded and $\exp(-\frac{1}{2} f_n^\top A_i \Sigma_w A_i^\top f_n)$ decays to zero.

- (iii) Decay of terms: The elements $\mu_{\phi_n}^j$, M_{ji} , and $\mu_{\phi_n}^i$ in (3.24) decay exponentially to zero due to $\exp(-\frac{1}{2} f_n^\top A_i \Sigma_w A_i^\top f_n)$. From (3.22), one can also observe that $c(i)$ decays exponentially to zero, while $r(i)$ exponentially converges to $\sigma_{v_i}^2$. Consequently, $\gamma(i)$ and $\gamma(j)$ also exponentially converge to $\sigma_{v_i}^2$ and $\sigma_{v_j}^2$, respectively.

Thus, the numerator of the series (3.24) decays exponentially as $\mathcal{O}(\exp(-\frac{1}{2} f_n^\top A_i \Sigma_w A_i^\top f_n))$, and therefore, the summation converges to a bounded value, i.e.,

$$\lim_{t \rightarrow \infty} \sum_{i=1}^t \frac{1}{\gamma(i)} (\mathbf{1}_{\{n\}}^\top s(i))^2 < \infty.$$

Consequently, there exists a unit vector $\hat{v} = \mathbf{1}_{\{n\}}$ and $n \geq 1$, such that the series converges, violating the condition (3.23). Therefore, $\lim_{t \rightarrow \infty} \mathcal{J}_B^*(t) > 0$, proving the Bayesian affine MMSE estimator is not consistent for Fourier bases and stochastically unstable LTI systems with Gaussian process noise w . $\square$

Proof of Proposition 3.4.3. Let $\mathcal{J}_B^*(\tau, T^\tau)$ denote the Bayesian MMSE (3.7b), where τ is the number of statistically independent output trajectories, each with the length of $(T_i + 1)$, namely, $y_i = [y_0, \dots, y_{T_i}]^\top$, $i \in \{1, \dots, \tau\}$. Thus, a total of $T^\tau = (T_1 + 1) + \dots + (T_\tau + 1)$

measurements are used. Since the error decreases monotonically with additional dependent measurements, as stated in Corollary 3.3.5, one can infer that

$$\mathcal{J}_B^*(\tau, T^\tau) \leq \mathcal{J}_B^*(\tau, \tau), \quad (3.25)$$

where $\mathcal{J}_B^*(\tau, \tau)$ represents the estimation error using τ statistically independent trajectories each with single measurements, $T^\tau = \tau$, where each trajectory i contains a single independent measurement, i.e., $y_i = y_0$. Using the equivalent representation of the optimal MMSE (3.11), we have

$$\mathcal{J}_B^*(\tau, \tau) = \text{tr}((\bar{\Sigma}_\theta^{-1} + \mu_\Phi(\tau)\bar{R}^{-1}(\tau)\mu_\Phi^T(\tau))^{-1}),$$

where $R(\tau) := M(\tau) + \Sigma_v(\tau)$. Consider the matrix definition and decompositions

$$\begin{cases} S := \mu_\Phi(\tau)\bar{R}^{-1}(\tau)\mu_\Phi^T(\tau), \\ \mu_\Phi(\tau) = [\mu_\phi^0(1), \dots, \mu_\phi^0(\tau)], \\ R(\tau) = \text{diag}(r(1), \dots, r(\tau)), \end{cases}$$

with $r(i) = M_{00}(i) + \sigma_{v_0}^2(i) > 0$ for batch i , and $M_{00}(i) = \text{tr}(\Sigma_\theta^{00}(i)(\Sigma_\theta + \mu_\theta\mu_\theta^T))$. The statistical independence between batches ensures $R(\tau)$ is diagonal, leading to

$$S = \sum_{i=1}^{\tau} \frac{1}{r(i)} \mu_\phi^0(i)\mu_\phi^{0T}(i).$$

Following the reasoning in the proof of Proposition 3.4.2, the necessary and sufficient condition for the convergence of the MMSE estimator with independent and identically distributed (i.i.d.) data is

$$\lim_{\tau \rightarrow \infty} \mathcal{J}_B^*(\tau, \tau) = 0 \iff \lim_{\tau \rightarrow \infty} \lambda_{\min} \left(\sum_{i=1}^{\tau} \frac{1}{r(i)} \mu_\phi^0(i)\mu_\phi^{0T}(i) \right) = \infty.$$

From the inequality (3.25), the above provides a sufficient condition for $\mathcal{J}_B^*(\tau, T^\tau)$ to converge to 0. $\square$

3.5 ACTIVE LEARNING

Active learning seeks to develop input signals that maximize information gain, thereby reducing estimation error. While our proposed affine MMSE estimator in Theorem 3.3.2 is optimal among all affine estimators, it can be coupled with an optimal input signal to improve estimation performance further. To this end, our approach leverages the analytical expression of the estimation error (3.7b), distinguishing it from most active learning methods that rely on maximizing information gain as a proxy for the estimation error, which is typically unavailable. Consequently, the optimal inputs can be determined independently of measurements, either a-priori or in real-time, by solving the following optimization problem:

$$u^* \in \underset{u \in \mathcal{U}}{\text{argmin}} \mathbb{E}[\|\theta - \hat{\theta}_B(y)\|^2] = \underset{u \in \mathcal{U}}{\text{argmin}} \mathcal{J}_B^*(u), \quad (3.26)$$

where $\mathcal{U}$ represents the input space, which may impose physical constraints on feasible inputs for estimation. Since only the second term of $\mathcal{J}_B^*$ in (3.7b) depends on $\mathbf{u}$ and is negative, the optimization problem (3.26) is equivalent to maximizing the second term of (3.7b). Nevertheless, we present this problem in its minimization form and solve it using an iterative first-order method, such as steepest descent or projected steepest descent when constraints are involved [16], while leveraging an analytical expression for its gradient. It should be noted that (3.26) is potentially non-convex due to how $\mathbf{u}$ influences matrices μ_Φ and $\mathbf{M}$. The iterative update rule for projected gradient descent is given by

$$\mathbf{u}^{k+1} = \mathcal{P}_{\mathcal{U}}[\mathbf{u}^k - \alpha_k \nabla_{\mathbf{u}} \mathcal{J}_B^*(\mathbf{u}^k)], \quad \nabla_{\mathbf{u}} \mathcal{J}_B^* = \left[\frac{\partial \mathcal{J}_B^*}{\partial \mathbf{u}_1}, \dots, \frac{\partial \mathcal{J}_B^*}{\partial \mathbf{u}_{(n_X + T n_{\mathbf{u}})}} \right]^\top, \quad (3.27)$$

where k represents the current iteration step, $\mathcal{P}_{\mathcal{U}}[\cdot]$ denotes the projection operator that maps the argument onto $\mathcal{U}$, α_k is a positive stepsize, and $\nabla_{\mathbf{u}} \mathcal{J}_B^*(\mathbf{u}^k)$ is the gradient of the cost function evaluated at $\mathbf{u}^k$. Various algorithms exist for selecting the stepsize α_k , including the standard approach of diminishing stepsize rules [16, p. 69]. Furthermore, since the explicit description of $\mathcal{J}_B^*$ is available in (3.7b) (aka. zero-order information), more sophisticated methods such as exact line search can also be employed; we refer interested readers to [16, Ch. 2] for further details. It is worth noting that these stepsize rules may require parameter tuning (e.g., a constant in diminishing stepsize methods) or involve computational overhead in line search techniques. To circumvent these possible limitations, one can also utilize the recent, easy-to-implement, adaptive stepsize from [92] defined as

$$\alpha_k = \min \left\{ \sqrt{1 + \beta_{k-1} \alpha_{k-1}}, \frac{\|\mathbf{u}^k - \mathbf{u}^{k-1}\|}{2 \|\nabla_{\mathbf{u}} \mathcal{J}_B^*(\mathbf{u}^k) - \nabla_{\mathbf{u}} \mathcal{J}_B^*(\mathbf{u}^{k-1})\|} \right\}, \quad \beta_k = \frac{\alpha_k}{\alpha_{k-1}}, \quad k \geq 1,$$

with initial conditions $\beta_0 = \infty$ and $\alpha_0 = 10^{-10}$. Finally, the gradient of the optimal Bayesian estimation error (3.7b) with respect to each $\mathbf{u}_i$ of the input vector is derived by applying the chain rule to $\mathcal{J}_B^*(\mathbf{u})$, resulting in

$$\frac{\partial \mathcal{J}_B^*}{\partial \mathbf{u}_i} = \text{tr}(\Psi^{*\top} \Psi^* \partial_{\mathbf{u}_i} \mathbf{M} + 2 \Psi^{*\top} (\Psi^* \mu_\Phi^\top - \mathbf{I}) \Sigma_\theta \partial_{\mathbf{u}_i} \mu_\Phi), \quad (3.28)$$

where Ψ^* is defined in (3.7a). The partial derivative terms $\partial_{\mathbf{u}_i} \mathbf{M}$ and $\partial_{\mathbf{u}_i} \mu_\Phi$ depend on the gradient of the respective DBS as follows:

- $\partial_{\mathbf{u}_i} \mu_\Phi = [\partial_{\mathbf{u}_i} \mu_\phi^0, \dots, \partial_{\mathbf{u}_i} \mu_\phi^T]$,
- the $(t+1, t'+1)$ element of $\partial_{\mathbf{u}_i} \mathbf{M}$ is $\partial_{\mathbf{u}_i} M_{tt'} = \text{tr}((\Sigma_\theta + \mu_\theta \mu_\theta^\top) \partial_{\mathbf{u}_i} \Sigma_\Phi^{tt'})$,
- the $(n+1)^{\text{th}}$ element of $\partial_{\mathbf{u}_i} \mu_\phi^t$ and $(m+1, n+1)^{\text{th}}$ element of $\partial_{\mathbf{u}_i} \Sigma_\Phi^{tt'}$ are

$$\partial_{\mathbf{u}_i} \mu_{\phi_n}^t = \frac{\partial \mathbb{E}[\phi_n(\mathbf{x}_t)]}{\partial \mathbf{u}_i}, \quad \partial_{\mathbf{u}_i} \Sigma_{\phi_{mn}}^{tt'} = \frac{\partial \mathbb{E}[\phi_m(\mathbf{x}_t) \phi_n^\top(\mathbf{x}_{t'})]}{\partial \mathbf{u}_i} - \partial_{\mathbf{u}_i} \mu_{\phi_m}^t \mu_{\phi_n}^{t'} - \mu_{\phi_m}^t \partial_{\mathbf{u}_i} \mu_{\phi_n}^{t'}.$$

For Fourier basis functions (3.13), with the explicit expressions of the DBS $(\mu_\phi^t, \Sigma_\Phi^{tt'})$ provided in Lemma 3.4.1, the partial derivatives $\partial_{\mathbf{u}_i} \mu_\phi^t$ and $\partial_{\mathbf{u}_i} \Sigma_\Phi^{tt'}$, are the following DBS gradients:

(i) **DBS mean gradient:** $\partial_{u_i} \mu_{\phi_0}^t = 0$, and for $n \geq 1$,

$$\partial_{u_i} \mu_{\phi_n}^t = \sum_{\ell \in \{-1, 1\}} j \ell f_n^\top A_t B \mathbf{1}_{\{i\}} \exp(j \langle A_t^\top \ell f_n, Bu \rangle) \varphi(\Sigma_w^{\frac{1}{2}} A_t^\top \ell f_n), \quad (3.29a)$$

(ii) **DBS covariance gradient:** $\partial_{u_i} \Sigma_{\phi_{mn}}^{tt'} = 0$ if $m = 0$ or $n = 0$. For $m, n \geq 1$,

$$\begin{aligned} \partial_{u_i} \Sigma_{\phi_{mn}}^{tt'} = \sum_{\ell, \ell' \in \{-1, 1\}} j (\ell f_m^\top A_t + \ell' f_n^\top A_{t'}) B \mathbf{1}_{\{i\}} \exp(j \langle A_t^\top \ell f_m + A_{t'}^\top \ell' f_n, Bu \rangle) \\ \times \left(\varphi\left(\Sigma_w^{\frac{1}{2}} (A_t^\top \ell f_m + A_{t'}^\top \ell' f_n)\right) - \varphi(\Sigma_w^{\frac{1}{2}} A_t^\top \ell f_m) \varphi(\Sigma_w^{\frac{1}{2}} A_{t'}^\top \ell' f_n) \right), \end{aligned} \quad (3.29b)$$

where A_t and $A_{t'}$ are defined analogously to (3.16), and $\mathbf{1}_{\{i\}}$ is the single-entry vector with 1 at index i and zero elsewhere, i.e.,

$$\mathbf{1}_{\{i\}} = [0, \dots, 0, \underset{\substack{\uparrow \\ i^{\text{th}}}}{1}, 0, \dots, 0]^\top.$$

In the case of Gaussian noise, it is sufficient to substitute the characteristic functions in the derived terms of (3.29) with their corresponding expressions in (3.18). The computational complexity of computing the gradient (3.28) for each element of input per iteration of the first-order method is equivalent to that of the optimal Bayesian affine MMSE estimator, which scales as $\mathcal{O}(T^3)$ when $N \ll T$. In the following section, we numerically demonstrate the reduction in estimation error achieved by applying this active learning technique.

3.6 NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of four estimators through numerical examples: two *approximate* regularized least-squares (RLS) linear estimators (DLS and MLS) from Subsection 3.3.1, the optimal Bayesian affine MMSE estimator (BMS) from Subsection 3.3.2, and its integration with active learning (BAL) from Subsection 3.5. The MATLAB library provided at “<https://github.com/sasanvakili/Bayesian4Wiener>” enables reproducibility of the experiments in this section. Using a marginally stable dynamical system with a true function in Fourier subspace, we observe the effects of increasing process noise uncertainty and demonstrate the inherent inconsistency discussed in Proposition 3.4.2. We compare estimators using the squared error criterion, $\|\theta - \hat{\theta}(y)\|^2$, over 10,000 simulations, employing identical realizations of w and v across all estimators for fair comparison. In the resulting plots, dashed lines represent the analytically computed mean squared error, $\mathbb{E}[\|\theta - \hat{\theta}(y)\|^2]$, for each method. Non-histogram plots display shaded areas representing the 20th to 80th percentile range of squared error, while histograms show probability densities.

Experiment setup. Our experiments are based on time-invariant Gaussian measurement noise $v_t \sim \mathcal{N}(0, \sigma_v^2 \mathbb{I})$, time-invariant Gaussian process noise $w_{t+1} \sim \mathcal{N}(0, \sigma_w^2 \mathbb{I})$, and an initial state $x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2 \mathbb{I})$, with $\sigma_{x_0}^2 = \sigma_w^2$. To examine the impact of process noise, we use three incremental variances, $\sigma_w^2 = \{0, 0.001, 0.01\}$, while maintaining a consistent measurement noise variance of $\sigma_v^2 = 0.01$ across all experiments. Our experimental design involves generating 100 independent samples each of θ , w , and v for various trajectory lengths T , resulting in 10,000 experiments. The dynamical system under study is a linear

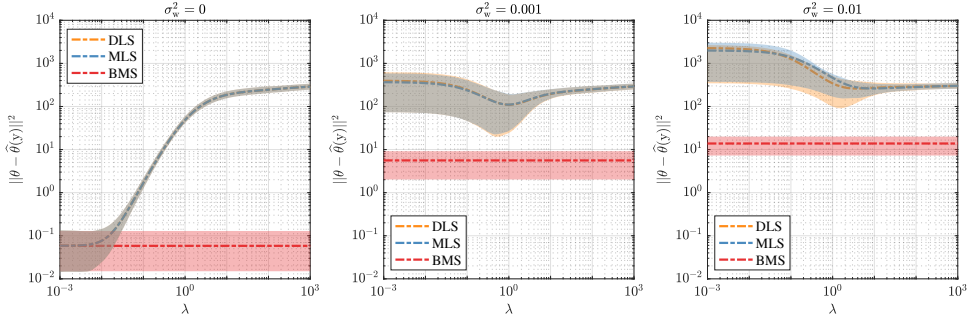


Figure 3.1: Squared errors of DLS, MLS, and BMS for different λ values.

time-invariant (LTI) kinematic model representing a robot moving in two dimensions. In this model, the system states $\mathbf{x}_t \in \mathbb{R}^2$ correspond to the robot's position, while the inputs $\mathbf{u}_t \in \mathbb{R}^2$ represent velocities in each dimension. The system dynamics are defined by $\mathbf{A}_t = \mathbf{I}$ and $\mathbf{B}_t = \Delta t \mathbf{I}$, with a sampling time $\Delta t = 0.1$. We use $\mu_{\mathbf{x}_0} = [3.2, 2.8]^\top$ and $\mathbf{u}_t = 4.5 \sum_{v \in Y} [\cos(vt\Delta t), \sin(vt\Delta t)]^\top$, $Y = \{3, 5, 10, 20, 100\}$, for DLS, MLS, and BMS experiments, as well as for initializing the projected steepest descent algorithm of BAL. Each dimension of $\mathbf{u}_t$ is constrained within $[-200, 200]$, representing the robot's achievable velocity range. The BAL estimator derives $\mathbf{u}^*$ within this input range as described in Subsection 3.5. To ensure consistent initial states across all estimators, BAL does not optimize the first element of $\mathbf{u}$, namely $\mu_{\mathbf{x}_0}$. Following the Fourier basis representation in (3.13), we employ 11 unknown parameters θ_n , i.e., $n = 0, \dots, 10$, with known frequency vectors $\mathbf{f}_n \in \mathbb{R}^2$. Specifically, $\mathbf{f}_0 = [0, 0]^\top$, $\mathbf{f}_n = [n\frac{2\pi}{10}, 0]^\top$ for $n = \{1, 2, 3\}$, and $\mathbf{f}_n = [(n-7)\frac{2\pi}{10}, \frac{2\pi}{6}]^\top$ for $n = \{4, \dots, 10\}$. The prior distributions of the unknown parameters follow a uniform distribution $\mathcal{U}(2, 8)$, implying $\mu_{\theta_n} = 5$ and $\sigma_{\theta_n}^2 = 3$, from which the true parameter values are drawn.

Benchmark 1 (Optimal Bayesian vs. RLS expected error). Our first numerical benchmark involves 10,000 simulations for a trajectory of $T = 100$, i.e., 101 total measurements, to tune the hyperparameter λ of DLS and MLS for three incremental process noise scenarios. The error of BMS remains constant across λ values, as it does not depend on this hyperparameter. Figure 3.1 illustrates that the squared error of DLS and MLS deviates increasingly from BMS as the process noise variance increases. When $\sigma_w^2 = 0$, as shown in the left plot of Figure 3.1, all three estimators perform comparably for small λ values. However, the middle and right plots of Figure 3.1 demonstrate that DLS and MLS significantly underperform compared to BMS when $\sigma_w^2 = \{0.001, 0.01\}$.

Benchmark 2 (Optimal Bayesian vs. optimal RLS error histogram). After tuning the hyperparameter λ from Benchmark 1 and determining the optimal hyperparameter λ^* for MLS in each process noise case for a trajectory length of $T = 100$, we compare the squared error difference between methods. The estimates are denoted as follows: MLS using λ^* as $\hat{\theta}_{\text{MLS}}(\mathbf{y}, \lambda^*)$, BMS as $\hat{\theta}_{\text{BMS}}(\mathbf{y})$, and BAL as $\hat{\theta}_{\text{BAL}}(\mathbf{y}, \mathbf{u}^*)$, where BAL utilizes optimal inputs $\mathbf{u}^*$. Figure 3.2 presents the probability density histogram of squared error differences between these pairs for 101 total measurements and across 10,000 simulations. Each histogram's area

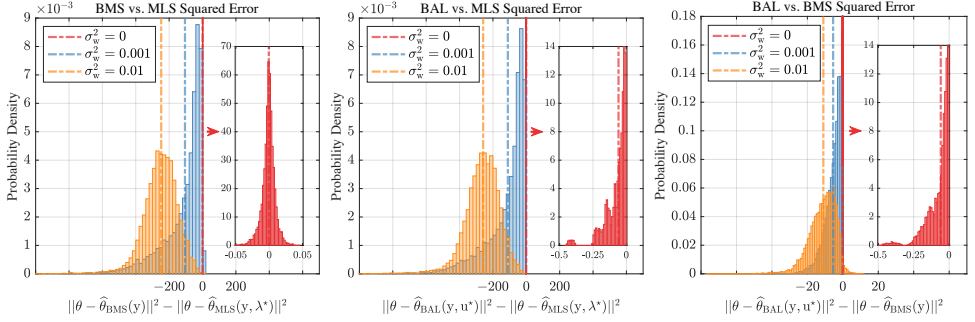


Figure 3.2: Distribution of squared error differences between pairs of estimators.

sums to unity, representing the probability density function, with dashed lines indicating the analytically computed mean of the differences. The red, blue, and orange distributions correspond to process noise variances of $\sigma_w^2 = 0, 0.001$, and 0.01 , respectively. As observed in the left plots, MLS shows less squared error than BMS in only 1.62% and 0.03% of cases when $\sigma_w^2 = 0.001$ and 0.01 , respectively. The middle plot demonstrates that MLS never outperforms BAL. These results indicate that BMS and BAL almost surely outperform MLS when process noise exists. Without process noise, BMS and MLS yield similar results, as evident by the smaller red histogram centered around zero on the left plot. The right plot shows that BAL with u^* consistently provides lower error than BMS with the input indicated in Experiment Setup when $\sigma_w^2 = 0$ or 0.001 , while BMS outperforms BAL in 1.7% of cases when $\sigma_w^2 = 0.01$.

Benchmark 3 (Impact of trajectory length and process noise). Building on our previous benchmarks, we extend our analysis to 10,000 simulations, each featuring trajectories of varying lengths, $T \in \{0, 4, 10, 13, 16, 20, 25, 32, 40, 50, 63, 79, 100\}$. Figure 3.3 compares the squared errors, shown as shaded areas, and analytically computed mean squared errors, represented by dashed lines, for DLS, MLS, BMS, and BAL across three process noise variances. For DLS and MLS, we employ the optimal hyperparameter λ^* , while BAL utilizes optimized input u^* , derived separately for each trajectory length. The results confirm that input optimization significantly reduces squared error when the measurement count equals the number of unknown parameters. In the presence of process noise, i.e., $\sigma_w^2 = \{0.001, 0.01\}$, DLS and MLS diverge as they fail to account for system states drift due to accumulated process noise. In contrast, Bayesian methods maintain robustness by precisely calculating the covariance matrices. Notably, the estimation error reduction of Bayesian estimators slows considerably with increasing statistically dependent measurements. This observation confirms the inconsistency described in Proposition 3.4.2, which arises from the marginal stability of the chosen dynamical system.

Benchmark 4 (Evaluation with multiple trajectories). Lastly, we address the slowing rate of estimation error reduction observed in Figure 3.3 of the previous benchmark by employing multiple batches of independent trajectories which demonstrates the consistency condition of the BAL estimator as outlined in Proposition 3.4.3. We conduct an experiment using 10,000 simulations, comparing three settings with different numbers of independent

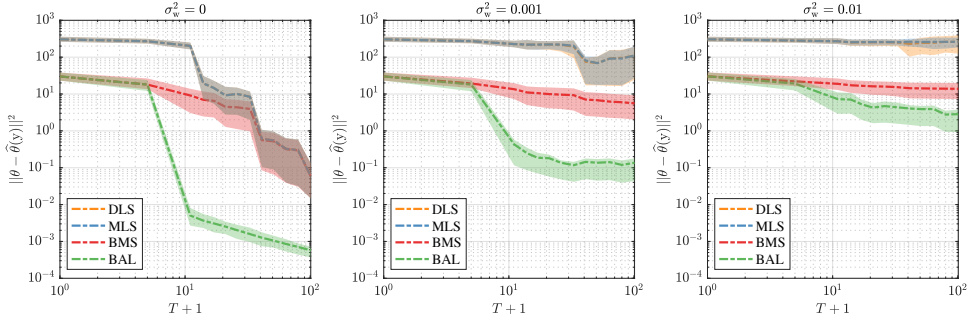


Figure 3.3: Squared errors of DLS, MLS, BMS, and BAL for varying trajectory length T .

trajectories: $\tau = \{1, 11, 101\}$. The same realizations of w and v were maintained across all cases. When $\tau = 101$, each trajectory consists of 1 sample; when $\tau = 11$, we have 10 trajectories of 10 samples each and 1 trajectory of 1 sample; and when $\tau = 1$, there is only 1 trajectory of 101 samples, representing the performance of the BAL estimator from the two previous benchmarks with $T = 100$. For each setting, optimal inputs u^* are obtained using (3.26), which includes μ_{x_0} for different trajectories except the first. The optimization of μ_{x_0} for the first batch is excluded to ensure that the BAL estimates for $\tau = 1$ are identical to those obtained from $T = 100$ in Benchmarks 2 and 3. Figure 3.4 compares the probability density histograms of the squared estimation error for all three settings under process noise variances $\sigma_w^2 = 0.001$ and 0.01 . Unlike Figure 3.2, the horizontal axes here use a logarithmic scale to better resolve differences in the error distributions. The $\sigma_w^2 = 0$ scenario is excluded because, in the absence of process noise, measurements across time steps become statistically independent, providing similar results for all three settings. Results show that the BAL estimator achieves the least estimation error with 101 independent samples, followed by 11 independent trajectories, and then 1 trajectory of 101 correlated samples. The comparison between the two plots demonstrates that the estimation error increases with process noise, highlighting its impact on estimation performance. This experiment confirms the consistency condition in Proposition 3.4.3.

Experimental analysis summary. Benchmark 1 and Benchmark 2 demonstrate the superiority of the Bayesian affine MMSE estimator, showing that it almost surely outperforms the *approximate* regularized least-squares method (cf. Figure 3.1 and Figure 3.2). Furthermore, when paired with active learning, the Bayesian affine MMSE estimator achieves the lowest estimation error (cf. Figure 3.2 and Figure 3.3). Benchmark 3 further investigates the impact of process noise in a marginally stable dynamical system, emphasizing that unbounded growth in process noise slows down estimation error reduction (cf. Figure 3.3). As the number of dependent data points increases, the information gained from each new measurement diminishes, preventing convergence to the true value even with infinite measurements. Benchmark 4 highlights the importance of using multiple statistically independent trajectories for accurate parameter estimation under process noise. It demonstrates convergence as the number of independent trajectories increases (cf. Figure 3.4).

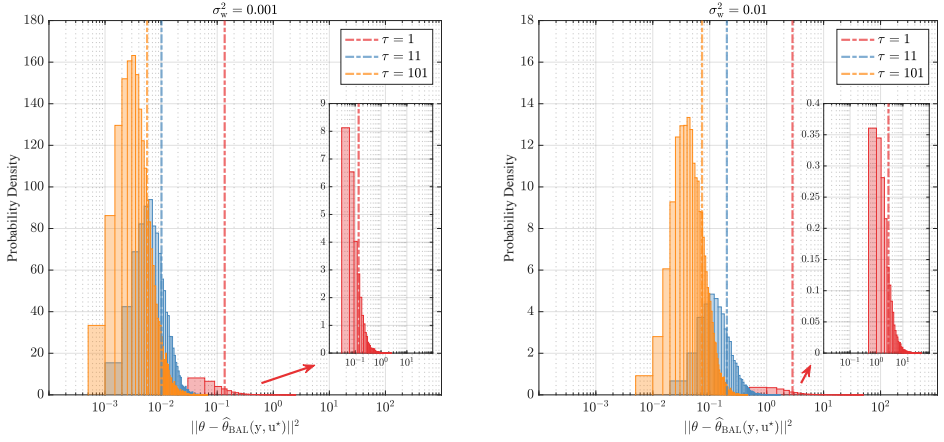


Figure 3.4: Distribution of squared errors for multiple statistically independent trajectories τ .

3.7 CONCLUSION

This chapter has developed a Bayesian affine MMSE estimator for nonlinear output functions in Wiener models with known linear state dynamics, characterized in terms of dynamic basis statistics (DBS) that encode how the stochastic state trajectory propagates through the basis functions. The estimator admits closed-form approximate posterior statistics, is Bayesian unbiased, and its error decreases monotonically with the trajectory length. We also established a condition under which the use of multiple statistically independent trajectories guarantees consistency. The closed-form error expression was used to synthesize informative input signals via an active learning scheme, thereby further reducing parameter estimation error. For Fourier bases, explicit DBS expressions were derived, revealing an inherent inconsistency in single-trajectory identification under stochastically unstable dynamics. Finally, numerical results showed that the proposed estimator consistently outperforms approximate regularized least-squares methods across various scenarios, positioning the affine MMSE with DBS framework as a principled and computationally tractable alternative to classical linear estimators in Wiener-type models.

4

NONLINEAR ESTIMATORS VIA FIXED-POINT CHARACTERIZATION

4

This chapter 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.

4.1 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 [87] and statistical inference [17]. 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 [51, 91]. 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 of randomness further complicate the inference process [126]. Consider a *known* discrete-time linear time-varying (LTV) dynamical system

$$\mathbf{x}_{t+1} = \mathbf{A}_t \mathbf{x}_t + \mathbf{B}_t \mathbf{u}_t + \mathbf{w}_{t+1}, \quad \mathbf{y}_t = h_\theta(\mathbf{x}_t) + \mathbf{v}_t, \quad (4.1)$$

where t is the time index, $\mathbf{x}_t \in \mathbb{R}^{n_x}$ denotes the system state vector, $\mathbf{A}_t \in \mathbb{R}^{n_x \times n_x}$ and $\mathbf{B}_t \in \mathbb{R}^{n_x \times n_u}$ are known system matrices, and $\mathbf{u}_t \in \mathbb{R}^{n_u}$ represents the sequence of known or controlled inputs. The process noise $\mathbf{w}_{t+1} \in \mathbb{R}^{n_x}$ is modeled as a zero-mean random vector with known covariance $\Sigma_{\mathbf{w}_{t+1}}$, i.e., $\mathbf{w}_{t+1} \sim \mathbb{P}(0, \Sigma_{\mathbf{w}_{t+1}})$. The observation $\mathbf{y}_t \in \mathbb{R}^{n_y}$ is generated via an *unknown* parameterized function $h_\theta : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y}$ corrupted by measurement noise $\mathbf{v}_t \in \mathbb{R}^{n_y}$ and $\mathbf{v}_t \sim \mathbb{P}(0, \Sigma_{\mathbf{v}_t})$. Throughout this work, and without loss of generality, we focus on the single-output case $n_y = 1$, since the extension to multi-output measurements ($\mathbf{y}_t \in \mathbb{R}^{n_y}$) is straightforward and essentially reduces to a set of independent single-output problems, see Remark 3.2.1 and [140, 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 $\mathbf{y}_t$ and known inputs $\mathbf{u}_t$. We further parameterize the output function $h_\theta : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ as a linear combination of known basis functions $\phi_n : \mathbb{R}^{n_x} \rightarrow \mathbb{C}$, i.e.,

$$h_\theta(\mathbf{x}_t) = \sum_{n=0}^N \theta_n \phi_n(\mathbf{x}_t) = \langle \phi(\mathbf{x}_t), \theta \rangle, \quad (4.2)$$

where the number of basis functions is $N+1$, the coefficient vector $\theta = [\theta_0, \dots, \theta_N]^\top$ contains the unknown parameters, and $\phi(\mathbf{x}_t) = [\phi_0(\mathbf{x}_t), \dots, \phi_N(\mathbf{x}_t)]^\top$ is the vector of basis functions evaluated at $\mathbf{x}_t$. 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 modeled as block-oriented structures consisting of a linear dynamical process followed by a static nonlinear observation model, commonly known as Wiener models [129]. In this setting, parameter estimation is inherently challenging because the system states, $\mathbf{x}_t$, are stochastic due to the process noise, $\mathbf{w}_t$, 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 $\mathbf{y}_t$ is, in general, not available in closed-form [40, 126]. This intractability fundamentally complicates Bayesian parameter estimation and necessitates the use of approximate inference techniques [17]. 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 [39, 136] as well as more tractable alternatives such as maximum a posteriori (MAP) estimation [26].

Fully Bayesian methods aim to characterize the entire posterior distribution, typically only approximately, using techniques such as Markov chain Monte Carlo (MCMC) [121, Ch. 6], sequential Monte Carlo (SMC) [2, 32, 101], and variational inference [18, 34]. MCMC algorithms construct a Markov chain that converges in distribution to the posterior of the unknown [122]. Classic examples include the Metropolis-Hastings (MH) algorithm [62, 97] and the Gibbs sampler [50]. 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) [49], and combinations of MCMC and SMC techniques [113], collectively referred to as particle Markov chain Monte Carlo (PMCMC) algorithms. By exploiting different theoretical properties of the SMC framework [32], the MH and Gibbs sampler algorithms can be extended to the particle marginal Metropolis-Hastings (PMMH) method [127] and the particle Gibbs (PG) algorithm [33], respectively. Although these methods provide reliable uncertainty quantification and improved performance, they are typically computationally demanding [35] and scale poorly to high-dimensional problems [6, 116]. In this work, the particle Gibbs with ancestor sampling (PGAS) algorithm [81], a prominent particle sampler for Bayesian learning of state-space models, is included as one of the benchmark methods for numerical evaluation.

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 [38, 127], 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 [53], EM guarantees a monotonic increase in the likelihood [51, 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 [63, Ch. 6], stochastic methods [37], and variational approaches [13]. Among approximate E-step methods, particle filters and kernels such as PGAS sample the smoothing distribution to approximate the likelihood [80, 127]. These approximations, while often effective, do not guarantee monotonic likelihood improvement and can lead to suboptimal modes or slow convergence [5]. Nevertheless, we use the EM algorithm [127] as a benchmark in our experiments.

A recent perspective focuses on affine estimators rather than approximating full posterior distributions, see Chapter 3 and [140]. 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 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 particular, when $h_\theta(\cdot)$ is expressed as a linear combination of known basis functions as in (4.2), the DBS encode the statistics of these basis functions evaluated at different time steps along the stochastic state trajectory. However, in the presence of a latent state variable, the optimal Bayesian MMSE estimator can become highly nonlinear and deviate significantly from the affine estimator. To illustrate how state uncertainty affects the estimation of the parameter vector θ , we consider the following simple example with a linear basis function $\phi(x_t) = x_t$ in (4.1), where θ is a scalar parameter inferred from a single measurement. Specifically, we focus on a one-dimensional dynamical system with $n_x = 1$ and only an initial-state prior $x_0 \sim \mathcal{P}(\mu_{x_0}, \Sigma_{x_0})$. In this setting, the empirical MMSE estimator is tractable and is taken to represent the optimal Bayesian MMSE estimator.

4

Example 4.1 (1-D linear state-space model). Let the output function h_θ in (4.1) be defined using a single linear basis function, $\phi_n(x) = x$, so that

$$y = \theta x + v, \quad (4.3)$$

where $x \in \mathbb{R}$ is the initial state with $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $\theta \in \mathbb{R}$ is an unknown scalar parameter, and $v \sim \mathcal{N}(0, \sigma_v^2)$ is measurement noise with $\sigma_v^2 = 0.16$. 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 4.1 shows the optimal affine and empirical MMSE estimators for θ as functions of the measurement y for two choices of initial-state statistics: $(\mu_x, \sigma_x^2) = (0.5, 0.05)$ (left plot) and $(\mu_x, \sigma_x^2) = (0.5, 4)$ (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 Theorem 3.3.2. 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.

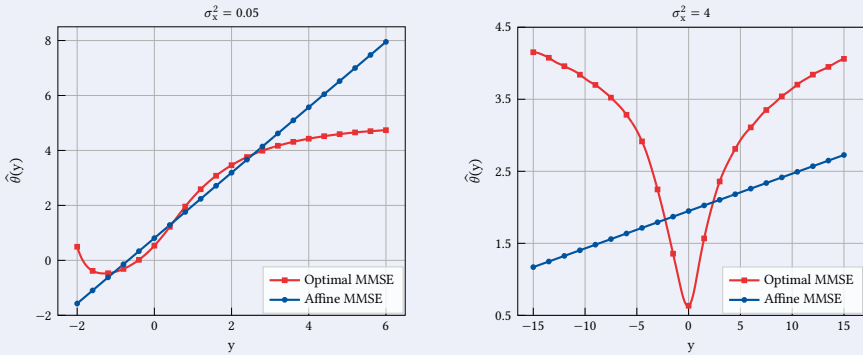


Figure 4.1: Optimal MMSE in (4.5) vs. affine MMSE in (4.10a) for estimating θ given y .

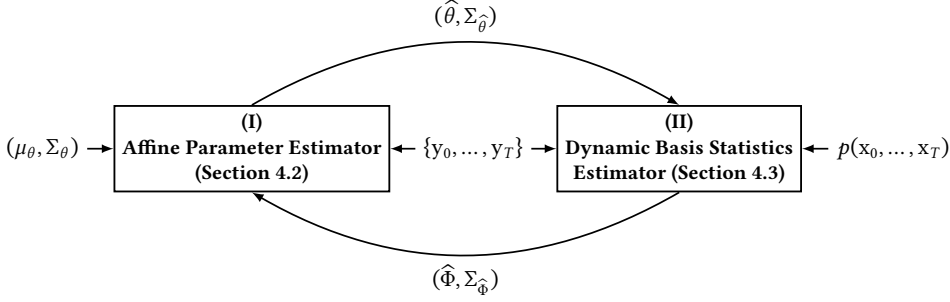


Figure 4.2: Architecture of the proposed nonlinear parameter estimator.

Contributions. In this chapter, we propose an analytical approach to construct a nonlinear estimator that improves upon the affine class. 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 4.2. As shown in the block diagram, the proposed estimator consists of two coupled components: block (I), described in Section 4.2, implements the affine MMSE parameter estimator, which takes as input the entire measurement sequence $\{y_0, \dots, y_T\}$, the prior $(\mu_\theta, \Sigma_\theta)$ on the unknown parameters, and DBS estimates $(\hat{\Phi}, \Sigma_{\hat{\Phi}})$ summarizing basis-function evaluations. This block produces the parameter estimates and their estimation-error covariance $(\hat{\theta}, \Sigma_{\hat{\theta}})$, which are used as input to the DBS estimator in block (II). Block (II), discussed in Section 4.3, implements the DBS estimator, which uses the parameter estimates $(\hat{\theta}, \Sigma_{\hat{\theta}})$, the same measurement sequence, and information about the state-trajectory distribution $p(x_0, \dots, x_T)$ to construct updated DBS statistics. This nonlinear estimator architecture 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 4.2, as detailed in Figure 4.3:

- **Basis-estimation approach (II-A):** This approach directly estimates basis-function evaluations over the entire trajectory. We derive a closed-form expression for the optimal affine MMSE estimator of these unknown basis evaluations (Lemma 4.3.1), 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 $p(x_0, \dots, x_T)$ to construct prior DBS statistics (μ_Φ, Σ_Φ) for the basis evaluations via Definition 4.1. This prior, together with the measurement sequence and the parameter-estimate statistics, is then processed by the affine basis estimator to obtain DBS estimates $(\hat{\Phi}, \Sigma_{\hat{\Phi}})$. 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 4.2. However, numerical experiments show that this approach can underperform the purely affine MMSE parameter estimator in

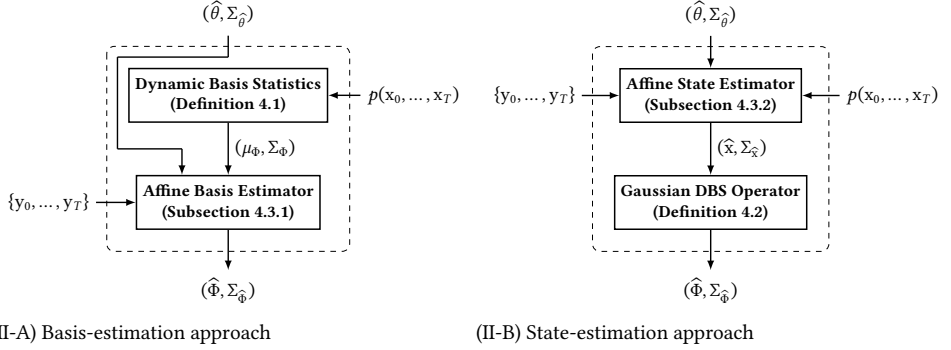


Figure 4.3: Architectures of the Dynamic Basis Statistics estimator.

certain regimes (cf. Figure 4.6), highlighting the limitations discussed in Remark 4.3.3 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 4.3.5). As shown in subfigure (II-B), the affine state estimator uses information about the state-trajectory distribution $p(x_0, \dots, x_T)$, along with the measurement sequence and the parameter-estimate statistics $(\hat{\theta}, \Sigma_{\hat{\theta}})$ in place of its prior, to compute state estimates and their estimation-error covariance. These state-estimate statistics $(\hat{x}, \Sigma_{\hat{x}})$ 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 4.2 to obtain DBS estimates $(\hat{\Phi}, \Sigma_{\hat{\Phi}})$. 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 4.2. 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 4.6, 4.7, and 4.8).

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 4.2.5 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 [140]. All affine estimators in DB-P and DS-P admit explicit expressions for Fourier basis functions, yielding fully explicit DBS formulas in this case [140]. To facilitate reproducibil-

ity, an open-source MATLAB library implementing all proposed methods 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 [147], the iterative smoothing scheme of the dual estimation method [63, Ch. 5], and in SMC-based approaches such as PGAS [81] and SMC-EM (particle EM) [127]. Both DKS and PGAS target the joint smoothing distribution $p(\theta, \{x_t\}_{t=0}^T | \{y_t\}_{t=0}^T)$ 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 $p(\{x_t\}_{t=0}^T | \theta, \{y_t\}_{t=0}^T)$, 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 4.9).

Organization. Section 4.2 formulates the MMSE estimation problem for the parameterized output function (4.2), reviews the affine parameter estimator of Chapter 3, and motivates the nonlinear parameter estimator structure. Section 4.3 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 4.4 presents numerical experiments comparing the proposed dual estimators with the affine parameter estimator and SMC-based methods. Detailed proofs of all mathematical statements are provided in Section 4.6.

Notation. Throughout this chapter, $\mathbb{R}$ 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., $x_{t,i}$ represents the i^{th} element of the vector x_t), while superscripts represent the time index of vector or matrix elements. The trace operator is denoted by $\text{tr}(\cdot)$, the transpose of a matrix A is denoted by $A^\top$, and $\text{diag}(A_1, \dots, A_k)$ represents a block-diagonal matrix with diagonal entries $A_1, \dots, A_k$. Given $A \in \mathbb{R}^{m \times n}$, a matrix with columns $a_1, \dots, a_n \in \mathbb{R}^m$, $\text{vec}(A)$ is the vector $[a_1^\top, \dots, a_n^\top]^\top \in \mathbb{R}^{mn}$. The inner product of two vectors x and y is given by $\langle x, y \rangle = x^\top y$, and the corresponding 2-norm is $\|x\| = \sqrt{\langle x, x \rangle}$. The inner product of two matrices $A, B \in \mathbb{R}^{m \times n}$ is defined as $\langle A, B \rangle = \text{tr}(A^\top B)$. The conic inequality $A \preceq B$ means that the matrix difference $B - A$ is positive semidefinite, i.e., $B - A \succeq 0$. The notation $\mathcal{P}(\mu, \Sigma)$ 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.”

4.2 BAYESIAN ESTIMATION FOR WIENER MODEL

In this section, we formulate the Bayesian MMSE estimation problem for the Wiener model and review the Bayesian affine MMSE estimator from Chapter 3. We then discuss the relevance of these results and their connections to the proposed nonlinear estimator.

4.2.1 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 (4.2) and state representation in (4.1), together with the sequence of measurements $y = [y_0, \dots, y_T]^T$, our objective is to design an estimator $\hat{\theta}(\cdot)$ that minimizes the following mean-squared error (MSE):

$$\min_{\hat{\theta}(\cdot) \in \mathcal{F}_\theta} \mathbb{E}[\|\theta - \hat{\theta}(y)\|^2], \quad (4.4)$$

where $\mathcal{F}_\theta := \{\hat{\theta}: \mathbb{R}^{T+1} \rightarrow \Theta\}$ 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 $\hat{\theta}$ in $\mathcal{F}_\theta$, the optimal Bayesian MMSE estimator $\hat{\theta}_{\text{opt}}$ for (4.4) coincides with the posterior mean of θ given the measurements [76, Ch. 4]. Accordingly, the parameter estimate $\hat{\theta}_{\text{opt}}$ and its error covariance $\Sigma_{\hat{\theta}_{\text{opt}}}$ are

$$\hat{\theta}_{\text{opt}} = \mathbb{E}[\theta|y] = \int_{\Theta} \theta p(\theta|y) d\theta, \quad \Sigma_{\hat{\theta}_{\text{opt}}} = \mathbb{E}[(\theta - \hat{\theta}_{\text{opt}})(\theta - \hat{\theta}_{\text{opt}})^T]. \quad (4.5)$$

However, this estimator is generally computationally intractable due to the complexity of the posterior distribution $p(\theta|y)$ 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 and serves as the foundation for the new contributions presented in this work.

4.2.2 FROM AFFINE TO NONLINEAR ESTIMATORS

To obtain a tractable alternative to the Bayesian MMSE estimator in (4.5), 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 4.2.1 (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 $\Sigma_{ij} := \mathbb{E}[(\xi_i - \mu_i)(\xi_j - \mu_j)^T]$ 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,

$$\mathcal{A}_\zeta := \left\{ \widehat{\xi}_1(\xi_2) = \Psi \xi_2 + \psi \mid \Psi \in \mathbb{R}^{n_1 \times n_2}, \psi \in \mathbb{R}^{n_1} \right\}. \quad (4.6)$$

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

$$\begin{cases} \widehat{\xi}_1(\xi_2) = \Psi^* \xi_2 + \psi^*, \\ \Sigma_{\widehat{\xi}_1} := \mathbb{E}[(\xi_1 - \widehat{\xi}_1(\xi_2))(\xi_1 - \widehat{\xi}_1(\xi_2))^\top] = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, \\ \mathcal{J}^* = \text{tr}(\Sigma_{\widehat{\xi}_1}), \end{cases} \quad (4.7a)$$

with the optimal coefficients given by

$$\Psi^* = \Sigma_{12} \Sigma_{22}^{-1}, \quad \psi^* = \mu_1 - \Psi^* \mu_2. \quad (4.7b)$$

The proof of this lemma is provided in Section 4.6. Finding the affine MMSE estimator for estimating θ from y using Lemma 4.2.1 requires the observation covariance Σ_y and the cross-covariance between the unknown parameters and the observation vector, $\Sigma_{\theta y}$. This task is nontrivial because the observation model (4.2) depends on the inner product of two unknown quantities, namely the basis-function evaluations of the unknown system states $\phi(x)$ and the unknown parameters θ . A central component in characterizing the affine estimator for the MMSE problem (4.4) 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 4.1 (Dynamic basis statistics). Let the state sequence $x = [x_0^\top, \dots, x_T^\top]^\top$ be a trajectory of (4.1) for $t \in \{0, \dots, T\}$, and let $\Phi(x) = [\phi(x_0), \dots, \phi(x_T)]$ be the matrix whose columns are the basis function vectors evaluated at each x_t as in (4.2). The dynamic basis statistics (DBS) collection with respect to the joint distribution of the state trajectory is

$$\mu_\Phi = [\mu_\phi^0, \dots, \mu_\phi^T], \quad \Sigma_\Phi = \begin{bmatrix} \Sigma_\phi^{00} & \dots & \Sigma_\phi^{0T} \\ \vdots & \ddots & \vdots \\ \Sigma_\phi^{T0} & \dots & \Sigma_\phi^{TT} \end{bmatrix}, \quad (4.8)$$

where μ_ϕ^t and $\Sigma_\phi^{tt'}$ are the mean of $\phi(x_t)$ and the covariance between $\phi(x_t)$ and $\phi(x_{t'})$, respectively, and $(\mu_\phi^t, \Sigma_\phi^{tt'})$ denotes the DBS defined in Definition 3.1, given by

$$\mu_\phi^t := \mathbb{E}[\phi(x_t)], \quad \Sigma_\phi^{tt'} := \mathbb{E}[\phi(x_t)\phi(x_{t'})^\top] - \mathbb{E}[\phi(x_t)]\mathbb{E}[\phi(x_{t'})]^\top. \quad (4.9)$$

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 4.2.2 (Affine MMSE parameter estimator [140, Thm. 3.2]). *The Bayesian MMSE estimator of Problem 4.4 within the affine class, analogous to (4.6), and its corresponding estimation-error covariance with the optimal cost, as functions of μ_Φ and Σ_Φ , are given by*

$$\begin{cases} \hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi) = \Psi_\theta^*(\mu_\Phi, \Sigma_\Phi)y + \psi_\theta^*(\mu_\Phi, \Sigma_\Phi), \\ \Sigma_{\hat{\theta}_{\text{af}}}(\mu_\Phi, \Sigma_\Phi) = \Sigma_\theta - \Sigma_\theta \mu_\Phi \left(\mu_\Phi^\top \Sigma_\theta \mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_v \right)^{-1} \mu_\Phi^\top \Sigma_\theta, \\ \mathcal{J}_\theta^*(\mu_\Phi, \Sigma_\Phi) = \text{tr}(\Sigma_{\hat{\theta}_{\text{af}}}(\mu_\Phi, \Sigma_\Phi)), \end{cases} \quad (4.10a)$$

where the coefficient matrices are defined as

$$\Psi_\theta^*(\mu_\Phi, \Sigma_\Phi) = \Sigma_\theta \mu_\Phi \left(\mu_\Phi^\top \Sigma_\theta \mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_v \right)^{-1}, \quad \psi_\theta^*(\mu_\Phi, \Sigma_\Phi) = \mu_\theta - \Psi_\theta^*(\mu_\Phi, \Sigma_\Phi) \mu_\Phi^\top \mu_\theta, \quad (4.10b)$$

$\mathcal{M} : \mathbb{R}^{((N+1)(T+1))^2} \rightarrow \mathbb{R}^{(T+1)^2}$ is a linear operator whose $(t, t')^{\text{th}}$ component is

$$\mathcal{M}^{tt'}(\Sigma_\Phi) = \langle \Sigma_\Phi^{t't}, (\Sigma_\theta + \mu_\theta \mu_\theta^\top) \rangle, \quad (4.10c)$$

and Σ_v is the covariance matrix of the measurement noise sequence $v = [v_0, \dots, v_T]^\top$, which is the diagonal matrix $\Sigma_v = \text{diag}(\sigma_{v_0}^2, \dots, \sigma_{v_T}^2)$, when they are mutually independent.

Remark 4.2.3 (Functional dependence). *In principle, $\hat{\theta}_{\text{af}}(\cdot)$ is a function of three arguments, $(y, \mu_\Phi, \Sigma_\Phi)$. However, the notation used in (4.10a), i.e., $\hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)$, emphasizes the dependence of the estimator on the DBS collection (μ_Φ, Σ_Φ) in the last two arguments. The estimator $\hat{\theta}_{\text{af}}(\cdot)$ 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 4.1, the proof of Theorem 4.2.2 follows directly from the generic affine MMSE estimator of Lemma 4.2.1 by setting $\xi_1 = \theta$ and $\xi_2 = y$ (see Section 3.3.3 for a complete proof). While $\hat{\theta}_{\text{af}}(\cdot)$ is provably optimal within the affine class, the proximity of its MSE to that of the Bayesian MMSE estimator, $\hat{\theta}_{\text{opt}}$, 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 $\hat{\theta}_{\text{opt}}$ if the prior on the unknown parameters, $\mathbb{P}(\mu_\theta, \Sigma_\theta)$, and the distribution of the measurement noise sequence, $\mathbb{P}(0, \Sigma_v)$, belong to an elliptical family [140, Rem. 3.1] (alternatively, see Remark 3.3.1). As observed in the simple case of Example 4.1, $\hat{\theta}_{\text{af}}(\cdot)$ can track $\hat{\theta}_{\text{opt}}$ well for certain priors on x and ranges of y , particularly when the state uncertainty is small. Let us revisit Example 4.1 to see how Theorem 4.2.2 specializes to an explicit expression for the affine estimator.

Example 4.1 (continued). In the setting of (4.3), the dynamical system is one-dimensional, and the observation model involves the linear basis function $\phi(x) = x$. In this case, the DBS from Definition 4.1 simplify to

$$\mu_\Phi = \mu_x, \quad \Sigma_\Phi = \sigma_x^2,$$

and the affine MMSE estimator and its estimation-error covariance from Theorem 4.2.2 are

$$\begin{aligned}\hat{\theta}_{\text{af}}(y; \mu_x, \sigma_x^2) &= \mu_\theta + \frac{\sigma_\theta^2 \mu_x}{\mu_x^2 \sigma_\theta^2 + \sigma_x^2 \sigma_\theta^2 + \sigma_x^2 \mu_\theta^2 + \sigma_v^2} (y - \mu_x \mu_\theta), \\ \Sigma_{\hat{\theta}_{\text{af}}}(\mu_x, \sigma_x^2) &= \sigma_\theta^2 - \frac{\sigma_\theta^4 \mu_x^2}{\mu_x^2 \sigma_\theta^2 + \sigma_x^2 \sigma_\theta^2 + \sigma_x^2 \mu_\theta^2 + \sigma_v^2}.\end{aligned}\quad (4.11)$$

However, as shown in Figure 4.1, the Bayesian MMSE estimator for Example 4.1 can become highly nonlinear as the uncertainty in the system state increases. Consequently, the MSE of the affine estimator in (4.11) may be far from that of $\hat{\theta}_{\text{opt}}$. More generally, for hidden Markov processes (HMPs), the MSE decomposes into two terms [42, 43]: 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 4.2.4 (Lower bound on the MSE of the affine estimator). *Consider the affine MMSE estimator $\hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)$ and its optimal error $\mathcal{J}_\theta^*(\mu_\Phi, \Sigma_\Phi)$ as defined in (4.10a). The following inequality then holds:*

$$\mathcal{J}_\theta^*(\mu_\Phi, \Sigma_\Phi) \geq \mathbb{E}_x [\mathcal{J}_\theta^*(\Phi(x), 0)] = \mathbb{E}_x \left[\min_{\hat{\theta}(\cdot) \in \mathcal{A}_\theta} \mathbb{E} [\|\theta - \hat{\theta}(y)\|^2 | x] \right], \quad (4.12)$$

where x is the state trajectory, $\Phi(x)$ is the matrix collection of basis functions as in Definition 4.1, the set $\mathcal{A}_\theta$ denotes the class of affine estimators of θ from y , analogous to (4.6), 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 Section 4.6. Notably, the term $\mathcal{J}_\theta^*(\Phi(x), 0)$ in (4.12) 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 4.2.4 suggests a fundamental performance limit attainable by affine estimators when the state trajectory is perfectly known. In practice, however, the affine MMSE estimator (4.10) operates with DBS collection (μ_Φ, Σ_Φ) that summarize the distribution of the latent state trajectory, so its MSE generally remains strictly above this perfect-state benchmark and may be far from the Bayesian MMSE $\hat{\theta}_{\text{opt}}$. 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 4.2.4. Consequently, this study proposes a nonlinear parameter estimator by composing the affine MMSE mapping (4.10) with data-dependent DBS estimates. Let $\hat{\Phi}(y)$ and $\hat{\Sigma}_\Phi(y)$ denote estimates of the DBS collection components μ_Φ and Σ_Φ constructed from the measurements y . Substituting these estimates into affine

MMSE expressions $\hat{\theta}_{\text{af}}(\mathbf{y}; \cdot, \cdot)$ and $\Sigma_{\hat{\theta}_{\text{af}}}(\cdot, \cdot)$ in (4.10a) yields the nonlinear estimator

$$\begin{cases} \hat{\theta}_{\text{nl}}(\mathbf{y}) = \hat{\theta}_{\text{af}}(\mathbf{y}; \hat{\Phi}(\mathbf{y}), \Sigma_{\hat{\Phi}}(\mathbf{y})), \\ \Sigma_{\hat{\theta}_{\text{nl}}}(\mathbf{y}) = \Sigma_{\hat{\theta}_{\text{af}}}(\hat{\Phi}(\mathbf{y}), \Sigma_{\hat{\Phi}}(\mathbf{y})). \end{cases} \quad (4.13)$$

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

Remark 4.2.5 (Plug-in interpretation). *The nonlinear estimator (4.13) can be interpreted as a plug-in construction with respect to the DBS statistics. Whereas the affine MMSE estimator (4.10a) treats $(\mu_{\Phi}, \Sigma_{\Phi})$ as known hyperparameters, the nonlinear variant substitutes their data-dependent estimates $(\hat{\Phi}(\mathbf{y}), \Sigma_{\hat{\Phi}}(\mathbf{y}))$ 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 [27]. Consequently, the independence assumptions underlying the affine derivation no longer hold exactly, since the effective DBS now depend on $\mathbf{y}$. The covariance $\Sigma_{\hat{\theta}_{\text{nl}}}(\mathbf{y})$ 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.

4.3 FIXED-POINT CHARACTERIZATION

The nonlinear parameter estimator (4.13) depends critically on accurate DBS estimates $(\hat{\Phi}(\mathbf{y}), \Sigma_{\hat{\Phi}}(\mathbf{y}))$ that improve parameter estimation performance beyond the affine counterpart. Since $(\mu_{\Phi}, \Sigma_{\Phi})$ are determined by the latent state sequence, we investigate two distinct strategies: (i) direct estimation of the basis-function evaluations $\phi(\mathbf{x})$ in (4.2) over the entire trajectory, and (ii) estimation of the internal latent system states $\mathbf{x}$ 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.

4.3.1 DUAL BASIS-PARAMETER ESTIMATION

The fundamental limit implied by the lower bound in Lemma 4.2.4 suggests that DBS estimates should satisfy $\hat{\Phi}(\mathbf{y}) \rightarrow \Phi(\mathbf{x})$ and $\Sigma_{\hat{\Phi}}(\mathbf{y}) \rightarrow 0$ to improve the estimation performance of the nonlinear estimator (4.13) compared to the affine counterpart (4.10a). This subsection implements the first strategy by estimating the basis-function evaluations $\phi(\mathbf{x})$ in (4.2) via an affine MMSE estimator, and then forming $\hat{\Phi}(\mathbf{y})$ and $\Sigma_{\hat{\Phi}}(\mathbf{y})$ from the resulting basis estimates and their estimation-error covariance. The following lemma is analogous to Theorem 4.2.2 and formalizes this construction as an immediate consequence of the generic affine MMSE estimator described in Lemma 4.2.1.

Lemma 4.3.1 (DBS estimation via affine MMSE basis estimator). *Let $\Phi(\mathbf{x})$ be the unknown basis-function matrix over the trajectory, with prior DBS collection $(\mu_{\Phi}, \Sigma_{\Phi})$ as in Definition 4.1, and define the stacked basis-function vector $\phi := \text{vec}(\Phi(\mathbf{x})) = [\phi^{\top}(\mathbf{x}_0), \dots, \phi^{\top}(\mathbf{x}_T)]^{\top}$.*

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 $\hat{\phi}_{\text{af}}(\mathbf{y}; \mu_\theta, \Sigma_\theta) = \Psi_\phi^*(\mu_\theta, \Sigma_\theta)\mathbf{y} + \psi_\phi^*(\mu_\theta, \Sigma_\theta)$, where

$$\Psi_\phi^*(\mu_\theta, \Sigma_\theta) = \Sigma_\Phi \mu_\Theta (\mu_\Theta^\top \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v)^{-1}, \quad \psi_\phi^*(\mu_\theta, \Sigma_\theta) = \bar{\phi} - \Psi_\phi^*(\mu_\theta, \Sigma_\theta) \mu_\Phi \mu_\theta, \quad (4.14a)$$

with $\bar{\phi} = \text{vec}(\mu_\Phi)$, $\mu_\Theta = \text{diag}(\mu_\theta^0, \dots, \mu_\theta^T)$ where $\mu_\theta^t = \mu_\theta$ for each $t \in \{0, \dots, T\}$, Σ_v identical to that in (4.10), and the linear operator $\mathcal{S} : \mathbb{R}^{(N+1)^2} \rightarrow \mathbb{R}^{(T+1)^2}$ having $(t, t')^{\text{th}}$ component

$$\mathcal{S}^{tt'}(\Sigma_\theta) = \langle \Sigma_\theta, (\Sigma_\phi^{t'} + \mu_\phi^t \mu_{\phi'}^{\top}) \rangle. \quad (4.14b)$$

Consequently, the resulting DBS estimates and the corresponding MMSE are

$$\begin{cases} \hat{\Phi}_{\text{af}}(\mathbf{y}; \mu_\theta, \Sigma_\theta) = [\hat{\phi}_{\text{af}}^0(\mathbf{y}; \mu_\theta, \Sigma_\theta), \dots, \hat{\phi}_{\text{af}}^T(\mathbf{y}; \mu_\theta, \Sigma_\theta)], \\ \Sigma_{\hat{\Phi}_{\text{af}}}(\mu_\theta, \Sigma_\theta) = \Sigma_\Phi - \Sigma_\Phi \mu_\Theta (\mu_\Theta^\top \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v)^{-1} \mu_\Theta^\top \Sigma_\Phi, \\ \mathcal{J}_\phi^*(\mu_\theta, \Sigma_\theta) = \text{tr}(\Sigma_{\hat{\Phi}_{\text{af}}}(\mu_\theta, \Sigma_\theta)), \end{cases} \quad (4.14c)$$

where $\hat{\phi}_{\text{af}}^t(\mathbf{y}; \mu_\theta, \Sigma_\theta)$ is the estimate of $\phi(\mathbf{x}_t)$.

Remark 4.3.2 (Observation covariance forms). The Bayesian affine MMSE estimators for the parameters in (4.10) and for the basis collection in (4.14) both rely on the observation covariance

$$\Sigma_y = \mu_\Phi^\top \Sigma_\theta \mu_\Phi + \mathcal{M}(\Sigma_\Phi) + \Sigma_v = \mu_\Theta^\top \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v, \quad (4.15)$$

where $\mathcal{M}(\Sigma_\Phi)$ and $\mathcal{S}(\Sigma_\theta)$ are defined in (4.10c) and (4.14b), respectively. The two expressions follow from the lifted observation model written either as $\mathbf{y} = \Phi^\top(\mathbf{x})\theta + \mathbf{v}$ in Chapter 3 or $\mathbf{y} = \mathbf{G}_\theta^\top \phi + \mathbf{v}$ in (4.37), where $\phi = \text{vec}(\Phi(\mathbf{x}))$ and $\mathbf{G}_\theta = \text{diag}(\theta^0, \dots, \theta^T)$ with $\theta^t = \theta$ for all $t \in \{0, \dots, T\}$ (see also the proof of Lemma 4.3.1 in Section 4.6). 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 4.3.1, see Section 4.6. In the context of Example 4.1, the DBS estimation admits an explicit expression via Lemma 4.3.1. In this setting, because the basis function is linear, basis evaluations and states coincide. Hence, the DBS estimates in (4.14c) for a single measurement is essentially an optimal affine MMSE state estimator.

Example 4.1 (continued). Given the linear basis function $\phi(\mathbf{x}) = \mathbf{x}$ in (4.3), the DBS estimator described in (4.14c) is equal to the optimal affine MMSE state estimator

$$\begin{aligned} \hat{\Phi}_{\text{af}}(\mathbf{y}; \mu_\theta, \sigma_\theta^2) &= \mu_x + \frac{\sigma_x^2 \mu_\theta}{\mu_\theta^2 \sigma_x^2 + \sigma_\theta^2 \sigma_x^2 + \sigma_\theta^2 \mu_x^2 + \sigma_v^2} (\mathbf{y} - \mu_x \mu_\theta), \\ \Sigma_{\hat{\Phi}_{\text{af}}}(\mu_\theta, \sigma_\theta^2) &= \sigma_x^2 - \frac{\sigma_x^4 \mu_\theta^2}{\mu_\theta^2 \sigma_x^2 + \sigma_\theta^2 \sigma_x^2 + \sigma_\theta^2 \mu_x^2 + \sigma_v^2}. \end{aligned} \quad (4.16)$$

Observe that the operators in (4.10c) and (4.14b) evaluate to

$$\mathcal{M}(\sigma_x^2) = \sigma_x^2(\sigma_\theta^2 + \mu_\theta^2), \quad \mathcal{S}(\sigma_\theta^2) = \sigma_\theta^2(\sigma_x^2 + \mu_x^2),$$

and that the denominators in (4.11) and (4.16) are equal, representing the observation covariance for a scalar measurement.

The DBS estimates in (4.14c) 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 4.2.5 by treating the parameter estimates $(\hat{\theta}(y), \Sigma_{\hat{\theta}}(y))$ as fixed prior moments,

$$\begin{cases} \hat{\Phi}_{\text{nl}}(y) = \hat{\Phi}_{\text{af}}(y; \hat{\theta}(y), \Sigma_{\hat{\theta}}(y)), \\ \Sigma_{\hat{\Phi}_{\text{nl}}}(y) = \Sigma_{\hat{\Phi}_{\text{af}}}(\hat{\theta}(y), \Sigma_{\hat{\theta}}(y)). \end{cases} \quad (4.17)$$

The mutual substitution of (4.17) and (4.13) 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 $(\hat{\theta}_{\text{nl}}^*, \Sigma_{\hat{\theta}_{\text{nl}}}^*, \hat{\Phi}_{\text{nl}}^*, \Sigma_{\hat{\Phi}_{\text{nl}}}^*)$ such that

$$\begin{cases} \hat{\theta}_{\text{nl}}^* = \hat{\theta}_{\text{af}}(y; \hat{\Phi}_{\text{nl}}^*, \Sigma_{\hat{\Phi}_{\text{nl}}}^*), \\ \Sigma_{\hat{\theta}_{\text{nl}}}^* = \Sigma_{\hat{\theta}_{\text{af}}}(\hat{\Phi}_{\text{nl}}^*, \Sigma_{\hat{\Phi}_{\text{nl}}}^*), \end{cases} \quad \begin{cases} \hat{\Phi}_{\text{nl}}^* = \hat{\Phi}_{\text{af}}(y; \hat{\theta}_{\text{nl}}^*, \Sigma_{\hat{\theta}_{\text{nl}}}^*), \\ \Sigma_{\hat{\Phi}_{\text{nl}}}^* = \Sigma_{\hat{\Phi}_{\text{af}}}(\hat{\theta}_{\text{nl}}^*, \Sigma_{\hat{\theta}_{\text{nl}}}^*), \end{cases} \quad (\text{DB-P})$$

where the functions $(\hat{\theta}_{\text{af}}(\cdot), \Sigma_{\hat{\theta}_{\text{af}}}(\cdot))$ and $(\hat{\Phi}_{\text{af}}(\cdot), \Sigma_{\hat{\Phi}_{\text{af}}}(\cdot))$ are the closed form solutions of the Bayesian affine MMSE defined in (4.10a) and (4.14c), 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 (4.11) and (4.16) for Example 4.1, we illustrate the dual basis-parameter (DB-P) estimator on the same data and compare it with the affine and empirical MMSE estimators.

Example 4.1 (continued). The dual basis-parameter (DB-P) estimator for θ given a single measurement, constructed using the affine MMSE parameter and basis estimators in (4.11) and (4.16), is shown in green in Figure 4.4. The algorithm used to construct the dual basis-parameter estimator (green curve) is described in subsection 4.3.3. As can be seen, the resulting nonlinearity closely follows the one obtained from the empirical MMSE estimator (red curve).

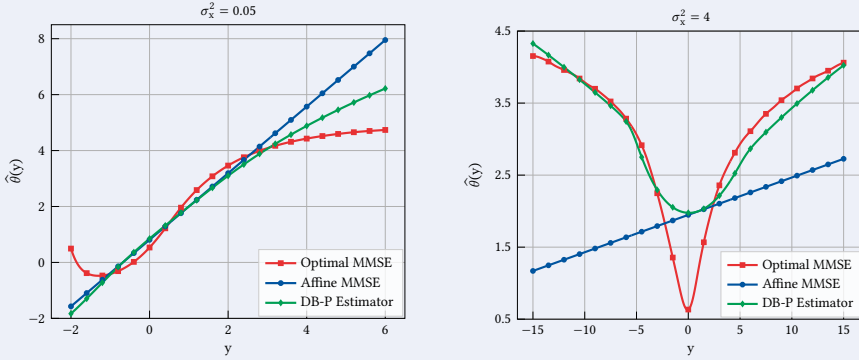


Figure 4.4: Dual basis-parameter (DB-P), optimal MMSE (4.5), and affine MMSE (4.10a) estimators.

In the above example, the basis function is 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.

Remark 4.3.3 (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 $\phi = \text{vec}(\Phi(x))$ 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, $(N + 1)(T + 1)$, typically exceeds the dimensionality of the state sequence x , which is $(n_x)(T + 1)$. 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 (4.1) 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.

4.3.2 DUAL STATE-PARAMETER ESTIMATION

In contrast to the previous approach, which directly estimates the basis collection as in (4.14c), this second approach constructs DBS estimates from the state estimate sequence $\hat{x}(y) = [\hat{x}_0^T(y), \dots, \hat{x}_T^T(y)]^T$ and its corresponding covariance $\Sigma_{\hat{x}}(y)$. This approach is particularly effective when the process noise is Gaussian, addressing a key limitation of the previous method discussed in Remark 4.3.3.

Assumption 4.1 (Gaussian process noise). *Throughout this subsection, the initial state and the process noise are drawn from a Gaussian distribution, i.e., $\mathbf{x}_0 \sim \mathcal{N}(\mu_{\mathbf{x}_0}, \Sigma_{\mathbf{x}_0})$ and $\mathbf{w}_{t+1} \sim \mathcal{N}(0, \Sigma_{\mathbf{w}_{t+1}})$.*

Under this assumption, the state trajectory admits a Gaussian prior distribution. We adopt the *lifted matrix* representation of the process model (4.1) as

$$\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{u} + \mathbf{w}), \quad (4.18)$$

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

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & 0 & 0 & \dots & 0 \\ \mathbf{A}_0 & \mathbf{I} & 0 & \dots & \vdots \\ \mathbf{A}_1\mathbf{A}_0 & \mathbf{A}_1 & \mathbf{I} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \prod_{i=0}^{T-1} \mathbf{A}_i & \prod_{i=1}^{T-1} \mathbf{A}_i & \dots & \mathbf{A}_{T-1} & \mathbf{I} \end{bmatrix}, \quad \mathbf{B} = \text{diag}(\mathbf{I}, \mathbf{B}_0, \dots, \mathbf{B}_{T-1}), \quad (4.19)$$

where $\prod_{i=t}^{t'} \mathbf{A}_i = \mathbf{A}_{t'} \dots \mathbf{A}_t$. The process noise vector $\mathbf{w} \sim \mathcal{N}(0, \Sigma_{\mathbf{w}})$ follows a block-diagonal covariance structure under temporal independence, $\Sigma_{\mathbf{w}} = \text{diag}(\Sigma_{\mathbf{x}_0}, \Sigma_{\mathbf{w}_1}, \dots, \Sigma_{\mathbf{w}_T})$. Temporal correlations may also be incorporated if the corresponding covariance matrix is known. With the lifted representation (4.18), the state trajectory prior becomes $\mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{B}\mathbf{u}, \mathbf{A}\Sigma_{\mathbf{w}}\mathbf{A}^\top)$, 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 4.1 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 4.2 (Gaussian DBS operator). *For a Gaussian state trajectory $\mathbf{v} \sim \mathcal{N}(\mu_{\mathbf{v}}, \Sigma_{\mathbf{v}})$, the DBS collection can be computed from only the mean and covariance as*

$$(\mu_{\Phi(\mathbf{v})}, \Sigma_{\Phi(\mathbf{v})}) = \Phi_{\#}(\mu_{\mathbf{v}}, \Sigma_{\mathbf{v}}), \quad (4.20)$$

where $\Phi_{\#} : \mathbb{R}^{n_{\mathbf{x}}(T+1)} \times \mathbb{R}^{(n_{\mathbf{x}}(T+1))^2} \rightarrow \mathbb{R}^{(N+1)(T+1)} \times \mathbb{R}^{((N+1)(T+1))^2}$ maps the first two moments of the state trajectory to the basis statistics as in Definition 4.1.

Given the linearity of the basis function in Example 4.1, the Gaussian DBS operator reduces to the identity map. To highlight its nontrivial behavior, we now consider another one-dimensional dynamical system in which the output function (4.2) 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 4.2 (1-D Wiener model with sinusoidal basis function). Consider the following sinusoidal nonlinearity in the measurement model:

$$y = \theta \sin(fx) + v. \quad (4.21)$$

where $x \in \mathbb{R}$ is a scalar variable representing the initial state, with $x \sim \mathbb{P}(\mu_x, \sigma_x^2)$, and f is a known frequency. The Gaussian DBS operator $\Phi_{\#}(\cdot, \cdot)$ in Definition 4.2 maps the statistics of x to the following DBS, evaluated using Lemma 3.4.1 and Example 3.1; see also [140, Lem. 4.1, Ex. 1]:

$$\begin{aligned} \mu_{\Phi} &= \exp\left(-\frac{1}{2}f^2\sigma_x^2\right) \sin(f\mu_x), \\ \Sigma_{\Phi} &= \frac{1}{2} - \frac{1}{2}\exp(-2f^2\sigma_x^2) \cos(2f\mu_x) - \exp(-f^2\sigma_x^2) \sin^2(f\mu_x). \end{aligned} \quad (4.22)$$

4

For the lifted prior in (4.18), the Gaussian DBS operator provides an equivalent computation of the basis statistics from Definition 4.1 via $(\mu_{\Phi}, \Sigma_{\Phi}) = \Phi_{\#}(\text{ABu}, \text{A}\Sigma_w\text{A}^{\top})$. Once state estimates $\hat{x}(y)$ and estimation-error covariance $\Sigma_{\hat{x}}(y)$ are obtained, we apply a plug-in interpretation analogous to Remark 4.2.5: the estimated mean and covariance are treated as fixed values of a Gaussian prior distribution for the subsequent DBS computation, i.e., $x \sim \mathcal{N}(\hat{x}(y), \Sigma_{\hat{x}}(y))$, analogous to an empirical Bayes procedure applied to the state trajectory. This procedure leads to data-driven DBS estimates via the Gaussian DBS operator,

$$(\hat{\Phi}(y), \Sigma_{\hat{\Phi}}(y)) = \Phi_{\#}(\hat{x}(y), \Sigma_{\hat{x}}(y)). \quad (4.23)$$

The affine MMSE state estimator is a natural choice for providing $\hat{x}(y)$ and $\Sigma_{\hat{x}}(y)$, 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 4.2.1 requires the cross-covariance Σ_{xy} between the state trajectory and observation sequence. Under Assumption 4.1, this cross-covariance can be explicitly derived using Stein's identity.

Remark 4.3.4 (Cross-covariance via Stein's identity). For a random vector $v \sim \mathcal{N}(\mu, \Sigma)$ and a differentiable function h with appropriate integrability conditions (sufficient for applying Fubini's theorem [98]), Stein's identity [135] states

$$\mathbb{E}[(v - \mu)h(v)] = \Sigma \mathbb{E}[\nabla_v h(v)]. \quad (4.24)$$

While this study applies Stein's identity under Gaussian noise, the identity extends to elliptical distributions [75], enabling broader applicability to process noise models. Additionally, extensions to other distributions such as Poisson and t -distributions exist [85], suggesting potential avenues for non-elliptical process noise distributions.

The following lemma leverages this identity to derive the closed-form affine MMSE state estimator.

Lemma 4.3.5 (Affine MMSE state estimator). *Under Assumption 4.1, suppose that for each basis function $\phi_n(\cdot)$ defined in (4.2), the gradient vector $\nabla_{x_i} \phi_n(x_t)$, for $t \in \{0, \dots, T\}$, has components that are continuous almost everywhere and satisfy the following condition:*

$$\nabla_{x_i} \phi_n(x_t) = \left[\frac{\partial \phi_n(x_t)}{\partial x_{t,1}}, \dots, \frac{\partial \phi_n(x_t)}{\partial x_{t,n_x}} \right]^\top, \quad \mathbb{E} \left[\left| \frac{\partial \phi_n(x_t)}{\partial x_{t,i}} \right| \right] < \infty, \quad n = 0, \dots, N, \quad i = 1, \dots, n_x, \quad (4.25)$$

where $x_{t,i}$ denotes the i^{th} element of the state vector $x_t \in \mathbb{R}^{n_x}$. Then, the affine MMSE estimator for the state trajectory (4.18), its estimation-error covariance, and the optimal MSE, all as functions of $(\mu_\theta, \Sigma_\theta)$, are given by

4

$$\begin{cases} \hat{x}_{\text{af}}(y; \mu_\theta, \Sigma_\theta) = \Psi_x^*(\mu_\theta, \Sigma_\theta)y + \psi_x^*(\mu_\theta, \Sigma_\theta), \\ \Sigma_{\hat{x}_{\text{af}}}(\mu_\theta, \Sigma_\theta) = A\Sigma_w A^\top - A\Sigma_w A^\top C^\top \mu_\Theta (\mu_\Theta^\top \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v)^{-1} \mu_\Theta^\top C A \Sigma_w A^\top, \\ \mathcal{J}_x^*(\mu_\theta, \Sigma_\theta) = \text{tr}(\Sigma_{\hat{x}_{\text{af}}}(\mu_\theta, \Sigma_\theta)), \end{cases} \quad (4.26a)$$

with the optimal coefficient matrices

$$\Psi_x^*(\mu_\theta, \Sigma_\theta) = A\Sigma_w A^\top C^\top \mu_\Theta (\mu_\Theta^\top \Sigma_\Phi \mu_\Theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v)^{-1}, \quad \psi_x^*(\mu_\theta, \Sigma_\theta) = ABu - \Psi_x^*(\mu_\theta, \Sigma_\theta) \mu_\Phi \mu_\theta, \quad (4.26b)$$

where $(\mu_\Phi, \Sigma_\Phi) = \Phi_\#(ABu, A\Sigma_w A^\top)$ is the DBS collection for the Gaussian state trajectory prior from Definition 4.2, $\mu_\Theta = \text{diag}(\mu_\theta^0, \dots, \mu_\theta^T)$ with $\mu_\theta^t = \mu_\theta$ for $t \in \{0, \dots, T\}$, the linear operator $\mathcal{S}(\cdot)$ is defined in (4.14b), and

$$C = \text{diag}(C_0, \dots, C_T), \quad C_t = \mathbb{E}[J_\phi(x_t)], \quad (4.26c)$$

with $J_\phi(x_t) = [\nabla_{x_i} \phi_0(x_t), \dots, \nabla_{x_i} \phi_N(x_t)]^\top$ being the Jacobian of basis functions evaluated at x_t .

A detailed proof is presented in Section 4.6. The explicit computation and tractability of the affine MMSE state estimator depend on the choice of basis functions, since the closed-form expression (4.26a) requires both the DBS (μ_Φ, Σ_Φ) and the expected Jacobian of the basis functions for constructing the matrix C in (4.26c). The sinusoidal basis of Example 4.2 satisfies the bounded expectation condition in (4.25), and the corresponding matrix C admits a closed-form expression, as discussed below.

Example 4.2 (continued). In the context of (4.21), the matrix C is given by

$$C = \mathbb{E} \left[\frac{\partial \phi(x)}{\partial x} \right] = \mathbb{E}[f \cos(fx)] = f \exp\left(-\frac{1}{2} f^2 \sigma_x^2\right) \cos(f\mu_x), \quad (4.27)$$

where $J_\phi(x)$ in (4.26c) reduces to a scalar derivative since x is a scalar variable. Given the explicit expressions for the DBS quantities in (4.22) and for C in (4.27),

the affine MMSE state estimator (4.26a) yields

$$\begin{aligned}\hat{x}_{\text{af}}(y; \mu_\theta, \Sigma_\theta) &= \mu_x + \frac{\sigma_x^2 C \mu_\theta}{\mu_\theta^2 \Sigma_\Phi + \sigma_\theta^2 \Sigma_\Phi + \sigma_\theta^2 \mu_\Phi^2 + \sigma_v^2} (y - \mu_\Phi \mu_\theta), \\ \Sigma_{\hat{x}_{\text{af}}}(\mu_\theta, \Sigma_\theta) &= \sigma_x^2 - \frac{\sigma_x^4 C^2 \mu_\theta^2}{\mu_\theta^2 \Sigma_\Phi + \sigma_\theta^2 \Sigma_\Phi + \sigma_\theta^2 \mu_\Phi^2 + \sigma_v^2}.\end{aligned}\quad (4.28)$$

Furthermore, the affine MMSE parameter and basis estimators for the setting of (4.21) follow their descriptions in Example 4.1, given in (4.11) and (4.16), respectively, after replacing the pair (μ_x, σ_x^2) in those expressions with the DBS quantities obtained in (4.22).

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 [140, Lem. 4.1, Ex. 1] (alternatively, see Lemma 3.4.1 and Example 3.1), relying on the lifted representation (4.18). 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 (4.23).

Remark 4.3.6 (Lifted plug-in prior for DBS estimates). *An equivalent representation of the affine MMSE state estimator in Lemma 4.3.5 is*

$$\hat{x}_{\text{af}}(y; \mu_\theta, \Sigma_\theta) = A(Bu + \zeta), \quad \Sigma_{\hat{x}_{\text{af}}}(\mu_\theta, \Sigma_\theta) = A(\Sigma_w - Z)A^\top,$$

with measurement-driven corrections

$$\begin{cases} \zeta = \Sigma_w A^\top C^\top \mu_\theta (\mu_\theta^\top \Sigma_\Phi \mu_\theta + S(\Sigma_\theta) + \Sigma_v)^{-1} (y - \mu_\Phi^\top \mu_\theta), \\ Z = \Sigma_w A^\top C^\top \mu_\theta (\mu_\theta^\top \Sigma_\Phi \mu_\theta + S(\Sigma_\theta) + \Sigma_v)^{-1} \mu_\theta^\top C A \Sigma_w. \end{cases}$$

Thus, the plug-in empirical Bayes Gaussian prior on x induced by the affine estimator can be written in lifted form as $x \sim \mathcal{N}(A(Bu + \zeta), A(\Sigma_w - Z)A^\top)$. For Fourier basis functions, this structure results in closed-form DBS estimates in (4.23) via the explicit characteristic-function formulas of Example 3.1.

The affine state estimator in (4.26a) 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 4.2.5 and construct a nonlinear state estimator by inserting the parameter estimates $(\hat{\theta}(y), \Sigma_{\hat{\theta}}(y))$ as if they were the true prior moments,

$$\begin{cases} \hat{x}_{\text{nl}}(y) = \hat{x}_{\text{af}}(y; \hat{\theta}(y), \Sigma_{\hat{\theta}}(y)), \\ \Sigma_{\hat{x}_{\text{nl}}}(y) = \Sigma_{\hat{x}_{\text{af}}}(\hat{\theta}(y), \Sigma_{\hat{\theta}}(y)). \end{cases}\quad (4.29)$$

Using (4.29) within the DBS estimates construction (4.23) for the nonlinear parameter estimator (4.13), and simultaneously feeding the resulting parameter estimates back into (4.29), 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 $(\hat{\theta}_{nl}^*, \Sigma_{\hat{\theta}_{nl}}^*, \hat{x}_{nl}^*, \Sigma_{\hat{x}_{nl}}^*)$ such that

$$\begin{cases} \hat{\theta}_{nl}^* = \hat{\theta}_{af}(y; \Phi_{\#}(\hat{x}_{nl}^*, \Sigma_{\hat{x}_{nl}}^*)), \\ \Sigma_{\hat{\theta}_{nl}}^* = \Sigma_{\hat{\theta}_{af}}(\Phi_{\#}(\hat{x}_{nl}^*, \Sigma_{\hat{x}_{nl}}^*)), \end{cases} \quad \begin{cases} \hat{x}_{nl}^* = \hat{x}_{af}(y; \hat{\theta}_{nl}^*, \Sigma_{\hat{\theta}_{nl}}^*) \\ \Sigma_{\hat{x}_{nl}}^* = \Sigma_{\hat{x}_{af}}(\hat{\theta}_{nl}^*, \Sigma_{\hat{\theta}_{nl}}^*) \end{cases} \quad (\text{DS-P})$$

where $\Phi_{\#}(\hat{x}_{nl}^*, \Sigma_{\hat{x}_{nl}}^*)$ returns the DBS estimates $(\hat{\Phi}(y), \Sigma_{\hat{\Phi}}(y))$ according to (4.23) and Definition 4.2, and the functions $(\hat{\theta}_{af}(\cdot), \Sigma_{\hat{\theta}_{af}}(\cdot))$ and $(\hat{x}_{af}(\cdot), \Sigma_{\hat{x}_{af}}(\cdot))$ are the closed form solutions of the Bayesian affine MMSE defined in (4.10a) and (4.26a), 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 4.3.3, the asymptotic performance of the parameter estimates with infinitely many samples depends critically on the properties of the underlying dynamical system [140, Prop. 4.2] (alternatively, see Proposition 3.4.2).

Remark 4.3.7 (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 Proposition 3.4.2. 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 4.4 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 4.2) 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 behavior in a concrete setting, we now revisit Example 4.2 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 4.2 (continued). For the parameter-estimation simulations, consider the observation model (4.21) with sinusoidal frequency $f = \frac{\pi}{6}$, a scalar latent state $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ (interpreted as the initial state of a dynamical system) with $\mu_x = 0.5$ and two choices of variance, $\sigma_x^2 = 0.05$ (left plot) and $\sigma_x^2 = 4$ (right plot), an unknown parameter $\theta \in \mathbb{R}$ with prior $\theta \sim \mathcal{U}(-1, 5)$, and measurement noise $v \sim \mathcal{N}(0, \sigma_v^2)$ with $\sigma_v^2 = 0.16$. Figure 4.5 shows the estimates of the unknown parameter θ as a function of the measurement value 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 (4.10a) (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.

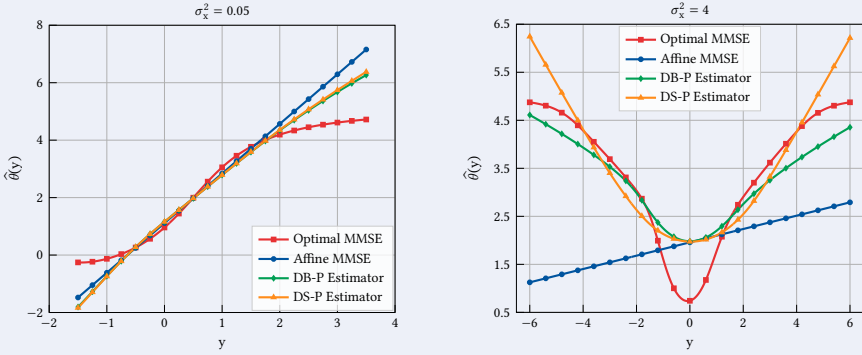


Figure 4.5: Dual state-parameter (DS-P), Dual basis-parameter (DB-P), optimal MMSE (4.5), and affine MMSE (4.10a) estimators.

4.3.3 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 ζ^* of an operator $\mathfrak{F}$ satisfying $\zeta^* = \mathfrak{F}(\zeta^*)$. We collect the estimates and their covariances into a tuple ζ . For the DB-P estimator, $\zeta = (\hat{\theta}_{\text{nl}}, \Sigma_{\hat{\theta}_{\text{nl}}}, \hat{\Phi}_{\text{nl}}, \Sigma_{\hat{\Phi}_{\text{nl}}})$, whereas for the DS-P estimator, $\zeta = (\hat{\theta}_{\text{nl}}, \Sigma_{\hat{\theta}_{\text{nl}}}, \hat{x}_{\text{nl}}, \Sigma_{\hat{x}_{\text{nl}}})$. For a given measurement sequence y , the corresponding fixed-point operators are defined as

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

respectively. A fixed point can be computed via the simultaneous (Jacobi-style) iteration $\zeta^{(k+1)} = \mathfrak{F}(\zeta^{(k)})$, whose pseudo-code for both variants is given in Algorithm 1. The following averaged update with $0 < \alpha \leq 1$ may improve stability, albeit with slower convergence:

$$\zeta^{(k+1)} = \zeta^{(k)} + \alpha(\mathfrak{F}(\zeta^{(k)}) - \zeta^{(k)}), \quad (4.30)$$

where $\alpha = 1$ recovers the Jacobi-style iteration used in Algorithm 1. Alternating (Gauss-Seidel-style) updates, which update the components of ζ sequentially, are also common for coupled fixed-point problems and may improve convergence [23, Ch. 7].

Remark 4.3.8 (Convergence condition). *The convergence of the sequence of iterates $\{\zeta^{(k)}\}_{k \geq 0}$ in (4.30) depends on the properties of the operator $\mathfrak{F}$. A sufficient condition is that $\mathfrak{F}$ is nonexpansive [12, 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 behavior.*

The iterative scheme in Algorithm 1 initializes the means and covariances of the parameters and DBS-related quantities from their respective priors. At iteration $k = 1$, the tuple $\zeta^{(1)}$ is obtained from a single pass of the Bayesian affine MMSE estimators for the unknown parameters and either the DBS or the system states, depending on the chosen variant. In the DB-P variant, the DBS is computed only at initialization from the prior of the system state $\mathbf{x}$, whereas in the DS-P variant, it is recomputed at every iteration by applying the Gaussian DBS operator $\Phi_{\#}(\cdot, \cdot)$ to the current state mean and covariance estimates. In the latter case, the matrix $\mathbf{C}$ in (4.26c) is evaluated only once, since it depends solely on the prior of the system state $\mathbf{x}$. The stopping criterion of Algorithm 1 is either a maximum number of iterations K or a threshold ϵ on the change in the parameter MSE cost $\mathcal{J}_{\theta}(\cdot, \cdot)$ in (4.10a); that is, the iteration terminates when $|\mathcal{J}_{\theta}^{(k)} - \mathcal{J}_{\theta}^{(k-1)}| < \epsilon$.

Remark 4.3.9 (Per-iteration computational complexity). *Both nonlinear estimators, dual basis-parameter (DB-P) and dual state-parameter (DS-P), have a per-iteration computational complexity of $\mathcal{O}(T^3)$ when $N, n_x \ll T$. This complexity is 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: $(\hat{\theta}_{\text{nl}}, \Sigma_{\hat{\theta}_{\text{nl}}})$ and **DB-P:** $(\hat{\Phi}_{\text{nl}}, \Sigma_{\hat{\Phi}_{\text{nl}}})$ or **DS-P:** $(\hat{\mathbf{x}}_{\text{nl}}, \Sigma_{\hat{\mathbf{x}}_{\text{nl}}})$

- 1: $(\hat{\theta}_{\text{nl}}^{(0)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(0)}) \leftarrow (\mu_{\theta}, \Sigma_{\theta})$ ▷ θ prior information
 - 2: $\mathcal{J}_{\theta}^{(0)} \leftarrow 0$ ▷ initial cost
 - 3: $(\hat{\Phi}_{\text{nl}}^{(0)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(0)}) \leftarrow (\mu_{\Phi}, \Sigma_{\Phi})$ ▷ DBS, Definition 4.1
 - 4: **for** $k = 1, 2, \dots, K$ **do**
 - 5: $(\hat{\theta}_{\text{nl}}^{(k)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(k)}) \leftarrow (\hat{\theta}_{\text{af}}(\mathbf{y}; \hat{\Phi}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(k-1)}), \Sigma_{\hat{\theta}_{\text{af}}}(\hat{\Phi}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(k-1)}))$ ▷ affine MMSE, (4.10a)
 - 6: $\mathcal{J}_{\theta}^{(k)} \leftarrow \mathcal{J}_{\theta}^*(\hat{\Phi}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(k-1)})$ ▷ associated cost
 - 7: **DB-P:** $(\hat{\Phi}_{\text{nl}}^{(k)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(k)}) \leftarrow (\hat{\Phi}_{\text{af}}(\mathbf{y}; \hat{\theta}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(k-1)}), \Sigma_{\hat{\Phi}_{\text{af}}}(\hat{\theta}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(k-1)}))$ ▷ affine MMSE, (4.14c)
 - 8: **DS-P:** $(\hat{\mathbf{x}}_{\text{nl}}^{(k)}, \Sigma_{\hat{\mathbf{x}}_{\text{nl}}}^{(k)}) \leftarrow (\hat{\mathbf{x}}_{\text{af}}(\mathbf{y}; \hat{\theta}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(k-1)}), \Sigma_{\hat{\mathbf{x}}_{\text{af}}}(\hat{\theta}_{\text{nl}}^{(k-1)}, \Sigma_{\hat{\theta}_{\text{nl}}}^{(k-1)}))$ ▷ affine MMSE, (4.26a)
 - 9: **DS-P:** $(\hat{\Phi}_{\text{nl}}^{(k)}, \Sigma_{\hat{\Phi}_{\text{nl}}}^{(k)}) \leftarrow \Phi_{\#}(\hat{\mathbf{x}}_{\text{nl}}^{(k)}, \Sigma_{\hat{\mathbf{x}}_{\text{nl}}}^{(k)})$ ▷ Gaussian DBS, Definition 4.2
 - 10: **if** $|\mathcal{J}_{\theta}^{(k)} - \mathcal{J}_{\theta}^{(k-1)}| < \epsilon$ **then** ▷ stopping criterion
 - 11: **break**
 - 12: **end if**
 - 13: **end for**
-

4.4 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 4.2.2), hereafter called A-MMSE, across two setups: ($N = 2, n_x = 10$) with more states than parameters, and ($n_x = 2, N = 10$) 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) [81] and SMC-EM [127]. 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 [14]. While modern methods such as nested sequential Monte Carlo (NSMC) [100] can improve SMC scalability in high dimensions, we restrict the benchmark comparisons to the second setup ($n_x = 2, N = 10$) for two reasons: (i) the first setup primarily illustrates a regime where the dual basis-parameter (DB-P) estimator underperforms the affine MMSE estimator, demonstrating the limitations of the elliptical-distribution approximation discussed in Remark 4.3.3; (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 ($n_x = 2, N = 10$) 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 ($A_t = I$) with a true function in the Fourier subspace

$$\begin{cases} \phi_0(x) = 1 & n = 0 \\ \phi_n(x) = \sum_{\ell \in \{-1, 1\}} \exp(j\ell f_n, x) & n \geq 1, \end{cases} \quad (4.31)$$

where $f_n \in \mathbb{R}^{n_x}$ 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: $v_t \sim \mathcal{N}(0, \sigma_v^2 I)$, $w_{t+1} \sim \mathcal{N}(0, \sigma_w^2 I)$, and $x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2 I)$, with $\sigma_{x_0}^2 = \sigma_w^2$. To assess the impact of process noise, we vary $\sigma_w^2 \in \{0.001, 0.01\}$ while holding the measurement-noise variance fixed at $\sigma_v^2 = 0.01$ across all experiments. The input trajectories are generated using the active-learning procedure from Section 3.5, 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 $\|\theta - \hat{\theta}(y)\|^2$ 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, $\mathbb{E}[\|\theta - \hat{\theta}(y)\|^2]$, for each method. Non-histogram plots include shaded regions indicating the 15th-85th percentile range of the squared error. Histograms display the probability density of $\log \|\theta - \hat{\theta}(y)\|^2$ 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 ($N = 2, n_x = 10$). This setup considers an LTI system with $n_x = 10$ states and $n_u = 2$ control inputs. The system dynamics are $A_t = I$ and $B_t = \Delta t[I, \dots, I]^\top$, 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, $\mu_{x_0} \sim \mathcal{N}(0, I)$ and the input sequence $u_t = 4.5 \sum_{v \in Y} [\cos(vt\Delta t), \sin(vt\Delta t)]^\top$, where $Y \in \{3, 5, 10, 20, 100\}$, is used to initialize the projected steepest descent algorithm in the active-learning procedure of Section 3.5, with each component constrained to $[-200, 200]$. The optimal input u^* is precomputed once and used for all methods evaluated under this configuration. The unknown parameters follow the Fourier basis representation in (4.31). 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 $f_n \sim \mathcal{N}(0, I)$ 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 $\sigma_{\theta_n}^2 = 3$. True parameter values are generated from this same prior.

Benchmark 1.1 (High-dimensional states). This benchmark uses Experiment setup 1 with trajectory length $T = 100$ (i.e., 101 measurements) and two process-noise variances $\sigma_w^2 \in \{0.001, 0.01\}$, with 10,000 simulations per configuration. Using identical realizations of (w, v) across methods, we compute parameter estimates $\hat{\theta}(y)$ from A-MMSE (4.10a) and the dual estimators DB-P and DS-P. Figure 4.6 shows histograms of $\log \|\theta - \hat{\theta}(y)\|^2$ 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 elliptical prior approximation used in the basis estimation (cf. Remark 4.3.3). With a scalar measurement at each time and a high state dimension ($n_x = 10$), 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 process noise is Gaussian.

Experiment setup 2 ($N = 10, n_x = 2$). This setup focuses on the complementary configuration of low-dimensional states and a larger number of parameters. The LTI system has two states, $x_t \in \mathbb{R}^2$, and shares the same structure as Experiment setup 1: $u_t \in \mathbb{R}^2$, $A_t = I$, and $B_t = \Delta t I$, with $\Delta t = 0.1$. The initial state is set to $\mu_{x_0} = [3.2, 2.8]^\top$. The active-learning procedure uses the same input initialization structure $u_t = 4.5 \sum_{v \in Y} [\cos(vt\Delta t), \sin(vt\Delta t)]^\top$, with $Y \in \{3, 5, 10, 20, 100\}$ and bounds $[-200, 200]$, and the resulting optimal trajectory u^* is used for all methods in this setup. Under the Fourier basis representation in (4.31), we use eleven unknown parameters θ_n ($n = 0, \dots, 10$) with known frequency vectors $f_n \in \mathbb{R}^2$ defined as: $f_0 = [0, 0]^\top$, $f_n = [n \frac{2\pi}{10}, 0]^\top$ for $n \in \{1, 2, 3\}$, and $f_n = [(n-7) \frac{2\pi}{10}, \frac{2\pi}{6}]^\top$ for $n \in \{4, \dots, 10\}$. The unknown parameters follow the prior $\mathcal{U}(2, 8)$, corresponding to $\mu_{\theta_n} = 5$ and $\sigma_{\theta_n}^2 = 3$, from which true parameter values are sampled.

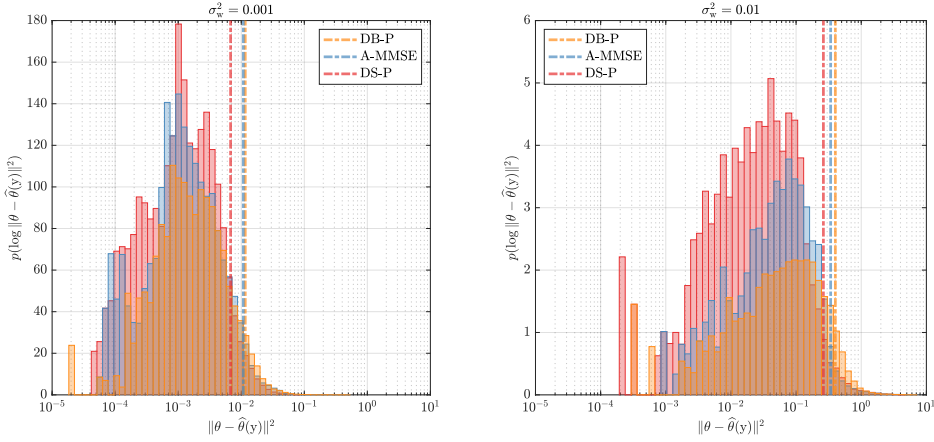


Figure 4.6: Parameter estimation squared-error distributions for Experiment setup 1.

Benchmark 2.1 (High-dimensional parameters). This benchmark uses Experiment setup 2 with trajectory length $T = 100$ (i.e., 101 measurements) and two process-noise variances $\sigma_w^2 \in \{0.001, 0.01\}$, with 10,000 simulations per configuration. Using identical realizations of (w, v) across methods, we compute parameter estimates $\hat{\theta}(y)$ from A-MMSE (4.10a) and the dual estimators DB-P and DS-P. Figure 4.7 shows histograms of $\log \|\theta - \hat{\theta}(y)\|^2$ 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 ($n_x = 2$) 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.

Benchmark 2.2 (Trajectory horizon). 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 4.8 compares the squared errors for A-MMSE (orange), DB-P (blue), and DS-P (red) under two process-noise variances $\sigma_w^2 \in \{0.001, 0.01\}$. Shaded regions show the 15th-85th percentile range of the squared error, while dashed lines indicate the empirical MSE. All estimators use the identical optimized input u^* 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 behavior aligns with the established limitations of A-MMSE for marginally stable systems (cf. Proposition 3.4.2), 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

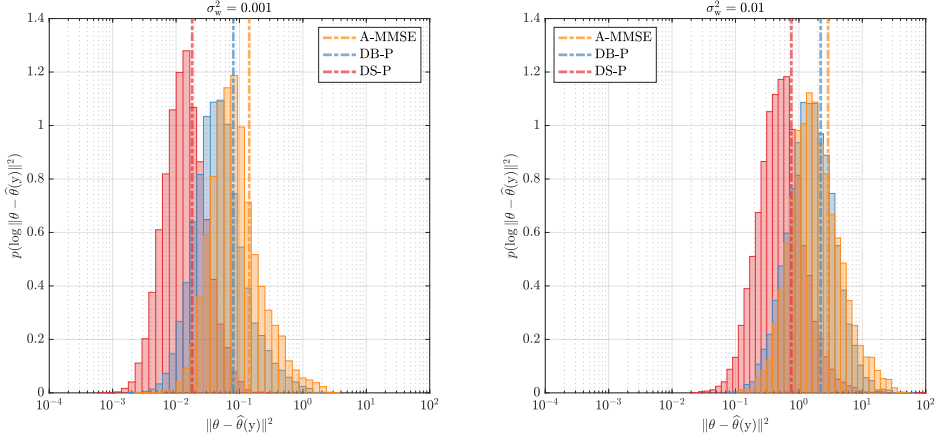


Figure 4.7: Parameter estimation squared-error distributions for Experiment setup 2.

grows. In the low process-noise case ($\sigma_w^2 = 0.001$), the red dashed line decreases steadily with increasing T . For higher process noise ($\sigma_w^2 = 0.01$), 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 85th percentile) of the squared-error distribution, as the central 15th-85th 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.

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 $\sigma_w^2 \in \{0.001, 0.01\}$. Identical realizations of (w, v) and the same optimized input u^* from the active-learning procedure are used for all three methods. In summary, the two SMC-based approaches are:

- (i) **PGAS** [81, 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 [81, Alg. 3] given θ^{k-1} , and (ii) sampling θ^k from the conditional posterior $p(\theta | x^k, y)$. 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 [77] as their conditional posterior, and the proposal is accepted or rejected via an MH step. The final estimate is the empirical mean of the retained samples, after discarding the first 2500 burn-in samples.
- (ii) **SMC-EM** [127, 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

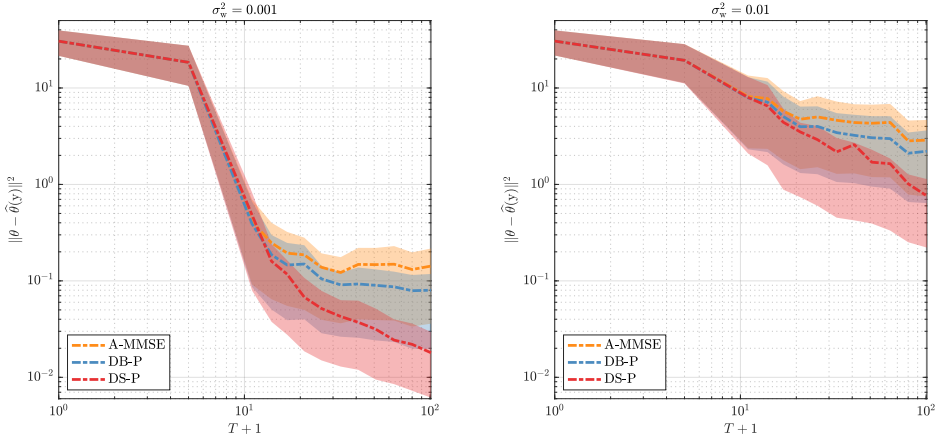


Figure 4.8: Parameter estimation squared-error vs. trajectory length T for Experiment setup 2.

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 $\mathbf{x}$; 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 [127, Alg. 2-3] is replaced by the Metropolis-Hastings forward-filtering backward-proposing (MH-FFBP) joint smoother [22, Fig. 5], which is an MCMC-based refinement of forward filtering backward smoothing methods [40, 57]. We use a standard particle filter with $N_F = 6000$ particles in the forward pass [22, 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 [22] 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 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 $\sigma_w^2 = 0.001$, the state-parameter variant (DS-P) of Algorithm 1 requires at most 56 iterations to satisfy the stopping criterion across all experiments. For $\sigma_w^2 = 0.01$, only 3 out of 10,000 experiments reach the maximum

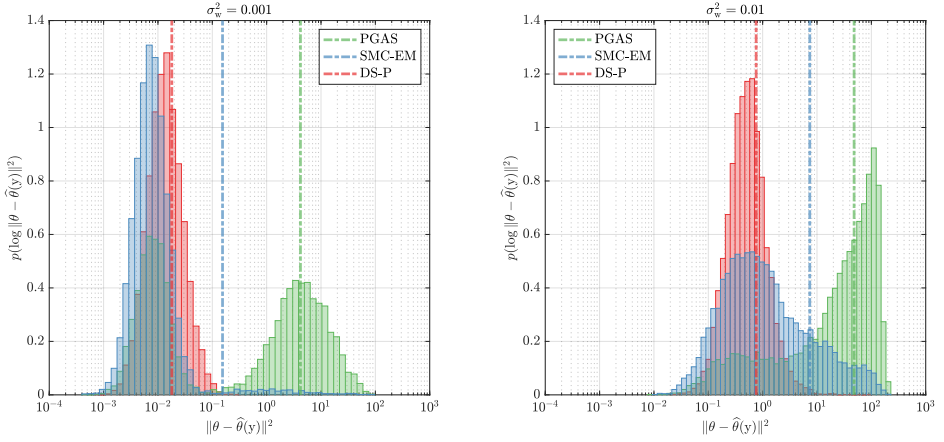


Figure 4.9: Squared-error distributions: dual state-parameter vs. SMC-based methods.

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 4.9 compares the distributions of $\log \|\theta - \hat{\theta}(y)\|^2$ 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 ($\sigma_w^2 = 0.001$), SMC-EM produces an error distribution that overlaps partially 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.

4.5 CONCLUSION

This chapter presented 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 refine their estimates iteratively. Two variants were 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 validated 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 4.8). 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.

4.6 TECHNICAL PROOFS

This section presents detailed proofs of the theoretical results introduced in this chapter.

Proof of Lemma 4.2.1. The proof follows the first-order optimality over the constants of the affine estimator (4.6). 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

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

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 $\psi^* = \mu_1 - \Psi\mu_2$. With this choice, applying the trace operator, and defining the covariances $\Sigma_{11} := \mathbb{E}[\Delta\xi_1\Delta\xi_1^\top]$, $\Sigma_{12} := \mathbb{E}[\Delta\xi_1\Delta\xi_2^\top]$, and $\Sigma_{22} := \mathbb{E}[\Delta\xi_2\Delta\xi_2^\top]$, where $\Delta\xi_1 = \xi_1 - \mu_1$ and $\Delta\xi_2 = \xi_2 - \mu_2$, the optimization reduces to minimizing

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

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

$$\left. \frac{\partial \mathcal{J}}{\partial \Psi} \right|_{\Psi=\Psi^*} = -2\Sigma_{12} + 2\Psi^*\Sigma_{22} = 0,$$

which yields $\Psi^* = \Sigma_{12}\Sigma_{22}^{-1}$. 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 Ψ^* and $\psi^* = \mu_1 - \Psi^*\mu_2$

into the estimation-error covariance $\Sigma_{\hat{\xi}_1} := \mathbb{E}[(\xi_1 - \Psi\xi_2 - \psi)(\xi_1 - \Psi\xi_2 - \psi)^\top]$, yields the expressions in (4.7a), which completes the proof. $\square$

Proof of Lemma 4.2.4. By the law of total expectation [28, Ch. 4],

$$\mathcal{J}_\theta^*(\mu_\Phi, \Sigma_\Phi) = \mathbb{E}[\|\theta - \hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)\|^2] = \mathbb{E}_x \left[\mathbb{E}[\|\theta - \hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)\|^2 | x] \right], \quad (4.32)$$

where $\hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)$ is the affine MMSE estimator with its optimal MSE $\mathcal{J}_\theta^*(\mu_\Phi, \Sigma_\Phi)$ as defined in (4.10a). Let $\mathcal{A}_\theta$ denote the class of affine estimators of θ from y , analogous to (4.6). Since $\hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi) \in \mathcal{A}_\theta$, the conditional MSE of this estimator cannot be smaller than the minimum conditional MSE over all affine estimators in $\mathcal{A}_\theta$ for each fixed x , i.e.,

$$\mathbb{E}_x \left[\mathbb{E}[\|\theta - \hat{\theta}_{\text{af}}(y; \mu_\Phi, \Sigma_\Phi)\|^2 | x] \right] \geq \mathbb{E}_x \left[\min_{\hat{\theta}(\cdot) \in \mathcal{A}_\theta} \mathbb{E}[\|\theta - \hat{\theta}(y)\|^2 | x] \right]. \quad (4.33)$$

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

$$\begin{cases} \hat{\theta}_{\text{af}}(y; \Phi(x), 0) = \operatorname{argmin}_{\hat{\theta}(\cdot) \in \mathcal{A}_\theta} \mathbb{E}[\|\theta - \hat{\theta}(y)\|^2 | x], \\ \mathcal{J}_\theta^*(\Phi(x), 0) = \mathbb{E}[\|\theta - \hat{\theta}_{\text{af}}(y; \Phi(x), 0)\|^2 | x]. \end{cases} \quad (4.34)$$

Combining (4.32), (4.33), and (4.34) arrives at the desired bound in (4.12). $\square$

Proof of Lemma 4.3.1. Define the stacked basis-function vector

$$\phi := \operatorname{vec}(\Phi(x)) = [\phi^\top(x_0), \dots, \phi^\top(x_T)]^\top, \quad (4.35)$$

where $\phi(x_t)$ is defined in (4.2). By Definition 4.1, the prior DBS (μ_Φ, Σ_Φ) induce the prior mean and covariance of ϕ as

$$\bar{\phi} := \mathbb{E}[\phi] = \mathbb{E}[\operatorname{vec}(\Phi(x))] = \operatorname{vec}(\mu_\Phi), \quad \mathbb{E}[(\phi - \bar{\phi})(\phi - \bar{\phi})^\top] = \Sigma_\Phi. \quad (4.36)$$

Next, introduce the lifted observation model over the trajectory. Stacking the sequence of measurements as $y = [y_0, \dots, y_T]^\top$ and using (4.2) yields

$$y = G_\theta^\top \phi + v, \quad (4.37)$$

where $G_\theta = \operatorname{diag}(\theta^0, \dots, \theta^T)$ with $\theta^t = \theta$ for all $t \in \{0, \dots, T\}$, and $v = [v_0, \dots, v_T]^\top$ has covariance matrix Σ_v . The affine MMSE estimator of ϕ given y follows from the generic affine MMSE estimator of Lemma 4.2.1 with $\xi_1 = \phi$ and $\xi_2 = y$, resulting in the form $\hat{\phi}_{\text{af}}(y) = \Psi_\phi^* y + \psi_\phi^*$, with coefficients (cf. (4.7b))

$$\Psi_\phi^* = \Sigma_{\phi y} \Sigma_y^{-1}, \quad \psi_\phi^* = \bar{\phi} - \Psi_\phi^* \mu_y, \quad (4.38)$$

where $\bar{\phi}$ is the mean of ϕ defined in (4.36), μ_y is the observation mean, $\Sigma_{\phi y}$ 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 $G_\theta = \mu_\theta + \Delta G_\theta$, where $\Delta\phi$ and ΔG_θ are zero-mean random quantities, and $\mu_\theta := \mathbb{E}[G_\theta] = \text{diag}(\mu_\theta^0, \dots, \mu_\theta^T)$ with $\mu_\theta^t = \mu_\theta$ for all $t \in \{0, \dots, T\}$. By construction, $\mathbb{E}[\Delta\phi\Delta\phi^\top] = \Sigma_\Phi$ and $\mathbb{E}[(\theta^t - \mu_\theta^t)(\theta^t - \mu_\theta^t)^\top] = \Sigma_\theta$. From the independence assumptions between ϕ , θ and $\mathbf{v}$ it follows that $\mathbb{E}[\Delta G_\theta^\top \Delta\phi] = 0$ and $\mathbb{E}[\Delta G_\theta^\top \Delta\phi \mathbf{v}^\top] = 0$. Using (4.37) together with these decompositions, and exploiting the zero-mean and independence properties, yields

$$\begin{aligned}\mu_y &= \mathbb{E}[(\mu_\theta + \Delta G_\theta)^\top (\bar{\phi} + \Delta\phi) + \mathbf{v}] = \mu_\theta^\top \bar{\phi} = \mu_\Phi^\top \mu_\theta, \\ \Sigma_{\phi y} &= \mathbb{E}[\Delta\phi ((\mu_\theta + \Delta G_\theta)^\top (\bar{\phi} + \Delta\phi) + \mathbf{v} - \mu_\theta^\top \bar{\phi})^\top] = \mathbb{E}[\Delta\phi \Delta\phi^\top] \mu_\theta = \Sigma_\Phi \mu_\theta, \\ \Sigma_y &= \mathbb{E}[(\mathbf{y} - \mu_\theta^\top \bar{\phi})(\mathbf{y} - \mu_\theta^\top \bar{\phi})^\top] = \mu_\theta^\top \Sigma_\Phi \mu_\theta + \mathbf{P} + \mathbf{Q} + \mathbf{S} + \mathbf{S}^\top + \Sigma_v,\end{aligned}\quad (4.39)$$

where $\mathbf{P} = \mathbb{E}[\Delta G_\theta^\top \Delta\phi \Delta\phi^\top \Delta G_\theta]$, $\mathbf{Q} = \mathbb{E}[\Delta G_\theta^\top \bar{\phi} \bar{\phi}^\top \Delta G_\theta]$ and $\mathbf{S} = \mathbb{E}[\Delta G_\theta^\top \bar{\phi} \Delta\phi^\top \Delta G_\theta]$. Consider the $(t, t')^\text{th}$ components of these matrices, denoted by $P_{tt'}$, $Q_{tt'}$, and $S_{tt'}$. 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

$$\begin{aligned}P_{tt'} &= \mathbb{E}[\Delta\theta^\top \Delta\phi(x_t) \Delta\phi^\top(x_{t'}) \Delta\theta] = \mathbb{E}[\text{tr}(\Delta\theta^\top \Delta\phi(x_t) \Delta\phi^\top(x_{t'}) \Delta\theta)] = \text{tr}(\Sigma_\theta \Sigma_\phi^{tt'}), \\ Q_{tt'} &= \mathbb{E}[\Delta\theta^\top \mu_\phi^t \mu_\phi^{t'} \Delta\theta] = \mathbb{E}[\text{tr}(\Delta\theta^\top \mu_\phi^t \mu_\phi^{t'} \Delta\theta)] = \text{tr}(\Sigma_\theta \mu_\phi^t \mu_\phi^{t'}), \\ S_{tt'} &= \mathbb{E}[\Delta\theta^\top \mu_\phi^t \Delta\phi^\top(x_{t'}) \Delta\theta] = \mathbb{E}[\text{tr}(\Delta\theta^\top \mu_\phi^t \Delta\phi^\top(x_{t'}) \Delta\theta)] = \text{tr}(\mathbb{E}[\Delta\theta \Delta\theta^\top] \mu_\phi^t \mathbb{E}[\Delta\phi^\top(x_{t'})]) = 0,\end{aligned}$$

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

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

Consequently, $\Sigma_y = \mu_\theta^\top \Sigma_\Phi \mu_\theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v$, which is invertible because $\Sigma_v \succ 0$. Using this Σ_y together with μ_y and $\Sigma_{\phi y}$ from (4.39) to substitute into (4.38) yields the closed-form expressions for $\hat{\phi}_\phi^*(\mu_\theta, \Sigma_\theta)$ and $\hat{\psi}_\phi^*(\mu_\theta, \Sigma_\theta)$ in (4.14a), both as functions of $(\mu_\theta, \Sigma_\theta)$. Hence the affine MMSE basis estimator can be written as $\hat{\phi}_{\text{af}}(y; \mu_\theta, \Sigma_\theta) = \Psi_\phi^*(\mu_\theta, \Sigma_\theta)y + \psi_\phi^*(\mu_\theta, \Sigma_\theta)$, 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 (4.7a). Since $\hat{\phi}_{\text{af}}(y; \mu_\theta, \Sigma_\theta)$ is stacked as in (4.35), its components $\hat{\phi}_{\text{af}}^t(y; \mu_\theta, \Sigma_\theta)$ directly define the column-wise DBS estimate $\hat{\Phi}_{\text{af}}(y; \mu_\theta, \Sigma_\theta)$, together with $\Sigma_{\hat{\Phi}_{\text{af}}}(\mu_\theta, \Sigma_\theta)$ in (4.14c), with the corresponding MMSE $\mathcal{J}_\phi^*(\mu_\theta, \Sigma_\theta) = \text{tr}(\Sigma_{\hat{\Phi}_{\text{af}}}(\mu_\theta, \Sigma_\theta))$. $\square$

Proof of Lemma 4.3.5. The affine MMSE estimator of $\mathbf{x}$ given $\mathbf{y}$ follows from the affine MMSE estimator of Lemma 4.2.1 with $\xi_1 = \mathbf{x}$ and $\xi_2 = \mathbf{y}$, resulting in the form $\hat{\mathbf{x}}_{\text{af}}(\mathbf{y}) = \Psi_{\mathbf{x}}^* \mathbf{y} + \psi_{\mathbf{x}}^*$, where the optimal coefficients (cf. (4.7b)) are $\Psi_{\mathbf{x}}^* = \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{y}}^{-1}$ and $\psi_{\mathbf{x}}^* = \mu_{\mathbf{x}} - \Psi_{\mathbf{x}}^* \mu_{\mathbf{y}}$. Moreover, the estimation-error covariance and optimal MSE (cf. (4.7a)) are $\Sigma_{\hat{\mathbf{x}}_{\text{af}}} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{y}}^{-1} \Sigma_{\mathbf{xy}}^\top$ and $\mathcal{J}^* = \text{tr}(\Sigma_{\hat{\mathbf{x}}_{\text{af}}})$, respectively. From (4.18), we have $\mu_{\mathbf{x}} = \mathbf{A} \mathbf{B} \mathbf{u}$ and $\Sigma_{\mathbf{x}} = \mathbf{A} \Sigma_{\mathbf{w}} \mathbf{A}^\top$. Furthermore, from (4.39) in the proof of Lemma 4.3.1, we have $\mu_{\mathbf{y}} = \mu_\Phi^\top \mu_\theta$ and $\Sigma_{\mathbf{y}} = \mu_\theta^\top \Sigma_\Phi \mu_\theta + \mathcal{S}(\Sigma_\theta) + \Sigma_v$.

Thus, it remains to compute the cross-covariance matrix Σ_{xy} between the state trajectory x and the observation sequence y . In particular,

$$\Sigma_{xy} = E[(x - ABu)(\Phi^T(x)\theta + v - \mu_{\Phi}^T\mu_{\theta})^T] = E[(x - ABu)\mu_{\theta}^T\Phi(x)], \quad (4.40)$$

where the observation model $y = \Phi^T(x)\theta + v$ is used as in Chapter 3, and the equality follows from algebraic expansion using the independence of θ from x (and hence from $\Phi(x)$), which yields $E[x\theta^T\Phi(x)] = E[x\mu_{\theta}^T\Phi(x)]$, together with $E[xv^T] = 0$ by independence of x and v . Consider the (t, t') th block of this cross-covariance matrix, $\Sigma_{xy}^{tt'}$, which captures the cross-covariance between x_t and $y_{t'}$. The state x_t can be written as $x_t = A_t(Bu + w)$, where A_t is the $(t + 1)$ th block-row of A as in (4.19), i.e.,

$$A_t = \begin{bmatrix} \prod_{i=0}^{t-1} A_i & \prod_{i=1}^{t-1} A_i & \dots & I & 0 & \dots & 0 \end{bmatrix}. \quad (4.41)$$

The (t, t') th block of the cross-covariance matrix can then be expressed as

$$\Sigma_{xy}^{tt'} = E[(x_t - A_tBu)\mu_{\theta}^T\phi(x_{t'})] = \sum_{n=0}^N E[(x_t - A_tBu)\phi_n(x_{t'})]\mu_{\theta_n}, \quad (4.42)$$

where $\phi(x_{t'})$ is the $(t' + 1)$ th column of the basis collection matrix $\Phi(x)$, according to Definition 4.1, and μ_{θ_n} denotes the n th component of the prior mean μ_{θ} for the unknown parameter vector θ . Under Assumption 4.1, $w \sim \mathcal{N}(0, \Sigma_w)$, hence x_t and $x_{t'}$ are both multivariate Gaussian random variables with $x_t \sim \mathcal{N}(A_tBu, A_t\Sigma_wA_t^T)$ and $x_{t'} \sim \mathcal{N}(A_{t'}Bu, A_{t'}\Sigma_wA_{t'}^T)$, where $A_{t'}$ is the $(t' + 1)$ th block-row of A analogous to (4.41). These variables can be represented as affine transformations of a standard Gaussian random vector $v \sim \mathcal{N}(0, I)$: $x_t = A_tBu + A_t\Sigma_w^{\frac{1}{2}}v$ and $x_{t'} = A_{t'}Bu + A_{t'}\Sigma_w^{\frac{1}{2}}v$. Under the regularity conditions (4.25) on $\phi_n(\cdot)$ (continuity almost everywhere and bounded expectation of partial derivatives) for all $n \in \{0, \dots, N\}$, define $g(v) := \phi_n(A_{t'}Bu + A_{t'}\Sigma_w^{\frac{1}{2}}v)$. By Stein's identity [135],

$$E[(x_t - A_tBu)\phi_n(x_{t'})] = E[A_t\Sigma_w^{\frac{1}{2}}vg(v)] = A_t\Sigma_w^{\frac{1}{2}}E[\nabla_v g(v)]. \quad (4.43)$$

Using the chain rule, $E[\nabla_v g(v)] = \Sigma_w^{\frac{1}{2}}A_{t'}^T E[\nabla_x \phi_n(x_{t'})]$, where $\nabla_x \phi_n(x_{t'})$ is the gradient of $\phi_n(\cdot)$ evaluated at $x_{t'}$. Substituting this into (4.43) gives

$$E[(x_t - A_tBu)\phi_n(x_{t'})] = A_t\Sigma_wA_{t'}^T E[\nabla_x \phi_n(x_{t'})]. \quad (4.44)$$

Finally, inserting (4.44) into (4.42), we obtain

$$\Sigma_{xy}^{tt'} = \sum_{n=0}^N A_t\Sigma_wA_{t'}^T E[\nabla_x \phi_n(x_{t'})]\mu_{\theta_n} = A_t\Sigma_wA_{t'}^T C_{t'}^T \mu_{\theta}, \quad (4.45)$$

where $C_{t'} = E[J_{\phi}(x_{t'})]$, and $J_{\phi}(x_{t'}) = [\nabla_x \phi_0(x_{t'}), \dots, \nabla_x \phi_N(x_{t'})]^T$ is the Jacobian matrix evaluated at $x_{t'}$. Assembling these blocks for all pairs (t, t') with $t, t' \in \{0, \dots, T\}$ results in the full form of the cross-covariance matrix (4.40),

$$\Sigma_{xy} = A\Sigma_wA^T\mu_{\theta}, \quad C = \text{diag}(C_0, \dots, C_T), \quad (4.46)$$

with $\mu_\Theta = \text{diag}(\mu_\theta^0, \dots, \mu_\theta^T)$ and $\mu_\theta^t = \mu_\theta$ for all $t \in \{0, \dots, T\}$. Using Σ_{xy} from (4.46), along with the previously derived means and covariances, in the generic affine MMSE formulas provides the closed-form expressions for $\Psi_x^*(\mu_\theta, \Sigma_\theta)$ and $\psi_x^*(\mu_\theta, \Sigma_\theta)$ in (4.26b), both as functions of $(\mu_\theta, \Sigma_\theta)$. Hence, $\hat{x}_{\text{af}}(y; \mu_\theta, \Sigma_\theta) = \Psi_x^*(\mu_\theta, \Sigma_\theta)y + \psi_x^*(\mu_\theta, \Sigma_\theta)$ explicitly emphasizes the dependence of the affine MMSE state estimator on the prior parameter statistics $(\mu_\theta, \Sigma_\theta)$. The corresponding estimation-error covariance $\Sigma_{\hat{x}_{\text{af}}}(\mu_\theta, \Sigma_\theta)$ and optimal MSE $\mathcal{J}_x^*(\mu_\theta, \Sigma_\theta)$ follow directly from (4.26a), completing the proof. $\square$

5

CONCLUSIONS

This thesis develops a hierarchy of Bayesian estimation methods for parameter identification in Wiener-type state-space models, where linear dynamics are observed through unknown static nonlinear functions corrupted by measurement noise. The research is motivated by practical challenges in autonomous underwater vehicle (AUV) bathymetric mapping, where vehicles must simultaneously navigate and map an unknown seabed topography under significant process and measurement uncertainty. The proposed methods address fundamental computational and statistical challenges that arise when latent system states, corrupted by correlated process noise, serve as inputs to unknown parametric observation functions.

5.1 SUMMARY OF CONTRIBUTIONS

The thesis contributions are organized according to a taxonomy that distinguishes between *one-shot optimization* and *fixed-point characterization* approaches for constructing parameter estimators. This taxonomy provides a unifying lens for understanding existing methods and for situating the novel contributions developed across the technical chapters. Across all chapters, the performance of the proposed estimators is assessed on synthetic dynamical systems with known ground truth, allowing comparison of statistical accuracy and computational efficiency.

TIME-VARYING PARAMETER ESTIMATORS FOR LINEAR SYSTEMS

Chapter 2 develops a one-shot optimization method for Bayesian maximum a posteriori (MAP) estimation of linear time-varying (LTV) observation maps. The central contribution is a tractable semidefinite programming (SDP) relaxation, formulated via linear matrix inequalities (LMIs), that provides a high-quality initialization for subsequent local refinement using quasi-Newton methods. This two-stage pipeline, including convex relaxation followed by nonlinear refinement, combines computational efficiency with improved statistical performance.

Extensive numerical experiments show that the proposed SDP-initialized gradient-based scheme attains a lower parameter mean-squared error (MSE) than both expectation-maximization (EM) and the dual Kalman smoother (DKS), representative methods from the

fixed-point characterization class, particularly in high process-noise regimes where the bilinear coupling between states and parameters is more pronounced. With computational complexity $\mathcal{O}(T^3)$ in the trajectory length T , the method compares favorably with EM, which in this setting lacks guaranteed monotonic likelihood improvement and requires repeated state smoothing in the E-step, as well as gradient-based optimization in the M-step due to the absence of a closed-form update. In this context, the SDP-based warm-start strategy is especially effective in mitigating sensitivity to the initialization and in reducing the risk of convergence to poor local optima for gradient-based refinements.

AFFINE ESTIMATORS AND ACTIVE LEARNING FOR WIENER MODELS

Chapter 3 develops a theoretical framework for Bayesian minimum mean-squared error (MMSE) estimation in parametric Wiener models with nonlinear basis-function observation maps. The central contribution is a closed-form optimal affine MMSE estimator that marginalizes the latent state trajectory via dynamic basis statistics (DBS), which capture the means and covariances of basis-function evaluations along the stochastic state trajectory. Specialization to Fourier basis functions yields computational tractability through explicit DBS expressions derived from the characteristic function of the state trajectory. The chapter establishes several theoretical properties of the affine MMSE estimator, including an inherent inconsistency of single-trajectory measurements with Fourier bases under stochastically unstable dynamics. It further provides a consistency condition under which the parameter estimates converge to their true values when multiple statistically independent trajectories are available, and their initial-state distributions yield an increasingly informative basis-function information matrix.

On the practical side, the chapter introduces an active learning algorithm that synthesizes input signals to minimize the analytical estimation error expression. By directly optimizing the closed-form MMSE criterion, the active learning procedure generates excitation signals that systematically reduce parameter uncertainty and improve the rate at which the estimation error decays. Numerical experiments demonstrate that the affine MMSE estimator, particularly when combined with optimized inputs, achieves over an order of magnitude lower parameter MSE values than regularized least-squares methods, which adopt a prediction-error minimization viewpoint and do not account for the contribution of state uncertainty to the observations. The Fourier specialization also links the results to random Fourier features, suggesting extensions to approximate kernel methods.

NONLINEAR ESTIMATORS VIA FIXED-POINT CHARACTERIZATION

Chapter 4 develops nonlinear dual-estimator architectures that extend the affine MMSE framework via fixed-point characterizations that couple two affine MMSE estimators: one for the unknown observation parameters and one for latent variables (basis-function evaluations or system states). This work marks a transition from the one-shot optimization paradigm of Chapters 2 and 3 to the fixed-point characterization end of the taxonomy, and introduces a deterministic moment-based alternative to iterative schemes such as EM, dual Kalman methods, and particle-based Markov chain Monte Carlo (PMCMC) techniques. Two variants are developed and analyzed:

- (i) **Dual basis-parameter (DB-P) estimator:** This architecture combines an affine basis estimator that directly estimates basis-function evaluations over the trajectory

with the affine parameter estimator from Chapter 3. The basis estimator provides DBS estimates that serve as inputs to the parameter estimator, while the parameter estimates in turn define the prior statistics for the basis estimator. The resulting fixed-point equations admit a Jacobi-style iteration that alternates between updating each component using the other's plug-in statistics.

- (ii) **Dual state-parameter (DS-P) estimator:** This architecture first computes affine MMSE state estimates and their estimation-error covariances, and then maps these state statistics through a Gaussian DBS operator to obtain basis statistics. By explicitly estimating the state dynamics over the entire trajectory rather than relying on higher-dimensional basis evaluations, the DS-P estimator more directly exploits the system structure and mitigates the limitations of the elliptical-distribution approximation inherent in the DB-P approach.

The dual architectures show how fixed-point characterizations can be constructed entirely from mean and covariance, providing a deterministic, moment-based alternative to likelihood-driven or sampling-based methods for joint state-parameter estimation. The DS-P estimator, in particular, illustrates that explicitly propagating state-estimate covariances through the DBS operator can alleviate inconsistency and correlation-induced degradation observed in purely affine parameter estimators for marginally stable systems.

Extensive Monte Carlo experiments across multiple configurations show that the DS-P estimator consistently achieves the lowest empirical parameter MSE among all methods tested, outperforming the purely affine MMSE estimator from Chapter 3, the DB-P estimator, sequential Monte Carlo (SMC)-based EM methods, and fully Bayesian particle Gibbs with ancestor sampling (PGAS). In the high-dimensional parameter regimes (low-dimensional states with many unknown parameters), the DS-P estimator yields typical reductions in parameter MSE on the order of 70–90% relative to the affine MMSE estimator and roughly an order of magnitude MSE reduction over the chosen SMC-based methods. This dual estimator further maintains error reduction as the trajectory length increases, in contrast to the affine MMSE baseline, whose convergence slows due to increasing measurement correlations in marginally stable systems and the inherent inconsistency discussed in Chapter 3. The fixed-point iterations of the DS-P estimator exhibit rapid convergence in practice and, when combined with a per-iteration computational complexity of $\mathcal{O}(T^3)$ in the trajectory length T , render it practically viable for moderate-horizon problems.

5.2 METHODOLOGICAL INSIGHTS

Beyond the individual chapter-level contributions, this thesis yields several methodological insights that span the specific estimators and models. These insights are discussed in detail in the remainder of this section.

THE ROLE OF DYNAMIC BASIS STATISTICS

A unifying conceptual thread throughout Chapters 3 and 4 is the central role of *dynamic basis statistics* (DBS) as a sufficient statistic for marginalizing the latent state trajectory in the parameter estimation problem. The DBS construction summarizes the contribution of state uncertainty to the observations through the first two moments of basis-function evaluations along the stochastic state trajectory. This moment-based representation enables closed-

form affine MMSE estimators while avoiding the computational burden of full posterior inference over high-dimensional state sequences.

The DBS framework allows one to decompose the parameter estimation error into two components: the irreducible error corresponding to perfect state knowledge and an additional cross-error term arising from state uncertainty. Chapter 4 formalizes this decomposition and establishes a lower bound on the affine MMSE error, suggesting that improved DBS estimates obtained through dual estimation can systematically reduce the gap between the achievable affine MSE and the fundamental perfect-state benchmark.

ONE-SHOT OPTIMIZATION VERSUS FIXED-POINT CHARACTERIZATION

The taxonomy introduced in Chapter 1 provides insight into the landscape of parameter estimation methods for dynamical systems. *One-shot optimization* methods (Chapters 2 and 3) formulate a single objective function for minimization with respect to parameter-only decision variables, either by marginalizing the latent states analytically (as in the affine MMSE estimator via DBS) or by jointly optimizing over an augmented state-parameter space. These methods benefit from clear optimality principles and, when tractable, admit efficient implementations. *Fixed-point characterization* methods (Chapter 4), in contrast, introduce two coupled estimators with separate loss functions that iteratively refine both state and parameter estimates. While this architecture introduces additional algorithmic complexity, it enables each estimator to exploit problem structure appropriate to its target variables. For instance, the dual state-parameter estimator leverages Gaussian linear dynamics in the state-estimation step, providing more accurate state statistics than would be obtained by directly estimating the nonlinear basis evaluations.

The synthetic experimental evidence in this thesis suggests that the choice between these paradigms depends on problem characteristics. For linear time-varying (LTV) observation maps (Chapter 2), the one-shot MAP formulation with SDP warm-start outperforms iterative dual methods. For nonlinear Wiener models with basis-function observations (Chapters 3 and 4), the one-shot affine MMSE estimator provides a strong baseline; however, the dual state-parameter estimator achieves further gains in this model class by iteratively refining state and parameter statistics.

COMPARISON WITH SAMPLING-BASED METHODS

The deterministic moment-based dual estimators developed in Chapter 4 occupy a distinct position relative to fully Bayesian sampling methods such as particle Gibbs with ancestor sampling (PGAS) and SMC-EM. Particle methods aim to characterize the full posterior distribution over states and parameters through Monte Carlo samples, providing uncertainty quantification beyond second-order moments. However, this generality comes at a substantial computational cost: particle filters and smoothers require careful tuning of multiple hyperparameters (particle numbers, resampling schemes, proposal distributions) [32] and scale poorly with state dimension due to particle degeneracy [14].

In contrast, the dual state-parameter (DS-P) estimator achieves lower parameter MSE values than PGAS and SMC-EM in the experiments of Chapter 4, with significantly reduced computational cost and required no hyperparameter tuning beyond prior specification. This advantage arises from three factors: (i) the closed-form affine structure avoids Monte Carlo approximation errors; (ii) the Gaussian assumption on process noise, while restrictive, enables accurate state estimates via affine MMSE smoothing; and (iii) the moment-based

representation remains computationally tractable in moderate state dimensions where particle methods would require prohibitively many particles. The DS-P estimator thus sacrifices distributional generality (restricting attention to Gaussian process noise and affine estimator classes) in exchange for computational tractability and, empirically, superior parameter estimation accuracy in the Wiener models with Fourier bases and Gaussian linear dynamics considered in Chapter 4.

IMPLICATIONS FOR NONLINEAR SYSTEM IDENTIFICATION

The methods developed in this thesis contribute to the broader field of nonlinear system identification by providing tractable Bayesian alternatives to classical techniques. Traditional identification methods for Wiener systems often employ prediction-error minimization or two-stage procedures that estimate the linear dynamics and static nonlinearity separately. These approaches typically do not provide principled uncertainty quantification for the estimated parameters, and their statistical efficiency can be suboptimal when process noise introduces significant state uncertainty. The Bayesian framework adopted throughout this thesis offers several advantages:

- (i) **Principled uncertainty quantification:** The affine and dual MMSE estimators provide explicit estimation-error covariances alongside point estimates, enabling risk-aware decision-making in downstream applications such as model predictive control or safety verification.
- (ii) **Prior incorporation:** The MAP (Chapter 2) and MMSE (Chapters 3-4) frameworks naturally accommodate prior information about parameters, which can encode physical constraints, regularization preferences, or knowledge from previous experiments.
- (iii) **Process noise accounting:** By explicitly modeling and marginalizing latent state uncertainty through DBS or SDP-based formulations, the proposed methods avoid potential estimation-error degradation observed in approximate prediction-error minimization approaches that neglect process-noise correlations.
- (iv) **Active learning integration:** The closed-form MMSE expressions enable gradient-based input design (Chapter 3), providing a principled approach to experimental design that maximizes information gain rather than relying on generic excitation signals such as pseudorandom binary sequences (PRBS).

Although the numerical evidence is based on synthetic case studies, these contributions position the thesis methods as valuable tools for applications that require both accurate parameter estimates and reliable uncertainty bounds, particularly when process noise cannot be neglected or prior information is available. Such settings arise naturally in several engineering domains. In AUV bathymetric mapping, the proposed methods can estimate the seabed observation model parameters directly from sonar measurements while accounting for vehicle navigation uncertainty, yielding depth maps with quantified reconstruction errors. In robotic manipulation and locomotion, where contact dynamics introduce non-negligible process noise and the robot's kinematics are known, the dual MMSE estimators provide a tractable approach to online model refinement with uncertainty-aware control. Finally, the Fourier basis specialization connects the proposed framework to kernel-based learning via random Fourier features, enabling nonparametric function

approximation with probabilistic error bounds that support the use of reliable confidence intervals of the learned model in safety verification applications.

5.3 FUTURE RESEARCH DIRECTIONS

This thesis formalizes a tractable Bayesian estimation in Wiener-type models and suggests several theoretical and practical directions for future work. The following directions summarize some of the most promising avenues for further investigation.

CONVERGENCE AND CONSISTENCY OF DUAL ESTIMATORS

The fixed-point iterations underlying the dual estimators converge reliably in the numerical experiments of Chapter 4, with rapid convergence observed for the dual state-parameter (DS-P) variant. The dual architecture induces a nonlinear mapping on the joint space of parameter and DBS statistics, and the iteration corresponds to a simultaneous (Jacobi-style) fixed-point scheme. However, formal convergence guarantees and conditions under which convergence to a fixed point is assured, remain to be established.

Analyzing convergence would require studying the spectral properties of the Jacobian of this mapping in a neighborhood of the fixed point. Sufficient conditions for contraction, such as the mapping being a strict contraction in an appropriate norm, would guarantee local convergence. Alternatively, if the dual system can be shown to decrease a Lyapunov-like function at each iteration (analogous to the monotonic likelihood increase in exact EM), global convergence results might be obtained. This analysis is complicated by the plug-in interpretation of the dual estimators, which treats estimated statistics as fixed hyperparameters and thereby violates the independence assumptions underlying the affine MMSE derivations.

The consistency properties of the dual state-parameter (DS-P) estimator also remain an open question. The affine MMSE estimator analyzed in Chapter 3 exhibits consistency for linear dynamical systems under suitable persistent-excitation conditions on the initial states of multiple statistically independent trajectories. The DS-P estimator modifies this affine MMSE baseline by iteratively refining both state and parameter estimates, effectively reducing the influence of process noise through improved state estimation. Intuitively, if the state estimation error remains bounded as the trajectory length increases, the DS-P parameter estimates should inherit the consistency properties of an affine estimator applied with multiple statistically independent trajectories.

Formalizing this intuition requires characterizing conditions under which the state estimation error remains bounded along the trajectory. For time-invariant detectable and stabilizable systems, standard Kalman filtering and smoothing results show that the state estimation error covariances converge to steady-state limits determined by the system dynamics and the relative magnitudes of the process and measurement noise covariances [1, 126]. However, the DS-P estimator uses parameter estimates as inputs to the state estimator, introducing an additional source of errors that couples the two estimation problems. A rigorous consistency analysis would need to account for this coupling, potentially employing techniques from stochastic approximation theory or establishing uniform error bounds on the state estimator as a function of parameter error. Such an analysis could provide asymptotic error rates and sample-complexity bounds for the dual estimators, complementing the finite-sample numerical evidence presented in Chapter 4.

RELAXING GAUSSIAN ASSUMPTIONS AND ROBUSTNESS

The dual state-parameter (DS-P) estimator in Chapter 4 assumes Gaussian process noise, which enables closed-form expressions for the affine MMSE state estimator via Stein's identity and the Gaussian DBS operator. This assumption is restrictive and may not hold in applications with heavy-tailed disturbances, multi-modal uncertainties, or non-Gaussian measurement noise. Extensions to elliptical distributions (which include Gaussian, Student- t , and other symmetric heavy-tailed families) are conceptually straightforward, as Stein's identity generalizes to elliptical models. However, broader relaxations to non-elliptical process noise would require fundamentally different tools. Moment-matching approximations (e.g., unscented transforms) or variational inference techniques could provide approximate Gaussian DBS under non-Gaussian dynamics, trading exactness for wider applicability. Assessing the robustness of the dual estimators to misspecified process-noise distributions is therefore an important direction for future work.

More generally, the Bayesian framework adopted throughout this thesis assumes known noise distributions and exact prior specifications. In practice, these assumptions may be violated due to model mismatch, unmodeled disturbances, or adversarial perturbations. Robust and distributionally robust variants of the proposed methods could provide performance guarantees under such uncertainties. One approach would employ distributionally robust optimization techniques, replacing exact noise distributions with ambiguity sets (e.g., Wasserstein balls around nominal models) and optimizing over the worst-case distribution in the set. The SDP relaxation techniques developed in Chapter 2 could potentially be adapted to incorporate distributional-robustness constraints, providing tractable conservative approximations to robust MAP problems. Another direction would be to incorporate robust loss functions, such as Huber or other M-estimator losses [67, 94] that downweight outliers, into the affine estimation framework, yielding robust minimum-risk estimators that may still admit fixed-point characterizations analogous to those developed for the dual architectures in Chapter 4.

SCALABILITY AND RECURSIVE IMPLEMENTATIONS

The per-iteration complexity of $\mathcal{O}(T^3)$ for both affine and dual MMSE estimators (where T is the trajectory length) arises from inverting the $(T + 1) \times (T + 1)$ observation covariance matrix in the affine MMSE updates. This cubic scaling limits applicability to large-horizon problems, such as the long trajectories encountered in extended AUV missions, where the computational burden becomes prohibitive. One approximate strategy is to split long trajectories into smaller *batches* and use the final estimate from each batch as the prior for the unknown variables in the subsequent batch. A more principled approach to reduce complexity is to develop recursive implementations, similar to recursive least-squares, that process measurements sequentially and update a fixed-size set of summary statistics rather than reprocessing the entire trajectory. Developing such recursive variants would require reformulating the batch DBS and parameter updates as filtering recursions, analogous to the relationship between Rauch-Tung-Striebel (RTS) smoothing and Kalman filtering [7]. This reformulation is non-trivial because the affine parameter estimator depends on the entire trajectory through the DBS, and recursively updating DBS requires propagating the joint statistics of basis-function evaluations over time.

The batch formulations developed in this thesis are designed for offline settings where all data are available before estimation begins. Many applications, including the AUV

bathymetric mapping problem, require online parameter updates as new measurements arrive. For the affine MMSE estimator, a recursive formulation would involve deriving a filtering recursion that updates parameter estimates and DBS as each new measurement is processed. For the dual estimators, a recursive implementation would require interleaving state filtering, DBS updates, and parameter updates at each time step before processing the subsequent measurement. Analyzing the tracking performance and stability of such online dual estimators is a natural avenue for future research.

NONPARAMETRIC OBSERVATION MODELS AND NONLINEAR DYNAMICS

The parametric model adopted in Chapters 3 and 4 assumes a known finite-dimensional basis set, with unknown coefficients to be estimated. While the connection to random Fourier features suggests a path toward approximating kernel methods, directly extending the affine MMSE framework to infinite-dimensional reproducing kernel Hilbert spaces (RKHS) remains an open problem. In particular, when the system states are stochastic, treating the observation function as a Gaussian process introduces an infinite-dimensional prior over a function evaluated along a random trajectory, which makes it nontrivial to identify infinite-dimensional statistics that play the role of DBS within an RKHS or kernel-based representation. In such settings, kernel-based representations, for example via mean embeddings of state-observation joint distributions, could be explored as an alternative way to construct moment-like summaries.

The linear dynamics assumption, on the other hand, is fundamental to the tractability of the DBS framework and to the affine MMSE state estimator in the dual state-parameter (DS-P) architecture. Many practical systems, however, exhibit nonlinear dynamics that cannot be adequately approximated by linearization. Extending the dual estimator framework to nonlinear dynamics would therefore be a natural direction for future work. One possible avenue is to employ a hierarchical construction, analogous in spirit to hierarchical empirical Bayes methods [27], in which the dual estimator is formulated by coupling the affine MMSE parameter estimator with an affine MMSE process-noise estimator, rather than with an affine state estimator. Such a structure would retain a linear dynamical model for the process noise while incorporating the nonlinear system dynamics into the construction of the affine estimators. Analyzing how to compute the corresponding DBS for this architecture remains an open research problem.

APPLICATION TO AUV BATHYMETRIC SLAM

Although the thesis is motivated by AUV bathymetric mapping, the numerical experiments throughout focus on synthetic dynamical systems with known ground truth to enable rigorous statistical evaluation of the proposed estimators. Demonstrating the performance of the dual state-parameter estimator on real bathymetric datasets, or on high-fidelity simulations that mimic operational conditions, remains an important validation step.

The lack of absolute position fixes (as GPS is unavailable underwater) necessitates joint estimation of the vehicle trajectory and map, a structure for which the dual state-parameter architecture and DBS framework are naturally suited. Integrating the dual state-parameter estimator into a bathymetric SLAM pipeline would build on the existing architecture by representing the seabed as a spatial field parameterized by suitable spatial basis functions and treating the vehicle trajectory as the latent state sequence. Since the seafloor is a spatially extended field over a two-dimensional domain, the basis-function parametrization

developed in this thesis provides a natural way to model bathymetry, albeit at a much larger spatial scale and with a correspondingly high-dimensional parameter vector.

Extending and validating this approach on realistic bathymetric SLAM scenarios would also require addressing other established challenges, such as nonlinear vehicle kinematics with a six-degree-of-freedom pose (position and orientation) and sonar measurements with complex error characteristics, including multipath effects, beam geometry, and outliers. Nonlinear vehicle dynamics could be handled by locally linearizing the motion model around a nominal trajectory and interpreting the resulting linearization error as additional process noise, thereby preserving an effective linear-Gaussian representation at the estimator level. In such a formulation, the DBS framework would capture the correlations between basis-function evaluations induced by the uncertain vehicle trajectory, providing a principled mechanism for joint estimation. Beyond bathymetry, the methods developed in this thesis are applicable to other robotic perception problems that involve mapping unknown spatial fields from noisy mobile sensor platforms.

5.4 CONCLUDING REMARKS

This thesis has developed a systematic approach to tractable Bayesian parameter estimation for Wiener-type state-space models, progressing from one-shot MAP and MMSE methods to nonlinear dual estimators characterized as fixed points of coupled affine MMSE components. The introduction of dynamic basis statistics (DBS) as a central organizing principle enables closed-form MMSE estimators that explicitly account for uncertainty in the latent system states, thereby avoiding the approximation errors inherent in methods that neglect or inadequately model process-noise correlations.

The SDP warm-start strategy for MAP estimation and the active learning framework for input design in the MMSE setting further illustrate how problem structure can be exploited to enhance both statistical performance and computational efficiency. The nonlinear MMSE estimator architecture highlights how an affine estimator class can be leveraged to maintain computational tractability through deterministic moment-based updates within its fixed-point scheme. The experimental evidence demonstrates that the proposed methods achieve superior accuracy of parameter estimation over representative baselines, with the nonlinear MMSE dual state-parameter estimator emerging as the most advanced and effective instance of the overall framework.

The thesis further opens several avenues for extending tractable Bayesian estimation to richer model classes, developing recursive implementations for online operation, and applying these methods to challenging real-world problems such as autonomous underwater vehicle navigation and mapping. The fixed-point characterization framework, in particular, offers a flexible architecture for coupling specialized estimators that can incorporate additional structure while retaining the computational advantages of moment-based propagation. Ultimately, the aim of this line of research is to enable autonomous systems to learn and adapt reliably in uncertain and dynamic environments, quantify their uncertainty, and make principled decisions under incomplete information. The methods developed in this thesis represent a step toward that aim, providing a foundation for future work that balances statistical rigor, computational feasibility, and practical applicability.

BIBLIOGRAPHY

REFERENCES

- [1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice Hall, 1979.
- [2] C. Andrieu, A. Doucet, and R. Holenstein, “Particle Markov chain Monte Carlo methods,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, pp. 269–342, 2010.
- [3] A. S. S. Azhar and H. N. Al-Duwaish, “Identification of Wiener model using radial basis functions neural networks,” in *International Conference on Artificial Neural Networks*, 2002, pp. 344–350.
- [4] E.-W. Bai, “An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems,” *Automatica*, vol. 34, pp. 333–338, 1998.
- [5] S. Balakrishnan, M. J. Wainwright, and B. Yu, “Statistical guarantees for the EM algorithm: From population to sample-based analysis,” *The Annals of Statistics*, vol. 45, pp. 77–120, 2017.
- [6] R. Bardenet, A. Doucet, and C. Holmes, “On Markov chain Monte Carlo methods for tall data,” *Journal of Machine Learning Research*, vol. 18, pp. 1–43, 2017.
- [7] T. D. Barfoot, *State Estimation for Robotics*. Cambridge University Press, 2017.
- [8] S. Barkby, S. Williams, O. Pizarro, and M. Jakuba, “Incorporating prior maps with bathymetric distributed particle SLAM for improved AUV navigation and mapping,” in *OCEANS*, 2009, pp. 1–7.
- [9] —, “Bathymetric SLAM with no map overlap using Gaussian processes,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1242–1248.
- [10] —, “Bathymetric particle filter SLAM using trajectory maps,” *The International Journal of Robotics Research*, vol. 31, pp. 1409–1430, 2012.
- [11] L. E. Baum and J. A. Eagon, “An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology,” *Bulletin of the American Mathematical Society*, vol. 73, pp. 360–363, 1967.
- [12] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. Springer, Cham, 2017.

- [13] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 453–463, 2003.
- [14] T. Bengtsson, P. Bickel, and B. 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*. Institute of Mathematical Statistics, 2008, pp. 316–335.
- [15] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.
- [16] D. P. Bertsekas, *Convex Optimization Algorithms*. Athena Scientific, 2015.
- [17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [18] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, pp. 859–877, 2017.
- [19] M. Bocquet and P. Sakov, "Joint state and parameter estimation with an iterative ensemble kalman smoother," *Nonlinear Processes in Geophysics*, vol. 20, pp. 803–818, 2013.
- [20] G. Bottegai, R. Castro-Garcia, and J. A. Suykens, "On the identification of Wiener systems with polynomial nonlinearity," in *Proceedings of the IEEE Conference on Decision and Control*, 2017, pp. 6475–6480.
- [21] M. Briers, A. Doucet, and S. Maskell, "Smoothing algorithms for state-space models," *Annals of the Institute of Statistical Mathematics*, vol. 62, pp. 61–89, 2010.
- [22] P. Bunch and S. Godsill, "Improved particle approximations to the joint smoothing distribution using Markov chain Monte Carlo," *IEEE Transactions on Signal Processing*, vol. 61, pp. 956–963, 2012.
- [23] R. L. Burden and J. D. Faires, *Numerical Analysis*, 9th ed. Brooks/Cole, 2010.
- [24] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, vol. 11, pp. 368–385, 1981.
- [25] G. Canepa, O. Bergem, and N. G. Pace, "A new algorithm for automatic processing of bathymetric data," *IEEE Journal of Oceanic Engineering*, vol. 28, pp. 62–77, 2003.
- [26] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. Springer, 2005.
- [27] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Chapman and Hall/CRC, 2000.
- [28] M. A. Carlton and J. L. Devore, *Probability with Applications in Engineering, Science, and Technology*. Springer, 2017.

- [29] A. L. Cedeño, R. A. González, R. Carvajal, and J. C. Agüero, “Identification of Wiener state-space models utilizing Gaussian sum smoothing,” *Automatica*, vol. 166, p. 111707, 2024.
- [30] Z. Chen, “Bayesian filtering: from Kalman filters to particle filters, and beyond,” *Statistics*, vol. 182, pp. 1–69, 2003.
- [31] A. Chiuso and G. Pillonetto, “System identification: A machine learning perspective,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 281–304, 2019.
- [32] N. Chopin and O. Papaspiliopoulos, *An Introduction to Sequential Monte Carlo*. Springer International Publishing, 2020.
- [33] N. Chopin and S. S. Singh, “On particle Gibbs sampling,” *Bernoulli*, vol. 21, p. 1855–1883, 2015.
- [34] J. Courts, A. G. Wills, T. B. Schön, and B. Ninness, “Variational system identification for nonlinear state-space models,” *Automatica*, vol. 147, p. 110687, 2023.
- [35] H.-D. Dau and N. Chopin, “On backward smoothing algorithms,” *The Annals of Statistics*, vol. 51, pp. 2145–2169, 2023.
- [36] M. Deistler, “System identification – general aspects and structure,” in *Model Identification and Adaptive Control: From Windsurfing to Telecommunications*. Springer, 2001, pp. 3–26.
- [37] B. Delyon, M. Lavielle, and E. Moulines, “Convergence of a stochastic approximation version of the EM algorithm,” *The Annals of Statistics*, vol. 27, pp. 94–128, 1999.
- [38] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, pp. 1–38, 1977.
- [39] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
- [40] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and Computing*, vol. 10, pp. 197–208, 2000.
- [41] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, 2nd ed. Oxford University Press, 2012.
- [42] Y. Ephraim and N. Merhav, “Hidden Markov processes,” *IEEE Transactions on Information Theory*, vol. 48, pp. 1518–1569, 2002.
- [43] —, “Lower and upper bounds on the minimum mean-square error in composite source signal estimation,” *IEEE transactions on Information Theory*, vol. 38, pp. 1709–1724, 2002.
- [44] V. V. Fedorov, *Theory of Optimal Experiments*. Academic Press, 1972.

- [45] I. O. Ferreira, L. C. d. Andrade, V. G. Teixeira, and F. C. M. Santos, "State of art of bathymetric surveys," *Boletim de Ciências Geodésicas*, vol. 28, p. e2022002, 2022.
- [46] Y. Firouz, S. Goutam, M. C. Soult, A. Mohammadi, J. Van Mierlo, and P. Van den Bossche, "Block-oriented system identification for nonlinear modeling of all-solid-state Li-ion battery technology," *Journal of Energy Storage*, vol. 28, p. 101184, 2020.
- [47] I. Florescu and C. A. Tudor, *Handbook of Probability*. John Wiley & Sons, 2013.
- [48] E. Galceran and M. Carreras, "Planning coverage paths on bathymetric maps for in-detail inspection of the ocean floor," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2013, pp. 4159–4164.
- [49] N. Galioto and A. A. Gorodetsky, "Bayesian system ID: optimal management of parameter, model, and measurement uncertainty," *Nonlinear Dynamics*, vol. 102, pp. 241–267, 2020.
- [50] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, pp. 398–409, 1990.
- [51] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. CRC Press, 2013.
- [52] M. Gevers, M. Caenepeel, and J. Schoukens, "Experiment design for the identification of a simple Wiener system," in *IEEE Conference on Decision and Control*, 2012, pp. 7333–7338.
- [53] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," *Technical Report*, vol. University of Toronto, pp. CRG-TR-96-2, 1996.
- [54] Z. Ghahramani and S. Roweis, "Learning nonlinear dynamical systems using an EM algorithm," in *Advances in Neural Information Processing Systems*, vol. 11, 1998.
- [55] A. Girard, C. E. Rasmussen, J. Quiñonero Candela, and R. Murray-Smith, "Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting," in *Advances in Neural Information Processing Systems*, vol. 15, 2002, pp. 529–536.
- [56] B. V. Gnedenko, *Theory of Probability*. Routledge, 2018.
- [57] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, vol. 99, pp. 156–168, 2004.
- [58] A. A. Gorji and M. B. Menhaj, "A neuro-based EM-particle smoothing algorithm for identification of nonlinear state space models," *IFAC Proceedings Volumes*, vol. 41, pp. 4042–4047, 2008.
- [59] A. Hagenblad, "Aspects of the identification of Wiener models," PhD Licentiate Thesis, Division of Automatic Control, Department of Electrical Engineering, Linköpings Universitet, 1999.

- [60] A. Hagenblad, L. Ljung, and A. Wills, “Maximum likelihood identification of Wiener models,” *Automatica*, vol. 44, pp. 2697–2705, 2008.
- [61] A. C. Harvey and S. Peters, “Estimation procedures for structural time series models,” *Journal of Forecasting*, vol. 9, pp. 89–108, 1990.
- [62] W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, vol. 57, pp. 97–109, 1970.
- [63] S. Haykin, *Kalman Filtering and Neural Networks*. John Wiley & Sons, 2004.
- [64] J. Hensman, N. Durrande, and A. Solin, “Variational Fourier features for Gaussian processes,” *Journal of Machine Learning Research*, vol. 18, pp. 1–52, 2018.
- [65] T. Horváth and D. Fodor, “Performance comparison of the general the dual and the joint extended Kalman filter on state estimation of Li-ion battery cells for BMS,” in *IEEE 21st International Power Electronics and Motion Control Conference*, 2024, pp. 1–6.
- [66] R. Hosseini and S. Sra, “An alternative to EM for Gaussian mixture models: Batch and stochastic Riemannian optimization,” *Mathematical Programming*, vol. 181, pp. 187–223, 2020.
- [67] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [68] M. E. Johnson, *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*. John Wiley & Sons, 1987.
- [69] R. E. Kalman and R. S. Bucy, “New results in linear filtering and prediction theory,” *Journal of Basic Engineering*, vol. 83, pp. 95–108, 1961.
- [70] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.
- [71] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, “On particle methods for parameter estimation in state-space models,” *Statistical Science*, vol. 30, pp. 328–351, 2015.
- [72] T. Kogut, A. Tomczak, A. Słowik, and T. Oberski, “Seabed modelling by means of airborne laser bathymetry data and imbalanced learning for offshore mapping,” *Sensors*, vol. 22, p. 3121, 2022.
- [73] S. J. Koopman and J. Durbin, “Fast filtering and smoothing for multivariate state space models,” *Journal of Time Series Analysis*, vol. 21, pp. 281–296, 2000.
- [74] K. E. Krasnosky and C. Roman, “A massively parallel implementation of Gaussian process regression for real time bathymetric modeling and simultaneous localization and mapping,” *Field Robotics*, vol. 2, pp. 940–970, 2022.

- [75] Z. Landsman and J. Nešlehová, “Stein’s lemma for elliptical random vectors,” *Journal of Multivariate Analysis*, vol. 99, pp. 912–927, 2008.
- [76] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Springer, 2008.
- [77] Y. Li and S. K. Ghosh, “Efficient sampling methods for truncated multivariate Normal and Student-t distributions subject to linear inequality constraints,” *Journal of Statistical Theory and Practice*, vol. 9, pp. 712–732, 2015.
- [78] Z. Li, Z. Peng, Z. Zhang, Y. Chu, C. Xu, S. Yao, Á. F. García-Fernández, X. Zhu, Y. Yue, A. Levers, J. Zhang, and J. Ma, “Exploring modern bathymetry: A comprehensive review of data acquisition devices, model accuracy, and interpolation techniques for enhanced underwater mapping,” *Frontiers in Marine Science*, vol. 10, p. 1178845, 2023.
- [79] K. Lindqvist and H. Hjalmarsson, “Identification for control: Adaptive input design using convex optimization,” in *IEEE Conference on Decision and Control*, 2001, pp. 4326–4331.
- [80] F. Lindsten, “An efficient stochastic approximation EM algorithm using conditional particle filters,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6274–6278.
- [81] F. Lindsten, M. I. Jordan, and T. B. Schön, “Particle Gibbs with ancestor sampling,” *Journal of Machine Learning Research*, vol. 15, pp. 2145–2184, 2014.
- [82] F. Lindsten, T. B. Schön, and M. I. Jordan, “Bayesian semiparametric Wiener system identification,” *Automatica*, vol. 49, pp. 2053–2063, 2013.
- [83] B. Liu, Y. Huang, H. Feng, J. Yu, J. Qiao, and X. Ju, “A review of AUV-based bathymetric SLAM technology,” *Ocean Engineering*, vol. 342, p. 122858, 2025.
- [84] F. Liu, X. Huang, Y. Chen, and J. A. Suykens, “Random features for kernel approximation: A survey on algorithms, theory, and beyond,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 7128–7148, 2021.
- [85] J. S. Liu, “Siegel’s formula via Stein’s identities,” *Statistics & Probability Letters*, vol. 21, pp. 247–251, 1994.
- [86] S. Liu, X. Zhang, L. Xu, and F. Ding, “Expectation-maximization algorithm for bilinear systems by using the Rauch–Tung–Striebel smoother,” *Automatica*, vol. 142, p. 110365, 2022.
- [87] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Prentice Hall PTR, 1999.
- [88] —, “Perspectives on system identification,” *Annual Reviews in Control*, vol. 34, pp. 1–12, 2010.
- [89] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

- [90] H. Ma, W. Sheng, S. Wang, and S. Wu, “A state-parameter joint estimation approach for distribution networks based on dual extended kalman filter,” *Sustainable Energy, Grids and Networks*, vol. 43, p. 101884, 2025.
- [91] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [92] Y. Malitsky and K. Mishchenko, “Adaptive gradient descent without descent,” in *International Conference on Machine Learning*, 2020, pp. 6702–6712.
- [93] H. Mania, M. I. Jordan, and B. Recht, “Active learning for nonlinear system identification with guarantees,” *Journal of Machine Learning Research*, vol. 23, pp. 1–30, 2022.
- [94] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*, 2nd ed. John Wiley & Sons, 2019.
- [95] A. McHutchon and C. E. Rasmussen, “Gaussian process training with input noise,” in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 182–190.
- [96] R. K. Mehra, “Synthesis of optimal inputs for multiinput-multioutput (Mimo) systems with process noise part i: Frequency-domain synthesis part ii: Time-domain synthesis,” in *Mathematics in Science and Engineering*. Elsevier, 1976, pp. 211–249.
- [97] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *The Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.
- [98] J. Mikusiński, “The Fubini theorem,” in *The Bochner Integral*. Springer, 1978, pp. 91–105.
- [99] S. A. Murawski, S. Gilbert, M. Hommeyer, Y. Liu, C. Lembke, S. Grasty, D. English, C. Hu, T. Dixon, and T. S. Chorsi, “Bathymetric mapping in the coastal zone: Approaches, challenges, and opportunities,” *Marine Technology Society Journal*, vol. 59, pp. 62–77, 2025.
- [100] C. Naesseth, F. Lindsten, and T. Schön, “Nested sequential Monte Carlo methods,” in *International Conference on Machine Learning*, 2015, pp. 1292–1301.
- [101] C. A. Naesseth, F. Lindsten, and T. B. Schön, “Elements of sequential Monte Carlo,” *Foundations and Trends in Machine Learning*, vol. 12, pp. 307–392, 2019.
- [102] I. Naim and D. Gildea, “Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients,” in *arXiv preprint arXiv:1206.6427*, 2012.
- [103] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks, Fuzzy Models, and Gaussian Processes*. Springer International Publishing, 2020.

- [104] A. T. Nelson, "Nonlinear estimation and modeling of noisy time-series by dual Kalman filter methods," PhD Thesis, Oregon Graduate Institute of Science and Technology, 2000.
- [105] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.
- [106] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *Mathematics of Operations Research*, vol. 48, pp. 1–37, 2023.
- [107] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006.
- [108] O. Nonnis, C. Maggi, P. Lanera, R. Proietti, A. Izzi, P. Antonelli, and M. Gabellini, "The management of information to identify the submarine cable route: The case study of Campania islands," *Journal of Coastal Research*, vol. 75, pp. 1002–1006, 2016.
- [109] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon, "Identification of nonlinear systems using polynomial nonlinear state space models," *Automatica*, vol. 46, pp. 647–656, 2010.
- [110] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, *Regularized System Identification: Learning Dynamic Models from Data*. Springer Nature, 2022.
- [111] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, pp. 81–93, 2010.
- [112] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, pp. 657–682, 2014.
- [113] M. K. Pitt, R. d. S. Silva, P. Giordani, and R. Kohn, "On some properties of Markov chain Monte Carlo simulation methods based on the particle filter," *Journal of Econometrics*, vol. 171, pp. 134–151, 2012.
- [114] A. Popovici, P. Zaal, and D. M. Pool, "Dual extended Kalman filter for the identification of time-varying human manual control behavior," in *AIAA Modeling and Simulation Technologies Conference*, 2017-3666.
- [115] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1177–1184, 2007.
- [116] B. Rajaratnam and D. 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.
- [117] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press Cambridge, 2006.
- [118] H. E. Rauch, "Solutions to the linear smoothing problem," *IEEE Transactions on Automatic Control*, vol. 8, pp. 371–372, 1963.

- [119] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, pp. 1445–1450, 1965.
- [120] R. S. Risuleo and H. Hjalmarsson, "Nonparametric models for Hammerstein-Wiener and Wiener-Hammerstein system identification," in *IFAC-PapersOnLine*, vol. 53, 2020, pp. 400–405.
- [121] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer, 2004.
- [122] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.
- [123] W. Rudin, *Fourier Analysis on Groups*. Interscience Publishers, 1962.
- [124] N. Sammaknejad, Y. Zhao, and B. Huang, "A review of the expectation maximization algorithm in data-driven process identification," *Journal of process control*, vol. 73, pp. 123–136, 2019.
- [125] N. Santitissadeekorn and C. Jones, "Two-stage filtering for joint state-parameter estimation," *Monthly Weather Review*, vol. 143, pp. 2028–2042, 2015.
- [126] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.
- [127] T. B. Schön, A. Wills, and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, vol. 47, pp. 39–49, 2011.
- [128] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented road map," *IEEE Control Systems Magazine*, vol. 39, pp. 28–99, 2019.
- [129] M. Schoukens and K. Tiels, "Identification of block-oriented nonlinear systems starting from linear approximations: A survey," *Automatica*, vol. 85, pp. 272–292, 2017.
- [130] B. Settles, *Active Learning*. Morgan & Claypool Publishers, 2012.
- [131] S. Shafieezadeh Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [132] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, pp. 253–264, 1982.
- [133] T. Söderström and P. Stoica, *System Identification*. Prentice Hall, 1989.
- [134] S. Sra and R. Hosseini, "Conic geometric optimization on the manifold of positive definite matrices," *SIAM Journal on Optimization*, vol. 25, pp. 713–739, 2015.
- [135] C. M. Stein, "Estimation of the mean of a multivariate Normal distribution," *The Annals of Statistics*, vol. 9, pp. 1135–1151, 1981.

- [136] S. T. Tokdar and R. E. Kass, "Importance sampling: A review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, pp. 54–60, 2010.
- [137] S. Tötterman and H. T. Toivonen, "Support vector method for identification of Wiener models," *Journal of Process Control*, vol. 19, pp. 1174–1181, 2009.
- [138] A. Tulsyan, B. Huang, R. B. Gopaluni, and J. F. Forbes, "Bayesian identification of non-linear state-space models: Part II-Error Analysis," *IFAC Proceedings Volumes*, vol. 46, pp. 631–636, 2013.
- [139] S. Vakili, M. Khosravi, P. M. Esfahani, and M. Mazo Jr, "Linear time-varying parameter estimation: Maximum a posteriori approach via semidefinite programming," *IEEE Control Systems Letters*, vol. 8, pp. 73–78, 2023.
- [140] S. Vakili, M. Mazo Jr, and P. M. Esfahani, "Optimal Bayesian affine estimator and active learning for the Wiener model," *arXiv preprint arXiv:2504.05490*, 2025.
- [141] S. Vakili, D. Woonings, P. Paruchuri, and P. M. Esfahani, "Nonlinear estimator: Dual bayesian affine estimators for parameter learning," *arXiv preprint arXiv:2606.10111*, 2026.
- [142] P. E. Valenzuela, J. Dahlin, C. R. Rojas, and T. B. Schön, "On robust input design for nonlinear dynamical models," *Automatica*, vol. 77, pp. 268–278, 2017.
- [143] P. E. Valenzuela, C. R. Rojas, and H. Hjalmarsson, "Optimal input design for nonlinear dynamic systems: a graph theory approach," in *IEEE Conference on Decision and Control*, 2013, pp. 5740–5745.
- [144] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, p. 1, 2021.
- [145] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory–Implementation–Applications*. Springer, 1996.
- [146] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conference on Learning Theory*, vol. 125, 2020, pp. 3487–3582.
- [147] E. Wan and A. Nelson, "Dual Kalman filtering methods for nonlinear prediction, smoothing and estimation," in *Advances in Neural Information Processing Systems*, vol. 9, 1996, pp. 793–799.
- [148] E. A. Wan, R. van der Merwe, and A. T. Nelson, "Dual estimation and the unscented transformation," in *Advances in Neural Information Processing Systems*, vol. 12, 1999.
- [149] L. Wang, H. Liu, H. Su, and J. Wang, "Bathymetry retrieval from optical images with spatially distributed support vector machines," *GIScience & Remote Sensing*, vol. 56, pp. 323–337, 2019.

- [150] R. B. Wynn, V. A. I. Huvenne, T. P. L. Bas, B. J. Murton, D. P. Connelly, B. J. Bett, H. A. Ruhl, K. J. Morris, J. Peakall, D. R. Parsons, E. J. Sumner, S. E. Darby, R. M. Dorrell, and J. E. Hunt, "Autonomous underwater vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience," *Marine Geology*, vol. 352, pp. 451–468, 2014.
- [151] B. Yan, M. Yin, and P. Sarkar, "Convergence of gradient EM on multi-component mixture of Gaussians," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [152] S. Yi and M. Zorzi, "Robust Kalman filtering under model uncertainty: The case of degenerate densities," *IEEE Transactions on Automatic Control*, vol. 67, pp. 3458–3471, 2021.
- [153] —, "Robust fixed-lag smoothing under model perturbations," *Journal of the Franklin Institute*, vol. 360, 2023.
- [154] J. Zambrano, J. Sanchis, J. M. Herrero, and M. Martínez, "A unified approach for the identification of Wiener, Hammerstein, and Wiener–Hammerstein models by using WH-EA and multistep signals," *Complexity*, vol. 2020, p. 7132349, 2020.
- [155] M. Zheng and Y. Ohta, "Bayesian positive system identification: Truncated Gaussian prior and hyperparameter estimation," *Systems & Control Letters*, vol. 148, p. 104857, 2021.

GLOSSARY

A-MMSE affine minimum mean-squared error parameter estimator.

AUV autonomous underwater vehicle.

BAL Bayesian affine minimum mean-squared error estimator with active learning.

BFGS Broyden–Fletcher–Goldfarb–Shanno.

BMS Bayesian affine minimum mean-squared error estimator.

DB-P dual basis-parameter.

DBS dynamic basis statistics.

DFT discrete Fourier transform.

DKS dual Kalman smoother.

DLS dead reckoning least squares.

DS-P dual state-parameter.

EM expectation-maximization.

GMM Gaussian mixture model.

GNSS global navigation satellite system.

GP Gaussian process.

GPS global positioning system.

HMP hidden Markov process.

LMI linear matrix inequality.

LTI linear time-invariant.

LTV linear time-varying.

MAP maximum a posteriori.

MCMC Markov chain Monte Carlo.

MH Metropolis–Hastings.

MH-FFBP Metropolis-Hastings forward-filtering backward-proposing.

ML maximum likelihood.

MLS Mean least squares.

MMSE minimum mean-squared error.

MSE mean-squared error.

NSMC nested sequential Monte Carlo.

PEM prediction error method.

PG particle Gibbs.

PGAS particle Gibbs with ancestor sampling.
PMCMC particle Markov chain Monte Carlo.
PMMH particle marginal Metropolis-Hastings.
PRBS pseudorandom binary sequence.

RBF radial basis function.
RKHS reproducing kernel Hilbert space.
RLS regularized least squares.
RTS Rauch–Tung–Striebel.

SDP semidefinite-programming.
SDP-GD semidefinite-programming-enhanced gradient descent.
SLAM simultaneous localization and mapping.
SMC sequential Monte Carlo.
SNR signal-to-noise ratio.

UKF unscented Kalman filter.

LIST OF PUBLICATIONS

1. **Sasan Vakili**, Mohammad Khosravi, Peyman Mohajerin Esfahani, and Manuel Mazo Jr., “Linear Time-Varying Parameter Estimation: Maximum A Posteriori Approach via Semi-definite Programming,” *IEEE Control Systems Letters* (LCSS), vol. 8, pp. 73 - 78, 2023.
 2. **Sasan Vakili**, Manuel Mazo Jr., and Peyman Mohajerin Esfahani, “Optimal Bayesian Affine Estimator and Active Learning for the Wiener Model,” *Transactions on Automatic Control* (TAC), under revision, 2025.
 3. **Sasan Vakili**, Daniël Woonings, Pradyumna Paruchuri, and Peyman Mohajerin Esfahani, “Nonlinear Bayesian Estimator for Parameter Learning: A Fixed-Point Characterization,” *Journal of Machine Learning Research* (JMLR), under review, 2026.
- Included in this thesis.

