Thesis for The Degree of Licentiate of Engineering

Semi-analytical MAP estimation of hidden dynamics in continuous-state POMDPs

Erik Karlsson Nordling

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden, 2026

Semi-analytical MAP estimation of hidden dynamics in continuous-state POMDPs

Erik Karlsson Nordling

Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Typeset with LaTeX
Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

Semi-analytical MAP estimation of hidden dynamics in continuous-state POMDPs
ERIK KARLSSON NORDLING
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

This thesis explores a problem in dynamical systems control with a finite action set. A common situation is that the underlying dynamics of the system is unknown and needs to be estimated based on observed data. However, when only partial information of the latent state space is available, and the latent state space is continuous, this task becomes exceedingly difficult. Unlike probabilistic sampling-based methods for solving this problem, a semi-analytical sampling-free procedure is examined.

We propose a representation framework based on weighted sums of basis functions for approximating a transition probability density function and show how this can be used with a truncated Gaussian prior probability measure in a semi-analytical algorithm for calculating the posterior log-probability of a represented transition PDF. We also provide efficient forward-backward algorithms for first- and second-order differentiation of this posterior which can either be used weight-wise to calculate a gradient and Hessian over the weights or functionally to calculate a representation of the functional first- and second-derivative.

Furthermore, semi-analytical optimization approaches for finding a maximum a posteriori estimate are examined, including a first- and a second-order function optimization method. These methods use a pricing oracle approach with active set iteration to dynamically grow the representation. Finally, the computational complexity of the various steps as well as theoretical convergence properties are analysed.

Keywords

Continuous-state POMDP, Dynamics model learning, Function optimization, Mixed-distribution representation, Forward-backward iteration

List of papers

This thesis is based on the work presented in the following manuscripts:

1. **Erik Karlsson Nordling**, Analytical Approaches for Posterior Estimation with Differentiation of Transition Dynamics in a POMDP. Work in progress. 2026.

2. **Erik Karlsson Nordling**, Semi-analytical Decomposition and Solution Strategies for Maximum A Posteriori Estimation of Transition Dynamics in a POMDP. Work in progress. 2026.

Acknowledgements

First and foremost, I would like to thank my supervisor Marina Axelson-Fisk for the excellent support you have given, not least with the proof-reading and the feedback to my ideas. I would also like to thank Erik Kristiansson and Umberto Picchini for the sober feedback on how to write a scientific article, Robert Feldt for the advice on how to implement algorithms efficiently and Ann-Brith Strömberg for the advice on optimization theory.

I would also like to thank my colleagues and fellow PhD students at the Department of Mathematical Sciences, from which I have learned much. Thanks especially to Selma Tabakovic for the many lunchtime discussions and to Isac Boström for the advice on how to deal with numerous administrative details.

Finally I would also like to thank my parents Margareta and Göran for the feedback and emotional support you have provided and I would also like to thank my sister Frida and brother-in-law Henric for the advice on how to structure my work and for taking care of the dogs in my hour of need.

Contents

1 Introduction

A fundamental challenge in autonomous decision-making is controlling a system when its underlying state cannot be observed directly. The Partially Observable Markov Decision Process (POMDP) provides a rigorous mathematical framework for these scenarios, maintaining a probability distribution over the possible states - or beliefs - based on a history of noisy observations [1, 2, 3]. However, a significant complication arises when the system operates in a continuous state space. In this setting, the transition function is not a simple matrix of probabilities, but a probability density function (PDF) on a continuous domain. Estimating this function from data is the primary problem addressed in this thesis.

Commonly used methods for estimating continuous state-space dynamics often rely on state-space discretization (such as fixed grids or adaptive tiling) [4, 5]. While straightforward to implement, these methods artificially cut the state space into pieces, resulting in discontinuous dynamics at the cell boundaries that rarely reflect physical reality. Alternatively, general function approximation methods like neural networks offer higher flexibility [6, 7, 8]. Yet, they lack the ability to be analytically integrated during updates, and require Monte Carlo trajectory sampling or similar statistical approaches. This introduces noise and makes it difficult to provide theoretical guarantees regarding the local uncertainty of the learned model.

This thesis addresses these limitations by proposing a non-parametric Bayesian framework for the estimation of transition dynamics in continuous state-space POMDPs. The core of the proposed method is a representation of the transition probability density function (PDF) through a weighted sum over a finite collection of separable basis functions. This choice allows for the analytical integration of state beliefs, effectively bypassing the need for stochastic sampling and enabling exact first- and second-order differentiation of the posterior log-probability.

The work conducted is divided into two primary parts:

- **Analytical foundations for representation and differentiation (Paper I):** We first establish a framework for calculating the maximum a posteriori (MAP) estimate of the transition dynamics. By using a Gaussian process (GP) prior, we encode regularity assumptions, such as smoothness, directly into the learning process [9]. We derive analytical algorithms for first- and second-order functional differentiation of the posterior, including the weight-gradient and Hessian. This provides the tools necessary for efficient weight optimization, such as the Expectation-Maximization algorithm [10].

- **Adaptive structural refinement and function optimization (Paper II):** Recognizing that a fixed set of basis functions may be insufficient for complex dynamics, we extend the framework to include adaptive model refinement. By employing a column generation procedure inspired by large-scale optimization theory, the model can dynamically add, split or remove basis functions to improve the function approximation based on data [11, 12]. This allows for convergence in the function space.

By unifying exact analytical inference with adaptive function optimization, this thesis provides a flexible, and analytically rigorous method for learning POMDP dynamics. The resulting framework preserves the strengths of classical Bayesian inference while offering the flexibility required to model the intricate dynamics of continuous-world systems.

2 Background

This section provides background for the models and methods used in the included papers. The main focus lies on how to represent and find the hidden transition dynamics of a partially observable Markov decision process (POMDP) when the state space is continuous.

2.1 Partially observable Markov decision processes

A common model used to represent sequential decision making under uncertainty is the Markov Decision Process (MDP) [13]. It models the state evolution of a system and where the transition depends on the current state and actions taken at each time point. The goal is to maximize some cumulative expected reward, usually by finding a decision policy which controls action decisions. An important property is the Markov property which states that the state change at each time step only depends on the current state and action. This means that all possible system dynamics can be expressed as an action-dependent stochastic mapping from the state space back onto itself. Different algorithms exist for finding a good policy for a given MDP, such as Value or Policy iteration algorithms [14, 7].

A Partially Observable Markov Decision Process (POMDP) is a type of MDP where the true state is not exactly know. Instead, we use a *belief*, which is a probability distribution over all possible states, to describe knowledge about it. The belief is updated by new observations, which complicates the problem since a policy must map entire state distributions or observation-action histories to actions instead of singular states [1].

When the system operates in a continuous domain (e.g., physical space, temperature, or velocity), the state space is continuous and the belief is a continuous probability density function. Updating this belief after taking an action and receiving an observation requires integrating over the entire state space using the system's transition dynamics. When the transition dynamics are unknown, estimating them from historical data is a mathematically ill-posed problem, heavily afflicted by the curse of dimensionality [15].

Roughly put, there are two approaches to handling this problem. One is to use strong assumptions on the system to constrain the dynamics to a well-defined form. This can result in a closed-form expression for estimating the transition dynamics and is very efficient, but with the drawback the the assumptions are strong which can result in a suboptimal solution. As such these are not considered further in this thesis. The other approach is to estimate the transition dynamics through function approximation.

2.2 Function approximation in POMDPs

Historically, continuous state spaces have been tackled by discretizing them into a grid of finite states [16]. While this allows the use of standard discrete POMDP algorithms, it assumes the system behaves uniformly within each cell and jumps discontinuously between them, which violates the physics of most real-world systems.

The alternative is function approximation methods where the transition dynamics are modelled as a continuous mathematical function [17]. Modern machine learning approaches, such as deep neural networks, excel at this by offering immense expressive power. However, they introduce a new problem: they cannot be integrated analytically. Updating the POMDP belief with a neural network transition model requires stochastic Monte Carlo sampling, which is computationally expensive and obfuscates theoretical guarantees about the model's uncertainty [8]. The theoretical gap this thesis addresses is the need for an analytical approach for continuous function approximation of transition dynamics with minimal assumptions on the form of the dynamics. The approach is related to mixed-Gaussian approaches, but without their non-separable limitations [18, 19].

2.3 Bayesian inference and Gaussian processes

When learning dynamics from limited data, multiple different transition functions might explain the observed trajectories equally well. Bayesian inference solves this by introducing a prior—a mathematical formulation of our assumptions before seeing the data [20]. The use of a prior also prevents overfitting since unreasonable dynamics are penalized, where the concept of a ”reasonable” dynamic is encoded in the prior.

For continuous functions, a powerful prior is the Gaussian Process (GP) [9]. Instead of putting a prior on a single parameter, a GP places a probability distribution over the infinite space of all possible functions. By carefully designing the mean and covariance kernel of the GP, one can encode physical regularities, such as requiring the learned transition dynamics to be smooth, rather than jagged or erratic.

There are other prior distributions for continuous functions as well. An important property of the Gaussian prior is that it corresponds to quadratic regularization on a posterior probability density.

2.4 Dynamic model growth and column generation

A fundamental dilemma in modelling is choosing the complexity of the model. A model that is too simple cannot capture the true dynamics; a model that is too complex can overfit to the data and become computationally paralysing.

In large-scale mathematical optimization, a technique called Column Generation is used to solve problems with millions of variables [11]. Instead of optimizing all variables at once, the algorithm starts with a small, manageable set of *active* variables. It then uses the derivatives of the problem to calculate the "price" or "reduced cost" of the unused *inactive* variables, adding them to the model only if they are proven to improve the objective. The benefit of this approach compared to using a fixed variable set is that only variables that are actually useful for solving the problem are added. This avoids unnecessary computations for variables that don't contribute enough to the solution. This thesis borrows and extends this concept to the infinite-dimensional setting, treating the building blocks of the transition function as variables that can be dynamically generated as the data demands.

3 Methods

This section details the specific mathematical methods developed in the included papers to solve the continuous POMDP learning problem. We begin by establishing the formal notation, followed by a description of the representation framework and the method choices made.

3.1 Mathematical notation and problem formulation

We begin by formalizing the components of the continuous-state POMDP[1, 2, 3]:

- A state space $\mathcal{S}$ which describes all possible states of the process. $\mathcal{S} \subset \mathbb{R}^N$ is assumed to be compact. For simplicity $\mathcal{S} = [0, 1]^N$ will be used. (The concepts here are applicable to any separable probability space.)
- An action space $\mathcal{A}_x \ \forall \ x \in \mathcal{S}$ which contains all possible actions for each state $x \in \mathcal{S}$. We assume that $\mathcal{A}_x = \mathcal{A} \ \forall \ x \in \mathcal{S}$, i.e. all possible actions are available in each state.
- An observation space $\mathcal{Z}$ that contains all possible observations of the process.
- An observation function $O : \mathcal{Z} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ that describes the probability for an observation given the current state and action. We assume that this is known and further that, $\forall \ z \in \mathcal{Z}, a \in \mathcal{A}$; $O(z|\cdot, a) : \mathcal{S} \to \mathbb{R}$ is bounded and measurable.
- A reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ which at each time step assigns a reward to each state.
- An initial state belief $b_{\text{prior}} : \mathcal{S} \to \mathbb{R}_+$ which is a probability density function of the state at time point 0 without knowledge of future observations.

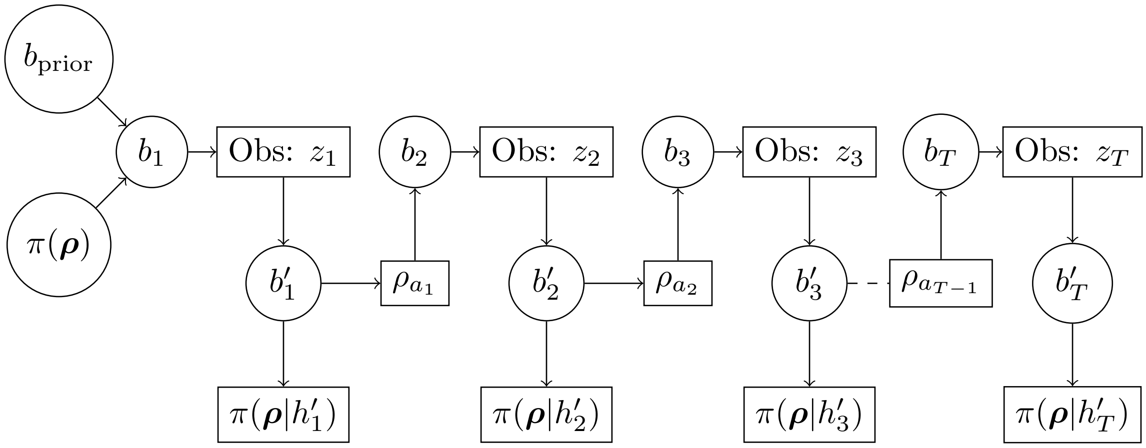

Figure 1: A graph describing the relations in the POMDP dynamics. b_t and b'_t are state beliefs at different stages of the process. "Obs: z_t" denotes obtaining observation z_t while "ρ_{a_t}" denotes state transition according to the dynamics $\boldsymbol{\rho}$ given action a_t. $\pi(\boldsymbol{\rho}|h'_t)$ denotes the marginal posterior probability density of the transition dynamics $\boldsymbol{\rho}$ given the history $h'_t = [z_1, a_1, z_2, ...a_{t-1}, z_t]$.

- A transition function $\pi : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ which describes the dynamics of the process. It is the conditional PDF of the next state given the current state and action. In some applications, the transition function depends on the current time as well but time invariance is assumed here.

Because the true state x is hidden, the system maintains beliefs, $b_t(x)$ and $b'_t(x)$, which are probability density functions over $\mathcal{S}$ representing the current knowledge of the state at time t. When an action a_t is executed and a new observation z_{t+1} is recorded, the belief is updated recursively using Bayes' rule:

$$b'_t(x) \propto O(z_{t+1}|x, a_t) b_t(x)$$
$$b_{t+1}(x) = \int_{\mathcal{S}} \rho_{a_t}(x|y) b'_t(y) dy$$

The core technical problem addressed in this thesis is that when $\rho_a(x|y)$ is unknown, evaluating this continuous integral numerically over multiple time steps becomes computationally intractable, making standard inference impossible.

3.2 Separable basis function representation

To solve the integration problem without resorting to discrete grids or stochastic sampling, this thesis represents the transition PDF using a finite collection of separable basis functions [17, 18]. We approximate the unknown transition dynamics for an action a as:

$$\tilde{\rho}_a(x|y) = \boldsymbol{\phi}_a^T(x) \Theta_a \boldsymbol{\phi'}_a(y) \tag{1}$$

where $\boldsymbol{\phi}_a(x)$ and $\boldsymbol{\phi'}_a(y)$ are vectors of basis functions (such as Bernstein polynomials or rectangular pulses) defined over the state space, and Θ_a is a matrix of optimization weights.

The choice of separable bases is the mathematical engine of this framework, representing a deliberate and critical trade-off. By algebraically decoupling the post-transition state x from the pre-transition state y, the continuous integral in the POMDP belief update is finitely separated over the basis functions. Consequently, the belief update collapses into a simple, finite matrix-vector multiplication:

$$\boldsymbol{\mu}_{t+1} = \Theta_{a_t} \boldsymbol{\mu'}_t$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$ are finite coefficient vectors. This entirely removes the need for Monte Carlo trajectory sampling. The inherent limitation, however, is that separable functions struggle to efficiently represent highly correlated, diagonal transitions in multi-dimensional spaces (e.g., a rotated 2D Gaussian kernel cannot be perfectly represented as a finite sum of separable 1D functions) [21]. We accepted this limitation because the computational advantage—specifically, avoiding an exponential explosion in the belief representation during sequential updates—far outweighs the loss of expressive freedom.

3.3 The Bayesian framework

In Paper 1, we establish the framework for the Maximum A Posteriori (MAP) estimation of the weight matrices Θ_a [20]. Because the transition dynamics exist in an infinite-dimensional function space, formulating a rigorous MAP estimate requires careful treatment of the prior measure [22]. To ensure the learned dynamics are physically plausible, we incorporate a Gaussian Process (GP) prior [9]. This prior is integrated into the log-posterior as a quadratic regularization term governed by an inverse covariance operator, Q^{-1}:

$$\ln \pi(\boldsymbol{\rho}) \propto -\frac{1}{2} \langle (\boldsymbol{\rho} - \boldsymbol{\beta}), \boldsymbol{Q}^{-1}(\boldsymbol{\rho} - \boldsymbol{\beta}) \rangle$$

Using a variant of the Forward-Backward algorithm [23], we derive the exact analytical first- and second-order functional derivatives (gradients and Hessians) of this posterior [24]. These exact derivatives allow us to use robust optimization algorithms, such as a pseudo-Newton Expectation-Maximization (EM) variant, to find the optimal weights [10]. This is described further in paper 1.

An alternative to the Gaussian Process prior would be the Dirichlet Process (DP), which is commonly used as a prior in discrete Bayesian POMDPs [25]. The motivation for the GP prior is as follows: The classical DP measure generates realizations that are zero almost everywhere, making it fundamentally unsuitable for representing continuous, non-negative transition PDFs without adding complex, unwanted smoothing layers. The GP prior, particularly when using a Sobolev-space inner product for Q^{-1} [26], naturally enforces continuous spatial regularity (smoothness) and translates elegantly into standard quadratic regularization during optimization.

Testing the pseudo-Newton EM weight optimization highlighted critical behavioural differences between basis families [27]. As shown in the results, Bernstein polynomials ensure infinite smoothness but converge slowly when the prior allows for the true transition dynamics to contain sharp edges[28]. Conversely, rectangular basis functions (piecewise constant) capture sharp shifts rapidly but sacrifice overall smoothness [29, 30]. The representation framework and Bayesian approach ultimately allows the modeller to explicitly select the basis family that best matches prior physical knowledge of the system, while the algorithm handles the optimal placement and scaling of those bases.

3.4 Adaptive refinement via column generation

Paper 2 addresses a fundamental limitation of the approach in Paper 1: guessing the correct basis functions beforehand is difficult. If the chosen bases are too broad, the model lacks precision; if they are too fine, computational costs explode.

Using the exact functional gradients and Hessians derived in Paper 1, we formulate a Column Generation subproblem [12]. Extending the principles of conditional gradient methods to infinite functional domains [31], the algorithm systematically evaluates candidate functions and calculates their "price", the mathematically expected improvement in the posterior log-probability [11].

If a candidate basis function proves sufficiently valuable, it is dynamically added to the representation. Furthermore, if the basis function family allows it, the algorithm can "split" an existing basis function (e.g., dividing a broad polynomial into narrower, higher-degree ones) to locally increase the resolution of the model only where the data demands it. The expected improvement generated by a split can be estimated via its price as well.

In paper 2 it was also shown that the adaptive refinement allows for convergence to a local MAP estimate in the functional sense, but a couple of mild assumptions on the problem and prior are necessary for this.

3.5 First-order vs second-order subproblem

Two different subproblems were examined. The principle behind each of them is to use the mentioned differentiation methods to estimate the local behaviour of the MAP-estimate in the *functional sense*, not limited to the current representation space. The first-order approach uses a linear estimate of the functional dynamics which is regularized by the prior precision Q^{-1}, while the second-order approach uses a quadratic estimate. Although similar in principle, they give rise to different challenges when used for pricing. The intuitive interpretation is that the first-order approach generates basis functions that aligns well with the descent direction in the function space, while the second order approach generates basis functions that aligns well with the functional Newton-step.

4 Overview of papers

4.1 Paper 1: Analytical Approaches for Posterior Estimation with Differentiation of Transition Dynamics in a POMDP

The first paper explores a number of different properties of the Bayesian a posteriori probability estimate of a POMDP transition pdf represented on the form in equation 1.

Focus lies on the following topics:

- Analytical (non-stochastic) calculation, up to a constant offset, of the posterior log-probability of a proposed transition PDF $\tilde{\boldsymbol{\rho}}$ given data h'_T. In other words $\ln \pi(\tilde{\boldsymbol{\rho}}|h'_T)$.
- Efficient algorithms for analytical differentiation of the posterior log-probability, specifically:
  - The functional derivative.
  - The functional second derivative.
  - The gradient w.r.t. the weights of the basis functions.
  - The Hessian w.r.t. the weights of the basis functions.

  The gradient and Hessian are useful for first- and second-order weight optimization algorithms respectively. The functional derivatives on the other hand allow for analysis of local properties in the functional space of transition dynamics, which is useful for the column generation described in paper 2.
- A variant of the EM-algorithm for finding optimal representation weights. This algorithm turns out to be straightforward to implement in the analytical framework.
- A comparison between rectangular (piecewise constant) and Bernstein polynomial (infinitely smooth) basis functions. The results show that the rectangular basis functions give a higher MAP value when allowed by the prior.

The computational complexity of the analytical calculations is also presented. Most notably: the complexity of differentiation is linear in terms of the data amount T with a small caveat for the functional second-derivative, as the complexity can be quadratic in T if the observations are continuous. However, this complexity becomes linear as well if it is evaluated via an inner product with a proposed transition dynamic.

4.2 Paper 2: Semi-analytical Decomposition and Solution Strategies for Maximum A Posteriori Estimation of Transition Dynamics in a POMDP

The second paper describes an algorithm designed to build and optimize the basis function representation based on data. The algorithm iterates between two phases:

- **Restricted master problem for weight optimization**

  In this step, the weights for the current set of basis functions are modified to maximize the posterior log-probability of the transition PDF $\ln \pi(\tilde{\boldsymbol{\rho}}|h'_T)$. The EM algorithm presented in paper 1 can be utilized, but other approaches are presented as well.

- **Column generation subproblem for dynamic basis function updates**

  In this step, basis functions are added to, removed from or split inside the current representation $\tilde{\boldsymbol{\rho}}$. The addition occurs via the column generation procedure described in section 3.5. As mentioned, both a first-order and a second-order approach is examined.

The computational complexity and outer convergence rates of various steps are examined and analysed. As mentioned in section 3.4, the algorithm can be shown to converge under mild assumptions.

5 Discussion

5.1 Comparison between rectangular and Bernstein basis function families

In the results in paper 1 the algorithms seem to converge towards different values for different degrees. Compared to the rectangular basis functions, the convergence of the Bernstein bases appears to be slower. This is likely due to the fact that the rectangular basis functions are less smooth and can more exactly approximate sharp peaks in the transition PDF. This highlights the fact that basis functions that are non-smooth should be used when allowed by the prior.

5.2 Comparison between the first- and second-order column generation

Staying with the analogy to the finite-dimensional setting, the second-order approach is expected to generate more useful basis functions since the quadratic estimate is a better approximation of the posterior density than the linear regularized one. The drawback is the on-average computational complexity per tested basis function. While both approaches have linear complexity in terms of the amount of data, the second-order approach scales on average quadratically with respect to the current size of the representation, while the first-order approach scales only linearly. As such it is not possible to conclude that any of the approaches converges faster than the other. What can be concluded is that the second-order approach is the suitable choice when the sparsity of the representation is more important than the convergence rate.

6 Concluding remarks and future work

6.1 Extension to continuous time

In paper 1, the possibility to extend the problem to the continuous time setting with time-discrete events for observations and actions is examined. The biggest difference is that a generator for a time-invariant transition operator is the object which needs to be estimated. The representation size of the belief will grow over time in this setting, which needs to be handled adequately. The biggest issue found is however how to enforce non-negativity of the transition PDFs while ensuring general enough approximation.

6.2 Dual relaxation of the non-negativity of the transition PDF

In both papers, it is argued that the only safe way to enforce non-negativity of a represented transition PDF $\tilde{\boldsymbol{\rho}}$ is to assume non-negative weights and non-negative basis functions. This severely limits which basis function families can be used and makes it difficult to represent sharply swinging transitions. A possible future direction presented in paper 2 is to relax this constraint with dual relaxation.

In one sense it just means delegating the non-negativity constraint to a dual variable, so the issue remains. In most classical optimization problems nothing would be gained from relaxing this constraint, but the key point is that *one can use different basis function families for the primal and dual functions.* Thus it is enough to use basis functions which fulfil the strong non-negativity requirement to approximate the dual while more general basis functions can be used to approximate the primal function.

For example, Fourier basis functions does not fulfil the strong requirement but one can still use this family to approximate the primal function if, for example, rectangular basis functions are used to approximate the dual function. Also, the dual basis functions need not fulfil any requirements, such as continuity, enforced through the GP prior.

6.3 Extension to the reward function and Bellman equation

One possible future direction of inquiry concerns how to use this approximation framework for finding a useful policy to control the system, for example through policy iteration.

In paper 1 it is shown that if a transition dynamics $\tilde{\boldsymbol{\rho}}$ is represented with the basis function formulation in equation 1, the related Bellman equation can be reduced to a finite-state POMDP Bellman equation where the dimension depends on the representation size. This is a further argument for finding a representation of $\tilde{\boldsymbol{\rho}}$ which is as sparse as possible.

However, the more interesting question is how to take the entire posterior distribution of transition dynamics $\boldsymbol{\rho}$ into account, not just single elements. What complicates the situation is the fact that the value function cannot simply be averaged over the transition PDF distribution. The policy will instead take as argument a joint belief over the state and the transition PDF. Although the Bellman equation looks similar, it cannot be reduced to a finite form in the same nice way. A different type of algorithm is thus necessary in this setting.

References

[1] George E Monahan. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. In: *Management science* 28.1 (1982), pp. 1–16.

[2] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. In: *Artificial Intelligence* 101.1 (1998), pp. 99–134. ISSN: 0004-3702.

[3] Karl Johan Åström. Optimal control of Markov processes with incomplete state information I. In: *Journal of mathematical analysis and applications* 10 (1965), pp. 174–205.

[4] Konstantinos Lolos et al. Adaptive State Space Partitioning of Markov Decision Processes for Elastic Resource Management. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. 2017, pp. 191–194.

[5] Ivan S.K. Lee and Henry Y.K. Lau. Adaptive state space partitioning for reinforcement learning. In: *Engineering Applications of Artificial Intelligence* 17.6 (2004), pp. 577–588. ISSN: 0952-1976.

[6] Ilsang Ohn and Yongdai Kim. Smooth Function Approximation by Deep Neural Networks with General Activation Functions. In: *Entropy* 21.7 (2019). ISSN: 1099-4300.

[7] Martin Riedmiller. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In: *Machine learning: ECML 2005: 16th European conference on machine learning, Porto, Portugal, October 3-7, 2005. proceedings 16*. Springer. 2005, pp. 317–328.

[8] Timothy P. Lillicrap et al. *Continuous control with deep reinforcement learning*. 2019. arXiv: `1509.02971 [cs.LG]`.

[9] A. M. Stuart. Inverse problems: A Bayesian perspective. In: *Acta Numerica* 19 (2010), pp. 451–559.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[11] Ricardo García, Angel Marín, and Michael Patriksson. Column generation algorithms for nonlinear optimization, I: Convergence analysis. In: *Optimization* 52.2 (2003), pp. 171–200.

[12] Renaud Chicoisne. Computational aspects of column generation for nonlinear and conic optimization: classical and linearized schemes. In: *Computational Optimization and Applications* 84.3 (2023), pp. 789–831.

[13] Martin L. Puterman. Chapter 8 Markov decision processes. In: *Stochastic Models*. Vol. 2. Handbooks in Operations Research and Management Science. Elsevier, 1990, pp. 331–434.

[14] Martijn van Otterlo and Marco Wiering. *Reinforcement learning : state-of-the-art*. Adaptation, learning, and optimization. Springer, 2012.

[15] Richard Bellman. *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[16] Stéphane Ross et al. A Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. In: *Journal of Machine Learning Research* 12.5 (2011).

[17] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 4532–4576.

[18] M. D. Buhmann. Radial basis functions. In: *Acta Numerica* 9 (2000), pp. 1–38.

[19] Nikolaos Kariotoglou et al. The linear programming approach to reach-avoid problems for Markov decision processes. In: *J. Artif. Int. Res.* 60.1 (Sept. 2017), pp. 263–285. ISSN: 1076-9757.

[20] Andrew Gelman et al. *Bayesian data analysis*. Chapman Hall, 1995.

[21] Ingo Steinwart and Andreas Christmann. Support vector machines. In: Springer Science & Business Media, 2008. Chap. 4.

[22] Masoumeh Dashti et al. MAP estimators and their consistency in Bayesian nonparametric inverse problems. In: *Inverse Problems* 29.9 (2013), p. 095017.

[23] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[24] Walter Rudin. *Functional analysis*. McGraw-Hill, 1991.

[25] Jayaram Sethuraman. A Constructive Definition of the Dirichlet Prior. In: *Statistica Sinica* 4 (Jan. 1994), pp. 639–650.

[26] 4 - The Sobolev Imbedding Theorem. In: *Sobolev Spaces*. Ed. by Robert A. Adams and John J.F. Fournier. Vol. 140. Pure and Applied Mathematics. Elsevier, 2003, pp. 79–134.

[27] Jorge Nocedal and Stephen J Wright. Numerical optimization. In: Springer, 2006. Chap. 3.

[28] George G. Lorentz. *Bernstein Polynomials*. 2nd. New York: Chelsea Publishing Company, 1986. ISBN: 9780828403238.

[29] Carl De Boor. *A Practical Guide to Splines*. New York: Springer-Verlag, 1978.

[30] Mark Ainsworth and J.Tinsley Oden. A posteriori error estimation in finite element analysis. In: *Computer Methods in Applied Mechanics and Engineering* 142.1 (1997), pp. 1–88. ISSN: 0045-7825.

[31] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. In: *Mathematical Programming* 152.1-2 (2015), pp. 75–112.