Technische Universiteit
Eindhoven
University of Technology

Control Systems,
Electrical Engineering Department

Den Dolech 2, 5612 AZ Eindhoven
P.O. Box 513, 5600 MB Eindhoven
The Netherlands
www.tue.nl

Date
18 September 2014

Data-driven Property Verification of Grey-box Systems with Bayesian Experiment Design

S.Haesaert, P.M.J.Van den Hof, A.Abate

Data-driven Property Verification of Grey-box Systems with Bayesian Experiment Design

S.Haesaert, P.M.J.Van den Hof, and A.Abate[*][†]

Sunday 5th October, 2014

Abstract

A measurement-based statistical verification approach is developed for systems with partly unknown dynamics. These grey-box systems are subjected to identification experiments that enable accepting or rejection system properties expressed in a linear-time logic. By encompassing the partial knowledge on the system dynamics into an a-priori defined uncertainty distribution over a parameterised model set, we can use a Bayesian framework that includes optimal experiment design and the computation of a confidence level on an ascertained property via bayesian parametric inference over the acquired experiments. Applied to physical systems, this work enables the data-driven verification of only partly known systems with controllable non-determinism and noisy output observations. An approximate solution for the Bayesian experiment design problem is given on a case study. This numerical case study is used to elucidate this data-driven and model-based verification technique, by considering the safety of a discrete-time, continuous-space dynamical system.

1 Introduction

The design of complex, high-tech, safety-critical systems such as autonomous vehicles, intelligent robots, and cyber-physical infrastructures demands guarantees on their correct and reliable behaviour. These guarantees can be attained by the use of formal methods [1]: within the computer science area, the verification of software and hardware via formal methods has successfully led to industrially-relevant applications. Carrying the promise of a decrease in design and implementation errors, the use of formal methods has already become the standard in the avionics, automotive, and railway industries [2]. These application domains have demanded the extension of finite state models from the computer sciences to physical and cyber-physical models.

Formal methods prove or verify the validity of formal specifications. These are specifications formulated in mathematically sound terms, for example by means of Linear-time Temporal Logic (LTL). The use of formal specifications

[*] S.Haesaert and P.M.J.Van den Hof are with the Faculty of Electrical Engineering, Eindhoven University of Technology, The Netherlands s.haesaert@tue.nl, p.m.j.vandenhof@tue.nl

[†] A. Abate is with the Department of Computer Science, Oxford University, UK alessandro.abate@cs.ox.ac.uk

to describe the desired behaviour is quite versatile, and allows expressing time-dependent properties in the human language. Examples of such specifications include reachability, obstacle avoidance, stability, and recurrence. The mathematical structure of these specifications enables the use of automatic verification methods over classes of models.

Much recent research has concerned the extension of formal methods from finite-state models, widely employed for software and hardware verification, to models of complex systems. Based on enumerable abstractions, the decidability of LTL properties over linear time-invariant (LTI) control models has been shown in [3]. Building on these results, recent applications can be found in robotics for symbolic motion planning [4]; in cyber-physical systems and hybrid systems for the specification of complex tasks interleaved discrete/continuous behavioural space [5]; and in the analysis of biological systems such as genetic regulatory networks [6]. The recent work in [7] extends standard Model Predictive Control (MPC) to LTL-MPC, which is the design of optimal controllers satisfying functional or reliability requirements written as LTL formulae.

The strength of formal techniques is limited by the underlying assumption of access to a model built from the full knowledge of the behaviour of the underlying system under study. In reality for most physical systems the exact dynamical behaviour is mechanistically known only in part, and the state of the model is not fully accessible. Lack of full knowledge of the system can be encompassed by structuring the available knowledge as an uncertainty distribution over a model parameterisation. Measurements of physical systems are usually available as time series of input and noisy output signals. Even when the system state is only partly observable via the output, system measurements carry information on its dynamics: as such, they add to the available knowledge on the system and can be used to refine and possibly decrease the uncertainty over the model.

The goal of this line of research is data-driven and model-based verification for partly unknown physical systems that are accessible via noisy input-output measurements. With focus on the properties of interest over the systems, we are not limited to a single type of safety/reliability specification, but in the automatisation of this method towards general formal specifications. Moreover, we investigate the computation of identification experiments that optimally excite the system with respect to specification of interest. We thus quite naturally assume that the system is available for experiments in an environment where we can change its input signal at will.

The area of system identification [21, 22, 23] investigates measurement-based modelling of physical systems. It handles the modelling of physical systems via noisy input-output measurements, with model dynamics over uncountable (continuous) state spaces. Input signals excite the system behaviour, which is observed via measurements, and which can be chosen to maximise the amount of information gained. As the optimal input typically depends on the knowledge of the true system, the literature distinguishes three approaches: an iterative approach, where an estimate of the nominal system is used to design the experiment at each stage; a min-max design that is robust to the worst-case situation; and a Bayesian design that uses the expected value of a criterion computed on the prior uncertainty distribution over the model [24]. For the identification of dynamical systems for classical applications towards controller synthesis the first approach predominates [21, 22, 23], whereas some work has been done on the robust experiment design using the min-max approach [25]. On the other hand the third approach, well known from Bayesian statistics [24], is not widely employed in the research domain of identification.

In this work we focus on the verification of systems modelled within a linearly-parameterised class of deterministic[1] input-output models using Bayesian identification and experiment design. The contribution shows that for a subset of LTL specifications the confidence on the validity of a formal property can be computed using Bayesian inference over a finite sample set (cf. Section 2). Since the performance of this data-driven method depends on the design of the experiment, we further define an optimality criterium that allows selecting an input using Bayesian experiment design. We display this approach on the safety verification of linearly-parameterised models (cf. Section 3), where a numerical case study of the Bayesian experiment design is discussed.

Related work

Recent work within the formal methods community also concerns the use of simulations and of measurements for verification. Statistical Model Checking (SMC) [8, 9, 10, 11] uses executions generated by the model to find statistical evidence for the verification of bounded logical properties, which are defined over finite executions. SMC can be applied to black-box systems, as adopted in [12, 10], which have a probability distribution that is not fully known and can only be model checked over a finite set of pre-computed executions. Though SMC is well applicable to physical systems with unknown dynamics, it is in general limited to state-observable and fully-stochastic systems. The presence of sets of inputs, disturbances, and unknown initial states or other forms of non-determinisms, are not natively incorporated into SMC, since they do not allow a quantification of the probability of properties. Extensions towards including non-determinism have been attained in [9, 13], where preliminary steps towards extending SMC to Markov decision processes are made. In comparison to SMC, [14, 15, 16, 17] efficiently use data drawn from an input-output, finite state Markov system, to learn the corresponding model and to verify it. These results are bound to finite-state models, making them less applicable to more complex systems. Similarly, [18, 19] use advanced machine learning techniques to infer finite-state Markov models from data over given logical formulas. SMC is used as a tool in [20] to compute the satisfaction probability of a formula over stochastic models with parametric uncertainty.

2 Data-driven and Model-based verification

Let us recapitulate the overall goal of this work: starting from available a-priori knowledge over system $\mathbf{S}$, iteratively and efficiently gather measurements until a specification ψ defined over the system is verified or falsified with a given confidence δ.

[1] In physics a ‘deterministic system’ has state evolutions which are not affected by randomness. Hence this definition should not be confused with the term non-determinism used in Computer Science for Markov models, which indicates a state distribution that is not fully determined by the previous state – and, as such, governed by external non-determinism such as an input signal.

System and Models

The system, denoted by $\mathbf{S}$ as in Figure 1, is measured in discrete time. An input signal $u(t) \in \mathbb{U}, t \in \mathbb{N}$, captures how the environment acts on the system. Similarly, the output $y_0(t) \in \mathbb{Y}$ indicates how the system interacts with the environment (namely, how it can be measured). The measurements $\tilde{y}(t)$ at $t \in \mathbb{N}$ of $y_0(t)$ are disturbed by the measurement noise $e(t)$.

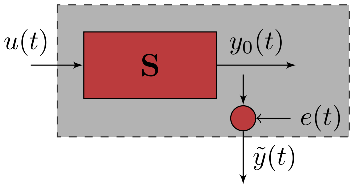

Figure 1: System $\mathbf{S}$ has input $u(t)$ and output $y_0(t)$. In the measurement setup, the measured output $\tilde{y}(t)$ includes the system output $y_0(t)$ and the measurement noise $e(t)$.

The behaviour of a deterministic system can be described by mathematical models as a (causal) relation between the system input and output. In most cases the knowledge of the behaviour of a system is only partial, making it impossible to represent the system by a "true" model. In such cases, a-priori available knowledge allows to construct a model set $\mathcal{G}$, with elements $\mathbf{M} \in \mathcal{G}$ representing possible mathematical models of $\mathbf{S}$. Let us denote a parameterisation of the model set $\mathcal{G}$ as the mapping $\mathbf{M}(\cdot) : \Theta \to \mathcal{G}$, from the parameters $\theta \in \Theta$ in the parameter set, which is a subset of a Euclidean space $\Theta \subset \mathbb{R}^d$, to the models $\mathbf{M}$ in $\mathcal{G}$. This allows for a parametrised expression of the model set as $\mathcal{G} = \{\mathbf{M}(\theta) | \theta \in \Theta\}$. The chosen parameterised model set is assumed to contain the "true" model denoted as $\mathbf{M}(\theta^0)$, $\theta^0 \in \Theta$, which exactly represents the behaviour of the system $\mathbf{S}$. Note that for any linear system there exists a set of linearly-parameterised models that is capable of representing it. Hence $\mathcal{G}$ encompasses the part of the behaviour that is mechanistically known. The remaining uncertainty about $\mathbf{M}(\theta^0)$ is structured as a distribution over the parameter set Θ. It is then the (unknown) model denoted by $\mathbf{M}(\theta^0) = \mathbf{S}$ that we would ideally like to formally model-check.

Whenever the lack of knowledge on the system behaviour impedes a formal verification step, it is still possible to collect data of the system by exciting it with an input sequence $\mathbf{u} = \begin{bmatrix} u(0) & u(1) & \dots & u(N_s - 1) \end{bmatrix}^T$, with N_s the length of the input sequence. This data set consists of samples drawn from the underlying physical system via the measurement setup, as depicted in Figure 1. It contains information on the system dynamics, even when the system is not fully observable and in particular when noisy observations $\tilde{y}(t)$ of the output $y_0(t)$ are measured. Classical noise characteristics deal with Gaussian white noise $e(t)$ that is additive to $y_0(t)$, i.e. $\tilde{y}(t) = y_0(t) + e(t)$. The measurement experiment consists of N_s input-output samples drawn from the system: let us denote the output samples obtained by exciting the system with the input $\mathbf{u}$ as $\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{y}(0) & \tilde{y}(1) & \dots & \tilde{y}(N_s - 1) \end{bmatrix}^T$. Since the collected data contains statistical information on the behaviour of the system, it is possible to refine the uncertainty distribution over the parameter space, as discussed in the second part of this section.

Properties

Starting from a finite set of atomic propositions $p_i \in AP$, $i = 1, \ldots, |AP|$, Linear-time Temporal Logic (LTL) [3] formulae are built recursively via the syntax $\psi ::= \text{true} \mid p \mid \neg\psi \mid \psi \wedge \psi \mid \psi \vee \psi \mid \bigcirc\psi \mid \psi \,\mathsf{U}\, \psi$. Let $\pi = \pi(0), \pi(1), \pi(2), \ldots \in \Sigma^{\mathbb{N}^+}$ be a word composed of letters from the alphabet $\Sigma = 2^{AP}$, let $\pi_t = \pi(t), \pi(t+1), \pi(t+2), \ldots$ be a subsequence of π, then the satisfaction relation between π and ψ, namely $\pi \vDash \psi$ (or equivalently $\pi_0 \vDash \psi$) is defined recursively over π_t and the LTL syntax as

$$\begin{array}{lll}
\text{(true)} & \pi_t \vDash \text{true} & \Leftrightarrow \text{true} \\
\text{(atomic proposition)} & \pi_t \vDash p & \Leftrightarrow \; p \in \pi(t) \\
\text{(negation)} & \pi_t \vDash \neg\psi & \Leftrightarrow \; \pi_t \nvDash \psi \\
\text{(conjunction)} & \pi_t \vDash \psi_1 \wedge \psi_2 & \Leftrightarrow \; \pi_t \vDash \psi_1 \text{ and } \pi_t \vDash \psi_2 \\
\text{(disjunction)} & \pi_t \vDash \psi_1 \vee \psi_2 & \Leftrightarrow \; \pi_t \vDash \psi_1 \text{ or } \pi_t \vDash \psi_2 \\
\text{(next)} & \pi_t \vDash \bigcirc\psi & \Leftrightarrow \; \pi_{t+1} \vDash \psi \\
\text{(until)} & \pi_t \vDash \psi_1 \,\mathsf{U}\, \psi_2 & \Leftrightarrow \; \exists\, i \in \mathbb{N} : \pi_{t+i} \vDash \psi_2, \\
 & & \quad \text{and } \forall j \in \mathbb{N} : \\
 & & \quad 0 \le j < i, \pi_{t+j} \vDash \psi_1
\end{array}$$

Of interest are formal properties encoded over the input-output behaviour of the system over the time horizon $t \geq 0$. Starting at an arbitrary time (say $t = 0$), the set of initial states of the system is given: we assume that this set encompasses the knowledge of past inputs and/or outputs of the system, and that it will be reflected in the parameterised models. The input signal is bounded and represents the possible external nondeterminism of the environment acting on the system. The output $y_0(t) \in \mathbb{Y}$ is labeled by a map $L : \mathbb{Y} \to \Sigma$, which assigns letters in the alphabet Σ to half spaces on the output, as $A_{p_i} y_0(t) \leq b_{p_i}$. In other words, sets of atomic propositions are associated to intervals over $\mathbb{Y} \subset \mathbb{R}$. A system, or equivalently the model that represents it, satisfies a property if all the words generated by the model verify that property. Since properties are encoded over the external (input-output) behaviour of the system, which is the behaviour of $\mathbf{M}(\theta^0)$, $\theta^0 \in \Theta$, we can equivalently assert that any property ψ is verified by the system, $\mathbf{S} \vDash \psi$, if and only if it is verified by the unknown model representing the system, namely $\mathbf{M}(\theta^0) \vDash \psi$. Let us underline that properties are defined over the behaviour of the system, and not over the noisy measurements $\tilde{y}(t)$ of the system. Let us define Θ_ψ to be the maximal feasible set of parameters, such that for every parameter in that set the property ψ holds, i.e. $\forall \theta \in \Theta_\psi : \mathbf{M}(\theta) \vDash \psi$. This set has been alternatively described [20] as the level set of a satisfaction function, however since we are working with deterministic models the satisfaction function would only take binary values.

System Verification in a Bayesian Framework

We now argue that the characterisation of a distribution over the parameter set Θ can be used to compute a confidence in the satisfaction relation over the system $\mathbf{S} \vDash \psi$. This distribution encompasses the current uncertainty over

$\mathbf{M}(\theta^0) = \mathbf{S}$, and can be characterised and refined using measurements of the system.s Therefore, it is possible to accept or reject $\mathbf{S} \vDash \psi$ by drawing data from the measurement set-up until a certain confidence level is achieved. Due to the dynamic interaction between input and output over $\mathbf{S}$, the necessary size of the data set to attain this confidence level depends on the chosen input data $\mathbf{u}$. This leads to an experiment design task: in order to optimise data efficiency, we structure the process of drawing and processing data as an iteration over 3 main stages, as represented schematically in Figure 2:

I. design (and perform) an experiment,

II. compute the corresponding parametric inference,

III. check if confidence in $\mathbf{S} \vDash \psi$ or in $\mathbf{S} \nvDash \psi$ is $> 1 - \delta$, else go to step I.

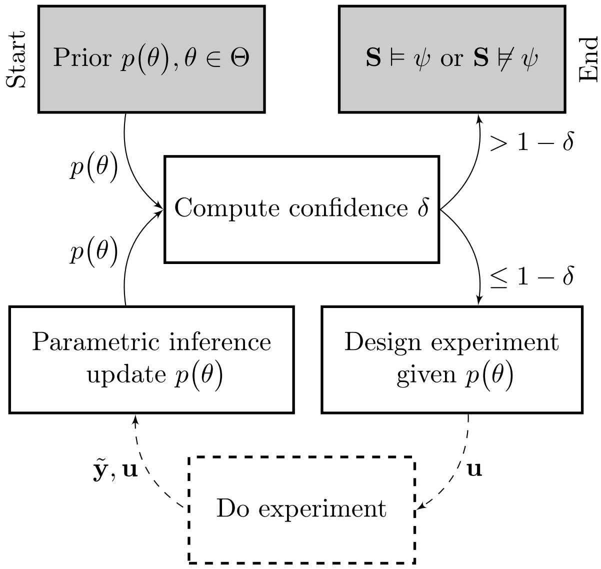
Start
Prior $p(\theta), \theta \in \Theta$
$\mathbf{S} \vDash \psi$ or $\mathbf{S} \nvDash \psi$
End
$p(\theta)$
$> 1 - \delta$
Compute confidence δ
$p(\theta)$
$\leq 1 - \delta$
Parametric inference update $p(\theta)$
Design experiment given $p(\theta)$
$\tilde{\mathbf{y}}, \mathbf{u}$
$\mathbf{u}$
Do experiment

Figure 2: The process of sequentially drawing data with the objective of verifying whether $\mathbf{S} \vDash \psi$ or not with a given confidence $1 - \delta$. Starting from a prior parameters distribution $p(\theta)$, the iteration comprising the computation of the confidence, the experiment design, the access to measurements and the parametric inference is initiated, and is continued until a confidence level of $1 - \delta$ is achieved. In the diagram the white blocks depict the key elements of the iteration, whereas the gray blocks represent the starting/ending parts.

With reference to Figure 2, the iteration is initiated with the construction of a distribution ('Prior'), which structures the initial limited knowledge by assigning a probability measure over the set of parameters. The first part in the iteration is to compute the confidence ('Compute confidence'), which allows leaving the iteration when the desired confidence level is achieved. On the contrary if the confidence level is not achieved, an identification experiment is designed to obtain more data from the system ('Design experiment'). The results of the experiment ('Do experiment') are then used to update the parameter distribution in 'Parametric inference'.

We employ conjugate priors [26] since they are closed under parametric inference and are in general quite expressive. Whenever the available knowledge is insufficient, that is whenever the confidence in accepting or rejecting a property is below the confidence threshold $1 - \delta$, the procedure will design and perform additional experiments.

The termination of this procedure hinges, over the parameter space, on the distance of the "true" model from the edge of the feasible set: as such, termination cannot be claimed if the true system does not satisfy or reject robustly the property of interest. This pathologic instance would reflect the case where an infinitesimal deviation of the true parameter affects the satisfaction or rejection of the property. In practice, as long as the uncertainty distribution $p\big(\theta\big)$ has a level set δ that converges to a single point (which corresponds to the true system), for infinitely long experiments and as far as the true system robustly satisfies or rejects the property of interest, termination of the procedure follows almost surely.

In the following, we first explain how the confidence in a property is computed, then we discuss the parametric inference step via Bayesian identification for a given set of data $\mathbf{u}, \tilde{\mathbf{y}}$. Building on these two stages it becomes possible to compute optimal experiments, as in the corresponding stage.

Confidence Computation (III)

Denote the maximal feasible set $\Theta_\psi \subset \Theta$, supposed to be known, such that $\forall \theta \in \Theta_\psi : \mathbf{M}(\theta) \vDash \psi$. The confidence in a specification ψ defined over the system $\mathbf{S}$ is computed based on the uncertainty distribution over Θ_ψ: given a prior uncertainty distribution $p\big(\theta\big)$, the confidence is computed as $\mathbf{P}\big(\Theta_\psi\big) = \int_{\Theta_\psi} p\big(\theta\big) d\theta$, whereas after an additional experiment and parametric inference (next paragraph), the a-posteriori uncertainty distribution $p\big(\theta|\tilde{\mathbf{y}}, \mathbf{u}\big)$ can be used to compute the confidence as

$$\mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}}, \mathbf{u}\big) = \int_{\Theta_\psi} p\big(\theta|\tilde{\mathbf{y}}, \mathbf{u}\big) d\theta \ . \tag{1}$$

Observe that according to Bayesian probability calculus for uncertainties [24], the confidence in a property becomes the measure of the uncertainty distribution.

Parametric Inference (II)

Given a prior distribution $p\big(\theta\big)$ and a data set $\tilde{\mathbf{y}}$ obtained by taking N_s measurements of $\tilde{y}(t)$, the a-posteriori uncertainty distribution $p\big(\theta|\tilde{\mathbf{y}}, \mathbf{u}\big)$ is based on parametric inference [24, 26] structured over the parameter set Θ as

$$p(\theta \mid \tilde{\mathbf{y}}, \mathbf{u}) = \frac{p(\tilde{\mathbf{y}}|\theta, \mathbf{u})p(\theta)}{\int_\Theta p(\tilde{\mathbf{y}}|\theta, \mathbf{u})p(\theta)d\theta}. \tag{2}$$

Remark 1. Considerations of computational complexity limit the use of (1) and (2), since their solution is seldom analytical. In Section 3 we choose a model set that is linearly-parameterised, and has Gaussian distributions for both the measurement noise and the prior, which allows for closed-form solutions of (2). In this case, polyhedral expressions of the maximal feasible set Θ_ψ can be found for a subset of BLTL properties, which leads to easily obtainable computations of (1).

The computation of (1) for more general problems would demand the use of either Monte-Carlo methods to solve the relevant integrals directly without computing the feasible set first, or of numerical approximation techniques to obtain an (upper/lower-)approximation of the feasible set. The use of Monte-Carlo methods allows for an empirical computation of the confidence if, for each sample of the uncertainty distribution (a model in the model set), the property is decidable. Note that this transfers the bottleneck of computational complexity to the verification properties over the models, that is to $\mathbf{M}(\theta) \vDash \psi$.

Numerical approximations can offer insight especially when enabled by the exploration of the parameter space in order to exploit the structure and properties of the models. The exploration of the parameters set to seek the validity of a formal property is also known as parameter synthesis: in [27, 28, 29] parameter synthesis for autonomous continuous-time systems is performed; whereas the robustness (with respect to parameter uncertainty) of a property over a continuous-time Markov decision process is analysed in [30, 20].

Experiment Design (I)

Every experiment contains statistical information on the behaviour of the system, which can be used to decide whether to accept or reject a specification over the system. The objective is to design an experiment $\mathbf{u}$ that optimally exploits the dynamic behaviour of the system and thus optimises the expected value of a criterion. In this subsection, the criterion for the Bayesian experiment design problem is the expected utility related to the acceptance or the rejection of a given specification of interest, based on the identification experiment. Let us define this criterion $J(\tilde{\mathbf{y}}, \mathbf{u})$ as a function of the measured output $\tilde{\mathbf{y}}$ and input $\mathbf{u}$ data, as $J : \mathbb{U}^{N_s} \times \mathbb{Y}^{N_s}$. The data realisation $\tilde{\mathbf{y}} \sim p(\tilde{\mathbf{y}}|\mathbf{u})$ has a probability density function which is conditioned on the input signal $\mathbf{u}$. Given a prior $p(\theta)$ the probability density distribution of the data can be expressed as

$$p(\tilde{\mathbf{y}}|\mathbf{u}) = \int_\Theta p(\tilde{\mathbf{y}}|\theta, \mathbf{u})p(\theta)d\theta\ , \tag{3}$$

where $p(\tilde{\mathbf{y}}|\theta, \mathbf{u})$ is the data distribution conditioned on the input $\mathbf{u}$ and on the parameter θ. The Bayesian experiment design problem optimises the expected value of the criterion J over the input signal $\mathbf{u}$ for a given prior $p(\theta)$, and is formulated as:

$$\max_{\mathbf{u}\in\mathcal{E}} \mathbf{E}\left[J(\tilde{\mathbf{y}},\mathbf{u}) \mid \tilde{\mathbf{y}} \sim p\big(\tilde{\mathbf{y}}|\mathbf{u}\big)\right], \tag{4}$$

where the set of allowed experiments $\mathcal{E}$ is defined as

$$\mathcal{E} = \{\mathbf{u} : u(t) \in \mathbb{U},\ \forall t = 0, \ldots N_s - 1\},$$

with $\mathbb{U}$ a bounded set, such as for instance $[-u_{\max}, u_{\max}],\ u_{\max} \in \mathbb{R}$.

In order to use the expected utility J related to the acceptance or rejection of a specification based on the identification experiment, consider that system $\mathbf{S}$ can be represented as $\mathbf{M}(\theta^0)$, with a nominal parameter θ^0. Although θ^0 is in general unknown, it can be perceived as a realisation of the uncertainty distribution over the parameters space, i.e. $\theta^0 \sim p\big(\theta\big)$, $\theta \in \Theta$. The acceptance or rejection of $\mathbf{S} \vDash \psi$ can be equivalently cast as the choice between hypothesis H_0: $\mathbf{M}(\theta^0) \vDash \psi$ and hypothesis H_1: $\mathbf{M}(\theta^0) \nvDash \psi$. This entails a decision which is valued with 1 when correct, and with 0 when incorrect. For a given choice of H_0 or H_1, and a nominal parameter θ^0, the utility is then a binary-valued function

$$\mathrm{ut}(H_i, \theta^0) = \begin{cases} 1 & \text{if} \begin{cases} H_0 & \text{and} & \mathbf{M}(\theta^0) \vDash \psi \\ H_1 & \text{and} & \mathbf{M}(\theta^0) \nvDash \psi, \end{cases} \\ 0 & \text{else.} \end{cases} \tag{5}$$

Note that ut has a 0 value when the chosen hypothesis is wrong, which is related in statistics to type I and type II error, respectively [31, page 514].

Conditional on a data set $\tilde{\mathbf{y}}$, the nominal parameter is distributed over the parameter space as $\theta^0 \sim p\big(\theta|\tilde{\mathbf{y}},\mathbf{u}\big)$. The expected utility of a decision H_i conditional on the data set is thus $\mathbf{E}\left[\mathrm{ut}(H_i,\theta^0) \mid \theta^0 \sim p\big(\theta|\tilde{\mathbf{y}},\mathbf{u}\big)\right]$. Note that the expected utility represents the confidence that $\mathbf{M}(\theta^0) \vDash \psi$ or $\mathbf{M}(\theta^0) \nvDash \psi$, and is a function of both the decision and the experiment $(\tilde{\mathbf{y}},\mathbf{u})$. Thus when deciding on H_0 or H_1, the expected utility is either $\mathrm{ut}(H_0;(\tilde{\mathbf{y}},\mathbf{u})) = \int_{\theta\in\Theta_\psi} p\big(\theta|\tilde{\mathbf{y}},\mathbf{u}\big)d\theta = \mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}},\mathbf{u}\big)$, or $\mathrm{ut}(H_1;(\tilde{\mathbf{y}},\mathbf{u})) = \int_{\theta\in\Theta\backslash\Theta_\psi} p\big(\theta|\tilde{\mathbf{y}},\mathbf{u}\big)d\theta = \mathbf{P}\big(\Theta\backslash\Theta_\psi|\tilde{\mathbf{y}},\mathbf{u}\big) = 1 - \mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}},\mathbf{u}\big)$. As a criterion, we then choose the expected utility maximised over the decision H_0 or H_1, namely

$$\begin{aligned} J(\tilde{\mathbf{y}},\mathbf{u}) &\doteq \max_{H_i} \mathrm{ut}(H_i;(\tilde{\mathbf{y}},\mathbf{u})) \\ &= \max\left\{\mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}},\mathbf{u}\big), \mathbf{P}\big(\Theta\backslash\Theta_\psi|\tilde{\mathbf{y}},\mathbf{u}\big)\right\}. \end{aligned} \tag{6}$$

3 Verification of Systems Representable by Linearly-Parameterised Models

In this section we provide a solution of the discussed new data-driven and model-based verification problem, according to the schematic process depicted in Figure 2, for single-input single-output systems in linearly-parameterised model sets, for a subset of properties expressed as bounded-horizon LTL formulae, and for a known Gaussian a priori uncertainty distribution and measurement noise. Under these restrictions, we obtain closed form solutions of (2) and convex sets for the feasible set Θ_ψ, which are then employed towards a relaxation of the criterion in (6) and a Monte Carlo solution to the Bayesian experiment design problem in (4). We practically justify the consideration of linearly-parameterised model sets since for any linear system there exists a set of linearly-parameterised models capable of representing it. Linearly-parameterised model sets such as orthonormal basis function parameterisations are able to represent a wide set of systems [32, Chapter 4 and 7],[33]. Models $\mathbf{M}$ within a linearly-parameterised model class $\mathcal{G}$ have the following state-space realisation:

$$\mathbf{M}(\theta): \quad \begin{cases} x(t+1) & = Ax(t) + Bu(t), \\ \hat{y}(t,\theta) & = \theta^T x(t) \end{cases} \tag{7}$$

$$\theta = \begin{bmatrix}\theta_1 & \dots & \theta_n\end{bmatrix}^T \in \Theta \subset \mathbb{R}^n,$$

and are (linearly) parameterised by $\theta \in \Theta$. We assume that the system has a representation $\mathbf{M}(\theta^0)$ in this model set, with unknown parameter θ^0, and has an output denoted as $y_0(t) = \hat{y}(t, \theta^0)$. It is assumed that the initial state of the system and of the model representing it is $x(0) = 0$, both in the identification experiment and for the verification of the property. The noise disturbance, $e(t)$, on the measurement $\tilde{y}(t) = y_0(t) + e(t)$ is assumed to be an additive zero-mean, white, Gaussian-distributed measurement noise with variance σ_e^2 that is uncorrelated with the input. The following theorem can be derived for properties defined on the model output $y_0(t)$ (the proof can be found in the Appendix).

Theorem 2. Consider a linearly-parameterised model set, a bounded polyhedron for the set of initial states $x(0) \in \mathbb{X}_0$, and inputs $u(t) \in \mathbb{U}$ for $t \geq 0$. For every specification ψ expressed within the LTL fragment $\psi := \sigma | \bigcirc\psi | \psi_1 \wedge \psi_2$, with $\sigma \in \Sigma$, the feasible set of parameters $\Theta_\psi = \{\theta \in \Theta : \mathbf{M}(\theta) \vDash \psi\}$ is a polyhedron.

Several observations can be made. Firstly, the number of half planes characterising the set Θ_ψ may quickly increase with the time bound of the LTL formula ψ (that is, with the repeated application of the $\bigcirc$ operator), and with the cardinality of the atomic propositions in the alphabet Σ. Secondly, the extension beyond the LTL fragment discussed above may lead to feasible sets that are in general not convex, and is therefore left for future work. The assumption on the initial state $x(0) = 0$ can be relaxed without consequences as long as the set of initial states of the parameterised models is independent of the parameterisation.

Recursive Parametric Inference

Let us denote the $(k+1)$-th iteration of the verification algorithm in Figure 2 as a combination of input design, experiment, and Bayesian identification starting from the prior knowledge gathered in the previous iterations. At the first iteration, the available knowledge is structured into a prior distribution $\mathcal{N}(\mu_0, R_0)$ over the parameter space, a multi-variate Gaussian with mean μ_0 and variance R_0. Employing Bayesian inference for the iterations of the identification procedure, the probability distributions in (2) and (3) can be computed recursively. At the $(k+1)$-th iteration the available knowledge is expressed as a prior $p(\theta) = \mathcal{N}(\mu_k, R_k)$ and in combination with data sets $\mathbf{u}$, $\tilde{\mathbf{y}}$ the distributions of interest are computed as

$$p(\tilde{\mathbf{y}} \mid \theta, \mathbf{u}) = \mathcal{N}\big(\Phi^T(\mathbf{u})\theta, I\sigma_e^2\big), \tag{8a}$$

$$p(\tilde{\mathbf{y}} \mid \mathbf{u}) = \mathcal{N}\big(\Phi^T(\mathbf{u})\mu_k, R_{\tilde{\mathbf{y}}}\big), \tag{8b}$$

$$R_{\tilde{\mathbf{y}}} \quad = \left[\sigma_e^2 + \Phi^T(\mathbf{u}) R_k \Phi(\mathbf{u})\right],$$

$$p\big(\theta \mid \tilde{\mathbf{y}}, \mathbf{u}\big) = \mathcal{N}\big(\mu_{k+1}, R_{k+1}\big), \tag{8c}$$
$$R_{k+1} = \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi^T(\mathbf{u})\right]^{-1},$$
$$\mu_{k+1} = R_{k+1}\left[R_k^{-1}\mu_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right],$$

with $\Phi(\mathbf{u}) = \begin{bmatrix} x(1) & \dots & x(N_s)\end{bmatrix} \in \mathbb{R}^{n\times N_s}$. In (8a), the distribution over the expected data $\tilde{\mathbf{y}} = \begin{bmatrix} y(1) & \dots & y(N_s)\end{bmatrix}^T$, conditioned on the parameter θ and the input sequence $\mathbf{u}$, can be computed through the distribution of the measurement noise. Its mean is a linear mapping of the input data to the matrix $\Phi(\mathbf{u})$. Marginalised over the prior distribution, this is the data distribution conditioned on the input alone, as per (8b). The posterior distribution $p\big(\theta \mid \tilde{\mathbf{y}}, \mathbf{u}\big)$ (8c) provides an expression for (2), and corresponds to the prior distribution for the $(k+2)$-th iteration.

Bayesian Θ_ψ-Optimal Experiment

We solve approximatively the optimisation problem related to experiment design via an empirical approximation of the objective function and an input parameterisation.

Consider the experiment design problem $\max_{\mathbf{u}\in\mathcal{E}} \mathbf{E}[J(\tilde{\mathbf{y}}, \mathbf{u}) \mid \tilde{\mathbf{y}} \sim p\big(\tilde{\mathbf{y}} \mid \mathbf{u}\big)]$ with $J(\tilde{\mathbf{y}}, \mathbf{u})$ the expected utility as given in (6). Note that the posterior distribution $p\big(\theta \mid \tilde{\mathbf{y}}, \mathbf{u}\big) = \mathcal{N}\big(\mu_{k+1}, R_{k+1}\big)$, hence $J(\tilde{\mathbf{y}}, \mathbf{u})$ depends on the measurements $\tilde{\mathbf{y}}$ only through μ_{k+1}, as in (8c). It follows that the optimisation problem can be written as an expected value over μ_{k+1} (instead of $\tilde{\mathbf{y}}$), reducing the complexity of the problem from the horizon of the data to the dimensionality of the parameterisation,

$$\max_{\mathbf{u}\in\mathcal{E}} \int\limits_{\Theta} \max\left\{\mathbf{P}\big(\Theta_\psi|\mu_{k+1}, \mathbf{u}\big), \mathbf{P}\big(\bar{\Theta}_\psi|\mu_{k+1}, \mathbf{u}\big)\right\}$$
$$\times\, p\big(\mu_{k+1}|\mathbf{u}\big)d\mu_{k+1}, \text{ with } \bar{\Theta}_\psi = \Theta \setminus \Theta_\psi \tag{9}$$
$$\text{s.t.} \qquad p\big(\mu_{k+1}|\mathbf{u}\big) = \mathcal{N}\big(\mu_k, R_k - R_{k+1}\big).$$

As an affine transformation of the measurements $\tilde{\mathbf{y}}$, the posterior mean μ_{k+1} is a random variable with a Gaussian distribution as $\mu_{k+1} = R_{k+1}\left[R_k^{-1}\mu_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right]$. Using this mean, a practical lower approximation of the maximisation inside the integral is found as $\mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}}, \mathbf{u}\big) = \int_{\Theta_\psi} p\big(\theta|\mu_{k+1}, \mathbf{u}\big)d\theta$ for $\mu_{k+1} \in \Theta_\psi$, and $1 - \mathbf{P}\big(\Theta_\psi|\tilde{\mathbf{y}}, \mathbf{u}\big)$ else. This provides a relaxed version of (9), expressed as

$$\max_{\mathbf{u}\in\mathcal{E}} \int\limits_{\Theta_\psi}\int\limits_{\Theta_\psi} p\big(\theta|\mu_{k+1}, \mathbf{u}\big)p\big(\mu_{k+1}|\mathbf{u}\big)d\theta d\mu_{k+1}$$
$$+ \int\limits_{\bar{\Theta}_\psi}\int\limits_{\bar{\Theta}_\psi} p\big(\theta|\mu_{k+1}, \mathbf{u}\big)p\big(\mu_{k+1}|\mathbf{u}\big)d\theta d\mu_{k+1}. \tag{10}$$

The combined distribution of θ and μ_{k+1}, denoted by variable $\underline{\theta} = \begin{bmatrix}\theta^T & \mu_{k+1}^T\end{bmatrix}^T$, has a normal distribution $p\big(\underline{\theta} \mid \mathbf{u}\big) = \mathcal{N}\big(\mu_{\underline{\theta}}, R\big)$, with mean $\mu_{\underline{\theta}}^T = \begin{bmatrix}\mu_k & \mu_k\end{bmatrix}^T$

and covariance matrix

$$R = \begin{bmatrix} R_k & (R_k - R_{k+1}) \\ (R_k - R_{k+1}) & (R_k - R_{k+1}) \end{bmatrix}.$$

Since the integral in (10) cannot in general be computed analytically, we can either compute it with an efficient numerical method or we can empirically approximate it.

Remark 3 (Numerical methods). Efficient numerical methods to compute the integral of multivariate densities over polytopes [34] are not an option. This is because these methods would approximate, depending on the mean of the prior, either the first term in (10) (the integral over $\Theta_\psi \times \Theta_\psi$) and neglect the second term, or the opposite. From a practical point of view this would make sense, since when the prior μ_k is in Θ_ψ we would expect the value of $\int_{\bar{\Theta}_\psi} \int_{\bar{\Theta}_\psi} p\big(\theta|\mu_{k+1}, \mathbf{u}\big) p\big(\mu_{k+1}|\mathbf{u}\big) d\theta d\mu_{k+1}$ to be very small. But it can be observed that in the case of $\mu_k \in \Theta_\psi$ the second term is strictly increasing for a decrease in the variance of the posterior density $p\big(\theta \mid \tilde{\mathbf{y}}\big)$, whereas the first term is not. In conclusion, we opt for the alternative use of a Monte-Carlo approximation of the objective.

Let ϵ be a dummy random variable with density distribution $\mathcal{N}\big(\underline{0}, I\big)$, which is independent of the decision variable $\mathbf{u}$. The value of the relaxed optimisation problem (10) can be approximated as

$$\hat{\mathbf{E}}J \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{(\Theta_\psi \times \Theta_\psi) \cup (\bar{\Theta}_\psi \times \bar{\Theta}_\psi)} \big(\mu_{\underline{\theta}} + \Lambda \epsilon_i\big), \tag{11}$$

with M realisations of $\epsilon_i \sim \mathcal{N}\big(\underline{0}, I\big)$ and $\Lambda\Lambda^T = R$. The realisations of $\mu_{\underline{\theta}} + \Lambda\epsilon_i$ have the same density distribution as $\underline{\theta} \sim \mathcal{N}\big(\mu_{\underline{\theta}}, R\big)$. Hence, for a given input $\mathbf{u}$, (11) is an unbiased estimate of (10) and it is also consistent, i.e., for $M \to \infty$ the estimated objective converges to the optimisation objective in (10) with probability 1. The N_s decision variables of $\mathbf{u}$ can be reduced by a parameterisation of the input signal $\mathbf{u}$ as $u(t) = \sum_{k=1}^{n_p} \beta_k \sin(\omega_k t + \alpha_k)$, with parameters $\alpha_k \in [0, 2\pi]$ and $\beta_k \in [0, \infty)$ for $k = 1, \ldots, n_p$ at predefined frequencies ω_k.

Case Study: Bounded-time Safety Verification of a Physical System

Consider a system $\mathbf{S}$ with input signals with support $u(t) \in \mathcal{U} = [-0.2,\ 0.2]$. For simplicity let us select a fixed initial state $x_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. Verify whether the output $y_0(t)$ remains within the interval $\mathcal{I} = \big[-0.5,\ 0.5\big]$, labeled as ι, for the next 4 time steps. Introduce accordingly the alphabet $\Sigma = \{\iota, \tau\}$ and the labelling map $L : L(y) = \iota, \forall y \in \mathcal{I},\ L(y) = \tau, \forall y \in \mathbb{Y} \setminus \mathcal{I}$. Now check whether the following finite-horizon LTL property holds: $\mathbf{S} \vDash \bigwedge_{i=1}^{4} (\bigcirc)^i \iota$. We assume that system $\mathbf{S}$ can be represented as an element of a model set $\mathcal{G}$ with transfer functions characterised by second-order Laguerre-basis functions [35] (a special case of orthonormal basis functions), which translates to the following

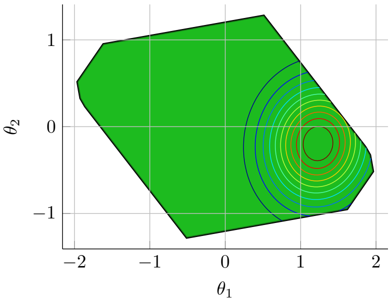

Figure 3: The green region in the parameter space $[\theta_1\ \theta_2]^T$ is the feasible set of the case study. The contour lines give the density function of a possible a-posteriori distribution over the parameter space (for confidence quantification).

parameterised state-space representation:

$$x(t+1) = \begin{bmatrix} a & 0 \\ 1-a^2 & a \end{bmatrix} x(t) + \begin{bmatrix} \sqrt{1-a^2} \\ (-a)\sqrt{1-a^2} \end{bmatrix} u(t),$$
$$\hat{y}(t,\theta) = \theta^T x(t)\ .$$

The coefficient a is chosen to be $a = 0.4$. We further consider, as prior available knowledge on the system, a distribution $p(\theta) = \mathcal{N}(\mu_k, R_k)$ on the model class, and a known variance $\sigma_e^2 = 0.5$ for the white additive measurement noise. The output of the computation of the feasible set is in Figure 3. We are interested in an experiment of length $N_s = 100$ with the input parameterised as a multi-sine with discrete frequencies $(\omega_0, 2\omega_0, \ldots, 5\omega_0)$ and fundamental frequency $\omega_0 = 2\pi/10$.

The results of the experiment design problem are given in Figure 4 and compared to the results for a classical D-optimal experiment design [36] on a single iteration of the verification algorithm in Figure 2. For D-optimality we minimise the determinant of R_{k+1}. Both the Θ_ψ- and D-optimal experiment designs have been performed for priors with several different mean values μ_k, and with a fixed variance of $R_k = 0.2I_{2\times2}$. Note that the D-optimal experiment is independent of the prior mean.

After designing an experiment $\mathbf{u}$, the optimisation objective $\mathbf{E}\left[J(\tilde{\mathbf{y}}, \mathbf{u}) \mid \tilde{\mathbf{y}} \sim p(\tilde{\mathbf{y}}|\mathbf{u})\right]$ is evaluated empirically. For this 400 data samples $\tilde{\mathbf{y}}$ are drawn from the distribution $p(\tilde{\mathbf{y}}|\mathbf{u})$, first drawing a parameterisation from the prior distribution $\theta \sim p(\theta)$, and subsequently performing an identification experiment. In Figure 4 the empirical evaluation of $\mathbf{E}\left[J(\tilde{\mathbf{y}}, \mathbf{u})\right]$ for both the Θ_ψ- and D- optimal experiment designs are plotted together with the attainable result without performing additional experiments $\max\{\mathbf{P}(\Theta_\psi), \mathbf{P}(\bar{\Theta}_\psi)\}$. Note that the figure displays in fact the values $1 - \mathbf{E}J$ and $1 - \max\{\mathbf{P}(\Theta_\psi), \mathbf{P}(\bar{\Theta}_\psi)\}$ for convenience, and also give the standard deviation of the empirical evaluations.

In Figure 4, the result shows that the empirical value of $\mathbf{E}[J]$ is higher for the Θ_ψ-optimal experiment design than for the D-optimal experiment design for all the given mean values. It can be observed that this is especially significant when $\max\{\mathbf{P}(\Theta_\psi), \mathbf{P}(\bar{\Theta}_\psi)\}$ is smaller. The authors have observed that in this case the posterior variances of the Θ_ψ- optimal experiment design tend to align

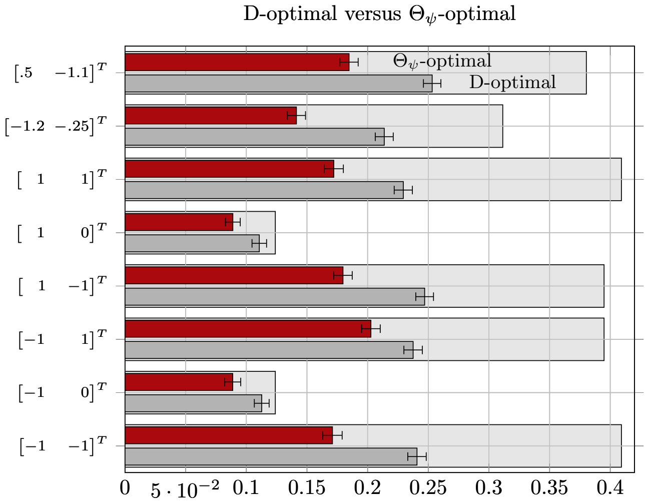

Figure 4: Empirical evaluation of $1 - \mathbf{E}[J]$ on the horizontal axis for the safety verification case study for both the Θ_ψ-optimal (red bar, top) and the D-optimal experiment design (grey bar, below). The wider light grey bar gives $1 - \max\{\mathbf{P}(\Theta_\psi), \mathbf{P}(\bar{\Theta}_\psi)\}$. The means of the priors in the experiment design are given on the vertical axis. On the bars for the empirical evaluation of $1 - \mathbf{E}[J]$, their standard deviation is also drawn on the bars by the symbol $\vdash\!\dashv$.

with the closest faces of the feasible region. For mean values that lie farther from the boundaries of the feasible region such as $[1 \ 0]^T$ and $[-1 \ 0]^T$, the $\max\{\mathbf{P}(\Theta_\psi), \mathbf{P}(\bar{\Theta}_\psi)\}$ is already quite big and the difference between the Θ_ψ- and D- optimal design is less significant. It can be concluded that the Θ_ψ-optimal experiment gives a significant improvement with respect to $\mathbf{E}[J]$ in comparison to the D-optimal solution.

4 Conclusions and Future Work

This paper has contributed to an original research line interested in data-driven and model-based property verification. This is studied first by formalising a Bayesian methodology to compute the confidence in a formal property using measurements of the system, which is facilitated by the derivation of a feasible set of parameters for linearly parameterised models; second, by defining a Bayesian experiment design problem, which selects the input with the goal of efficient verification of formal properties.

This contribution is only a first step towards a practically useful application of formal methods techniques for the verification and design of safety-critical systems, of which no exact models are known. The present work builds heavily on underlying assumptions such as the knowledge of the exact model structure (which is limited to linearly parameterised models) and of the exact noise dynamics.

Extending this work to more general model structures and noise dynamics is of interest, alongside considering more complex linear-time properties. Further, future work will consider the use of this theory as a practical tool for property optimisation via controller synthesis.

5 acknowledgement

We would like to acknowledge the support of the Netherlands Organisation for Scientific Research and of the Dutch Institute of Systems and Control.

References

[1] E. M. E. Clarke, “The birth of model checking,” in 25 Years Model Checking (O. Grumberg and H. Veith, eds.), vol. 5000 of LNCS, pp. 1–26, Springer, 2008.

[2] M. Y. M. Vardi, “From Philosophical to Industrial Logics,” in Proc. 3rd Indian Conf. Log. Its Appl., ICLA ’09, (Berlin, Heidelberg), pp. 89–115, Springer-Verlag, 2009.

[3] P. Tabuada and G. J. Pappas, “Model Checking LTL over Controllable Linear Systems Is Decidable,” Hybrid Syst. Comput. Control, vol. 2623, pp. 498–513, 2003.

[4] C. Belta, A. Bicchi, M. Egerstedt, E. Frazzoli, E. Klavins, and G. J. Pappas, “Symbolic planning and control of robot motion,” Robot. Autom. IEEE Trans., vol. 14, no. 1, pp. 61–70, 2007.

[5] N. Saeedloei and G. Gupta, “A logic-based modeling and verification of CPS,” ACM SIGBED Rev., vol. 8, pp. 31–34, June 2011.

[6] B. Yordanov, G. Batt, and C. Belta, “Model Checking Discrete-Time Piecewise Affine Systems: Application to Gene Networks,” in Eur. Control Conf., 2007.

[7] E. Aydin Gol and C. Belta, “An additive cost approach to optimal Temporal Logic control,” Control Conf. (ACC), 2014, 2014.

[8] H. L. S. Younes and R. G. Simmons, “Probabilistic verification of discrete event systems using acceptance sampling,” Comput. Aided Verif., 2002.

[9] D. Henriques, J. G. Martins, P. Zuliani, A. A. Platzer, and E. M. Clarke, “Statistical Model Checking for Markov Decision Processes,” in Quant. Eval. Syst. (QEST), 2012 Ninth Int. Conf., pp. 84–93, Ieee, Sept. 2012.

[10] K. Sen, M. Viswanathan, and G. Agha, “Statistical model checking of black-box probabilistic systems,” in Comput. Aided Verif. (R. Alur and D. Peled, eds.), vol. 3114 of LNCS, pp. 399–401, Springer, Springer Berlin Heidelberg, 2004.

[11] A. Legay, B. Delahaye, and S. Bensalem, “Statistical Model Checking: An Overview,” in Runtime Verif., vol. 6418 of LNCS, pp. 122–135, Springer Berlin Heidelberg, 2010.

[12] H. L. S. Younes, "Probabilistic verification for "black-box" systems," Comput. Aided Verif., pp. 253–265, 2005.

[13] A. Legay and S. Sedwards, "Lightweight Monte Carlo Algorithm for Markov Decision Process Verification," CoRR, p. 17, Oct. 2013.

[14] K. Sen, M. Viswanathan, and G. Agha, "Learning continuous time Markov chains from sample executions," in First Int. Conf. Quant. Eval. Syst. 2004. QEST 2004. Proceedings., pp. 146–155, IEEE, Sept. 2004.

[15] H. Mao, Y. Chen, M. Jaeger, T. D. Nielsen, K. G. Larsen, and B. Nielsen, "Learning Markov Decision Processes for Model Checking," Electron. Proc. Theor. Comput. Sci., vol. 103, pp. 49–63, Dec. 2012.

[16] H. Mao and M. Jaeger, "Learning and Model-Checking Networks of I/O Automata.," in Proc. Fourth Asian Conf. Mach. Learn., 2012.

[17] Y. Chen and T. T. D. Nielsen, "Active Learning of Markov Decision Processes for System Verification," Mach. Learn. Appl. (, vol. 2, pp. 289–294, Dec. 2012.

[18] E. Bartocci, L. Bortolussi, and G. Sanguinetti, "Learning temporal logical properties discriminating ECG models of cardiac arrhytmias," arXiv preprint arXiv:1312.7523, 2013.

[19] L. Bortolussi and G. Sanguinetti, "Learning and Designing Stochastic Processes from Logical Constraints," in Quant. Eval. Syst. (K. Joshi, M. Siegle, M. Stoelinga, and P. R. D'Argenio, eds.), vol. 8054 of LNCS, pp. 89–105, Springer Berlin Heidelberg, 2013.

[20] L. Bortolussi and G. Sanguinetti, "Smoothed model checking for uncertain continuous time markov chains," CoRR, vol. abs/1402.1450, 2014.

[21] H. Hjalmarsson, "From experiment design to closed-loop control," Automatica, vol. 41, pp. 393–438, Mar. 2005.

[22] X. Bombois, G. Scorletti, M. Gevers, P. M. J. Van den Hof, and R. Hildebrand, "Least costly identification experiment for control," Automatica, vol. 42, pp. 1651–1662, Oct. 2006.

[23] H. Hjalmarsson, "System Identification of Complex and Structured Systems," Eur. J. Control, vol. 15, pp. 275–310, Jan. 2009.

[24] D. V. Lindley, "The philosophy of statistics," J. R. Stat. Soc. Ser. D (The Stat., vol. 49, no. 3, pp. 293–337, 2011.

[25] C. R. Rojas, J. S. Welsh, G. C. Goodwin, and A. Feuer, "Robust optimal experiment design for system identification," Automatica, vol. 43, pp. 993–1008, June 2007.

[26] V. Peterka, "Bayesian Approach to System Identification," Trends Prog. Syst. Identif., 1981.

[27] G. Frehse, S. K. S. Jha, and B. H. Krogh, "A counterexample-guided approach to parameter synthesis for linear hybrid automata," Hybrid Syst. Comput. Control, vol. 4981, pp. 187–200, 2008.

[28] G. Batt, C. Belta, and R. Weiss, “Model checking genetic regulatory networks with parameter uncertainty,” Hybrid Syst. Comput. Control, vol. 4416, pp. 61–75, 2007.

[29] T. T. Henzinger and H. Wong-Toi, Using HyTech to synthesize control parameters for a steam boiler, vol. 1165 of LNCS. Springer Berlin Heidelberg, 1996.

[30] L. Brim, M. Ceska, S. Drazan, and D. Safranek, “On Robustness Analysis of Stochastic Biochemical Systems by Probabilistic Model Checking,” arXiv Prepr. arXiv1310.4734, 2013.

[31] K. S. Shanmugan and A. M. Breipohl, Random Signals: Detection, Estimation and Data Analysis. Wiley, 1988.

[32] P. S. C. Heuberger, P. M. J. Van den Hof, and B. Wahlberg, Modelling and identification with rational orthogonal basis functions. Springer, 2005.

[33] P. M. J. Van den Hof, P. S. C. Heuberger, and J. Bokor, “System identification with generalized orthonormal basis functions,” Automatica, vol. 31, no. 12, pp. 1821–1834, 1995.

[34] L. Blackmore and M. Ono, “Convex chance constrained predictive control without sampling,” in Proc. AIAA Guid. Navig. Control Conf., 2009.

[35] P. S. C. Heuberger, P. M. J. den Hof, and O. H. Bosgra, “A generalized orthonormal basis for linear dynamical systems,” , IEEE Trans., vol. 40, pp. 451–465, Mar. 1995.

[36] M. Gevers, X. Bombois, and R. Hildebrand, “Optimal experiment design for open and closed-loop system identification,” Communications in Information and Systems, pp. 1–24, 1944.

[37] C. Belta, L. Habets, and V. Kumar, “Control of multi-affine systems on rectangles with applications to hybrid biomolecular networks,” in Decision and Control, 2002, Proceedings of the 41st IEEE Conference on, vol. 1, pp. 534–539, IEEE, 2002.

A Feasible Set

In this appendix we prove Theorem 2, which states that under certain conditions a fragment of BLTL defined over linearly parameterised models results in feasible polytopes in the parameter space. Let $\otimes$ denote the Kronecker product. Consider the input set $\mathbb{U}$ to be the convex hull of U, i.e. $\text{conv}(U) = \mathbb{U}$. Let the model set be given as $\mathbf{M}(\theta) = (A, B, \theta^T, d)$ where d can be either a constant, a parameter or zero. In the derivation we treat it as a parameter. Thus the parameterisation is $\left[\theta^T, d\right] \in \mathbb{R}^{n+1}$. We specifically chose this realisation of the model as the mapping from input to state does not depend on the parameterisation. This gives a initial condition which is not dependent on the parameterisation.

Lemma 4. For every specification ψ composed from syntax fragment $\psi := p|\bigcirc\psi|\psi_1 \wedge \psi_2$ and for every $\theta \in \Theta$: the model $\mathbf{M}(\theta) = (A, B, \theta^T, d)$ with state x satisfies the specification ψ, denoted $< \mathbf{M}(\theta), x > \vDash \psi$, if and only if

$$\begin{bmatrix}\big((I_{n_\psi \times n_\psi} \otimes x^T)N_\psi + K_\psi\big) & D_\psi\end{bmatrix}\begin{bmatrix}\theta \\ d\end{bmatrix} < b_\psi. \tag{12}$$

The matrices $N_\psi \in \mathbb{R}^{nn_\psi \times n}, K_\psi \in \mathbb{R}^{n_\psi \times n}, D_\psi \in \mathbb{R}^{n_\psi \times 1}, b_\psi \in \mathbb{R}^{n_\psi \times 1}$ in the above satisfaction relation have dimensions in function of the parametrisation $dim(\theta) = n$ and a property dependent 'dimension' n_ψ and are given by induction over the syntax of the specification.

The step by step built up of $(N_\psi, K_\psi\, D_\psi, b_\psi)$ for syntax $\psi := p|\bigcirc\psi|\psi_1 \wedge \psi_2$ is given. For any atomic propositions $< \mathbf{M}(\theta), x > \vDash p_i \Leftrightarrow A_{p_i} y < b_{p_i}$ with $A_{p_i} \in \mathbb{R}$. the matrices are $N_{p_i} = 1_{|U|} \otimes (A_{p_i} \otimes I_{n\times n}) \in \mathbb{R}^{n\times n}$, $K_{p_i} = O_{|U|\times n} \in \mathbb{R}^{1\times n}$, $D_{p_i} = U \otimes A_{p_i} \in \mathbb{R}^{|U|\times 1}$, and $b_{p_i} = 1_{|U|} \otimes b_{p_i} \in \mathbb{R}^{|U|\times 1}$.
The next operation $\bigcirc\psi_1$ with matrices $(N_{\psi_1}$,K_{ψ_1},D_{ψ_1},$b_{\psi_1})$ gives

$$\begin{aligned}
N_{\bigcirc\psi_1} &= \mathbf{1}_{|U|} \otimes \big((I_{n_{\psi_1}\times n_{\psi_1}} \otimes A^T)N_{\psi_1}\big),\\
K_{\bigcirc\psi_1} &= U \otimes \big((I_{n_{\psi_1}\times n_{\psi_1}} \otimes B^T)N_{\psi_1}\big) + \mathbf{1}_{|U|} \otimes K_{\psi_1},\\
D_{\bigcirc\psi_1} &= \mathbf{1}_{|U|} \otimes D_{\psi_1} \text{ and } b_{\bigcirc\psi_1} = \mathbf{1}_{|U|} \otimes b_{\psi_1}
\end{aligned}$$

with dimensions $nn_{\psi_1}|U| \times n$, $n_{\psi_1}|U| \times n$, $n_{\psi_1}|U| \times 1$ and $n_{\psi_1}|U| \times 1$.
The and operation $\psi_1 \wedge \psi_2$ for $(N_{\psi_1}$, K_{ψ_1},D_{ψ_1},$b_{\psi_1})$ and $(N_{\psi_2}$, K_{ψ_2},D_{ψ_2},$b_{\psi_2})$ with $n_{\psi_1\wedge\psi_2} = (n_{\psi_1} + n_{\psi_2})$ gives

$$\begin{aligned}
N_{\psi_1\wedge\psi_2} &= \begin{bmatrix} N_{\psi_1}\\ N_{\psi_2}\end{bmatrix}, K_{\psi_1\wedge\psi_2} = \begin{bmatrix} K_{\psi_1}\\ K_{\psi_2}\end{bmatrix},\\
D_{\psi_1\wedge\psi_2} &= \begin{bmatrix} D_{\psi_1}\\ D_{\psi_2}\end{bmatrix}, \text{ and } \; b_{\psi_1\wedge\psi_2} = \begin{bmatrix} b_{\psi_1}\\ b_{\psi_2}\end{bmatrix}
\end{aligned}$$

with dimensions $nn_{\psi_1\wedge\psi_2} \times n$, $n_{\psi_1\wedge\psi_2} \times n$, $n_{\psi_1\wedge\psi_2} \times 1$, and $n_{\psi_1\wedge\psi_2} \times 1$.

The interpretation of (12) for a formula defined over a model with initial condition $x(0) \in \mathbb{X}_0 = \text{conv}(X_0)$ is given through the following lemma.

Lemma 5 (Feasible set of parameters, Θ_ψ). Given a satisfaction relation $< \mathbf{M}(\theta), x(0) > \vDash \psi$, expressed as (12) $\forall x(0) \in \mathbb{X}_0$ the maximal feasible set of parameters Θ_ψ is

$$\left\{\theta \in \Theta : \bigwedge_{x_i \in X_0} \big((I_{n_\psi\times n_\psi} \otimes x_i^T)N_\psi + K_\psi\big)\,\theta + D_\psi d < b_\psi\right\}.$$

The set Θ_ψ is a polyhedron

Theorem 2. Theorem 2 is proved by combining Lemma 4 and Lemma 5. In the former, we state that there exists a multi-affine satisfaction relation in function of θ for every formula in our basic set of formulae. The latter states that given a multi-affine satisfaction relation the feasible set Θ_ψ is a polyhedron set in the parameter space. □

The statements in Lemma 4 and Lemma 5 are proved in the following.

Lemma 4. We derive the multi-affine inequalities for the next operator and the and operator;
Firstly for the next operator, i.e. $\bigcirc\psi$, suppose that there exists suitable matrices $N_\psi \in \mathbb{R}^{nn_\psi \times n}$, $K_\psi \in \mathbb{R}^{n_\psi \times n}$, $D_\psi \in \mathbb{R}^{n_\psi \times 1}$ and $b_\psi \in \mathbb{R}^{n_\psi \times 1}$ representing the satisfaction of property ψ as an inequality, then for $\bigcirc\psi$

$$\begin{aligned}
&< \mathbf{M}(\theta), x(t) > \models \bigcirc\psi \\
&\Leftrightarrow \forall u(t) \in \mathbb{U} : \\
&\quad \left[\left(I_{n_\psi \times n_\psi} \otimes x^T(t+1)\right) N_\psi + K_\psi \quad D_\psi\right] \begin{bmatrix}\theta \\ d\end{bmatrix} \leq b_\psi, \\
&\Leftrightarrow \forall u(t) \in \mathbb{U} : \left[\left(I_{n_\psi \times n_\psi} \otimes x^T(t)\right)\left(I_{n_\psi \times n_\psi} \otimes A^T\right) N_\psi\right. \\
&\quad + \left(u(t) I_{n_\psi \times n_\psi} \otimes B^T\right) N_\psi + K_\psi \quad \left. D_\psi\right] \begin{bmatrix}\theta \\ d\end{bmatrix} \leq b_\psi.
\end{aligned}$$

Since the above is an affine function in $u(t)$, the image of every $u(t) \in \operatorname{conv}(U) = \mathbb{U}$ can be expressed as a convex combination of the values at the vertices $\{u_{\min}, u_{\max}\}$ [37]. The input is of dimensionality 1. Denote the matrix with as elements the set U by $U = \begin{bmatrix}u_{\min} & u_{\max}\end{bmatrix}^T$, and introduce the vector $\mathbf{1}_{|U|} = \begin{bmatrix}1 & 1\end{bmatrix}^T$, then an equivalent expression is

$$\begin{aligned}
\Leftrightarrow &\left[\mathbf{1}_{|U|} \otimes \left(I_{n_\psi \times n_\psi} \otimes x^T(t)\right)\left(I_{n_\psi \times n_\psi} \otimes A^T\right) N_\psi\right. \\
&\left. + U \otimes \left(I_{n_\psi \times n_\psi} \otimes B^T\right) N_\psi + \mathbf{1}_{|U|} \otimes K_\psi\right] \theta \\
&+ (\mathbf{1}_{|U|} \otimes D_\psi) d \leq \mathbf{1}_{|U|} \otimes b_\psi.
\end{aligned}$$

Having obtained $K_{\bigcirc\psi}$, $D_{\bigcirc\psi}$, and $b_{\bigcirc\psi}$, now rewrite the first term to obtain $N_{\bigcirc\psi}$

$$\begin{aligned}
&\mathbf{1}_{|U|} \otimes \left(I_{n_\psi \times n_\psi} \otimes x^T(t)\right)\left(I_{n_\psi \times n_\psi} \otimes A^T\right) N_\psi \\
&= \left(I_{|U| \times |U|} \mathbf{1}_{|U|}\right) \otimes \left(I_{n_\psi \times n_\psi} \otimes x^T(t)\right)\left(I_{n_\psi \times n_\psi} \otimes A^T\right) N_\psi \\
&= \left(I_{|U| n_\psi \times |U| n_\psi} \otimes x^T(t)\right)\left(\mathbf{1}_{|U|} \otimes \left(I_{n_\psi \times n_\psi} \otimes A^T\right) N_\psi\right).
\end{aligned}$$

Secondly, derive the matrices for the and operator, $\psi_1 \wedge \psi_2$. If there exists suitable matrices (N_{ψ_1}, K_{ψ_1}, D_{ψ_1}, b_{ψ_1}) and (N_{ψ_2}, K_{ψ_2},D_{ψ_2}, b_{ψ_2}) for formulae ψ_1 and ψ_2, then there also exists matrices $N_{\psi_1 \wedge \psi_2}$, $K_{\psi_1 \wedge \psi_2}$, $D_{\psi_1 \wedge \psi_2}$, and $b_{\psi_1 \wedge \psi_2}$

$$\begin{aligned}
&< \mathbf{M}(\theta), x(t) > \models \psi_1 \wedge \psi_2 \\
&\Leftrightarrow \bigwedge_{i \in \{1,2\}} \left[\left(I_{n_{\psi_i} \times n_{\psi_i}} \otimes x^T(t)\right) N_{\psi_i} + K_{\psi_i} \quad D_{\psi_i}\right] \begin{bmatrix}\theta \\ d\end{bmatrix} \leq b_{\psi_i} \\
&\Leftrightarrow \left[\left(I \otimes x^T(t)\right) \begin{bmatrix}N_{\psi_1} \\ N_{\psi_2}\end{bmatrix} + \begin{bmatrix}K_{\psi_1} \\ K_{\psi_2}\end{bmatrix} \quad \begin{bmatrix}D_{\psi_1} \\ D_{\psi_2}\end{bmatrix}\right] \begin{bmatrix}\theta \\ d\end{bmatrix} \leq \begin{bmatrix}b_{\psi_1} \\ b_{\psi_2}\end{bmatrix}
\end{aligned}$$

with $n_{\psi_1 \wedge \psi_2} = n_{\psi_1} + n_{\psi_2}$. This concludes the proof. □

Lemma 5. The matrix valued function

$$\left[\left((I_{n_\psi \times n_\psi} \otimes x^T(0)) N_\psi + K_\psi\right) \quad D_\psi\right] \begin{bmatrix}\theta \\ d\end{bmatrix}$$

is linear in $x^T(0)$ therefore its value at an initial condition $x(0) \in \mathbb{X}_0$ is a convex combination of the function values at the vertices X_0 of $\mathbb{X}_0$. Thus the satisfaction

relation $< \mathbf{M}(\theta), x(0) > \models \psi$ represented by the multi-affine inequality holds uniformly over $x(0) \in \mathbb{X}_0$ if and only if it holds for the vertices of $\mathbb{X}_0$.
This gives a set of affine inequalities in θ thus the maximal feasible set is a polyhedron. □

B Probability Distributions

For a given θ, the measurements $\tilde{y}(t)$ $t = 1, 2, \ldots$ are realised as

$$\begin{aligned} x(t+1) &= A^t x(0) + \sum_{k=0}^{t-1} A^k Bu(t-1-k) \quad x(0) \sim p(x_0) \\ y_0(t) &= \theta_0^T x(t), \\ \tilde{y}(t) &= y_0(t) + e(t), \quad e(t) \sim \mathcal{N}(0, \sigma_e^2). \end{aligned}$$

Their probability density distribution at time $t = 1$ can be defined as

$$\begin{aligned} & p(\tilde{y}(1)|u(0), \theta) \\ &= \int_{\mathbb{X}_0} \frac{1}{\sigma_e \sqrt{2\pi}} \exp\left[-\frac{(\theta^T Ax(0) + \theta^T Bu(0) - \tilde{y}(1))^2}{2\sigma_e^2}\right] p(x_0)dx_0. \end{aligned} \tag{13}$$

Remember that $\theta^T Ax(0) + \theta^T Bu(0)$ is equal to $\hat{y}(1, \theta)$ (implicitly parameterized on x_0). Therefore the conditional probability is derived as the probability density function of the Gaussian, zero-mean, additive measurement noise $e(t) = \hat{y}(t, \theta) - \tilde{y}(t)$ with variance σ_e^2. Denote the vector of N_s measurements as $\tilde{\mathbf{y}} = \begin{bmatrix} y(0) & y(1) & y(2) & \ldots & y(N_s) \end{bmatrix}^T$. Since the measurement noise is white it follows that the probability density distribution of $\tilde{\mathbf{y}}$, given a sequence of inputs $\mathbf{u} = \{u(t)\}^{N_s}$, is obtained via the product of a Gaussian distribution as in (13) and marginalised over the auxiliary variable x_0

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{u}, \theta) = &\quad \textstyle\int_{\mathbb{X}_0} \prod_{t=1}^{N_s} p(\tilde{y}(t)|\theta) p(x_0) dx_0, \\ = &\quad \textstyle\int_{\mathbb{X}_0} \frac{1}{\sqrt{\sigma_e^{2N_s} (2\pi)^{N_s}}} \exp\left[-\frac{\sum_{t=1}^{N_s} (\hat{y}(t,\theta) - \tilde{y}(t))^2}{2\sigma_e^2}\right] p(x_0) dx_0. \end{aligned}$$

Note that for a given initial condition $x(0)$, $p(x_0) = \delta_{x(0)}(x_0)$ the marginalisation over x_0 falls away. Let us assume that the initial condition is indeed given. The distribution over the data $\tilde{\mathbf{y}}$ is simplified using $\sum_{t=1}^{N_s} (\hat{y}(t, \theta) - \tilde{y}(t))^2 = \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)^T \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)$ and the vector function $\Phi(\mathbf{u}) = \begin{bmatrix} x(1) & x(2) & \ldots & x(N_s) \end{bmatrix}$,

$$p(\tilde{\mathbf{y}}|\mathbf{u}, \theta) = \frac{1}{\sqrt{\sigma_e^{2N_s} (2\pi)^{N_s}}} \exp\left[-\frac{\left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)^T \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)}{2\sigma_e^2}\right].$$

Compute the posterior distribution $p(\theta \mid \tilde{\mathbf{y}}, \mathbf{u}) = \frac{p(\tilde{\mathbf{y}}|\mathbf{u},\theta) p(\theta)}{\int_\Theta p(\tilde{\mathbf{y}}|\mathbf{u},\theta) p(\theta) d\theta}$ with the prior, $p(\theta) = \mathcal{N}(\theta_k, R_k)$. First we simplify the multiplication with the obtained data distribution

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{u}, \theta) p(\theta) = &\frac{1}{\sqrt{|R_k| \sigma_e^{2N_s} (2\pi)^{N_s + n}}} \exp\left[-\frac{1}{2}(\theta - \theta_k)^T R_k^{-1} (\theta - \theta_k)^T\right] \\ &\times \exp\left[-\frac{\left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)^T \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)}{2\sigma_e^2}\right]. \end{aligned}$$

The matrix summation in the exponents can be written as

$$\begin{aligned} &\sigma_e^{-2} \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)^T \left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right) + (\theta - \theta_k)^T R_k^{-1} (\theta - \theta_k)^T \\ &= \begin{bmatrix} -1 \\ \theta \end{bmatrix}^T \begin{bmatrix} \tilde{\mathbf{y}}^T \sigma_e^{-2} \tilde{\mathbf{y}} + \theta_k^T R_k \theta_k & \sigma_e^{-2} \tilde{\mathbf{y}}^T \Phi^T(\mathbf{u}) + \theta_k^T R_k^{-1} \\ \sigma_e^{-2} \Phi(\mathbf{u}) \tilde{\mathbf{y}} + R_\theta^{-1} \theta_k & \Phi(\mathbf{u}) \sigma_e^{-2} \Phi^T(\mathbf{u}) + R_k^{-1} \end{bmatrix} \begin{bmatrix} -1 \\ \theta \end{bmatrix}. \end{aligned}$$

Therefore the a-posteriori will have a Gaussian distribution with mean θ_{k+1} and variance R_{k+1}

$$\begin{aligned} R_{k+1} &= \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1}, \\ \theta_{k+1} &= R_{k+1}\left[R_k^{-1}\theta_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right], \end{aligned}$$

since we can rewrite the matrices as

$$\begin{gathered} \begin{bmatrix} -1 \\ \theta \end{bmatrix}^T \begin{bmatrix} {\theta_{k+1}}^T \\ I \end{bmatrix} R_{k+1}^{-1} \begin{bmatrix} {\theta_{k+1}}^T \\ I \end{bmatrix}^T \begin{bmatrix} -1 \\ \theta \end{bmatrix} \\ +\tilde{\mathbf{y}}^T\sigma_e^{-2}\tilde{\mathbf{y}} + \theta_k^T R_k \theta_k - \theta_{k+1}^T R_{k+1}^{-1}\theta_{k+1}. \end{gathered}$$

In this summation the terms dependent on θ are separated from the terms independent of θ. The independent terms are related to the unlikeliness of the data based on the prior knowledge. When computing the a-posteriori distribution they are cancelled by the normalisation. Thus for a Gaussian prior, the a-posteriori probability distribution equals $p\big(\theta|\tilde{\mathbf{y}}, \mathbf{u}\big) = \mathcal{N}\big(\theta_{k+1}, R_{k+1}\big)$.

$$\begin{aligned} p(\theta|\tilde{\mathbf{y}}, \mathbf{u}) &= \frac{1}{\sqrt{|R_{k+1}|(2\pi)^n}} \exp\left[-\frac{1}{2}\left(\theta - \theta_{k+1}\right)^T R_{k+1}^{-1}\left(\theta - \theta_{k+1}\right)\right], \\ \text{with} & \\ R_{k+1} &= \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1}, \\ \theta_{k+1} &= R_{k+1}\left[R_k^{-1}\theta_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right]. \end{aligned}$$

Let us now compute the distribution of the data conditioned on $\mathbf{u}$ only, and marginalised over θ

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{u}) &= \int_\Theta p(\tilde{\mathbf{y}}|\mathbf{u}, \theta)p(\theta)d\theta \\ p(\tilde{\mathbf{y}}|\mathbf{u}) &= \int_\Theta p(\tilde{\mathbf{y}}|\mathbf{u}, \theta)\frac{1}{\sqrt{|R_k|(2\pi)^n}} \exp\left[-\frac{1}{2}(\theta - \theta_k)^T R_k^{-1}(\theta - \theta_k)^T\right] d\theta \\ &= \int_\Theta \frac{1}{\sqrt{\sigma_e^{2N_s}|R_k|(2\pi)^{N_s+n}}} \exp\left[-\frac{1}{2}(\theta - \theta_k)^T R_k^{-1}(\theta - \theta_k)^T\right] \\ &\qquad \times \exp\left[-\frac{\left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)^T\left(\Phi^T(\mathbf{u})\theta - \tilde{\mathbf{y}}\right)}{2\sigma_e^2}\right] d\theta \\ &= \underbrace{\int_\Theta \frac{1}{\sqrt{|R_{k+1}|(2\pi)^n}} \exp\left[-\frac{1}{2}\left(\theta - \theta_{k+1}\right)^T R_{k+1}^{-1}\left(\theta - \theta_{k+1}\right)\right] d\theta}_{=1} \\ &\qquad \times \sqrt{\frac{|R_{k+1}|}{\sigma_e^{2N_s}|R_k|(2\pi)^{N_s}}} \\ &\qquad \times \exp\left[-\frac{1}{2}\left(\tilde{\mathbf{y}}^T\sigma_e^{-2}\tilde{\mathbf{y}} + \theta_k^T R_k\theta_k - \theta_{k+1}^T R_{k+1}^{-1}\theta_{k+1}\right)\right]. \end{aligned}$$

The term in the exponent can be written as a matrix multiplication by substituting $\theta_{k+1} = R_{k+1}\left[R_k^{-1}\theta_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right]$,

$$\begin{aligned} &\sigma_e^{-2}\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} + \theta_k^T R_k^{-1}\theta_k \\ &- \left[R_k^{-1}\theta_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right]^T R_{k+1}\left[R_k^{-1}\theta_k + \sigma_e^{-2}\Phi(\mathbf{u})\tilde{\mathbf{y}}\right] \\ &= \begin{bmatrix} -1 \\ \tilde{\mathbf{y}} \end{bmatrix}^T \times \\ &\begin{bmatrix} \theta_k^T R_k^{-1}\theta_k - \theta_k^T R_k^{-1} R_{k+1} R_k^{-1}\theta_k & \theta_k^T R_k^{-1} R_{k+1}\sigma_e^{-2}\Phi(\mathbf{u}) \\ \sigma_e^{-2}\Phi(\mathbf{u})^T R_{k+1} R_k^{-1}\theta_k & \sigma_e^{-2} - \sigma_e^{-4}\Phi(\mathbf{u})^T R_{k+1}\Phi(\mathbf{u}) \end{bmatrix} \\ &\qquad\qquad\qquad\qquad \times \begin{bmatrix} -1 \\ \tilde{\mathbf{y}} \end{bmatrix}. \end{aligned}$$

Consider the matrix inverse lemma and simplify the lower right corner of the block matrix

$$\begin{aligned}
&\sigma_e^{-2} - \sigma_e^{-4}\Phi(\mathbf{u})^T R_{k+1}\Phi(\mathbf{u}) \\
&= \sigma_e^{-2} - \sigma_e^{-4}\Phi(\mathbf{u})^T \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1} \Phi(\mathbf{u}) \\
&= \left\langle \text{ use } (A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \right\rangle \\
&= \left\langle \text{ with } A = \sigma_e^2,\ B = \Phi(\mathbf{u})^T,\ C = \Phi(\mathbf{u}),\ D = -R_k^{-1} \right\rangle \\
&= \left[\sigma_e^2 + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right]^{-1} .
\end{aligned}$$

Denote the above matrix as $R_{\tilde{\mathbf{y}}}^{-1}$, i.e. $R_{\tilde{\mathbf{y}}} := \left[\sigma_e^2 + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right]$. Using the dual lemma for the left upper term for $A = R_k^{-1}$, $B = \Phi(\mathbf{u})$, $C = \Phi(\mathbf{u})^T$, $D = -\sigma_e^2 I$ gives

$$\begin{aligned}
&\theta_k^T R_k^{-1}\theta_k - \theta_k^T R_k^{-1} R_{k+1} R_k^{-1}\theta_k = \theta_k^T R_k^{-1}(R_k - R_{k+1})R_k^{-1}\theta_k \\
&= \theta_k^T R_k^{-1}(R_k - \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1})R_k^{-1}\theta_k \\
&= \theta_k^T R_k^{-1}(R_k - R_k - R_k\Phi(\mathbf{u})(-\sigma_e^{-2} - \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u}))^{-1}\Phi(\mathbf{u})^T R_k) \\
&\quad \times R_k^{-1}\theta_k \\
&= \theta_k^T \left[\Phi(\mathbf{u})(\sigma_e^2 + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u}))^{-1}\Phi(\mathbf{u})^T\right] \theta_k \\
&= \theta_k^T \Phi(\mathbf{u}) R_{\tilde{\mathbf{y}}}^{-1} \Phi(\mathbf{u})^T \theta_k \ .
\end{aligned}$$

Hence the matrices inside the exponent of the density function become

$$\begin{bmatrix} -1 \\ \tilde{\mathbf{y}} \end{bmatrix}^T \times \begin{bmatrix} \theta_k^T \Phi(\mathbf{u}) R_{\tilde{\mathbf{y}}}^{-1} \Phi(\mathbf{u})^T \theta_k & \theta_k^T R_k^{-1} R_{k+1} \sigma_e^{-2} \Phi(\mathbf{u}) \\ \sigma_e^{-2}\Phi(\mathbf{u})^T R_{k+1} R_k^{-1}\theta_k & R_{\tilde{\mathbf{y}}} \end{bmatrix} \times \begin{bmatrix} -1 \\ \tilde{\mathbf{y}} \end{bmatrix} .$$

Now we want to extract the mean value $\mu_{\tilde{\mathbf{y}}}$ of the data sequence $\tilde{\mathbf{y}}$ from the off-diagonal terms

$$\begin{aligned}
&R_{\tilde{\mathbf{y}}}^{-1}\mu_{\tilde{\mathbf{y}}} = \sigma_e^{-2}\Phi(\mathbf{u})^T R_{k+1} R_k^{-1}\theta_k \\
&\mu_{\tilde{\mathbf{y}}} = \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \Phi(\mathbf{u})^T R_{k+1} R_k^{-1}\theta_k \\
&= \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \Phi(\mathbf{u})^T \left[R_k^{-1} + \sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1} R_k^{-1}\theta_k \\
&= \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \Phi(\mathbf{u})^T \left[I + \sigma_e^{-2} R_k \Phi(\mathbf{u})\Phi(\mathbf{u})^T\right]^{-1} \theta_k \\
&= [\text{ Note that : } (I + AB)^{-1} = I - A(I + BA)^{-1}B] \\
&= \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \Phi(\mathbf{u})^T \times \\
&\left[I - \sigma_e^{-2} R_k \Phi(\mathbf{u}) \left(I + \sigma_e^{-2}\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right)^{-1} \Phi(\mathbf{u})^T\right] \theta_k \\
&= \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \times \\
&[\Phi(\mathbf{u})^T\theta_k - \sigma_e^{-2}\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u}) \left(I + \sigma_e^{-2}\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right)^{-1} \Phi(\mathbf{u})^T\theta_k] \\
&= \sigma_e^{-2} R_{\tilde{\mathbf{y}}} \times \\
&[I - \sigma_e^{-2}\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u}) \left(I + \sigma_e^{-2}\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right)^{-1}]\Phi(\mathbf{u})^T\theta_k \\
&= [\text{Push through Identity } A(I + A)^{-1} = I - (I + A)^{-1}] \\
&\mu_{\tilde{\mathbf{y}}} = R_{\tilde{\mathbf{y}}}^{-1} \left[\left(\sigma_e^2 I_{N_s \times N_s} + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right)^{-1}\right] \Phi(\mathbf{u})^T\theta_k = \Phi(\mathbf{u})^T\theta_k
\end{aligned}$$

Thus

$$\begin{aligned}
p(\tilde{\mathbf{y}}|\mathbf{u}) &= \sqrt{\frac{|R_{k+1}|}{\sigma_e^{2N_s}(2\pi)^{N_s}|R_k|}} \exp\left[-\frac{1}{2}(\tilde{\mathbf{y}} - \Phi(\mathbf{u})^T\theta_k)^T R_{\tilde{\mathbf{y}}}^{-1}(\tilde{\mathbf{y}} - \Phi(\mathbf{u})^T\theta_k)\right] \\
&=^* \mathcal{N}\left(\Phi(\mathbf{u})^T\theta_k, \left[\sigma_e^2 + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right]\right)
\end{aligned}$$

The last equality $=^*$ follows from

$$\begin{aligned}
\frac{\sigma_e^{2N_s}|R_k|}{|R_{k+1}|} &= |R_{\tilde{\mathbf{y}}}| \\
\sigma_e^{2N_s}|R_k|\left|R_k^{-1}+\sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T\right| &= \left|\sigma_e^2+\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right| \\
\sigma_e^{2N_s}\left|I+\sigma_e^{-2}\Phi(\mathbf{u})\Phi(\mathbf{u})^T R_k\right| &= \left|\sigma_e^2+\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right| \\
&= \langle\text{Sylvester's determinant theorem}\rangle \\
\left|\sigma_e^2+\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right| &= \left|\sigma_e^2+\Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right|
\end{aligned}$$

$$\begin{aligned}
&\text{Gaussian prior} \rightarrow \text{ Gaussian data distribution} \\
&p(\tilde{\mathbf{y}}|\mathbf{u}) = \sqrt{\frac{1}{(2\pi)^{N_s}|R_{\tilde{\mathbf{y}}}|}}\exp\left[-\frac{1}{2}(\tilde{\mathbf{y}}-\Phi(\mathbf{u})^T\theta_k)^T R_{\tilde{\mathbf{y}}}^{-1}(\tilde{\mathbf{y}}-\Phi(\mathbf{u})^T\theta_k)\right],
\end{aligned}$$

with $R_{\tilde{\mathbf{y}}} = \left[\sigma_e^2 I_{N_s\times N_s} + \Phi(\mathbf{u})^T R_k \Phi(\mathbf{u})\right].$