Journal of Machine Learning Research 10 (2009) 515-554 Submitted 12/07; Revised 8/08; Published 2/09

Identification of Recurrent Neural Networks by Bayesian Interrogation Techniques

Barnabás Póczos[*] POCZOS@CS.UALBERTA.CA
András Lőrincz ANDRAS.LORINCZ@ELTE.HU
Department of Information Systems, Eötvös Loránd University
Pázmány P. sétány 1/C, Budapest H-1117, Hungary

Editor: Zoubin Ghahramani

Abstract

We introduce novel online Bayesian methods for the identification of a family of noisy recurrent neural networks (RNNs). We present Bayesian active learning techniques for stimulus selection given past experiences. In particular, we consider the unknown parameters as stochastic variables and use A-optimality and D-optimality principles to choose optimal stimuli. We derive myopic cost functions in order to maximize the information gain concerning network parameters at each time step. We also derive the A-optimal and D-optimal estimations of the additive noise that perturbs the dynamical system of the RNN. Here we investigate myopic as well as non-myopic estimations, and study the problem of simultaneous estimation of both the system parameters and the noise. Employing conjugate priors our derivations remain approximation-free and give rise to simple update rules for the online learning of the parameters. The efficiency of our method is demonstrated for a number of selected cases, including the task of controlled independent component analysis.

Keywords: active learning, system identification, online Bayesian learning, A-optimality, D-optimality, infomax control, optimal design

1. Introduction

When studying systems in *interactive* and *online* fashion, it is of high relevance to facilitate fast information gain during the interaction (Fedorov, 1972; Cohn, 1994). As an example, consider experiments aiming at the description of the receptive field of different neurons. These experiments look for those stimuli that maximize the response of the given neuron (deCharms et al., 1998; Földiák, 2001). Neurons, however, might change due to the investigation, so the minimization of interaction is highly desired. Different techniques have been developed to speed up the identification procedure. One approach searches for stimulus distribution that maximizes mutual information between stimulus and response (Machens et al., 2005). A recent technique assumes that the unknown system belongs to the family of generalized linear models (Lewi et al., 2007) and treats the parameters as probabilistic variables. Then the goal is to find the optimal stimuli by maximizing mutual information between the parameter set and the system's response.

This example motivates our interest in active learning (MacKay, 1992; Cohn et al., 1996; Fukumizu, 1996; Sugiyama, 2006) of noisy recurrent artificial neural networks (RNNs), when we have the freedom to interrogate the network and to measure its responses.

∗. Present address: Department of Computing Science, University of Alberta, Athabasca Hall, Edmonton, Canada, T6G 2E8

In active learning, the training set may be modified by the learning process itself based on the progress experienced so far. The goal of this modification is to maximize the expected improvement of the precision of the estimation. This idea can for example be used to improve generalization capability in regression and classification tasks or to better estimate hidden parameters. Theoretical concepts have been formulated in the fields of Optimal Experimental Design, or Optimal Bayesian Design (Kiefer, 1959; Fedorov, 1972; Steinberg and Hunter, 1984; Toman and Gastwirth, 1993; Pukelsheim, 1993).

Although active learning is in the focus of current research interest, some relevant theoretical issues are still unresolved. While there are promising studies showing that active learning may outperform uniform sampling under certain conditions (Freund et al., 1997; Seung et al., 1992), in other cases it has been proven that active learning has no advantage over non-adaptive algorithms. For example, this is the case in compressed sensing (Castro et al., 2006a) and also for certain function classes in the area of function approximation (Castro et al., 2006b). Even more problematic is the observation that active learning heuristics may be less efficient than uniform sampling in some situations (Schein, 2005).

There are several forms of active learning. The most relevant difference is in the definition of the value of information. One of the simplest heuristics is the Uncertainty Sampling (US): US suggests that in regression or in classification tasks one should choose those training examples, which have the largest uncertainty in the value of the function or in the label of the class, respectively (Lewis and Catlett, 1994; Lewis and Gale, 1994; Cohn et al., 1996). Although several US versions exist with different measure of the uncertainty itself, they all lack robustness. The Query by Committee method improves upon robustness (Seung et al., 1992; Freund et al., 1997): the committee of a few models are trained on the existing training set and the next query points are selected to reduce the disagreement among these models. The method of Roy and McCallum (2001) minimizes the direct error, that is, it tries to choose training points to minimize the expected classification error directly.

In the literature there are other approaches, including decision theory based methods. The original ideas were worked out in Raiffa and Schlaifer (1961) and Lindley (1971). The objective in this method family is to choose the design such that the predicted value of a given utility function become maximal. Numerous utility functions have been proposed. For example, if we aim to estimate the unknown parameter $\boldsymbol{\theta}$, then one possible direction is the minimization of, for example, the entropy or the standard deviation of the posterior distribution. If we minimize the entropy then we arrive at the D-optimality principle (Bernardo, 1979; Stone, 1959). This principle is equivalent to the information maximization method (also known as infomax principle) of Lewi et al. (2007). If we intend to minimize the standard deviation then the result is the A-optimality principle (Duncan and DeGroot, 1976). A special case is called the c-optimality principle (Chaloner, 1984) when the goal is to estimate a linear projection of parameter $\boldsymbol{\theta}$ ($\mathbf{c}^T\boldsymbol{\theta}$). There exist a number of other methods, called alphabetical optimality and utility functions. For a review see, for example, Chaloner and Verdinelli (1995). Although the original ideas belong to the field of optimal experimental design, they have appeared also in active learning recently (MacKay, 1992; Tong and Koller, 2000; Schein and Ungar, 2007).

Today, active learning is present almost in all fields of machine learning and there are many popular applications on diverse areas, including Gaussian Processes (Krause and Guestrin, 2007), Artificial Neural Networks (Fukumizu, 2000), Support Vector Machines (Tong and Koller, 2001b), Generalized Linear Models (Bach, 2007; Lewi et al., 2007), Logistic Regression (Schein, 2005),

learning the parameters and structure of Bayes nets (Tong and Koller, 2000, 2001a) and Hidden Markov Models (Anderson and Moore, 2005).

Our framework is similar to the generalized linear model (GLM) approach used by Lewi et al. (2007): we would like to choose interrogating, or '*control*' inputs in order to (i) identify the parameters of the network and (ii) estimate the additive noise efficiently. From now on, we use the terms *control* and *interrogation* interchangeably; control is the conventional expression, whereas the word interrogation expresses our aims better. We apply online Bayesian learning (Opper and Winther, 1999; Solla and Winther, 1999; Honkela and Valpola, 2003; Ghahramani, 2000). For Bayesian methods, prior updates often lead to intractable posterior distributions such as a mixture of exponentially numerous distributions. Here, we show that, for the model studied in this paper, computations are both tractable and approximation-free. Further, the emerging learning rules are simple. We also show that different stimuli are needed for the same RNN model depending on whether the goal is to estimate the weights of the RNN or the additive perturbation (referred to as 'driving noise').

In this article we investigate the D-optimality, as well as the A-optimality principles. To the best of our knowledge, neither of them has been applied to the typical non-spiking stochastic artificial recurrent neural network model that we treat here.

The contribution of this paper can be summarized as follows: We use A-optimality and D-optimality principles and derive cost functions and algorithms for (i) the learning of parameters of the stochastic RNN and (ii) the estimation of its driving noise. We show that, (iii) using the D-optimality interrogation technique, these two tasks are incoherent in the myopic (i.e., single step look-ahead) control scheme: signals derived from this principle for parameter estimation are sub-optimal (basically the worst possible) for the estimation of the driving noise and vice versa. (iv) We show that for the case of noise estimation task the two principles, that is, A- and D-optimality principles result in the same cost function. (v) For the A-optimality case, we derive equations for the joined estimation of the noise and the parameters. On the contrary, we show also that (vi) D-optimality cannot be applied on the same joined task. For the case of noise estimation, (vii) a non-myopic multiple step look-ahead heuristics is introduced and we demonstrate its applicability through numerical experiments.

The paper is structured as follows: In Section 2 we introduce our model. Section 3 concerns the Bayesian equations of the RNN model. In Section 4 optimal control for parameter identification is derived from the D-optimality principle. Section 5 is about the same task, but using the A-optimality principle instead. Section 6 deals with our second task, when the goal is the estimation of the driving noise of the RNN. Here we treat the D-optimality principle. Section 7 is about the same problem, but for the A-optimality principle. We combine the two tasks for both optimality principles in Section 8 and consider the cost functions for the joined estimation of the parameters and the driving noise. All of these considerations concern myopic algorithms. In Section 9 a non-myopic heuristics is introduced for the noise estimation task. Section 10 contains our numerical experiments for a number of cases, including independent component analysis. The paper ends with a short discussion and some conclusions (Section 11). Technical details of the derivations can be found in the Appendix.

2. The Model

Let $P(\mathbf{e}) = \mathcal{N}_{\mathbf{e}}(\mathbf{m}, \mathbf{V})$ denote the probability density of a normally distributed stochastic variable $\mathbf{e}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{V}$. Let us assume that we have d simple computational units called '*neurons*' in a recurrent neural network:

$$\mathbf{r}_{t+1} \quad = \quad g\left(\sum_{i=0}^{I} \mathbf{F}_i \mathbf{r}_{t-i} + \sum_{j=0}^{J} \mathbf{B}_j \mathbf{u}_{t+1-j} + \mathbf{e}_{t+1}\right), \tag{1}$$

where $\{\mathbf{e}_t\}$, the driving noise of the RNN, denotes temporally independent and identically distributed (i.i.d.) stochastic variables and $P(\mathbf{e}_t) = \mathcal{N}_{\mathbf{e}_t}(\mathbf{0}, \mathbf{V})$, $\mathbf{r}_t \in \mathbb{R}^d$ represents the observed activities of the neurons at time t. Let $\mathbf{u}_t \in \mathbb{R}^c$ denote the control signal at time t. The neural network is formed by the weighted delays represented by matrices $\mathbf{F}_i$ ($i = 0, \ldots, I$) and $\mathbf{B}_j$ ($j = 0, \ldots, J$), which connect neurons to each other and also the control components to the neurons, respectively. Control can also be seen as the means of interrogation, or the stimulus to the network (Lewi et al., 2007). We assume that function $g : \mathbb{R}^d \to \mathbb{R}^d$ in (1) is known and invertible. The computational units, the neurons, sum up weighted previous neural activities as well as weighted control inputs. These sums are then passed through identical non-linearities according to Eq. (1). Our goal is to estimate the parameters $\mathbf{F}_i \in \mathbb{R}^{d \times d}$ ($i = 0, \ldots, I$), $\mathbf{B}_j \in \mathbb{R}^{d \times c}$ ($j = 0, \ldots, J$) and the covariance matrix $\mathbf{V}$, as well as the driving noise $\mathbf{e}_t$ by means of the control signals.

In artificial neural network terms, (1) is in the form of *rate code models*. This is the typical form for RNNs, but there are methods to approximate rate code description with spike codes and vice versa. For the case of RNNs, the best is to compare Liquid State Machine, a spike code model of Maass et al. (2002) with the Echo State Network, the corresponding rate code model of Jaeger (2001). Rate code, very crudely, is the low pass filtered spike code, whereas spike code can be seen as the response of integrate-and-fire neurons. We show that analytic cost functions emerge for the rate code RNN model. Due to the applied conjugate priors, we can calculate the high dimensional integrals involved in our derivations, and hence these derivations remain approximation-free and give rise to simple update rules.

3. Bayesian Approach

Here we embed the estimation task into the Bayesian framework. First, we introduce the following notations: $\mathbf{x}_{t+1} = [\mathbf{r}_{t-I}; \ldots; \mathbf{r}_t; \mathbf{u}_{t-J+1}; \ldots; \mathbf{u}_{t+1}]$, $\mathbf{y}_{t+1} = g^{-1}(\mathbf{r}_{t+1})$, $\mathbf{A} = [\mathbf{F}_I, \ldots, \mathbf{F}_0, \mathbf{B}_J, \ldots, \mathbf{B}_0] \in \mathbb{R}^{d \times m}$. With these notations, model (1) reduces to a linear equation

$$\mathbf{y}_t \quad = \quad \mathbf{A}\mathbf{x}_t + \mathbf{e}_t. \tag{2}$$

In order to estimate the unknown quantities (parameter matrix $\mathbf{A}$, noise $\mathbf{e}_t$ and its covariance matrix $\mathbf{V}$) in an online fashion, we rely on Bayes' method. We assume that prior knowledge is available and we update our posteriori knowledge on the basis of the observations. Control will be chosen at each instant to provide maximal expected information concerning the quantities we have to estimate. Starting from an arbitrary prior distribution of the parameters the posterior distribution needs to be computed. This latter distribution, however, can be highly complex, so approximations are applied. For example, assumed density filtering, when the computed posterior is projected to simpler distributions, has been suggested (Boyen and Koller, 1998; Minka, 2001; Opper and Winther,

1999). In order to avoid approximations, we apply the method of conjugated priors (Gelman et al., 2003). For matrix $\mathbf{A}$ we assume a matrix valued normal distribution prior.

For the case of D-optimality principle, we shall use the inverted Wishart (IW) distribution as our prior for covariance matrix $\mathbf{V}$. This is the most general known conjugate prior distribution for the covariance matrix of a normal distribution at present. For A-optimality, however, we keep the derivations simple and assume that the covariance matrix has diagonal structure. In turn, we replaced the IW assumption on the prior with the distribution of the Product of Inverted Gammas (PIG).

We define the normally distributed matrix valued stochastic variable $\mathbf{A} \in \mathbb{R}^{d\times m}$ by using the following quantities: $\mathbf{M} \in \mathbb{R}^{d\times m}$ is the expected value of $\mathbf{A}$. $\mathbf{V} \in \mathbb{R}^{d\times d}$ is the covariance matrix of the rows, and $\mathbf{K} \in \mathbb{R}^{m\times m}$ is the so-called precision parameter matrix that we shall modify in accordance with the Bayesian update. Matrix $\mathbf{K}$ contains the estimations of the 'Bayesian trainer' about the precision of parameters in $\mathbf{A}$. Informally, matrix $\mathbf{K}$ behaves as the inverse of a covariance matrix. Upon each observation, matrix $\mathbf{K}$ is updated. The larger the eigenvalues of this matrix, the smaller the variance ellipsoids of the posteriori estimations are.

Both $\mathbf{K}$ and $\mathbf{V}$ are positive semi-definite matrices. The density function of the stochastic variable $\mathbf{A}$ is defined as:

$$\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K}) = \frac{|\mathbf{K}|^{d/2}}{|2\pi\mathbf{V}|^{m/2}} \exp(-\frac{1}{2} tr((\mathbf{A}-\mathbf{M})^T \mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K})),$$

where tr, $|\cdot|$, and superscript T denote the trace operation, the determinant, and transposition, respectively (see, e.g., Gupta and Nagar, 1999; Minka, 2000). We assume that $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a positive definite matrix and $n > 0$. Using these notations, the density of the Inverted Wishart distribution with parameters $\mathbf{Q}$ and n is as follows (Gupta and Nagar, 1999):

$$I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n) = \frac{1}{Z_{n,d}} \frac{1}{|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{Q}}{2}\right|^{n/2} \exp(-\frac{1}{2} tr(\mathbf{V}^{-1}\mathbf{Q})),$$

where $Z_{n,d} = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((n+1-i)/2)$ and $\Gamma(.)$ denotes the gamma function.

Similarly, let $\mathbf{V} = diag(\mathbf{v}) \in \mathbb{R}^{d\times d}$ diagonal covariance matrix with $0 < \mathbf{v} \in \mathbb{R}^d$ diagonal values. With the slight abuse of notation we will use later the $\mathbf{v} = diag(\mathbf{V}) \in \mathbb{R}^d$ term, too. Then the density of PIG is defined as

$$\mathcal{PIG}_{\mathbf{V}}(\alpha,\beta) = \prod_{i=1}^{d} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{-\alpha_i-1} \exp(-\frac{\beta_i}{v_i}),$$

where $\alpha_i > 0$ and $\beta_i > 0$ are the shape and scale parameters respectively.

Now, one can rewrite model (2) as follows:

$$P(\mathbf{A}|\mathbf{V}) = \mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K}), \quad (3)$$

$$P(\mathbf{e}_t|\mathbf{V}) = \mathcal{N}_{\mathbf{e}_t}(\mathbf{0},\mathbf{V}), \quad (4)$$

$$P(\mathbf{y}_t|\mathbf{A},\mathbf{x}_t,\mathbf{V}) = \mathcal{N}_{\mathbf{y}_t}(\mathbf{A}\mathbf{x}_t,\mathbf{V}), \quad (5)$$

and $P(\mathbf{V}) = \mathcal{PIG}_{\mathbf{V}}(\alpha,\beta)$ or $P(\mathbf{V}) = I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)$ depending on whether we want to use A- or D-optimality.

4. D-Optimality Approach for Parameter Learning

Let us compute the D-optimal parameter estimation strategy for our RNN given by (1) and rewritten into (3)-(5). Let us introduce two shorthands; $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{V}\}$, and $\{\mathbf{x}\}_i^j = \{\mathbf{x}_i, \ldots, \mathbf{x}_j\}$. We choose the control value in (1) at each instant to provide maximal expected information concerning the unknown parameters. Assuming that $\{\mathbf{x}\}_1^t$, $\{\mathbf{y}\}_1^t$ are given, according to the infomax principle our goal is to compute

$$\arg\max_{\mathbf{u}_{t+1}} I(\boldsymbol{\theta}, \mathbf{y}_{t+1}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t), \tag{6}$$

where $I(a,b;c)$ denotes the mutual information of stochastic variables a and b for fixed parameters c. Let $H(a|b;c)$ denote the conditional entropy of variable a conditioned on variable b and for fixed parameter c. Note that

$$I(\boldsymbol{\theta}, \mathbf{y}_{t+1}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) = H(\boldsymbol{\theta}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) - H(\boldsymbol{\theta}|\mathbf{y}_{t+1}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t),$$

holds (Cover and Thomas, 1991) and $H(\boldsymbol{\theta}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) = H(\boldsymbol{\theta}; \{\mathbf{x}\}_1^t, \{\mathbf{y}\}_1^t)$ is independent from $\mathbf{u}_{t+1}$, hence our task is reduced to the evaluation of the following quantity:

$$\begin{aligned} &\arg\min_{\mathbf{u}_{t+1}} H(\boldsymbol{\theta}|\mathbf{y}_{t+1}; \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) = \\ &= \arg\min_{\mathbf{u}_{t+1}} -\int d\mathbf{y}_{t+1} P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) \int d\boldsymbol{\theta} P(\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) \log P(\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}). \end{aligned} \tag{7}$$

In order to solve this minimization problem we need to evaluate $P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t)$, the posterior $P(\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1})$, and the entropy of the posterior, that is $\int d\boldsymbol{\theta} P(\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) \log P(\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1})$, where $P(a|b)$ denotes the conditional probability of variable a given condition b. The main steps of these computations are presented below.

Assume that the *a priori* distributions $P(\mathbf{A}|\mathbf{V}, \{\mathbf{x}\}_1^t, \{\mathbf{y}\}_1^t) = \mathcal{N}(\mathbf{A}|\mathbf{M}_t, \mathbf{V}, \mathbf{K}_t)$ and $P(\mathbf{V}|\{\mathbf{x}\}_1^t, \{\mathbf{y}\}_1^t) = I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}_t, n_t)$ are known. Then the posterior distribution of $\boldsymbol{\theta}$ is:

$$\begin{aligned} P(\mathbf{A}, \mathbf{V}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) &= \frac{P(\mathbf{y}_{t+1}|\mathbf{A}, \mathbf{V}, \mathbf{x}_{t+1}) P(\mathbf{A}|\mathbf{V}, \{\mathbf{x}\}_1^t, \{\mathbf{y}\}_1^t) P(\mathbf{V}|\{\mathbf{x}\}_1^t, \{\mathbf{y}\}_1^t)}{P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t)}, \\ &= \frac{\mathcal{N}_{\mathbf{y}_{t+1}}(\mathbf{A}\mathbf{x}_{t+1}, \mathbf{V}) \mathcal{N}_{\mathbf{A}}(\mathbf{M}_t, \mathbf{V}, \mathbf{K}_t) I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}_t, n_t)}{\int_{\mathbf{A}} \int_{\mathbf{V}} \mathcal{N}_{\mathbf{y}_{t+1}}(\mathbf{A}\mathbf{x}_{t+1}, \mathbf{V}) \mathcal{N}_{\mathbf{A}}(\mathbf{M}_t, \mathbf{V}, \mathbf{K}_t) I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}_t, n_t)}. \end{aligned}$$

This expression can be rewritten in a more useful form: let $\mathbf{K} \in \mathbb{R}^{m \times m}$ and $\mathbf{Q} \in \mathbb{R}^{d \times d}$ be positive definite matrices. Let $\mathbf{A} \in \mathbb{R}^{d \times m}$, and let us introduce the density function of the matrix valued Student-t distribution (Kotz and Nadarajah, 2004; Minka, 2000) as follows:

$$\mathcal{T}_{\mathbf{A}}(\mathbf{Q}, n, \mathbf{M}, \mathbf{K}) = \frac{|\mathbf{K}|^{d/2}}{\pi^{dm/2}} \frac{Z_{n+m,d}}{Z_{n,d}} \frac{|\mathbf{Q}|^{n/2}}{|\mathbf{Q} + (\mathbf{A} - \mathbf{M})\mathbf{K}(\mathbf{A} - \mathbf{M})^T|^{(m+n)/2}}.$$

Now, we need the following lemma:

Lemma 4.1

$$\begin{aligned} \mathcal{N}_{\mathbf{y}}(\mathbf{A}\mathbf{x}, \mathbf{V}) \mathcal{N}_{\mathbf{A}}(\mathbf{M}, \mathbf{V}, \mathbf{K}) I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}, n) = \mathcal{N}_{\mathbf{A}}((\mathbf{M}\mathbf{K} + \mathbf{y}\mathbf{x}^T)(\mathbf{x}\mathbf{x}^T + \mathbf{K})^{-1}, \mathbf{V}, \mathbf{x}\mathbf{x}^T + \mathbf{K}) \times \\ \times I\mathcal{W}_{\mathbf{V}}\left(\mathbf{Q} + (\mathbf{y} - \mathbf{M}\mathbf{x})(1 - \mathbf{x}^T(\mathbf{x}\mathbf{x}^T + \mathbf{K})^{-1}\mathbf{x})(\mathbf{y} - \mathbf{M}\mathbf{x})^T, n+1\right) \times \\ \times \mathcal{T}_{\mathbf{y}}\left(\mathbf{Q}, n, \mathbf{M}\mathbf{x}, 1 - \mathbf{x}^T(\mathbf{x}\mathbf{x}^T + \mathbf{K})^{-1}\mathbf{x}\right). \end{aligned}$$

The proof can be found in the Appendix.

Using this lemma, we can compute the posterior probabilities. We introduce the following quantities:

$$\begin{aligned}
\gamma_{t+1} &= 1-\mathbf{x}_{t+1}^T(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1}\mathbf{x}_{t+1}, && (8)\\
n_{t+1} &= n_t+1,\\
\mathbf{M}_{t+1} &= (\mathbf{M}_t\mathbf{K}_t+\mathbf{y}_{t+1}\mathbf{x}_{t+1}^T)(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1}, && (9)\\
\mathbf{Q}_{t+1} &= \mathbf{Q}_t+(\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})\,\gamma_{t+1}\,(\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})^T. && (10)
\end{aligned}$$

For the posterior probabilities we have determined that

$$\begin{aligned}
P(\mathbf{A}|\mathbf{V},\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= \mathcal{N}_{\mathbf{A}}(\mathbf{M}_{t+1},\mathbf{V},\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t), && (11)\\
P(\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}_{t+1},n_{t+1}), && (12)\\
P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t}) &= \mathcal{T}_{\mathbf{y}_{t+1}}(\mathbf{Q}_t,n_t,\mathbf{M}_t\mathbf{x}_{t+1},\gamma_{t+1}).
\end{aligned}$$

Now we are in the position to compute the entropy of the posterior distribution of $\boldsymbol{\theta}=\{\mathbf{A},\mathbf{V}\}$ using the following lemma:

Lemma 4.2 *The entropy of a stochastic variable with density function* $P(\mathbf{A},\mathbf{V}) = \mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\,I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)$ *assumes the form* $-\frac{d}{2}\ln|\mathbf{K}|+(\frac{m+d+1}{2})\ln|\mathbf{Q}|+f_{1,1}(d,n)$*, where* $f_{1,1}(d,n)$ *depends only on* d *and* n*.*

The proof can be found in the Appendix.

Lemmas 4.1 and 4.2 lead to the following corollary:

Corollary 4.3 *For the entropy of a stochastic variable with posterior distribution* $P(\mathbf{A},\mathbf{V}|\mathbf{x},\mathbf{y})$ *it holds that*

$$H(\mathbf{A},\mathbf{V};\mathbf{x},\mathbf{y}) = -\frac{d}{2}\ln|\mathbf{x}\mathbf{x}^T+\mathbf{K}|+f_{1,1}(d,n)+(\frac{m+d+1}{2})\ln|\mathbf{Q}+(\mathbf{y}-\mathbf{M}\mathbf{x})\gamma(\mathbf{y}-\mathbf{M}\mathbf{x})^T|.$$

We note that the following lemma also holds:

Lemma 4.4

$$\int \mathcal{T}_{\mathbf{y}}(\mathbf{Q},n,\boldsymbol{\mu},\gamma)\ln|\mathbf{Q}+(\mathbf{y}-\boldsymbol{\mu})\gamma(\mathbf{y}-\boldsymbol{\mu})^T|d\mathbf{y}$$

is independent from both $\boldsymbol{\mu}$ *and* γ*,*

and thus we can compute the conditional entropy expressed in (7):

Lemma 4.5

$$H(\mathbf{A},\mathbf{V}|\mathbf{y};\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x})H(\mathbf{A},\mathbf{V};\mathbf{x},\mathbf{y})d\mathbf{y} = -\frac{d}{2}\ln|\mathbf{x}\mathbf{x}^T+\mathbf{K}|+f_{1,2}(\mathbf{Q},n,m),$$

where $f_{1,2}(\mathbf{Q},n,m)$ *depends only on* $\mathbf{Q}$*,* n *and* m*.*

Collecting all the terms, we arrive at the following *intriguingly simple* expression

$$\begin{aligned}\mathbf{u}_{t+1}^{opt} &= \arg\min_{\mathbf{u}_{t+1}} \int p\left(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t\right) H(\mathbf{A},\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t,\mathbf{y}_{t+1}) d\mathbf{y}_{t+1}, \\ &= \arg\min_{\mathbf{u}_{t+1}} -\frac{d}{2}\ln|\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t| = \arg\max_{\mathbf{u}_{t+1}} \mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}, \end{aligned} \tag{13}$$

where

$$\mathbf{x}_{t+1} \doteq [\mathbf{r}_{t-I};\ldots;\mathbf{r}_t;\mathbf{u}_{t-J+1};\ldots;\mathbf{u}_{t+1}],$$

and we used that $|\mathbf{x}\mathbf{x}^T+\mathbf{K}| = |\mathbf{K}|(1+\mathbf{x}^T\mathbf{K}^{-1}\mathbf{x})$ according to the Matrix Determinant Lemma (Harville, 1997). We assume a bounded domain $\mathcal{U}$ for the control, which is necessary to keep the maximization procedure of (13) finite. This is, however, a reasonable condition for all practical applications. So,

$$\mathbf{u}_{t+1}^{opt} = \arg\max_{\mathbf{u}_{t+1}\in\mathcal{U}} \mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}, \tag{14}$$

In what follows D-optimal control will be referred to as *'infomax interrogation scheme'*. The steps of our algorithm are summarized in Table 1.

Control Calculation
$\mathbf{u}_{t+1} = \arg\max_{\mathbf{u}\in\mathcal{U}} \hat{\mathbf{x}}_{t+1}^T\mathbf{K}_t^{-1}\hat{\mathbf{x}}_{t+1}$
where $\hat{\mathbf{x}}_{t+1} = [\mathbf{r}_{t-I};\ldots;\mathbf{r}_t;\mathbf{u}_{t-J+1};\ldots;\mathbf{u}_t;\mathbf{u}]$
set $\mathbf{x}_{t+1} = [\mathbf{r}_{t-I};\ldots;\mathbf{r}_t;\mathbf{u}_{t-J+1};\ldots;\mathbf{u}_t;\mathbf{u_{t+1}}]$
Observation
observe $\mathbf{r}_{t+1}$, and let $\mathbf{y}_{t+1} = g^{-1}(\mathbf{r}_{t+1})$
Bayesian update
$\mathbf{M}_{t+1} = (\mathbf{M}_t\mathbf{K}_t + \mathbf{y}_{t+1}\mathbf{x}_{t+1}^T)(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1}$
$\mathbf{K}_{t+1} = \mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t$
$n_{t+1} = n_t+1$
$\gamma_{t+1} = 1-\mathbf{x}_{t+1}^T(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1}\mathbf{x}_{t+1}$
$\mathbf{Q}_{t+1} = Q_t + (\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})\gamma_{t+1}(\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})^T$

Table 1: Pseudocode of the algorithm

Computation of the inverse $(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1}$ in Table 1 can be simplified considerably by the following recursion: let $\mathbf{P}_t = \mathbf{K}_t^{-1}$, then according to the Sherman-Morrison formula (Golub and Van Loan, 1996)

$$\mathbf{P}_{t+1} = (\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t)^{-1} = \mathbf{P}_t - \frac{\mathbf{P}_t\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{P}_t}{1+\mathbf{x}_{t+1}^T\mathbf{P}_t\mathbf{x}_{t+1}}. \tag{15}$$

In this expression matrix inversion disappears and only a real number is inverted instead.

5. A-Optimality Approach for Parameter Learning

The D-optimality principle aims to minimize the expected posteriori entropy of the parameters. A-optimality principle differs; it measures the uncertainty by means of the variance and not the entropy. Thus, instead of (7), the A-optimal objective function is as follows:

$$\mathbf{u}_{t+1}^{opt} = \arg\min_{\mathbf{u}_{t+1}} \int d\mathbf{y}_{t+1} P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t) trVar[\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}]. \tag{16}$$

where $Var[\boldsymbol{\theta}|\mathcal{F}]$ denotes the conditional covariance matrix of $\boldsymbol{\theta}$ given the condition $\mathcal{F}$.

To keep the calculations simple, in this case we use $\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha}_t,\boldsymbol{\beta}_t)$ prior distribution for the covariance matrix instead of $I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}_t,n_t)$. Using the notations of (8)-(10), the posterior distributions assume the following forms:

Lemma 5.1

$$\begin{aligned} P(\mathbf{A}|\mathbf{V},\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= \mathcal{N}_{\mathbf{A}}(\mathbf{M}_{t+1},\mathbf{V},\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T+\mathbf{K}_t), \qquad (17)\\ P(\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= \mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1}),\\ P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t}) &= \prod_{i=1}^{d}\mathcal{T}_{(\mathbf{y}_{t+1})_i}\left((\boldsymbol{\beta}_t)_i,2(\boldsymbol{\alpha}_t)_i,(\mathbf{M}_t\mathbf{x}_{t+1})_i,\frac{\gamma_{t+1}}{2}\right), \qquad (18)\end{aligned}$$

where we used the shorthands

$$\begin{aligned}(\boldsymbol{\alpha}_{t+1})_i &= (\boldsymbol{\alpha}_t)_i+1/2,\\ (\boldsymbol{\beta}_{t+1})_i &= (\boldsymbol{\beta}_t)_i+((\mathbf{y}_{t+1})_i-(\mathbf{M}_t\mathbf{x}_{t+1})_i)^2\frac{\gamma_{t+1}}{2}. \qquad (19)\end{aligned}$$

The proof can be found in the Appendix.

Given that $P(\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1})$ belongs to the $\mathcal{PIG}$ family we can calculate the quantity $Var(\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1})$ (Gelman et al., 2003):

$$tr\left(Var[\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}]\right) = \sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{((\boldsymbol{\alpha}_{t+1})_i-1)^2((\boldsymbol{\alpha}_{t+1})_i-2)}.$$

We will need the following lemma:

Lemma 5.2

$$\begin{aligned} tr\left(Var[\mathbf{A}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}]\right) &= tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}\right)E[tr\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}],\\ &= tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}\right)\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1}.\end{aligned}$$

The proof can be found in the Appendix.

Now we can elaborate on the A-optimal cost function for parameter estimation (16):

$$\begin{aligned} &\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})trVar[\boldsymbol{\theta}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}] = \qquad (20)\\ =\ &\int d\mathbf{y}_{t+1}\prod_{i=1}^{d}\mathcal{T}_{(\mathbf{y}_{t+1})_i}\left((\boldsymbol{\beta}_t)_i,2(\boldsymbol{\alpha}_t)_i,(\mathbf{M}_t\mathbf{x}_{t+1})_i,\frac{\gamma_{t+1}}{2}\right)\times\\ &\times\left(tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}\right)\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1}+\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{((\boldsymbol{\alpha}_{t+1})_i-1)^2((\boldsymbol{\alpha}_{t+1})_i-2)}\right),\\ =\ &tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}\right)f_{2,1}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_t)+f_{2,2}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_t),\end{aligned}$$

where $f_{2,1}$ and $f_{2,2}$ depend only on α_{t+1}, and β_t. Here we used (18), (19) and Lemma 4.4.

Applying again the Sherman-Morrison formula (15) and the fact that $tr[\mathbf{K}_t^{-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}] = tr[\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}]$, we arrive at the following expression for A-optimal parameter estimation:

$$\mathbf{u}_{t+1}^{opt} = \arg\max_{\mathbf{u}_{t+1}\in\mathcal{U}} \frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}, \tag{21}$$

which is a hyperbolic programming task.

We can conclude that while in the D-optimality case the task is to minimize expression $|(\mathbf{K}_t + \mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}|$, the A-optimality principle is concerned with the minimization of $tr[(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}]$.

6. D-Optimality Approach for Noise Estimation

One might wish to compute the optimal control for estimating noise $\mathbf{e}_t$ in (1), instead of the identification problem above. Based on (1) and because

$$\mathbf{e}_{t+1} = \mathbf{y}_{t+1} - \sum_{i=0}^{I}\mathbf{F}_i\mathbf{r}_{t-i} - \sum_{j=0}^{J}\mathbf{B}_j\mathbf{u}_{t+1-j}, \tag{22}$$

one might think that the best strategy is to use the optimal infomax control of Table 1, since it provides good estimations for parameters $\mathbf{A} = [\mathbf{F}_I,\ldots,\mathbf{F}_0,\mathbf{B}_J,\ldots,\mathbf{B}_0]$ and so for noise $\mathbf{e}_t$.

Another—and different—thought is the following. At time $t+1$, let the estimation of the noise be $\hat{\mathbf{e}}_{t+1} = \mathbf{y}_{t+1} - \sum_{i=0}^{I}\hat{\mathbf{F}}_i^t\mathbf{r}_{t-i} - \sum_{j=0}^{J}\hat{\mathbf{B}}_j^t\mathbf{u}_{t+1-j}$, where $\hat{\mathbf{F}}_i^t$ (i=0,...,I), and $\hat{\mathbf{B}}_j^t$ (j=0,...,J) denote the estimations of $\mathbf{F}$ and $\mathbf{B}$ respectively.

Using (22), we have that

$$\mathbf{e}_{t+1} - \hat{\mathbf{e}}_{t+1} = \sum_{i=0}^{I}(\mathbf{F}_i - \hat{\mathbf{F}}_i^t)\mathbf{r}_{t-i} + \sum_{j=0}^{J}(\mathbf{B}_j - \hat{\mathbf{B}}_j^t)\mathbf{u}_{t+1-j}. \tag{23}$$

This hints that the control should be $\mathbf{u}_t = \mathbf{0}$ for all times in order to get rid of the error contribution of matrix $\mathbf{B}_j$ in (23).

Straightforward D-optimality considerations oppose the utilization of objective (6) for the present task. One can optimize, instead, the following quantity:

$$\arg\max_{\mathbf{u}_{t+1}} I(\mathbf{e}_{t+1},\mathbf{y}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t).$$

In other words, for the estimation of the noise we want to design a control signal $\mathbf{u}_{t+1}$ such that the next output is the best from the point of view of greedy optimization of mutual information between the next output $\mathbf{y}_{t+1}$ and the noise $\mathbf{e}_{t+1}$. It is easy to show that this task is equivalent to the following optimization problem:

$$\arg\min_{\mathbf{u}_{t+1}} \int d\mathbf{y}_{t+1} P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t) H(\mathbf{e}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}), \tag{24}$$

where $H(\mathbf{e}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) = H(\mathbf{A}\mathbf{x}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1})$, because $\mathbf{e}_{t+1} = \mathbf{y}_{t+1} - \mathbf{A}\mathbf{x}_{t+1}$.

In practice, we perform this optimization in an appropriate domain $\mathcal{U}$. After some mathematical calculation we can prove that the D-optimal interrogation scheme for noise estimation gives rise to the following control:

Lemma 6.1

$$\mathbf{u}_{t+1}^{opt} = \arg \min_{\mathbf{u}_{t+1} \in \mathcal{U}} \mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}. \tag{25}$$

The proof of this lemma can be found in the Appendix.

It is worth noting that this D-optimal cost function for noise estimation and the D-optimal cost function derived for parameter estimation in (13) are not compatible with each other. Estimating one of them quickly will necessarily delay the estimation of the other.

We shall show later (Section 9) that for large t values, expression (25) gives rise to control values close to $\mathbf{u}_t = \mathbf{0}$.

7. A-Optimality Approach for Noise Estimation

Instead of (24), our task is to compute the following quantity:

$$\arg\min_{\mathbf{u}_{t+1}} \int d\mathbf{y}_{t+1} P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t)\, tr\left(Var[\mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}]\right). \tag{26}$$

We will apply the identity

$$\mathcal{N}_{\mathbf{A}\mathbf{x}_{t+1}}\left(\mathbf{M}_{t+1}\mathbf{x}_{t+1}, \mathbf{V}, \left(\mathbf{x}_{t+1}^T \mathbf{K}_{t+1}^{-1} \mathbf{x}_{t+1}\right)^{-1}\right) \mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1}) =$$

$$= \prod_{i=1}^{d} \mathcal{T}_{(\mathbf{A}\mathbf{x}_{t+1})_i}\left((\boldsymbol{\beta}_{t+1})_i, 2(\boldsymbol{\alpha}_{t+1})_i, (\mathbf{M}_{t+1}\mathbf{x}_{t+1})_i, \left(\mathbf{x}_{t+1}^T \frac{\mathbf{K}_{t+1}^{-1}}{2} \mathbf{x}_{t+1}\right)^{-1}\right) \times$$

$$\times \mathcal{PIG}_{\mathbf{V}}\left(\boldsymbol{\alpha}_{t+1} + 1, \boldsymbol{\beta}_{t+1} + diag[(\mathbf{y}_{t+1} - \mathbf{M}_t\mathbf{x}_{t+1}) \frac{\gamma_{t+1}}{2} (\mathbf{y}_{t+1} - \mathbf{M}_t\mathbf{x}_{t+1})^T]\right),$$

which can be proven by using Lemma A.1. We can simplify (26) by noting that

$$tr\left(Var[\mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}]\right) = tr\left(Var[\mathbf{A}\mathbf{x}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}]\right).$$

We also take advantage of the fact that

$$Var_{\mathbf{V}}[E[\mathbf{A}\mathbf{x}_{t+1}|\mathbf{V}, \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}]] = Var_{\mathbf{V}}[\mathbf{M}_{t+1}\mathbf{x}_{t+1}] = 0,$$

and proceed as

$$\begin{aligned} E_{\mathbf{V}}[trVar(\mathbf{A}\mathbf{x}_{t+1}|\mathbf{V}, \{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1})] &= E_{\mathbf{V}}[tr(\mathbf{V} \otimes \mathbf{x}_{t+1}^T \mathbf{K}_{t+1}^{-1} \mathbf{x}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}], \\ &= tr(E[\mathbf{V}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}])\mathbf{x}_{t+1}^T \mathbf{K}_{t+1}^{-1} \mathbf{x}_{t+1}, \\ &= \mathbf{x}_{t+1}^T \mathbf{K}_{t+1}^{-1} \mathbf{x}_{t+1} \sum_{i=1}^{d} \frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i - 1}, \end{aligned}$$

where $\otimes$ denotes the Kronecker product. The law of total variance says that

$$Var[\mathbf{A}\mathbf{x}] = Var[E[\mathbf{A}\mathbf{x}|\mathbf{V}]] + E[Var[\mathbf{A}\mathbf{x}|\mathbf{V}]],$$

and hence

$$tr Var[\mathbf{A}\mathbf{x}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}] = \mathbf{x}_{t+1}^T \mathbf{K}_{t+1}^{-1} \mathbf{x}_{t+1} \sum_{i=1}^{d} \frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i - 1}.$$

There is another way to arrive at the same result. One can apply (37) with Lemma A.1 and use the fact that the covariance matrix of a $\mathcal{T}_{\mathbf{x}}(\beta,\alpha,\boldsymbol{\mu},\mathbf{K})$ distributed variable is $\frac{\beta\mathbf{K}^{-1}}{\alpha-2}$ (Gelman et al., 2003). That is, we have

$$trVar[\mathbf{A}\mathbf{x}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}] = \sum_{i=1}^{d}\left(\frac{(\boldsymbol{\beta}_{t+1})_i\left(\mathbf{x}_{t+1}^T\frac{\mathbf{K}_{t+1}^{-1}}{2}\mathbf{x}_{t+1}\right)}{2(\boldsymbol{\alpha}_{t+1})_i-2}\right).$$

Now, we can proceed as

$$\begin{aligned}\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})tr\left(Var[\mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}]\right) &= \qquad (27)\\ \int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})tr(E[\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}])\mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1} &= \\ \mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1} &= \\ \mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\sum_{i=1}^{d}f_4((\boldsymbol{\alpha}_{t+1})_i,(\boldsymbol{\beta}_t)_i),\end{aligned}$$

where we used (18), (19) and Lemma 4.4 again. Applying the Sherman-Morrison formula one can see that the task is the same as in (25).

8. Joint Parameter and Noise Estimation

So far we wanted to optimize the control in order to speed-up learning of either the parameters of the dynamics or the noise. In this section we investigate the A- and D-optimality principles for the joint parameter and noise estimation task.

8.1 A-optimality

According to the A-optimality principle, the joined objective for parameter and noise estimation is given as:

$$\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})\,trVar[vec(\mathbf{A}),diag(\mathbf{V}),\mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}].$$

By means of (20), (27) and Lemma 5.2, it is equivalent to:

$$\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t})E[tr\mathbf{V}|\mathbf{x}_1^{t+1},\mathbf{y}_1^{t+1}]\,tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}+\mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\right).$$

From here, one can prove the following lemma in a few steps:

Lemma 8.1 *The A-optimality principle in the joined parameter and noise estimation task gives rise to the following choice for control:*

$$\mathbf{u}_{t+1}^{opt} = \arg\max_{\mathbf{u}_{t+1}\in\mathcal{U}}\frac{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}. \qquad (28)$$

The proof can be found in the Appendix.

Thus, the task is a hyperbolic programming task, similar to (21).

8.2 D-optimality

One of the most salient differences between A-optimality and D-optimality is that for D-optimality we have

$$H(X,Y) = H(X|Y) + H(Y),$$

however, for A-optimality the corresponding equation does not hold in general, because:

$$trVar(X,Y) \neq E_Y[trVar(X|Y)] + trVar(Y).$$

An implication—as we shall see below—is that we cannot use the D-optimality principle for the joint parameter and noise estimation task. For D-optimality our cost function would be

$$\arg\min_{\mathbf{u}_{t+1}} \int d\mathbf{y}_{t+1} P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^t) H(\mathbf{A}, \mathbf{V}, \mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}),$$

but the following equality holds:

$$H(\mathbf{A}, \mathbf{V}, \mathbf{e}_{t+1}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) = H(\mathbf{A}, \mathbf{V}|\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) + H(\mathbf{e}_{t+1}|\mathbf{A}, \mathbf{V}\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}),$$

and since $\mathbf{e}_{t+1} = \mathbf{y}_{t+1} - \mathbf{A}\mathbf{x}_{t+1}$, therefore the last term $H(\mathbf{e}_{t+1}|\mathbf{A}, \mathbf{V}\{\mathbf{x}\}_1^{t+1}, \{\mathbf{y}\}_1^{t+1}) = -\infty$. The first term is a finite real number, thus we can conclude that the D-optimality cost function is constant $-\infty$, and therefore the D-optimality principle does not suit the joint parameter and noise estimation task.

9. Non-myopic Optimization

Until now, we considered myopic methods for the optimization of control, that is, we aimed to determine the optimum of the objective only for the next step. In this section, we show a non-myopic heuristics for the noise estimation task (25).

The optimization of the derived objective function, $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$, is simple, provided that $\mathbf{K}_t$ is fixed during the optimization of $\mathbf{u}_{t+1}$. If so, then the optimization task is quadratic. To see this, let us partition matrix $\mathbf{K}_t$ as follows:

$$\mathbf{K}_t = \begin{pmatrix} \mathbf{K}_t^{11} & \mathbf{K}_t^{12} \\ \mathbf{K}_t^{21} & \mathbf{K}_t^{22} \end{pmatrix},$$

where $\mathbf{K}_t^{11} \in \mathbb{R}^{d \times d}$,$\mathbf{K}_t^{21} \in \mathbb{R}^{m-d \times d}$, $\mathbf{K}_t^{22} \in \mathbb{R}^{m-d \times m-d}$. It is easy to see that if domain $\mathcal{U}$ in (25) is large enough then

$$\mathbf{u}_{t+1}^{opt} = (\mathbf{K}_t^{22})^{-1} \mathbf{K}_t^{21} \mathbf{r}_t. \tag{29}$$

It occurs, however, that the objective $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$ may be improved by considering multiple-step lookaheads. In this case matrix $\mathbf{K}_t$ can be subject to changes in $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$, because it depends on previous control inputs $\mathbf{u}_1, \ldots, \mathbf{u}_t$ derived from previous optimization steps.

We propose a two-step heuristics for the long-term minimization of expression $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$. During the first τ-step long stage, we focus only on the minimization of the quantity $|\mathbf{K}_t^{-1}|$. Then, if this quantity $|\mathbf{K}_t^{-1}|$ becomes small, we start the second stage: we consider $\mathbf{K}_t^{-1}$ as given and search for control that minimizes quantity $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$, so we now apply the rule of (29). Thus,

this method 'sacrifices' the first τ steps in order to achieve smaller costs later; this heuristic optimization is non-myopic. More formally, we use the strategy of Table 1 in the first τ steps in order to decrease quantity $|\mathbf{K}_t^{-1}|$ fast. Then after this τ-steps, we switch to the control method of (29). This will decrease the cost function (25) further. We will call this non-greedy interrogation heuristics introduced for noise estimation 'τ-*infomax noise interrogation*'. This non-myopic heuristics admits that parameter estimation of the dynamics is the prerequisite of noise estimation, because improper parameter estimation makes apparent noise, and thus the heuristics sacrifices τ steps for parameter estimation.

In Section 10 we will empirically show that using this non-myopic strategy, after τ steps we can achieve smaller cost values in (25)—as well as better performance in parameter estimation—than using the greedy competitors. The compromise is that in the first τ steps the performance of the non-myopic control can be worse than that of the other control methods.

We note that in the τ-infomax noise interrogation, for large switching time τ and for large t values, $|\mathbf{K}_t^{22}|$ will be large, and hence—according to (29)—the optimal $\mathbf{u}_t$ for interrogation will be close to $\mathbf{0}$. (In Section 10 we will show this empirically.) A reasonable approximation of the 'τ-infomax noise interrogation' is to use the control given in Table 1 for τ steps and to switch to *zero-interrogation* onwards. This scheme will be called the 'τ-*zero interrogation*' scheme.

10. Numerical Illustrations

We illustrate by numerical simulations the power of A- and D-optimizations.

10.1 Generated Data

This section provides numerical experiments for parameter and noise estimations on artificially generated toy problems.

10.1.1 PARAMETER ESTIMATION

We investigated the parameter estimation capability of the D- and A-optimal interrogation. Matrix $\mathbf{F} \in \mathbb{R}^{d\times d}$ has been generated as a random orthogonal matrix multiplied by 0.9 so that the magnitudes of its eigenvalues remained below 1. Random matrix $\mathbf{B} \in \mathbb{R}^{d\times c}$ was generated from standard normal distribution. Elements of the diagonal covariance matrix $\mathbf{V}$ of noise $\mathbf{e}_t$ were generated from the uniform distribution over $[0,1]$. The process is stable under these conditions.

To study whether or not the D- and A-optimal interrogations are able to estimate the true parameters we measured the averages of the squared deviations of the true matrices $\mathbf{F}$ and $\mathbf{B}$ and the means of their posterior estimations, respectively. The square roots of these estimations are the mean squared errors (MSE). One might use other options to measure performance. For example, L_2 norm could be replaced by the L_1 norm and the variance of the posterior estimations could also be added as the complementary information for the bias.

We examined the following strategies: (i) D-optimal control of Table 1 with $\mathcal{U} = [-\delta,\delta]^c$, which defines a c-dimensional hypercube. The value of δ was set to 50. (ii) A-optimal control of (21) with the same $\mathcal{U}$, (iii) zero control: $\mathbf{u}_t = \mathbf{0} \in \mathbb{R}^c\ \forall t$, (iv) random control: $u_t \in [-\delta,\delta]^c$ generated randomly from the uniform distribution in the c-dimensional hypercube, (v) control defined by (25) for noise estimation, called 'noise control', (vi) 25-zero control and (vii) 75-zero control defined in Section 9.

For solving the quadratic problem of (14) and (25) we used a subspace trust-region procedure, which is based on the interior-reflective Newton method described by Coleman and Li (1996). Its implementation is available in the Matlab Optimization toolbox. However, the optimization task in (21) is more involved: Generally, the optimization of a constrained hyperbolic programming task is quite difficult. We tried the gradient ascent method, but its convergence appeared to be very slow and we got poor results. In this case, it was more efficient to apply a simplex method as follows: we know that the optimal solution of (21) lies at the boundary of $\mathcal{U}$. Thus, we chose one corner of hypercube $\mathcal{U}$ randomly with uniform distribution and moved greedily to the neighboring corners with the best improvement in the objective. This procedure was iterated until convergence. The method was efficient for our special simple optimization domain.

We investigated two distinct cases. In the first case we set $d = 10 < c = 40$; the dimension of the observations is smaller than the dimension of the control. By contrast, in the other case we set $d = 40 > c = 10$. Results are shown in Fig. 1 (a-b) and in Fig. 2 (a-b). We separated the MSE values of matrices **B** and **F**. According to the figure, zero control may give rise to early and sudden drops in the MSE values of matrix **F**. Not surprisingly, however, zero control is unable to estimate matrix **B**. For both types of matrices as well as for $d < c$ and for $d > c$, the D-optimal procedure produced the smallest MSE after about 50 online estimations, but the A-optimal method reached very similar levels only a few iterations later. As can be expected, τ-zero control, which is identical to D-optimal control in the first τ steps fell behind D-optimal control after τ since it changes the objective and estimates the noise and not the parameters afterwards.

For statistical significance studies, we introduced the concept of average correlation curves. We use Fig. 1 to explain this concept. There are 7 curves in Fig. 1 each representing the averages of 25 computer runs. Error bars make the curves incomprehensible and they hide the correlations that may be present between the errors. We note that the relative order of the curves is of interest for us. However, it is possible that in each run the relative order of the curves was the same and the overlap of the error bars—which originates from the large differences between the individual runs—hides this important piece of information. We treat this problem as follows. In each time instant $1 \leq t \leq 250$ and for all $1 \leq i < j \leq 25$ we compute the empirical (linear, or rank) correlation of the 7 curves of the i^{th} and j^{th} experiment and take the average of the $25 \times 24/2 = 300$ values. The most significant case gives rise to 1 for each of the 300 correlations, that is, the 25 experiments agree in the height of the curves at that time instant, or in their relative orderings for the case of rank correlation. If there is any single experiment out of 25 that produces different heights or orders then the average correlation becomes smaller than 1. For randomly chosen curves the average correlation is 0. Results can be seen in Fig. 1 (c-d) and Fig. 2 (c-d) for linear Pearson and for Kendal rank correlations, respectively. The curves demonstrate that after about 50 steps, the correlations, in particular the linear correlation is almost 1. This means that the curves behaved similarly in a considerable portion of the experiments. The slightly different picture shown by the linear correlation and the rank correlation could be due to the fact that the performance of the A and D-optimal control is very similar after some time, and their ordering may change often, thus giving rise to changes in the ranks in different experiments.

10.1.2 NOISE ESTIMATIONS

In Section 7 we showed that A- and D-optimality principles result in the same cost function.

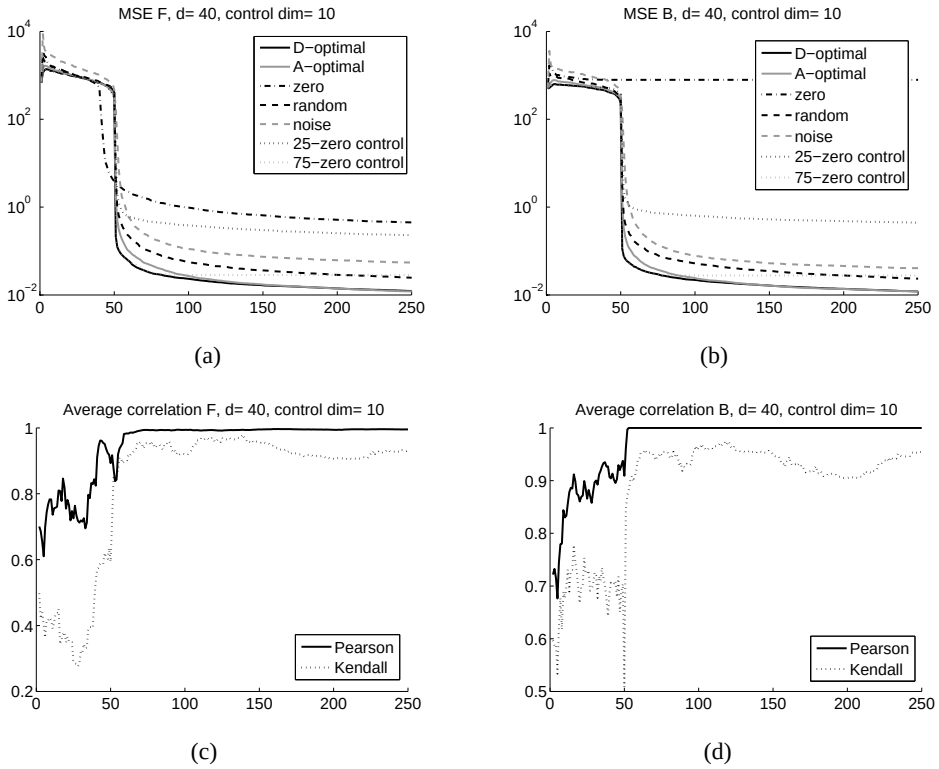


Figure 1: Mean Square Error of the estimated parameters for different control strategies and the significance of the curves. Magnitude of MSE as a function of time is averaged for 25 runs. Dimension of the control is 10. $\mathbf{F} \in \mathbb{R}^{40\times 40}$, $\mathbf{B} \in \mathbb{R}^{40\times 10}$. (a): MSE of the estimated matrix $\hat{\mathbf{F}}$. (b): MSE of the estimated matrix $\hat{\mathbf{B}}$. (c): The average correlation curves for the estimation of $\mathbf{F}$. (d): The average correlation curves for the estimation of $\mathbf{B}$. For details see the text.

We investigated the noise estimation capability of the interrogation in (25) for four cases. The first set of experiments illustrates that the estimation of driving noise $\mathbf{e}_t$ for large τ values barely differs if we replace the τ-*infomax noise interrogation* with the τ-*zero interrogation* scheme. Parameters were the same as above and the MSE of the noise estimation was computed. Results are shown in Fig. 3: for the case of $\tau = 21$, cost function (25) of the τ-zero interrogation is higher than that of τ-infomax interrogation. However, for values $\tau = 51$ and 81 the performances of the two schemes are approximately identical. Given that τ-zero and τ-infomax noise interrogation behave similarly for large τ values, we compare the τ-zero interrogation scheme with other schemes in our numerical experiments.

In the second experiment we investigated the problem of noise estimation on a toy problem. Parameters were set as in Section 10.1.1, and the following strategies were compared: zero control, infomax control, random control and τ-zero control for different τ values. Results are shown in Fig. 4. It is clear from the figure that neither the zero control, nor the infomax (D-optimal) control of Table 1 work for this case. If we want to have minimal MSE in approximately τ steps then the best strategy is to apply the τ-zero strategy, that is, the strategy of Table 1 up to τ steps and then to switch to zero control. Note, however, that parameter estimation requires to keep control values non-negligible forever (Table 1).

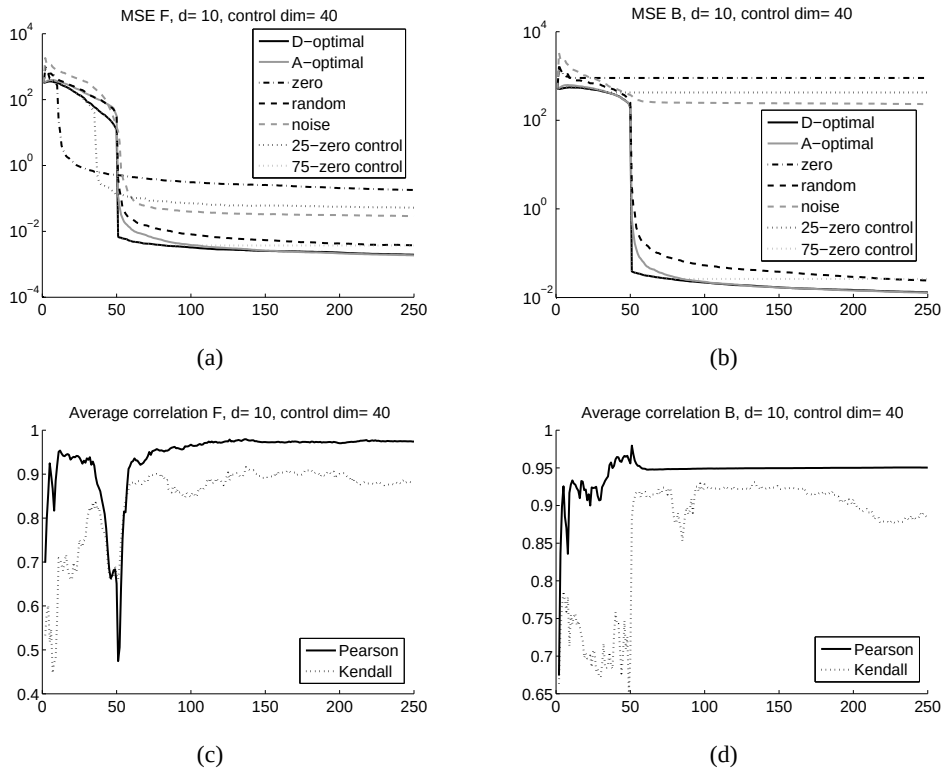
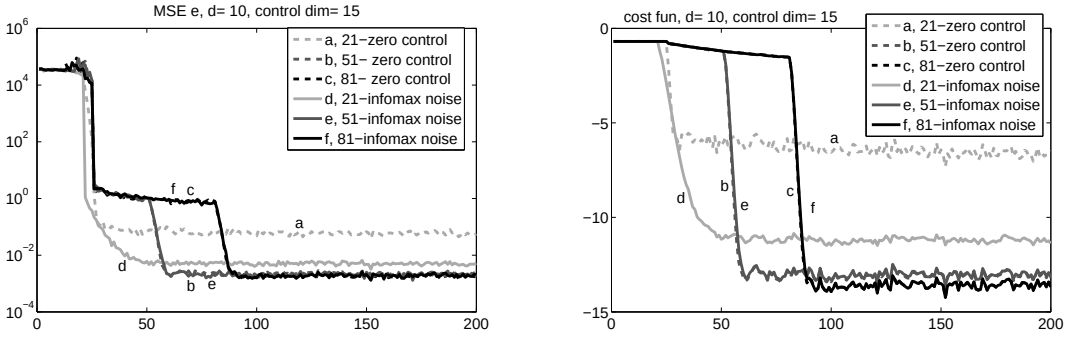


Figure 2: Mean Square Error of the estimated parameters for different control strategies and the significance of the curves. Magnitude of MSE as a function of time is averaged for 25 runs. Dimension of the control is 40. $\mathbf{F} \in \mathbb{R}^{10\times 10}$, $\mathbf{B} \in \mathbb{R}^{10\times 40}$. (a): MSE of the estimated matrix $\hat{\mathbf{F}}$. (b): MSE of the estimated matrix $\hat{\mathbf{B}}$. (c): The average correlation curves for the estimation of $\mathbf{F}$. (d): The average correlation curves for the estimation of $\mathbf{B}$. For details see the text.

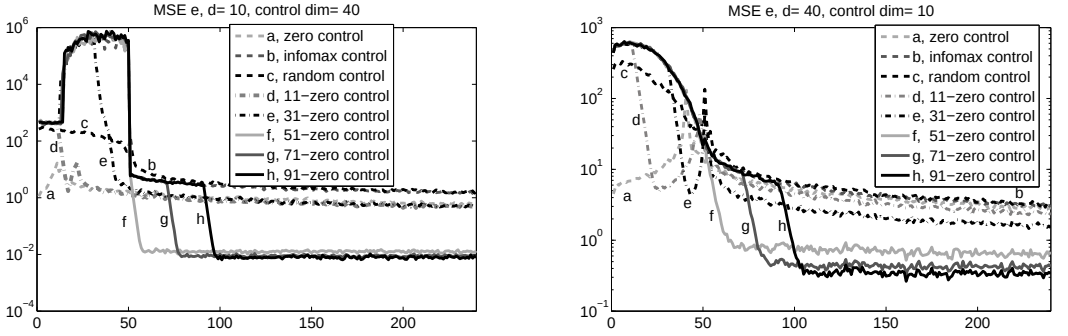
In the third experiment we used numerical tools to support our the arguments we made in Section 9. We investigate the D-optimal, the zero, the random, and the greedy noise control of (25), as well as the 71-zero and 101-zero controls. Results show that if we may sacrifice the first τ steps, then the non-myopic $\tau-$*zero* control gives rise to the smallest MSE for the estimated noise and the smallest values for the cost function (25) *after* τ steps considering all studied control methods. Figure 5a shows the MSE of the estimated driving noise, whereas Fig. 5b depicts the cost $\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}$. Figure 5c is about the time dependence of $\log|\mathbf{K}_t|$ that supports our argument in Section 9, namely, it may be worth to sacrifice steps at the beginning to quickly decrease $|\mathbf{K}_t^{-1}|$ (i.e., to decrease $\log|\mathbf{K}_t|$) in order to estimate (25) efficiently later. The problem we studied was the same as before, except that $d = 25$ and $c = 25$ were applied.

The fourth experiment illustrates the efficiency of the approximation of the noise for the case when our assumptions on $\mathbf{e}_t$ are not fulfilled. Here noise $\mathbf{e}_t$ was neither Gaussian nor i.i.d. 'Noise' $\mathbf{e}_t$ was chosen as equidistant points smoothly 'walking' along a 3 dimensional spiral curve as a function of time (Fig. 6a). Dimensions of observation and control were 3 and 15, respectively. Results are shown in Fig. 6. Neither random control, nor infomax interrogation of Table 1 (Fig. 6c), nor zero control (Fig. 6d) could produce reasonable estimation. However, the τ-zero interrogation scheme



(a) (b)

Figure 3: Comparing τ-infomax noise and τ-zero interrogations. The curves are averaged for 50 runs. Dimension of the control is 15 and the dimension of the observation is 10. (a): MSE of the estimated noise (b): Cost function as given in (25). 'τ-infomax noise' (τ-zero) means that up to step number τ strategy of Table 1 applies and then the control of Eq. (29) is followed.



(a) (b)

Figure 4: Mean Square Error of the estimated noise for different control strategies. Magnitude of MSE as a function of time is averaged for 20 runs. (a): Dimension of the control is 40. $\mathbf{F} \in \mathbb{R}^{40\times 40}$, $\mathbf{B} \in \mathbb{R}^{40\times 10}$. (b): Dimension of the control is 10. $\mathbf{F} \in \mathbb{R}^{10\times 10}$, $\mathbf{B} \in \mathbb{R}^{10\times 40}$. '$\tau$-zero' means that up to step number τ the strategy illustrated in Table 1 was applied and then zero control followed.

produced a good approximation for large enough τ values (Fig. 6e). Details of this illustration are shown in Fig. 6f.

10.1.3 JOINT PARAMETER AND NOISE ESTIMATIONS

In Section 8 we showed that the objective of the D-optimality principle is constant for the joined parameter and noise estimation task. However, A-optimality principle provides sensible cost function (Eq. (28)). Unfortunately, it leads to a hyperbolic programming task. This optimization is hard in most cases. One can estimate the complexity of the objective by inspecting lower dimensional cases. We show the negative logarithm of (28) for the 2 dimensional case for different $\mathbf{K}$ matrices (Fig. 7). In one of the cases the null vector corresponds to the minimum, whereas in the other case the minimum is at a boundary point of the optimization domain. Also, the cost functions appear to be flat in a large part of their domains, rendering gradient based methods ineffective.

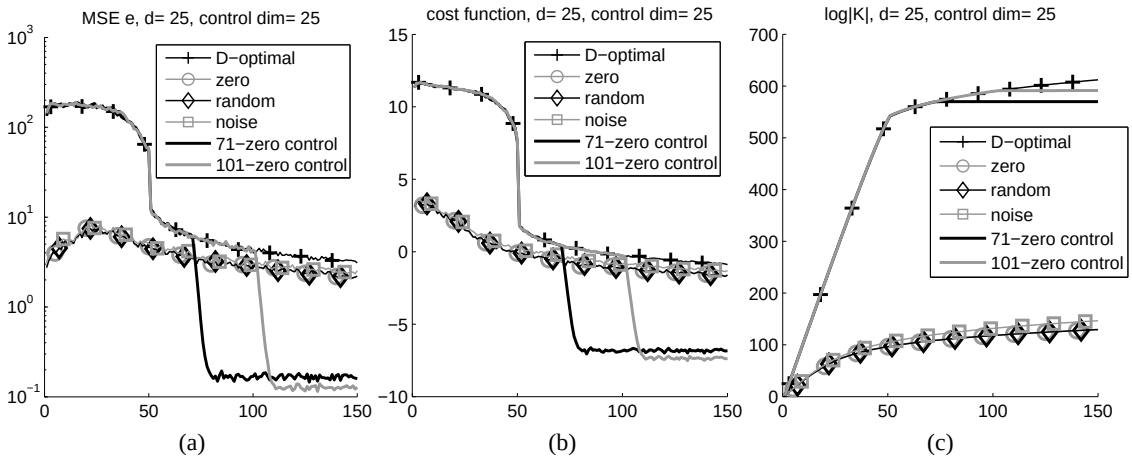


Figure 5: Empirical study on non-myopic controls for noise estimation. We sacrifice the first τ steps to achieve better MSE and smaller cost function. The curves are averaged over 25 runs. The dimension of the control and the dimension of the observation is 25. (a): MSE of noise estimation for different control strategies. (b): $\mathbf{x}_{t+1}^T \mathbf{K}_t^{-1} \mathbf{x}_{t+1}$ cost function for different control strategies. (c): $\log |\mathbf{K}_t|$ function for different control strategies.

Studies were conducted on the problem family of 10.1.1 for observation dimension $d = 15$ and control dimension $c = 30$. For the optimization, we modified the simplex method that we used for the hyperbolic task before. The single difference is that upon convergence, the best value was compared with the value of the objective at the 0 point and we chose the better one for control. We have compared this strategy with the parameter estimation strategy of the D-optimality principle, with zero control strategy, with random control strategy, and with $\tau - zero$ control for several τ values. Results are shown in Fig. 8. The figure indicates that control derived from the A-optimality principle (28) provides superior MSE results at approximately 45 iterations and then onwards compared to the other *myopic* techniques, however its performance was slightly worse than the *non-myopic* τ-zero control for τ values larger than an appropriate threshold.

Inspecting the optimal control series of the winner, we found that the algorithm chooses control values from the boundaries of the hypercube in the first 45 or so iterations. Then up to about 130 iterations it is switching between zero control and controls on the boundaries, but eventually it uses zero controls only. That is, the A-optimality principle is able to detect the need for the switch from high control values (to determine the parameters of the dynamics) to zero control values for noise estimation. This automated switching behavior is a special advantage of the A-optimality principle.

10.2 Controlled Independent Component Analysis

In this section we study the usefulness of our methods for auto-regressive (AR) hidden processes with independent driving sources and we would like to find the independent causes that drive the processes. This task belongs to independent component analysis (ICA) (Jutten and Hérault, 1991; Comon, 1994; Hyvärinen et al., 2001). Informally, we assume that our sources are doubly covered: they are the driving noise processes of AR processes which can not be directly observed due to the mixing with an unknown matrix. We will study our methods for this particular example and we

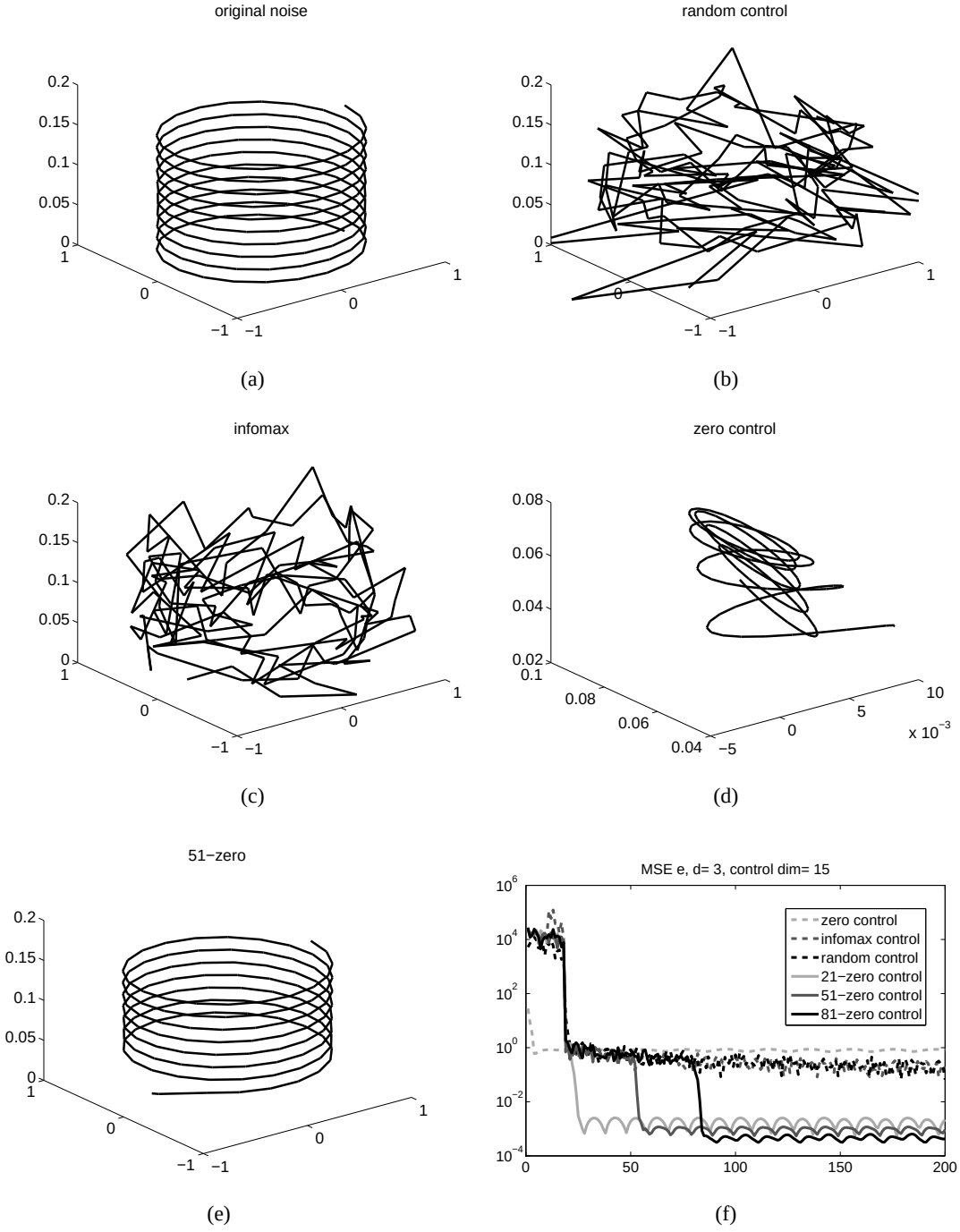
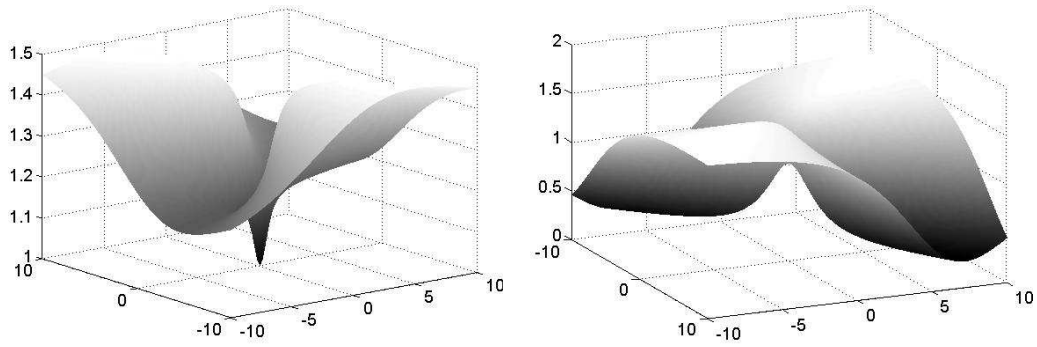


Figure 6: Different control strategies for non i.i.d. noise. (a): original noise. (b-e): estimated noise using random, infomax, zero, 51-zero strategy, respectively. (f): MSE for different control strategies. In (b-e), estimations of the first 51 time steps are not shown.



(a) (b)

Figure 7: Negative logarithm of the objective function for the joint parameter and noise estimation task for different **K** matrices. (a) the minimum point is in zero, (b) the minimum point is on the boundary.

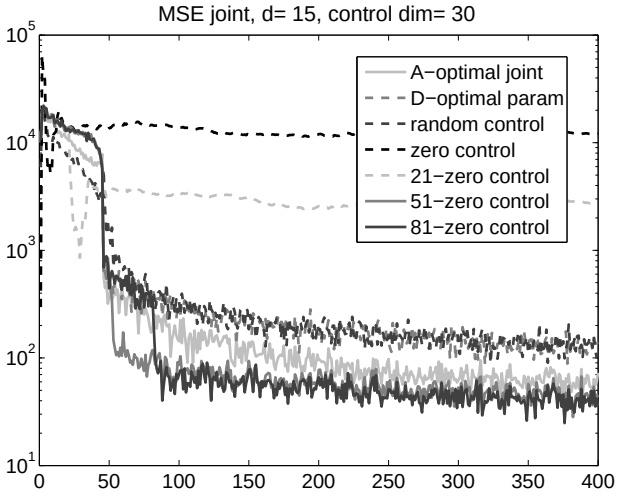


Figure 8: MSE of the joint parameter and noise estimation task. Comparisons between joint parameter and noise estimation using A-optimality principle, parameter estimation using D-optimality principle, random control, zero control and $\tau - zero$ control for different τ values. MSE values are averaged for 20 experiments.

.

assume that the processes can be exogenously controlled. Such processes are called ARX processes where X stands for letter x of the word eXogenous.

The 'classical' ICA task is as follows: we are given temporally i.i.d. signals $\mathbf{e}_t \in \mathbb{R}^d$ $(t = 1,2,\ldots,T)$ with statistically independent coordinates. We are unable to measure them directly, but their mixture $\mathbf{r}_t = \mathbf{C}\mathbf{e}_t$ is available for observation, where $\mathbf{C} \in \mathbb{R}^{d\times d}$ is an unknown invertible matrix. The task is to measure the observable signals and to estimate both mixing matrix $\mathbf{C}$ and sources $\mathbf{e}_t$.

There are several generalizations of this problem. Hyvärinen (1998) has introduced an algorithm to solve the ICA task even if the hidden sources are AR processes, whereas Szabó and Lőrincz (2008) generalized this problem for ARX processes in the following way: Processes $\tilde{\mathbf{e}}_t \in \mathbb{R}^d$ are given and they are statistically independent for the different coordinates and are temporally i.i.d signals. They generate ARX process $\mathbf{s}_t$ by means of parameters $\tilde{\mathbf{F}} \in \mathbb{R}^{d\times d}, \tilde{\mathbf{B}} \in \mathbb{R}^{d\times c}$:

$$\mathbf{s}_{t+1} = \sum_{i=0}^{I} \tilde{\mathbf{F}}_i \mathbf{s}_{t-i} + \sum_{j=0}^{J} \tilde{\mathbf{B}}_j \mathbf{u}_{t+1-j} + \tilde{\mathbf{e}}_{t+1}. \tag{30}$$

We assume that ARX process $\mathbf{s}_t$ can not be observed directly, but its mixture

$$\mathbf{r}_t = \mathbf{C}\mathbf{s}_t \tag{31}$$

is observable, where mixing matrix $\mathbf{C} \in \mathbb{R}^{d\times d}$ is invertible, *but unknown*. Our task is to estimate the original independent processes also called sources, noises or 'causes', that is, $\tilde{\mathbf{e}}_t$, the hidden process $\mathbf{s}_t$ and mixing matrix $\mathbf{C}$ from observations $\mathbf{r}_t$. It is easy to see that (30) and (31) can be rewritten into the following form

$$\mathbf{r}_{t+1} = \sum_{i=0}^{I} \mathbf{C}\tilde{\mathbf{F}}_i\mathbf{C}^{-1}\mathbf{r}_{t-i} + \sum_{j=0}^{J} \mathbf{C}\tilde{\mathbf{B}}\mathbf{u}_{t+1-j} + \mathbf{C}\tilde{\mathbf{e}}_{t+1}. \tag{32}$$

Using notations $\mathbf{F}_i = \mathbf{C}\tilde{\mathbf{F}}_i\mathbf{C}^{-1}$, $\mathbf{B}_j = \mathbf{C}\tilde{\mathbf{B}}_j$, $\mathbf{e}_{t+1} = \mathbf{C}\tilde{\mathbf{e}}_{t+1}$, (32) takes the form of the model (1) that we are studying with function g being the identity matrix. The only difference is that in ICA tasks $\mathbf{e}_t$ is assumed to be non-Gaussian, whereas in our derivations we always used the Gaussian assumption. In our studies, however, we found that the different control methods can be useful for non-Gaussian noise, too. Furthermore, the Central Limit Theorem says that the mixture of the variables $\tilde{\mathbf{e}}_t$, that is, $\mathbf{C}\tilde{\mathbf{e}}_t$ approximates Gaussian distributions, provided that the number of mixed variables is large enough.

In our numerical experiments we studied the following special case:

$$\mathbf{r}_{t+1} \quad = \quad \mathbf{F}\mathbf{r}_t + \mathbf{B}\mathbf{u}_{t+1} + \mathbf{C}\mathbf{e}_{t+1},$$

where the dimension of the noise was 3, the dimension of the control was 15. Matrices $\mathbf{F}$ and $\mathbf{B}$ were generated the same way as before, matrix $\mathbf{C}$ was a randomly chosen orthogonal mixing, noise sources $\mathbf{e}_{t+1}$ were chosen from the benchmark tasks of the fastICA toolbox[1] (Hyvärinen, 1999). We compared 5 different control methods (zero control, D-optimal control developed for parameter estimation, random control, A-optimal control developed for joint estimation of parameters and noise, as well as the τ-zero control with τ=81 that we developed for noise estimation). Comparisons are executed by first estimating the noise ($\mathbf{C}\mathbf{e}_{t+1}$) for times $T = 1,\ldots,1000$ and then applying the JADE ICA algorithm (Cardoso, 1999) for the estimation of the noise components ($\mathbf{e}_{t+1}$). Estimation was executed in each fiftieth steps, but only for the preceding 300 elements of the time series.

The quality of separation is evaluated by means of the Amari-error (Amari et al., 1996) as follows. Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be the estimated demixing matrix, and let $\mathbf{G} := \mathbf{W}\mathbf{C} \in \mathbb{R}^{d\times d}$. In case

1. Found at `http://www.cis.hut.fi/projects/ica/fastica/`.

of perfect separation the matrix **G** is a scaled permutation matrix. The Amari-error evaluates the quality of the separation by measuring the 'distance' of matrix **G** from permutation matrices:

$$r(\mathbf{G}) = \frac{1}{2d(d-1)} \sum_{i=1}^{d} \left(\frac{\sum_{j=1}^{d} |G_{ij}|}{\max_j |G_{ij}|} - 1 \right) + \frac{1}{2d(d-1)} \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{d} |G_{ij}|}{\max_i |G_{ij}|} - 1 \right).$$

The Amari-error $r(G)$ has the property that $0 \leq r(\mathbf{G}) \leq 1$, and $r(\mathbf{G}) = 0$ if and only if **G** is a permutation matrix.

Results are shown in Fig. 9. In short, τ-zero control performs slightly better than the joint parameter and noise estimation using the A-optimality principle. We note that for A-optimality design one does not have to worry about the duration of the parameter estimation. The performance of the other methods were considerably worse, especially for early times.

Most importantly, we found that if we use Bayesian methods for noise separation in ARX problems then it is worth to interrogate the system actively to improve the efficiency of the estimation.

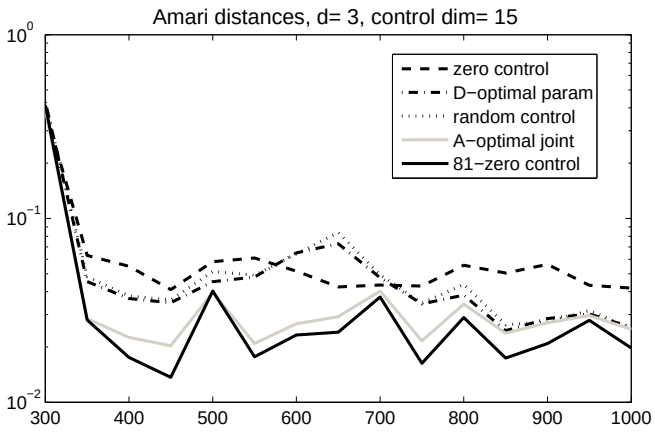


Figure 9: ARX-ICA experiment. Amari-error as a function of time for different control methods. Curves show the means of 100 experiments.

10.3 Model of the Furuta Pendulum

This section is concerned with more realistic simulations and investigate the robustness of our approach. We use a model of the Furuta pendulum (e.g., Yamakita et al., 1995) as our example, In this case, conditions of the theorems are not fulfilled and the task—in our formulation—can not be represented with a few matrices. In this simulation, we studied the D-optimality principle and compared it with the random control method. We were interested in the parameter estimation task in this example.

The two-segment Furuta pendulum problem (e.g., Yamakita et al., 1995; Gäfvert, 1998) was used. The pendulum has two links. Configuration of the pendulum is determined by the length of the links and by two angles. Dynamics of the pendulum are also determined by the different masses, that is, the masses of the links and the mass of the end effector as well as by the two motors, which are able to rotate the horizontal link and the swinging link in both directions. The angles of the horizontal and the swinging links are denoted by ϕ and θ, respectively (Fig. 10). Parameters

Name of parameter	Value	Unit	Notation
Angle of swinging link		rad	θ
Angle of horizontal link		rad	ϕ
Mass of horizontal link	0.072	kg	m_a
Mass of vertical link	0.00775	kg	m_p
Mass of the weight	0.02025	kg	M
Length of horizontal link	0.25	m	l_a
Length of vertical link	0.4125	m	l_p
Coulomb friction	0.015	Nm	τ_S
Coulomb stiction	0.01	Nm	τ_C
Maximal rotation speed for both links	2	$\frac{\text{rotation}}{s}$	
Approx. zero angular speed for swinging link	0.02	$\frac{rad}{s}$	$\dot{\phi}_\varepsilon$
Time intervals between interrogations	100	ms	
Maximum control value	0.05	Nm	δ

Table 2: Parameters of the Physical Model

of computer illustrations are provided in Table 2 for the sake of reproducibility. The state of the pendulum is given by ϕ, θ, $\dot{\phi}$ and $\dot{\theta}$. The magnitude of angular speeds $\dot{\phi}$ and $\dot{\theta}$ was restricted to 2 rotations/s, that is, to the interval $[-2\frac{\text{rot}}{\text{s}}, 2\frac{\text{rot}}{\text{s}}]$. For the equations of the dynamics and the details of the parameters, see, for example, the related technical report (Gäfvert, 1998).

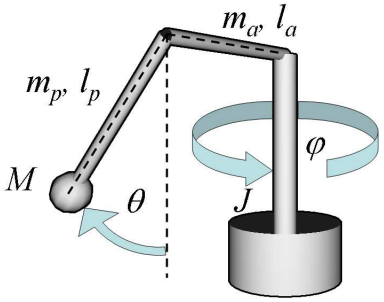


Figure 10: Furuta pendulum and notations of the different parameters. m: mass; l: length, M: mass of the end effector, subscript a: horizontal link, subscript p: swinging link, ϕ: angle of horizontal link, θ: angle of swinging link

.

The pendulum is a continuous dynamical system that we observe in discrete time steps. Furthermore, we assume that our observations are limited; we have only 144 low resolution and overlapping sensors for observing angles ϕ and θ. In each time step these sensors form our $\mathbf{r}(t) \in \mathbb{R}^{144}$ observations, which were simulated as follows: Space of angles ϕ and θ is $[0, 2\pi) \times [0, 2\pi)$, which we divided into $12 \times 12 = 144$ squared domains of equal sizes. There is a Gaussian sensor at the center of each domain. Each sensor gives maximal response 1 when angles θ and ϕ of the pendulum are in the center of the respective sensor, whereas the response decreased according to the Gaussian function. For example, for the i^{th} $(1 \leq i \leq 144)$ sensor characterized by angles θ_i, ϕ_i response y_i scaled as $y_i = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\theta-\theta_i)^2+(\phi-\phi_i)^2}{2\sigma^2})$ and the value of σ was set to 1.58 in radians. Sensors

were crude but noise-free; no noise was added to the sensory outputs. The inset at label 4 of Fig. 11 shows the outputs of the sensors in a typical case. Sensors satisfied periodic boundary conditions; if sensor S was centered around zero degree in any of the directions, then it sensed both small (around 0 radian) and large (around 2π radian) angles. We note that the outputs of the 144 domains are arranged for purposes of visualization; the underlying geometry of the sensors is hidden for the learning algorithm.

We observed these $\mathbf{r}_t \in \mathbb{R}^{144}$ quantities and then calculated the $\mathbf{u}_{t+1} \in \mathbb{R}^2$ D-optimal control using the algorithm of Table 1, where we approximated the pendulum with the model $\tilde{\mathbf{r}}_{t+1} = \mathbf{F}\mathbf{r}_t + \mathbf{B}\mathbf{u}_{t+1}$, $\mathbf{F} \in \mathbb{R}^{144\times144}$, $\mathbf{B} \in \mathbb{R}^{144\times2}$. Components of vector $\mathbf{u}_{t+1}$ controlled the 2 actuators of the angles separately. Maximal magnitude of each control signal was set to 0.05 Nm. Clearly we do not know the best parameters for $\mathbf{F}$ and $\mathbf{B}$ in this case, so we studied the prediction error and the number of visited domains instead. This procedure is detailed below.

First, we note that the angle of the swinging link and the angular speeds are important from the point of view of the prediction of the dynamics, whereas the angle of the horizontal link can be neglected. Thus, for the investigation of the learning process, we used the 3D space determined by $\dot{\phi}, \theta$ and $\dot{\theta}$. As was mentioned above, angular speeds were restricted to the $[-2\frac{\text{rot}}{\text{s}}, 2\frac{\text{rot}}{\text{s}}]$ domain. We divided each angular speed domain into 12 equal regions. We also used the 12-fold division of angle θ. Counting the domains, we had $12 \times 12 \times 12 = 1,728$ rectangular block shaped domains. Our algorithm provides estimations for $\hat{\mathbf{F}}_t$ and $\hat{\mathbf{B}}_t$ in each instant. We can use them to compute the predicted observation vector $\hat{\mathbf{r}}_{t+1} = \hat{\mathbf{F}}_t\mathbf{r}_t + \hat{\mathbf{B}}_t\mathbf{u}_{t+1}$. An example is shown in inset at label 4 of Fig. 11. We investigated the $\|\mathbf{r}_{t+1} - \hat{\mathbf{r}}_{t+1}\|$ prediction error (see Fig. 11) *cumulated over these domains* as follows. For each of the 1,728 domain, we set the initial error value at 30, a value somewhat larger than the maximal error we found in the computer runs. Therefore the cumulated error at start was $1,728 \times 30 = 51,840$.

The D-optimal algorithm does two things simultaneously: (i) it explores new domains, and (ii) it decreases the errors in the domains already visited. Thus, we measured the cumulated prediction errors during learning and corrected the estimation at each step. So, if our cumulated error estimation at time t was $e(t) = \sum_{k=1}^{1,728} e_k(t)$ and the pendulum entered the i^{th} domain at time $t+1$, then we set $e_k(t+1) = e_k(t)$ for all $k \neq i$ and $e_i(t+1)$ at $e_i(t+1) = \|\mathbf{r}_{t+1} - \hat{\mathbf{r}}_{t+1}\|$. Then we computed the new cumulated prediction error, that is, $e(t+1) = \sum_{k=1}^{1,728} e_k(t+1)$.

We compared the random and the D-optimality interrogation schemes. We show two sets of figures, Figs. 12a and 12b, as well as Figs. 12c and 12d. The upper set depicts the results for the full set of the 1,728 domains. It is hard for the random control to guide the pendulum to the upper domain, so we also investigated how the D-optimal control performs here. We computed the performance for cases when the swinging link was above vertical, that is for 864 domains (Figs. 12c and 12d).

For the full domain the number of visited domains is 456 (26%) and 818 (47%) for the random control and the D-optimal control, respectively after 5,000 control steps (Fig. 12a). The error drops by 13,390 (26%) and by 24,040 (46%), respectively (Fig. 12b). While the D-optimal controlled pendulum visited more domains and achieved smaller errors, the domain-wise estimation error is about the same for the domains visited; both methods gained about 29.4% per domains.

We can compute the same quantities for the upper domains as well. The number of visited upper domains is 9 and 114 for the random control and for the D-optimal control, respectively (Fig. 12c). The decrease of error is 265 and 3,342, respectively (Fig. 12d). In other words, D-optimal control gained 29.3% in each domain on average, whereas random control, on average, gained 29.4 %,

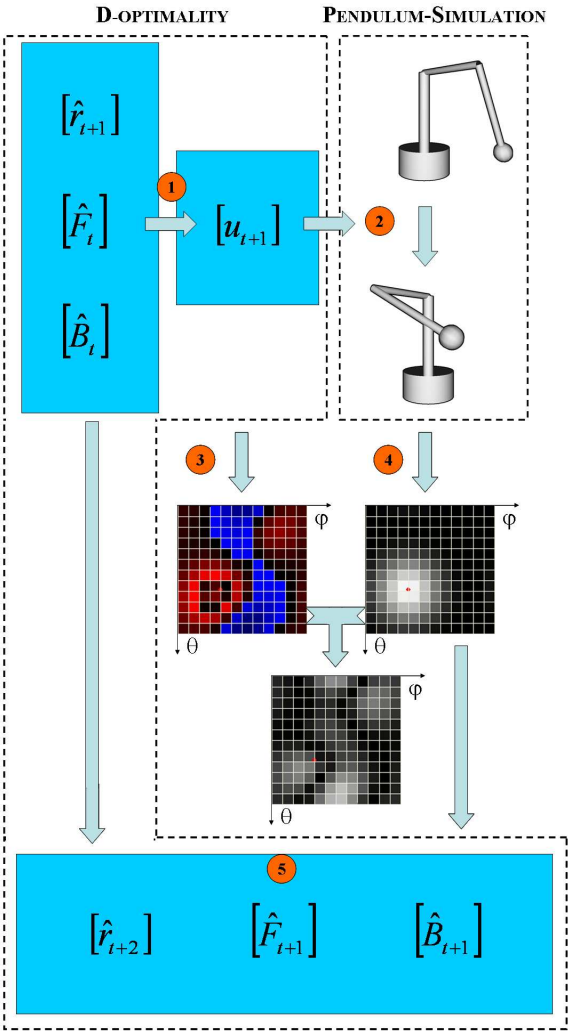


Figure 11: Scheme of D-optimal interrogation. (1) Control $\mathbf{u}_{t+1}$ is computed from D-optimal principle, (2) control acts upon the pendulum, (3) signals predicted before control step, (4) sensory information after control step. Difference between (3) and (4) is used for the computation of the cumulated prediction error. (5) Parameters were updated according to the pseudocode of Table 1. For more details, see text.

which are very close to the previous values in both cases. In this experiment D-optimal control gains more information concerning the system to be identified by visiting new domains.

This observation is further emphasized by the following data: The D-optimal algorithm discovered 37 new domains in the last 500 steps of the 5,000 step experiment. Out of these 37 domains, 20 (17) were discovered in the lower (upper) domain. By contrast, the random algorithm discovered 9 domains, out of which 5 (4) was in the lower (upper) domain. That is, D-optimality principle has a similar (roughly fourfold) lead in both the upper and lower domains, although the complexity of the task is different and the relative number of available volumes is also different in these two domains.

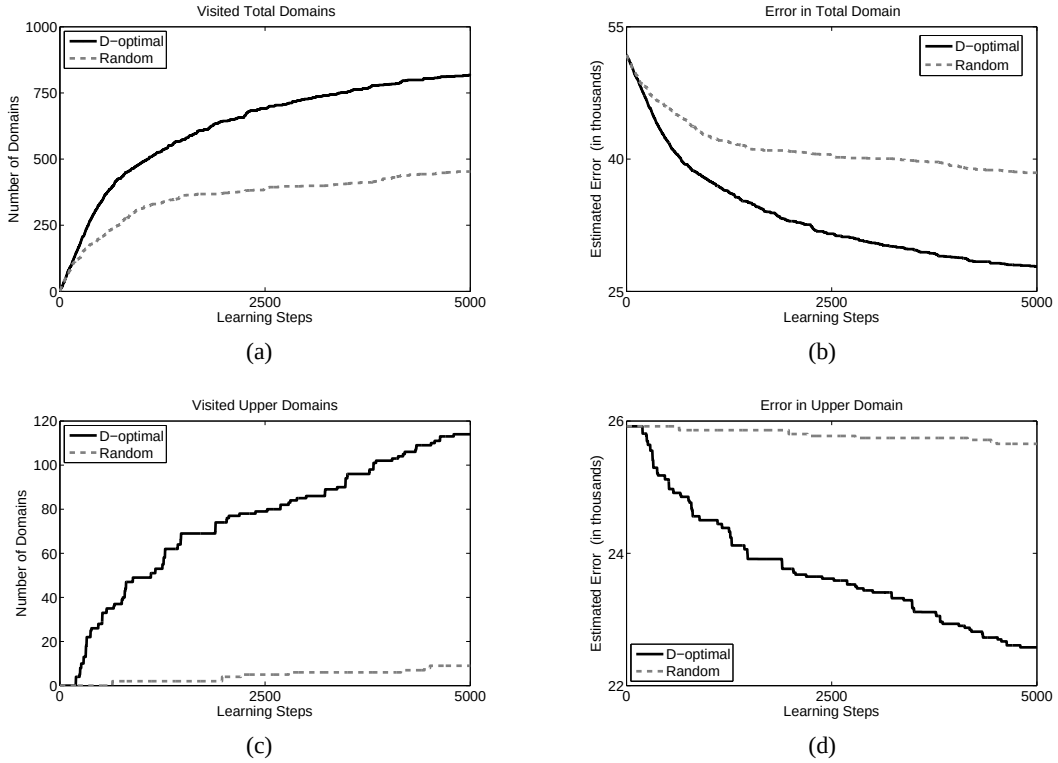


Figure 12: Furuta experiments driven by random and D-optimality controls. Solid (dotted) line: D-optimal (random) case. (a-b): Number of domains is 1728. (a): Visited domains, (b): upper bound for cumulated estimation error in all domains, (c-d): Number of domains is 864. (c): visited domains for swing angle above horizontal, (d): upper bound for cumulated estimation error for domains with swing angle above vertical. For more details, see text.

11. Discussion and Conclusions

We have treated the identification problem of recurrent neural networks as defined by the model detailed in (1). We applied active learning to solve this task. In particular, we studied the learning properties of the online A-optimality and D-optimality principles for parameter and noise estimations. We note that the D-optimal interrogation scheme is also called InfoMax control in the literature by Lewi et al. (2007). This name originates from the cost function that optimizes the mutual information.

In the generalized linear model (GLM) used by Lewi et al. (2007) $\mathbf{r}_{t+1}$ is drawn from an exponential family distribution with link function g and

$$E[\mathbf{r}_{t+1}] = g\left(\sum_{i=0}^{I}\mathbf{F}_i\mathbf{r}_{t-i} + \sum_{j=0}^{J}\mathbf{B}_j\mathbf{u}_{t+1-j}\right)$$

expected value. This model can be rewritten as

$$\mathbf{r}_{t+1} = g\left(\sum_{i=0}^{I}\mathbf{F}_i\mathbf{r}_{t-i} + \sum_{j=0}^{J}\mathbf{B}_j\mathbf{u}_{t+1-j}\right) + \mathbf{e}_{t+1}, \tag{33}$$

where $\{\mathbf{e}_t\}$ is a special noise process with $\mathbf{0} \in \mathbb{R}^d$ mean. The elements of this error series are independent of each other, but usually they are *not* identically distributed. The authors modeled spiking neurons and assumed that the main source of the noise is this spiking, which appears at the output of the neurons and adds linearly to the neural activity. They investigated the case in which the observed quantity $\mathbf{r}_t$ had a Poisson distribution. Unfortunately, in this model Bayesian equations become intractable and the estimation of the posterior may be corrupted, because the distribution is projected to the family of normal distributions at each instant. A serious problem with this approach is that the extent of the information loss caused by this approximation is not known. Our stochastic RNN model

$$\mathbf{r}_{t+1} = g\left(\sum_{i=0}^{I}\mathbf{F}_i\mathbf{r}_{t-i} + \sum_{j=0}^{J}\mathbf{B}_j\mathbf{u}_{t+1-j} + \mathbf{e}_{t+1}\right),$$

differs only slightly from the GLM model of (33), but it has considerable advantages, as we discuss it later. Note that the two models assume the same form if function g is the identity matrix and if the noise distribution is normal.

Our model is very similar to the well-studied non-linear Wiener (Celka et al., 2001) and Hammerstein (Pearson and Pottmann, 2000; Abonyi et al., 2000) systems. The Hammerstein model develops according to the following dynamics

$$\mathbf{r}_{t+1} = \sum_{i=0}^{I}\mathbf{F}_i\mathbf{r}_{t-i} + \sum_{j=0}^{J}\mathbf{B}_j g(\mathbf{u}_{t+1-j}) + \mathbf{e}_{t+1}.$$

The dynamics of the Wiener system is

$$\mathbf{r}_{t+1} = g\left(\sum_{i=0}^{I}\mathbf{F}_i g^{-1}(\mathbf{r}_{t-i}) + \sum_{j=0}^{J}\mathbf{B}_j\mathbf{u}_{t+1-j} + \mathbf{e}_{t+1}\right),$$

where we assumed that function g is invertible.

The Wiener and the Hammerstein systems have found applications in a broad range of areas, including financial predictions to the modeling of chemical processes. These models are special cases of non-linear ARX (NARX) models (Billings and Leontaritis, 1981). They are popular, because they belong to the simplest non-linear systems. Using block representation, they are simply the compositions of a static non-linear function and a dynamic ARX system. As a result, their properties can be investigated in a relatively simple manner and they are still able to model a large class of sophisticated non-linear phenomena.

Interesting comparisons between Wiener and Hammerstein systems can be found in Bai (2002), Aguirre et al. (2005), and Haber and Unbehauen (1990). We note that our Bayesian interrogation methods can be easily transferred to both the Wiener and to the Hammerstein systems.

Bayesian designs of different kinds were derived for the linear regression problem by Verdinelli (2000):

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \\ P(\mathbf{e}) &= \mathcal{N}_{\mathbf{e}}(0, \sigma^2\mathbf{I}).\end{aligned} \tag{34}$$

This problem is similar to ours ((3)-(5)), but while the goal of Verdinelli (2000) was to find an optimal design for the explanatory variables $\boldsymbol{\theta}$, we were concerned with the parameter ($\mathbf{X}$ in 34) and

	Parameter	Noise	Joint
D-optimal	$\max_{\mathbf{u}_{t+1}\in\mathcal{U}} \mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}$	$\min_{\mathbf{u}_{t+1}\in\mathcal{U}} \mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}$	N/A
A-optimal	$\max_{\mathbf{u}_{t+1}\in\mathcal{U}} \frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}$	$\min_{\mathbf{u}_{t+1}\in\mathcal{U}} \mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}$	$\max_{\mathbf{u}_{t+1}\in\mathcal{U}} \frac{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}$

Table 3: Cost functions in the parameter, noise, and in the joined parameter noise estimation task.

the noise (**e**) estimation tasks. In Verdinelli's paper inverted gamma prior and vector-valued normal distribution were assumed on the isotropic noise and on the explanatory variables, respectively. By contrast, we were interested in the matrix-valued coefficients and in general, non-isotropic noises. We used matrix-valued normal distribution for the coefficients, and in the D-optimal case we applied inverted Wishart distribution for the covariance matrix of the noise. Due to the properties of the inverted Wishart distribution, the noise covariance matrix is not restricted to the isotropic form. However, in the A-optimal case, we kept the derivations simple and we applied product of inverted gamma distributions for the covariance matrix as the conjugate prior.

The Bayesian online learning framework allowed us to derive analytic results for the myopic optimization of the parameters as well as the driving noise. In the *D-optimal* case the optimal interrogation strategies for parameter (14) and noise estimation (25) appeared in attractive, intriguingly simple quadratic forms. We have shown that these two tasks are incompatible with each other. Parameter and noise estimations require the maximization and the minimization of expression $\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}$, respectively. We have shown also that D-optimality can not be applied to the joined estimation of parameters and noise, because the corresponding cost function is constant.

For the *A-optimality* principle, we found that the objective of the noise estimation task is identical to that of the D-optimality principle. In this case, we were able to derive sensible results for the joined estimation of parameters and noise, and received hyperbolic optimization task. We also received a similar hyperbolic optimization task for the parameter estimation problem. The optimization of this task is non-trivial. For a simple hyper-cube domain we put forth a heuristics based on the simplex method. The different cost functions are summarized in Table 3.

We found empirically in the *parameter learning* task that the different objectives—that is, the minimization of $|(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}|$ and $tr[(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}]$, the results of the D-optimality and A-optimality principles, respectively—exhibit similar performances (Fig. 1, Fig. 2). However, D-optimality has slightly better performance and it is easier to compute, since we need to solve a quadratic problem only as opposed to a hyperbolic one. One of the reasons for the similar performance could be that the corresponding matrices are positive definite and thus they are diagonally dominant. In this case both the trace and the determinant are dominated by diagonal elements of the matrices.

In the *noise estimation* task, however, cost functions of the A- and D-optimality principles are the same. The main difficulty here is the non-myopic optimization of this cost function (Fig. 3, Fig. 4, Fig. 5, Fig. 6). The problem of non-greedy optimization of the full task has been left open for both the A-optimality and the D-optimality principles. For the noise estimation task, we suggested a heuristic solution that we called τ-infomax noise interrogation. Numerical experiments served to show that τ-infomax noise interrogation overcomes several other estimation strategies. The novel τ-infomax noise interrogation uses the D-optimal interrogation of Table 1 up to τ-steps, and applies the noise estimation control detailed in (25) afterwards. This heuristics decreases the estimation

error of the coefficients of matrices **F** and **B** up to time τ and thus—upon turning off the explorative D-optimization—tries to minimize the estimation error of the value of the noise at time $\tau+1$. We introduced the τ-zero interrogation scheme and showed that it is a good approximation of the τ-infomax noise scheme for large τ values.

In the *joint noise parameter estimation* task the D-optimal principle leads to a meaningless constant valued cost function. The A-optimal objective is a hyperbolic programming task, which is difficult to optimize. Still, its myopic optimization gave us better results than the myopic D-optimal objective derived for parameter estimation only. Interestingly, myopic A-optimal interrogations behaved similarly to the non-myopic τ-zero control (D-optimal parameter estimation up to τ steps, then zero control) with emergent automatic τ selection. However, these myopic results were somewhat worse than the non-myopic results from the τ-zero interrogation, when τ was larger than an appropriate threshold (Fig. 8).

In the field of active-learning, non myopic optimization is one of the most challenging tasks so most studies are concerned with greedy optimization only. Still, greedy optimization has produced valuable results in many cases. A few studies are concerned with non-greedy optimization of active learning, but only for certain special cases (Krause and Guestrin, 2007; Rangarajan et al., 2007). This field is in a rather early stage at the moment.

We illustrated the working of the algorithm on artificial databases both for the parameter estimation problem and for the noise estimation task. In the first set of experiments the database satisfied the conditions of our theorems. We studied the robustness of the algorithm and studied the noise estimation problem for a situation in which the conditions of our theorems were not satisfied, namely when the noise was neither Gaussian nor i.i.d. In particular, we studied the ARX problem family, when the hidden driving sources are non-Gaussian and statistically independent i.i.d. processes. We found that active control can speed-up the learning process. Function g was the identity function in these experiments.

We have also started to characterize the algorithm for a problem closer to reality. We chose the two-link Furuta pendulum task in these studies. We used a crude discretization for the pendulum, where the underlying dynamics and the low-dimensional nature of the problem are both hidden. Clearly we could express neither the ideal parameters for this case nor the noise that arose as a result of the disctretization. Thus we have studied the volume explored by the D-optimality method as well as the magnitude of the prediction errors.

The pendulum problem demonstrated that D-optimality maximizes mutual information by exploring new areas without significant compromise in the precision of estimation in the visited domains. The discovery rate is in favor of the D-optimality algorithm, which has a significant lead in both the frequently visited and the rarely visited domains, although the task is different and the relative number of available volumes is also different in these domains.

Our method treats the identification problem of non-linear ARX systems. We plan to generalize the method to NARMAX systems in the future. Such systems have several application fields, including financial modeling and the modeling of gas turbines (Chiras et al., 2001). One may try to apply these principles in a broad variety of fields, including selective laser chemistry Rabitz (2003), or the analysis of brain signals 'controlled' by visual inputs, for example, in brain-computer interfaces Vaughan et al. (2003).

Finally, it seems desirable to determine the conditions under which the algorithms derived from the optimality principles are both consistent and efficient. The tractable form of our approximation-free results is promising in this respect.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. We are grateful to György Flórea and László A. Jeni for running the pendulum simulations, and to Gábor Szirtes for his valuable comments. This research has been supported by the EC NEST 'Perceptual Consciousness: Explication and Testing' grant under contract 043261. Opinions and errors in this manuscript are the author's responsibility, they do not necessarily reflect the opinions of the EC or other project members.

Appendix A.

In this section we provide the technical details of our derivations.

A.1 Proof of Lemma 4.1

It is easy to show that the following equations hold:

$$\begin{aligned}\mathcal{N}_{\mathbf{y}}(\mathbf{Ax},\mathbf{V})\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K}) &= \mathcal{N}_{\mathbf{A}}(\mathbf{M}^{+},\mathbf{V},\mathbf{xx}^{T}+\mathbf{K})\mathcal{N}_{\mathbf{y}}(\mathbf{Mx},\mathbf{V},\gamma),\\ \mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\,I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n) &= I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}+\mathbf{H},n+m)\,\mathcal{T}_{\mathbf{A}}(\mathbf{Q},n,\mathbf{M},\mathbf{K}),\end{aligned}$$

where $\mathbf{M}^{+}=(\mathbf{MK}+\mathbf{yx}^{T})(\mathbf{xx}^{T}+\mathbf{K})^{-1}$, $\gamma=1-\mathbf{x}^{T}(\mathbf{xx}^{T}+\mathbf{K})^{-1}\mathbf{x}$, $\mathbf{H}=(\mathbf{A}-\mathbf{M})\mathbf{K}(\mathbf{A}-\mathbf{M})^{T}$ for the sake of brevity. Then we have

$$\mathcal{N}_{\mathbf{y}}(\mathbf{Mx},\mathbf{V},\gamma)\,I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)=I\mathcal{W}_{\mathbf{V}}(\mathbf{Q}+(\mathbf{y}-\mathbf{Mx})\gamma(\mathbf{y}-\mathbf{Mx})^{T},n+1)\mathcal{T}_{\mathbf{y}}(\mathbf{Q},n,\mathbf{Mx},\gamma),$$

and the statement of the lemma follows.

A.2 Proof of Lemma 4.2

Let $vec(\mathbf{A})$ denote a vector of dm dimensions where the $(d(i-1)+1)^{th},\ldots,(id)^{th}$ $(1\le i\le m)$ elements of this vector are equal to the elements of the i^{th} column of matrix $\mathbf{A}\in\mathbb{R}^{d\times m}$ in the appropriate order. Let $\otimes$ denote the Kronecker-product. It is known that for $P(\mathbf{A})=\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})$, $P(vec(\mathbf{A}))=\mathcal{N}_{vec(\mathbf{A})}(vec(\mathbf{M}),\mathbf{V}\otimes\mathbf{K}^{-1})$ holds (Minka, 2000). Using the well-known formula for the entropy of a multivariate and normally distributed variable (Cover and Thomas, 1991) and applying the relation $|\mathbf{V}\otimes\mathbf{K}^{-1}|=|\mathbf{V}|^{m}/|\mathbf{K}|^{d}$, we have that

$$H(\mathbf{A};\mathbf{V})=\frac{1}{2}\ln|\mathbf{V}\otimes\mathbf{K}^{-1}|+\frac{dm}{2}\ln(2\pi e)=\frac{m}{2}\ln|\mathbf{V}|-\frac{d}{2}\ln|\mathbf{K}|+\frac{dm}{2}\ln(2\pi e).$$

By exploiting certain properties of the Wishart distribution, we can compute the entropy of distribution $I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)$. The density of the Wishart distribution is defined by

$$\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)=\frac{1}{Z_{n,d}}|\mathbf{V}|^{(n-d-1)/2}\left|\frac{\mathbf{Q}^{-1}}{2}\right|^{n/2}\exp\left(-\frac{1}{2}tr(\mathbf{VQ}^{-1})\right).$$

Let Ψ denote the digamma function, and let $f_{1,3}(d,n)=-\sum_{i=1}^{d}\Psi(\frac{n+1-i}{2}))-d\ln 2$. Replacing $\mathbf{V}^{-1}$ with $\mathbf{S}$, we have for the Jacobian that $|\frac{d\mathbf{V}}{d\mathbf{S}}|=|\frac{d\mathbf{S}^{-1}}{d\mathbf{S}}|=|\mathbf{S}|^{-(d+1)}$ (Gupta and Nagar, 1999). To proceed

we use that $E_{\mathcal{W}_{\mathbf{S}}(\mathbf{Q},n)}\mathbf{S} = n\mathbf{Q}$, and $E_{\mathcal{W}_{\mathbf{S}}(\mathbf{Q},n)} \ln|\mathbf{S}| = \ln|\mathbf{Q}| - f_{1,3}(d,n)$, (Beal, 2003) and substitute them into $E_{I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)} \ln|\mathbf{V}|$, and $E_{I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)} tr(\mathbf{Q}\mathbf{V}^{-1})$:

$$\begin{aligned}
E_{I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)} \ln|\mathbf{V}| &= \int \frac{1}{Z_{n,d}} \frac{1}{|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Q})\right) \ln|\mathbf{V}| d\mathbf{V}, \\
&= -\int \frac{1}{Z_{n,d}} |\mathbf{S}|^{(d+1)/2} \left|\frac{\mathbf{S}\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{S}\mathbf{Q})\right) \ln|\mathbf{S}||\mathbf{S}|^{-d-1} d\mathbf{S}, \\
&= -\int \frac{1}{Z_{n,d}} |\mathbf{S}|^{(n-d-1)/2} \left|\frac{\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{S}\mathbf{Q})\right) \ln|\mathbf{S}| d\mathbf{S}, \\
&= -E_{\mathcal{W}_{\mathbf{S}}(\mathbf{Q}^{-1},n)} \ln|\mathbf{S}|, \\
&= \ln|\mathbf{Q}| + f_{1,3}(d,n). \qquad (35)
\end{aligned}$$

One can also show that

$$\begin{aligned}
E_{I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)} tr(\mathbf{Q}\mathbf{V}^{-1}) &= \int \frac{1}{Z_{n,d}} \frac{1}{|\mathbf{V}|^{(d+1)/2}} \left|\frac{\mathbf{V}^{-1}\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Q})\right) tr(\mathbf{Q}\mathbf{V}^{-1}) d\mathbf{V}, \\
&= \int \frac{1}{Z_{n,d}} |\mathbf{S}|^{(d+1)/2} \left|\frac{\mathbf{S}\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{S}\mathbf{Q})\right) tr(\mathbf{Q}\mathbf{S}) |\mathbf{S}|^{-d-1} d\mathbf{S}, \\
&= \int \frac{1}{Z_{n,d}} |\mathbf{S}|^{(n-d-1)/2} \left|\frac{\mathbf{Q}}{2}\right|^{n/2} \exp\left(-\frac{1}{2}tr(\mathbf{S}\mathbf{Q})\right) tr(\mathbf{Q}\mathbf{S}) d\mathbf{S}, \\
&= E_{\mathcal{W}_{\mathbf{S}}(\mathbf{Q}^{-1},n)} tr(\mathbf{Q}\mathbf{S}), \\
&= tr(\mathbf{Q}\mathbf{Q}^{-1}n) = nd. \qquad (36)
\end{aligned}$$

We calculate the entropy of stochastic variable $\mathbf{V}$ with distribution $I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)$. It follows from Eq. (35) and Eq. (36) that

$$\begin{aligned}
H(\mathbf{V}) &= -E_{I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n)} \left[-\ln(Z_{n,d}) + \frac{n}{2}\ln\left|\frac{\mathbf{Q}}{2}\right| - \frac{n+d+1}{2}\ln|\mathbf{V}| - \frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Q})\right], \\
&= \ln(Z_{n,d}) - \frac{n}{2}\ln\left|\frac{\mathbf{Q}}{2}\right| + \frac{n+d+1}{2}\left[\ln|\mathbf{Q}| - \sum_{i=1}^{d}\Psi(\frac{n+1-i}{2})) - d\ln 2\right] + \frac{nd}{2}, \\
&= \frac{d+1}{2}\ln|\mathbf{Q}| + f_{1,4}(d,n),
\end{aligned}$$

where $f_{1,4}(d,n)$ depends only on d and n.

Given the results above, we complete the computation of entropy $H(\mathbf{A},\mathbf{V})$ as follows:

$$\begin{aligned}
H(\mathbf{A},\mathbf{V}) &= H(\mathbf{A}|\mathbf{V}) + H(\mathbf{V}) = H(\mathbf{V}) + \int d\mathbf{V}\, I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n) H(\mathbf{A};\mathbf{V}), \\
&= \int d\mathbf{V}\, I\mathcal{W}_{\mathbf{V}}(\mathbf{Q},n) \left(\frac{m}{2}\ln|\mathbf{V}| - \frac{d}{2}\ln|\mathbf{K}| + \frac{dm}{2}\ln(2\pi e)\right) + H(\mathbf{V}), \\
&= -\frac{d}{2}\ln|\mathbf{K}| + \frac{dm}{2}\ln(2\pi e) + \frac{m}{2}[\ln|\mathbf{Q}| + f_{1,3}(d,n)] + \frac{d+1}{2}\ln|\mathbf{Q}| + f_{1,4}(d,n), \\
&= -\frac{d}{2}\ln|\mathbf{K}| + (\frac{m+d+1}{2})\ln|\mathbf{Q}| + f_{1,1}(d,n).
\end{aligned}$$

This is exactly what was claimed in Lemma 4.2.

A.3 Proof of Lemma 5.1

The proof is analogous to the D-optimality case. We need the following lemma:

Lemma A.1 *Let* $\mathbf{V}$ *diagonal positive definite matrix. Let* $\mathbf{X}_{(i,:)}$ *denote the* i^{th} *row of matrix* $\mathbf{X}$*. Let* $\eta_i = (\mathbf{A}_{(i,:)} - \mathbf{M}_{(i,:)})\mathbf{K}(\mathbf{A}_{(i,:)} - \mathbf{M}_{(i,:)})^T, \forall i = 1,\ldots,d.\ \boldsymbol{\eta} \in \mathbb{R}^d$*. Then the following statement holds:*

$$\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha}+m/2,\boldsymbol{\beta}+\boldsymbol{\eta}/2)\prod_{i=1}^{d}\mathcal{T}_{\mathbf{A}_{(i,:)}}(\beta_i,2\alpha_i,\mathbf{M}_{(i,:)},\frac{\mathbf{K}}{2}).$$

Proof:

$$\begin{aligned}
&\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \\
&= \frac{|\mathbf{K}|^{d/2}}{|2\pi\mathbf{V}|^{m/2}}\exp(-\frac{1}{2}tr(\mathbf{V}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{K}(\mathbf{A}-\mathbf{M})^T))\prod_{i=1}^{d}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}v_i^{-\alpha_i-1}\exp(-\frac{\beta_i}{v_i}),\\
&= \frac{|\mathbf{K}|^{d/2}}{\prod_{i=1}^{d}(2\pi)^{m/2}v_i^{m/2}}\exp(-\frac{1}{2}\sum_{i=1}^{d}\frac{1}{v_i}\eta_i)\prod_{i=1}^{d}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}v_i^{-\alpha_i-1}\exp(-\frac{\beta_i}{v_i}),\\
&= \frac{|\mathbf{K}|^{d/2}}{\prod_{i=1}^{d}(2\pi)^{m/2}}\prod_{i=1}^{d}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}v_i^{-\alpha_i-m/2-1}\exp(-\frac{\beta_i+\eta_i/2}{v_i}),\\
&= \frac{|\mathbf{K}|^{d/2}}{\prod_{i=1}^{d}(2\pi)^{m/2}}\prod_{i=1}^{d}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\frac{(\beta_i+\eta_i/2)^{\alpha_i+m/2}}{(\beta_i+\eta_i/2)^{\alpha_i+m/2}}\frac{\Gamma(\alpha_i+m/2)}{\Gamma(\alpha_i+m/2)},v_i^{-\alpha_i-m/2-1}\exp(-\frac{\beta_i+\eta_i/2}{v_i}),\\
&= \frac{|\mathbf{K}|^{d/2}}{\prod_{i=1}^{d}(2\pi)^{m/2}}\prod_{i=1}^{d}\frac{\beta_i^{\alpha_i}\Gamma(\alpha_i+m/2)/\Gamma(\alpha_i)}{(\beta_i+\eta_i/2)^{\alpha_i+m/2}}\prod_{i=1}^{d}\frac{(\beta_i+\eta_i/2)^{\alpha_i+m/2}}{\Gamma(\alpha_i+m/2)}v_i^{-\alpha_i-m/2-1}\exp(-\frac{\beta_i+\eta_i/2}{v_i}),\\
&= \prod_{i=1}^{d}\frac{|\mathbf{K}|^{1/2}}{(2\pi)^{m/2}}\frac{\Gamma(\alpha_i+m/2)}{\Gamma(\alpha_i)}\frac{\beta_i^{\alpha_i}}{(\beta_i+\eta_i/2)^{\alpha_i+m/2}}\mathcal{PIG}_{\mathbf{v}}(\boldsymbol{\alpha}+m/2,\boldsymbol{\beta}+\boldsymbol{\eta}/2),\\
&= \prod_{i=1}^{d}\frac{|\frac{\mathbf{K}}{2}|^{1/2}}{\pi^{m/2}}\frac{\Gamma(\alpha_i+m/2)\beta_i^{\alpha_i}}{\Gamma(\alpha_i)(\beta_i+(\mathbf{A}_{(i,:)}-\mathbf{M}_{(i,:)})\frac{\mathbf{K}}{2}(\mathbf{A}_{(i,:)}-\mathbf{M}_{(i,:)})^T)^{\alpha_i+m/2}}\mathcal{PIG}_{\mathbf{v}}(\boldsymbol{\alpha}+m/2,\boldsymbol{\beta}+\boldsymbol{\eta}/2),\\
&= \mathcal{PIG}_{\mathbf{v}}(\boldsymbol{\alpha}+m/2,\boldsymbol{\beta}+\boldsymbol{\eta}/2)\prod_{i=1}^{d}\mathcal{T}_{\mathbf{A}_{(i,:)}}(\beta_i,2\alpha_i,\mathbf{M}_{(i,:)},\frac{\mathbf{K}}{2}).
\end{aligned}$$

Lemma A.2

$$\begin{aligned}
\mathcal{N}_{\mathbf{y}}(\mathbf{Ax},\mathbf{V})\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathcal{N}_{\mathbf{A}}((\mathbf{MK}+\mathbf{yx}^T)(\mathbf{xx}^T+\mathbf{K})^{-1},\mathbf{V},\mathbf{xx}^T+\mathbf{K})\times\\
\times\mathcal{PIG}_{\mathbf{V}}\left(\boldsymbol{\alpha}+1/2,\boldsymbol{\beta}+diag\,(\mathbf{y}-\mathbf{Mx})\,\frac{(1-\mathbf{x}^T(\mathbf{xx}^T+\mathbf{K})^{-1}\mathbf{x})}{2}\,(\mathbf{y}-\mathbf{Mx})^T\right)\times\\
\times\prod_{i=1}^{d}T_{y_i}\left(\beta_i,2\alpha_i,(\mathbf{Mx})_i,\frac{1-\mathbf{x}^T(\mathbf{xx}^T+\mathbf{K})^{-1}\mathbf{x}}{2}\right).
\end{aligned}$$

Proof:

$$\begin{aligned}
\mathcal{N}_{\mathbf{y}}(\mathbf{Ax},\mathbf{V})\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathcal{N}_{\mathbf{A}}((\mathbf{MK}+\mathbf{yx}^T)(\mathbf{xx}^T+\mathbf{K})^{-1},\mathbf{V},\mathbf{xx}^T+\mathbf{K})\times\\
\mathcal{N}_{\mathbf{y}}((\mathbf{Mx},\mathbf{V},1-\mathbf{x}^T(\mathbf{xx}^T+\mathbf{K})^{-1}\mathbf{x})\times\\
\mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\beta}).
\end{aligned}$$

A.4 Proof of Lemma 5.2

We can see that $Var_{\mathbf{V}}[E[\mathbf{A}|\mathbf{V},\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}]] = Var_{\mathbf{V}}[\mathbf{M}_{t+1}] = 0$, and

$$\begin{aligned} E_{\mathbf{V}}[trVar(\mathbf{A}|\mathbf{V},\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1})] &= E_{\mathbf{V}}[tr(\mathbf{V}\otimes(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1})|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}], \\ &= tr(E[\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}])tr[(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}], \\ &= tr[(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}]\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1}, \end{aligned}$$

where we used that $P(\mathbf{V}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) = \mathcal{PIG}_{\mathbf{V}}(\boldsymbol{\alpha}_{t+1},\boldsymbol{\beta}_{t+1})$.

The law of total variance says that $Var[\mathbf{A}] = Var[E[\mathbf{A}|\mathbf{V}]] + E[Var[\mathbf{A}|\mathbf{V}]]$, hence

$$trVar[\mathbf{A}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}] = tr(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}\sum_{i=1}^{d}\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1}.$$

A.5 Proof of Lemma 6.1

To compute (24) we need the following lemma (Minka, 2000):

Lemma A.3 *If* $P(\mathbf{A}) = \mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})$*, then* $P(\mathbf{Ax}) = \mathcal{N}_{\mathbf{Ax}}\left(\mathbf{Mx},\mathbf{V},\left(\mathbf{x}^T\mathbf{K}^{-1}\mathbf{x}\right)^{-1}\right)$.

Applying this lemma and using (11) we have that

$$P(\mathbf{Ax}_{t+1}|\mathbf{V},\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) = \mathcal{N}_{\mathbf{Ax}_{t+1}}\left(\mathbf{M}_{t+1}\mathbf{x}_{t+1},\mathbf{V},\left(\mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\right)^{-1}\right). \tag{37}$$

We introduce the notations

$$\begin{aligned} \tilde{K}_{t+1} &= \left(\mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\right)^{-1} \in \mathbb{R}, \\ \lambda_{t+1} &= 1+(\mathbf{Ax}_{t+1}-\mathbf{M}_{t+1}\mathbf{x}_{t+1})^T(\tilde{K}_{t+1}\mathbf{Q}_{t+1}^{-1})(\mathbf{Ax}_{t+1}-\mathbf{M}_{t+1}\mathbf{x}_{t+1}) \in \mathbb{R}. \end{aligned} \tag{38}$$

Exploit the fact that

$$\mathcal{N}_{\mathbf{A}}(\mathbf{M},\mathbf{V},\mathbf{K})\,\mathcal{IW}_{\mathbf{V}}(\mathbf{Q},n) = \mathcal{IW}_{\mathbf{V}}(\mathbf{Q}+\mathbf{H},n+m)\,\mathcal{T}_{\mathbf{A}}(\mathbf{Q},n,\mathbf{M},\mathbf{K}),$$

and use (12) for the posterior distribution (37) and get

$$\begin{aligned} P(\mathbf{Ax}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= \mathcal{T}_{\mathbf{Ax}_{t+1}}\left(\mathbf{Q}_{t+1},n_{t+1},\mathbf{M}_{t+1}\mathbf{x}_{t+1},\tilde{K}_{t+1}\right), \\ &= \pi^{-d/2}|\tilde{K}_{t+1}^{-1}\mathbf{Q}_{t+1}|^{-1/2}\frac{\Gamma(\frac{n_{t+1}+1}{2})}{\Gamma(\frac{n_{t+1}+1-d}{2})}\lambda_{t+1}^{\frac{n_{t+1}+1}{2}}. \end{aligned}$$

The Shannon-entropy of this distribution according to Zografos and Nadarajah (2005) can be written as:

$$H(\mathbf{Ax}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) = f_{3,1}(d,n_{t+1}) + \frac{d}{2}\log|\tilde{K}_{t+1}^{-1}| + \log|\mathbf{Q}_{t+1}|,$$

where

$$f_{3,1}(d,n_{t+1}) = -\log\frac{\Gamma(\frac{n_{t+1}+1}{2})}{\pi^{d/2}\Gamma(\frac{n_{t+1}+1-d}{2})} + \frac{n_{t+1}+1}{2}\left(\Psi\left(\frac{n_{t+1}+1}{2}\right) - \Psi\left(\frac{n_{t+1}+1-d}{2}\right)\right).$$

Using the notations introduced in (10) and in (38), the above expressions can be transcribed as follows:

$$\begin{aligned} H(\mathbf{A}\mathbf{x}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1}) &= f_{3,1}(d,n_{t+1})-\frac{d}{2}\log|\tilde{K}_{t+1}|+\log|\mathbf{Q}_{t+1}|, \\ &= f_{3,1}(d,n_{t+1})+\frac{d}{2}\log|\mathbf{x}_{t+1}^T(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}\mathbf{x}_{t+1}|+\log|\mathbf{Q}_{t+1}|, \\ &= f_{3,1}(d,n_{t+1})+\frac{d}{2}\log|\mathbf{x}_{t+1}^T(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}\mathbf{x}_{t+1}|+ \\ &\quad +\log|\mathbf{Q}_t+(\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})\gamma_{t+1}(\mathbf{y}_{t+1}-\mathbf{M}_t\mathbf{x}_{t+1})^T|. \end{aligned}$$

Now, we are in a position to calculate (24) by applying Lemma 4.4 as before. We get that

$$\begin{aligned} &\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t)H(\mathbf{e}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^{t+1})= \\ &\quad = f_{3,2}(\mathbf{Q}_t,n_{t+1})+\frac{d}{2}\log|\mathbf{x}_{t+1}^T(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}\mathbf{x}_{t+1}|, \end{aligned}$$

where $f_{3,2}(\mathbf{Q}_t,n_{t+1})$ depends only on $\mathbf{Q}_t$, and n_{t+1}. We can proceed as follows

$$\begin{aligned} \arg\max_{\mathbf{u}_{t+1}} I(\mathbf{e}_{t+1},\mathbf{y}_{t+1};\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t) &= \arg\min_{\mathbf{u}_{t+1}}\log|\mathbf{x}_{t+1}^T(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T)^{-1}\mathbf{x}_{t+1}|, \\ &= \arg\min_{\mathbf{u}_{t+1}}\log\left|\mathbf{x}_{t+1}^T\left(\mathbf{K}_t^{-1}-\frac{\mathbf{K}_t^{-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right)\mathbf{x}_{t+1}\right|, \\ &= \arg\min_{\mathbf{u}_{t+1}}\log\left|\frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right|, \\ &= \arg\min_{\mathbf{u}_{t+1}}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}. \end{aligned}$$

This is exactly what we were to prove.

A.6 Proof of Lemma 8.1

One needs to compute the value of the integral

$$\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t)E[tr\mathbf{V}|\mathbf{x}_1^{t+1},\mathbf{y}_1^{t+1}]tr\left(\left(\mathbf{K}_t+\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\right)^{-1}+\mathbf{x}_{t+1}^T\mathbf{K}_{t+1}^{-1}\mathbf{x}_{t+1}\right).$$

However,

$$\begin{aligned} &\int d\mathbf{y}_{t+1}P(\mathbf{y}_{t+1}|\{\mathbf{x}\}_1^{t+1},\{\mathbf{y}\}_1^t)E[tr\mathbf{V}|\mathbf{x}_1^{t+1},\mathbf{y}_1^{t+1}] \\ =\int d\mathbf{y}_{t+1}\prod_{i=1}^d \mathcal{T}_{(\mathbf{y}_{t+1})_i}&\left((\boldsymbol{\beta}_t)_i,2(\boldsymbol{\alpha}_t)_i,(\mathbf{M}_t\mathbf{x}_{t+1})_i,\frac{\gamma_{t+1}}{2}\right)\sum_{i=1}^d\frac{(\boldsymbol{\beta}_{t+1})_i}{(\boldsymbol{\alpha}_{t+1})_i-1}, \end{aligned}$$

and depends only on the values of α_{t+1} and β_t as a result of (19) and Lemma 4.4, and is independent of the value of $\mathbf{x}_{t+1}$. Thus, we arrive at the minimization of the following expression:

$$
\begin{aligned}
tr\left[\left(\mathbf{K}_t^{-1}-\frac{\mathbf{K}_t^{-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right)+\mathbf{x}_{t+1}^T\left(\mathbf{K}_t^{-1}-\frac{\mathbf{K}_t^{-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right)\mathbf{x}_{t+1}\right] \\
=\left[tr(\mathbf{K}_t^{-1})-tr\frac{\mathbf{K}_t^{-1}\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}+\frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right], \\
=\left[tr(\mathbf{K}_t^{-1})-\frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}+\frac{\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\right], \\
=\frac{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}{1+\mathbf{x}_{t+1}^T\mathbf{K}_t^{-1}\mathbf{x}_{t+1}}\ .
\end{aligned}
$$

References

J. Abonyi, R. Babuska, A. Ayala-Botto, A. Szeifert, and L. Nagy. Identification and control of nonlienar systems using Hammerstein-models. *Ind. Eng. Chem. Res*, 39:4302–4314, 2000.

L. A. Aguirre, M. C. S. Coelho, and M. V. Correa. On the interpretation and practice of dynamical differences between Hammerstein and Wiener models. In *IEE. Proc. of Control Theory Application*, volume 152, pages 349–354, 2005.

S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, Cambridge, MA, 1996. MIT Press.

B. Anderson and A. Moore. Active learning for hidden Markov models: objective functions and algorithms. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 9–16, New York, NY, USA, 2005. ACM.

F. R. Bach. Active learning for misspecified generalized linear models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 65–72, Cambridge, MA, 2007. MIT Press.

E. W. Bai. A blind approach to the Hammerstein-Wiener model identification. *Automatica*, 38: 967–979, 2002.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

S. A. Billings and I. J. Leontaritis. Identifiaction of nonlinear systems using parametric estimation techniques. In *IEE. Proc. of Control and its Applications*, pages 183–187, 1981.

X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 33–42, 1998.

J. F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11 (1):157–192, 1999.

R. Castro, J. Jaupt, and R. Nowak. Compressed sensing vs. active learning. In *ICASSP 2006, IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, 2006a.

R. Castro, R. Willett, and R. Nowak. Faster rates in regression via active learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 179–186. MIT Press, Cambridge, MA, 2006b.

P. Celka, N. J. Bershad, and J. Vesin. Stochastic gradient identification of polynomial Wiener systems: Analysis and application. *IEEE Trans. Signal Process.*, 49(2):301–313, 2001.

K. Chaloner. Optimal bayesian experimental design for linear models. *The Annals of Statistics*, 12 (1):283–300, 1984.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statist. Sci.*, 10:273–304, 1995.

N. Chiras, C. Evans, and D. Rees. Nonlinear gas turibne modeling using NARMAX structures. *IEEE Trans. Instrum. Meas.*, 50(4):893–898, 2001.

D. A. Cohn. Neural network exploration using optimal experiment design. In *Advances in Neural Information Processing Systems*, volume 6, pages 679–686, 1994.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058, 1996.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

R. C. deCharms, D. T. Blake, and M. M. Merzenich. Optimizing sound features for cortical neurons. *Science*, 280:1439–1444, 1998.

G. Duncan and M. H. DeGroot. A mean squared error approach to optimal design theory. In *Proceedings of the 1976 Conference on Information: Sciences and Systems*, pages 217–221. The Johns Hopkins University, 1976.

V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

P. Földiák. Stimulus optimization in primary visual cortex. *Neurocomputing*, 38–40:1217–1222, 2001.

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

K. Fukumizu. Active learning in multilayer perceptrons. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 295–301. The MIT Press, 1996.

K. Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–26, 2000.

M. Gäfvert. Modelling the Furuta pendulum. Technical report ISRN LUTFD2/TFRT–7574–SE, Department of Automatic Control, Lund University, Sweden, April 1998.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2nd erdition, 2003.

Z. Ghahramani. Online variational Bayesian learning, 2000. Slides from talk presented at NIPS 2000 workshop on Online Learning.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins, Baltimore, MD, 3rd ed. edition, 1996.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*, volume 104 of *Monographs and Surveys in Pure and Applied Mathematics*. Chapman and Hall/CRC, 1999.

R. Haber and H. Unbehauen. Structure identification of nonlinear dynamic systems– a survey on input/output approaches. *Automatica*, 26(4):651–677, 1990.

D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, 1997.

A. Honkela and H. Valpola. On-line variational Bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 803–808, 2003.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, (10):626–634, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley, New York, 2001. URL `http://www.cis.hut.fi/projects/ica/book/`.

A. Hyvärinen. Independent component analysis for time-dependent stochastic processes. In *Proc. of ICANN'98, International Conference on Artificial Neural Networks, Skövde, Sweden*, pages 541–546, 1998.

H. Jaeger. Short term memory in echo state networks. GMD Report, 152, Fraunhofer AIS, 2001. http://publica.fraunhofer.de/starweb/pub08/en/index.htm.

C. Jutten and J. Hérault. Blind separation of sources: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, 21: 272–304, 1959.

S. Kotz and S. Nadarajah. *Multivariate T-Distributions and Their Applications*. Cambridge University Press, 2004.

A. Krause and C. Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 449–456, New York, NY, USA, 2007. ACM.

J. Lewi, R. Butera, and L. Paninski. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Proceedings of ICML-94, 11th International Conference on Machine Learning*, pages 148–156, New Brunswick, US, 1994. Morgan Kaufmann Publishers, San Francisco, US.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.

D. V. Lindley. *Bayesian Statistics: A Review*. SIAM, 1971.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14:2531–2560, 2002.

C. K. Machens, T. Gollisch, O. Kolesnikova, and A. V. M. Herz. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47:447–456, 2005.

D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

T. Minka. Bayesian linear regression, 2000. MIT Media Lab note.

T. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, MIT Media Lab, MIT, 2001.

M. Opper and O. Winther. A Bayesian approach to online learning. In *Online Learning in Neural Networks*. Cambridge University Press, 1999.

R. K. Pearson and M. Pottmann. Gray-box identification of block-oriented nonlnear models. *Journal of Process Control*, 10:301–315, 2000.

F. Pukelsheim. *Optimal Design of Experiments*. John Wiley & Sons, 1993.

H. Rabitz. Shaped laser pulses as reagents. *Science*, 299:525–527, 2003.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Boston, MIT Press, 1961.

R. Rangarajan, R. Raich, and A. O. Hero. Optimal sequential energy allocation for inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1:67–78, 2007.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conference on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

A. I. Schein. *Active Learning for Logistic Regression*. PhD thesis, University of Pennsylvania, 2005.

A. I. Schein and L. H. Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Computational Learning Theory*, 1992.

S. Solla and O. Winther. Optimal perceptron learning: An online Bayesian approach. In *Online Learning in Neural Networks*. Cambridge University Press, 1999.

D. M. Steinberg and W.G. Hunter. Experimental design: review and comment. *Technometrixs*, 26: 71–97, 1984.

M. Stone. Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist*, 30(1):55–70, 1959.

M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *The Journal of Machine Learning Research*, 7:141–166, 2006.

Z. Szabó and A. Lőrincz. Towards independent subspace analysis in controlled dynamical systems. ICA Research Network International Workshop, 2008.

B. Toman and J. L. Gastwirth. Robust Bayesian experimental design and estimation for analysis of variance models using a class of normal mixtures. *Journal of statistical planning and inference*, 35(3):383–398, 1993.

S. Tong and D. Koller. Active learning for parameter estimation in Bayesian networks. In *Advances in Neural Information Processing Systems*, pages 647–653, 2000.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001a.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, pages 45–66, 2001b.

T. M. Vaughan, W. J. Heetderks, L. J. Trejo, W. Z. Rymer, M. Weinrich, M. M. Moore, A. Kübler, B. H. Dobkin, N. Birbaumer, E. Donchin, E. W. Wolpaw, and J. R. Wolpaw. Brain-computer interface technology: a review of the Second International Meeting. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11:94–109, 2003.

I. Verdinelli. A note on Bayesian design for the normal linear model with unknown error variance. *Biometrika*, 87:222–227, 2000.

M. Yamakita, M. Iwashiro, Y. Sugahara, and K. Furuta. Robust swing-up control of double pendulum. In *American Control Conference*, volume 1, pages 290–295, 1995.

K. Zografos and S. Nadarajah. Expressions for Rényi and Shannon entropies for multivariate distributions. *Statistics and Probability Letters*, 71(1):71–84, 2005.