2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Metrics for Bayesian Optimal Experiment Design Under Model Misspecification

Tommie A. Catanach & Niladri Das

Abstract— **The conventional approach to Bayesian decision-theoretic experiment design involves searching over possible experiments to select a design that maximizes the expected value of a specified utility function. The expectation is over the joint distribution of all unknown variables implied by the statistical model that will be used to analyze the collected data. Utility functions define experiments' objectives; a common utility function is information gain. This article introduces an expanded framework for experimental design, where we go beyond the traditional Expected Information Gain criteria. We introduce Expected General Information Gain which measures robustness to the model discrepancy, and Expected Discriminatory Information to quantify how well an experiment can detect model discrepancy. The functionality of the framework is showcased through its application to a scenario involving a linearized spring mass damper system and an F-16 model where the model discrepancy is taken into account while doing Bayesian optimal experiment design.**

I. Introduction

For science and engineering systems there are often many choices of experiments to run, or data to collect, in order to infer information. Each of these choices has different costs in terms of time, money, or other resources. A common approach to designing the experiment stems from the field of Bayesian optimal experimental design (BOED). This approach uses the rigor of the Bayesian paradigm and information theory to formalize the design of experiments and treats it as an optimization problem. The aim is to maximize a utility function that captures the value of a particular experimental design. This utility function, typically the Expected Information Gain (EIG), depends on the posterior distribution sampled over many hypothetical realizations of plausible datasets from the experiment. However, for real applications, where there is model discrepancy, EIG is not the only measure of information we should consider.

In this work, we introduce two additional criteria that measure notions of robustness of a design. The first criterion, Expected Generalized Information Gain (EGIG), captures the expected information gained (or lost) when an experimenter uses a model with discrepancy. The second criterion, Expected Discriminatory Information (EDI) reflects whether the information gained from an experiment would be sufficient to discriminate between the model and an alternative. The EGIG-based design seeks to mitigate discrepancy while the EDI-based one seeks to only detect it. With these criteria, we aim to correct pathological issues in BOED and advance the BOED literature, which has only a few works concerning the robustness of BOED.

Tommie A. Catanach (tacatan@sandia.gov) and Niladri Das (corresponding author, ndas@sandia.gov), are with Sandia National Laboratories, Livermore, CA 94550, USA.

In [1] a Bayesian linear regression example is shown where the system is analysed without considering model discrepancies. There, not only is the parameter under-estimated but the posterior credible intervals are not even close to covering the true parameter value, which is alarming. In practice, despite the theoretical elegance and optimal performance for accurate models, BOED may encounter significant issues if our model is not properly specified. This means that there is no value of $\boldsymbol{x}^\star$ for which $p(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}^\star, \boldsymbol{d})$ corresponds to the true distribution for $p(\boldsymbol{y}|\boldsymbol{d})$, as noted in references [2] and [3]. Although model misspecification is a common problem in Bayesian settings, BOED methods are especially vulnerable because they use the model not only to fit data, but also to generate new data. The main issue is that Bayesian approaches only account for uncertainty in the model parameters, not in the model's correctness, which can lead to disastrous outcomes where BOED continuously queries similar designs and produces low-quality datasets. Eliminating misspecification entirely is unrealistic, particularly in a general BOED context. However, a lot of work can be done to improve our comprehension and management of it. Presently, there is only a limited amount of research that covers both the theoretical [4],[5],[6], and [7] and empirical implications of misspecification [8], and very little has been done to examine the specific mechanisms that can lead to failures. This is where our EGIG and EDI metrics play an important role in evaluating the model robustness and identifying modeling failures. Some Bayesian-adjacent approaches that call out the need for robustness and optimality in design are [9] and [10]. Most notably, [9] considers robust sensor placement for linear dynamical systems under asymptotic D-optimal design.

Outline: Section II introduces the model and key concepts, Section III presents the BOED criteria, Section IV studies EGIG and EDI for two examples systems, and Section V provides discussions.

II. Modeling and Key Concepts

A. System Description

We will study BOED in the context of simplified models, specifically stationary discrete-time linear processes driven

by Gaussian noise. We define the state vector as $\boldsymbol{x}_t \in \mathbb{R}^n$,

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t, \quad t = 1, 2, \ldots . \tag{1}$$

$\boldsymbol{A}$ is an $n \times n$ transition matrix and $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ is the process noise where $\boldsymbol{Q} \succeq 0$. We assume $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. For simplicity, unless specified, we will take $\boldsymbol{\mu_0} = 0$.

The observation equation is,

$$\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{2}$$

where the measurements are $\boldsymbol{y}_t \in \mathbb{R}^s$, $\boldsymbol{H}$ is the measurement matrix and $\boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ where $\boldsymbol{R} \succ \boldsymbol{0}$. The random vectors $\{\boldsymbol{x}_0, \boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_t, \boldsymbol{v}_1, ..., \boldsymbol{v}_t\}$ are all assumed to be independent.

From this general case, we will study two simplifications. First, we consider a system without dynamics (or equivalently a single time step of the system), corresponding to $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}$. Second, we will study the system after it has converged to its stationary distribution, assuming that $\boldsymbol{A}$ is asymptotically stable. In this case, if t is sufficiently large, we have that $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_L)$, where $\boldsymbol{\Sigma_L}$ is the solution to the discrete Lyapunov equation, $\boldsymbol{\Sigma_L} = \boldsymbol{A}\boldsymbol{\Sigma_L}\boldsymbol{A^T} + \boldsymbol{Q}$.

B. Bayesian Inference

In Bayesian inference, to rigorously update our beliefs about $\boldsymbol{X}$ with observation data $\boldsymbol{Y}$, we apply Bayes' theorem,

$$p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right) = \frac{p\left(\boldsymbol{Y} \mid \boldsymbol{X}\right) p\left(\boldsymbol{X}\right)}{p\left(\boldsymbol{Y}\right)}. \tag{3}$$

The prior $p\left(\boldsymbol{X}\right)$ reflects our initial beliefs about $\boldsymbol{X}$ while $p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right)$ is our posterior (after observations) belief. The likelihood, $p\left(\boldsymbol{Y} \mid \boldsymbol{X}\right)$ is the probability of observing $\boldsymbol{Y}$ given a state $\boldsymbol{X}$, while $p\left(\boldsymbol{Y}\right)$ is the overall probability of observing the data given our prior (called the evidence). Often we are interested in measuring how informative is the data. To do this we measure our change in belief, i.e., the information gain, using the Kullback–Leibler (KL) divergence,

$$\mathrm{D_{KL}}\Big[p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right) || p\left(\boldsymbol{X}\right)\Big] = \int p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right) \log \frac{p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right)}{p\left(\boldsymbol{X}\right)} d\boldsymbol{X} \tag{4}$$

For the Gaussian case where $p\left(\boldsymbol{X} \mid \boldsymbol{Y}\right) \sim \mathcal{N}\left(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}\right)$ and $p\left(\boldsymbol{X}\right) \sim \mathcal{N}\left(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}\right)$ the KL divergence is,

$$\frac{1}{2}\Bigg(\mathrm{Tr}\left[\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\Sigma_1}\right] - n + \left(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}\right)^T \boldsymbol{\Sigma_0^{-1}} \left(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}\right) + \log \frac{\mid \boldsymbol{\Sigma_0} \mid}{\mid \boldsymbol{\Sigma_1} \mid}\Bigg). \tag{5}$$

The KL divergence can be generalized using a more expressive, yet still information theoretically valid, measure of information [11] defined over three distributions, $r(\boldsymbol{X})$, $p(\boldsymbol{X})$, and $q(\boldsymbol{X})$, given by:

$$\mathcal{I}_{r(\boldsymbol{X})}\Big[p(\boldsymbol{X}) \;||\; q(\boldsymbol{X})\Big] = \int r\left(\boldsymbol{X}\right) \log \frac{p\left(\boldsymbol{X}\right)}{q\left(\boldsymbol{X}\right)} d\boldsymbol{X}. \tag{6}$$

The interpretation of this form of information is that we want to measure a change in belief (e.g., information gained or lost) when updating from $q\left(\boldsymbol{X}\right)$ to $p\left(\boldsymbol{X}\right)$ in the view of $r\left(\boldsymbol{X}\right)$. The view defines our reference frame for assessing changes in information. Typically, both $r\left(\boldsymbol{X}\right)$ and $p\left(\boldsymbol{X}\right)$ would represent the posterior with $q\left(\boldsymbol{X}\right)$ as the prior, thus recovering the regular KL divergence expression. However, in the case where there is model discrepancy, $r\left(\boldsymbol{X}\right)$ could be the unknown posterior from the true model, while $p\left(\boldsymbol{X}\right)$ could the inferred posterior from the model with discrepancy. Therefore, we could measure whether inference with the model discrepancy is still getting close to the correct result. We note that unlike KL divergence this measure can be negative, meaning that $q\left(\boldsymbol{X}\right)$ provides more information about $r\left(\boldsymbol{X}\right)$ than $p\left(\boldsymbol{X}\right)$ does.

For the case where, $r(\boldsymbol{X})$, $p(\boldsymbol{X})$, and $q(\boldsymbol{X})$ are all described by multivariate Gaussians,

$$\mathcal{I}_{r(\boldsymbol{X})}\Big[p(\boldsymbol{X}) \;||\; q(\boldsymbol{X})\Big] = \frac{1}{2}\Bigg(\mathrm{Tr}\Big[(\boldsymbol{\Sigma_q^{-1}} - \boldsymbol{\Sigma_p^{-1}})\boldsymbol{\Sigma_r}\Big] - \left(\boldsymbol{\mu_r} - \boldsymbol{\mu_p}\right)^T \boldsymbol{\Sigma_p^{-1}} \left(\boldsymbol{\mu_r} - \boldsymbol{\mu_p}\right) + \left(\boldsymbol{\mu_r} - \boldsymbol{\mu_q}\right)^T \boldsymbol{\Sigma_q^{-1}} \left(\boldsymbol{\mu_r} - \boldsymbol{\mu_q}\right) + \log \frac{\mid \boldsymbol{\Sigma_q} \mid}{\mid \boldsymbol{\Sigma_p} \mid}\Bigg). \tag{7}$$

This uses the fact that eq. 6 can be expressed as the difference of two KL divergences and employ eq. 5.

C. Bayesian Filtering

For a Markov process where the state $\boldsymbol{x_t}$ only depends on $\boldsymbol{x_{t-1}}$ and the observation $\boldsymbol{y_t}$ only depends on $\boldsymbol{x_t}$ we can simplify the inference problem for the state $\boldsymbol{x_t}$ given a time series of observations $\boldsymbol{Y_t} = \{\boldsymbol{y_0} \ldots \boldsymbol{y_t}\}$ as

$$p\left(\boldsymbol{x_t} \mid \boldsymbol{Y_t}\right) = \frac{p\left(\boldsymbol{y_t} \mid \boldsymbol{x_t}\right) p\left(\boldsymbol{x_t} \mid \boldsymbol{Y_{t-1}}\right)}{p\left(\boldsymbol{Y} \mid \boldsymbol{Y_{t-1}}\right)}. \tag{8}$$

Using this, the Bayesian filter for the system described by eq.1-2, is the Kalman filter,

$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{A}\boldsymbol{\mu}_{t-1|t-1} \tag{9}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{A}\boldsymbol{\Sigma}_{t-1|t-1}\boldsymbol{A}^T + \boldsymbol{Q} \tag{10}$$

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1}) \tag{11}$$

$$\boldsymbol{\Sigma}_{t|t} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H})\boldsymbol{\Sigma}_{t|t-1}, \tag{12}$$

where $\boldsymbol{K}_t = \boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T\boldsymbol{S_t^{-1}}$ is the Kalman gain matrix, $\boldsymbol{S_t} = \boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T + \boldsymbol{R}$ is the predictive uncertainty, and $\boldsymbol{I}$ is the identity matrix. Considering a single time step, the *a-priori* estimator of $\boldsymbol{x}_t$ is $\boldsymbol{\mu}_{t|t-1}$ with covariance $\boldsymbol{\Sigma}_{t|t-1}$. The *a-posteriori* estimator of $\boldsymbol{x}_t$ is $\boldsymbol{\mu}_{t|t}$ with covariance $\boldsymbol{\Sigma}_{t|t}$. Therefore, the prior, posterior, and evidence are:

$$p(\boldsymbol{x}_t) \sim \mathcal{N}(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}), \tag{13}$$

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_t, \boldsymbol{d}) \sim \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}), \tag{14}$$

$$p(\boldsymbol{y}_t \mid \boldsymbol{d}) \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{\mu}_{t|t-1}, \boldsymbol{S_t}). \tag{15}$$

As we can see from eq. 9-12, only the means $\boldsymbol{\mu}$ depend on the observations $\boldsymbol{y}$. Thus, when $\boldsymbol{A}$ is asymptotically stable we can find the stationary distribution of $\boldsymbol{\Sigma}_{t|t}$. We define $\boldsymbol{\Sigma}_{t|t-1} \to \boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}_{t|t} \to \boldsymbol{\Sigma}_D$ as $t \to \infty$. We first use the discrete time algebraic Riccati equation (DARE) to give,

$$\boldsymbol{\Gamma} = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T + \boldsymbol{Q} - \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^T + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{A}^T, \quad (16)$$

and then solve for $\boldsymbol{\Sigma}_D$ via:

$$\boldsymbol{\Sigma}_D = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^T + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{\Gamma}. \quad (17)$$

D. Bayesian Optimal Experimental Design

In BOED, the first step to modeling the problem is to define a utility function $U(\boldsymbol{d})$ that gives the value of performing an experiment at $\boldsymbol{d} \in \mathcal{D}$. The set $\mathcal{D}$ spans the space of possible designs. In Bayesian design, the utility is a function of the posterior distribution $p(\boldsymbol{X} \mid \boldsymbol{d}, \boldsymbol{Y})$. The utility function is maximized to find the optimal design $\boldsymbol{d}^*$, i.e. $\boldsymbol{d}^* = \operatorname{argmax}_{\boldsymbol{d}\in\mathcal{D}} U(\boldsymbol{d})$. The choice of the utility function $U(\boldsymbol{d})$ is crucial, as different functions will usually lead to different optimal designs [12]. A principled choice often used in BOED is mutual information. This is the information gained about $\boldsymbol{X}$ by taking measurements, $\boldsymbol{Y}$, according to design $\boldsymbol{d}$. This is just the KL divergence from prior to posterior, $\mathrm{D_{KL}}\Big[p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d})||p(\boldsymbol{X})\Big]$, eq. 4.

However at the point of choosing $\boldsymbol{d}$, we do not have measurements. So to assess the effectiveness of design $\boldsymbol{d}$, we take the expected KL divergence over plausible datasets $p(\boldsymbol{Y} \mid \boldsymbol{d})$. This utility function is known as Expected Information Gain (EIG) and is defined as,

$$\begin{aligned}\mathrm{EIG}(\boldsymbol{d}) &= \mathbb{E}_{p(\boldsymbol{Y} \mid \boldsymbol{d})}\Big[\mathrm{D_{KL}}(p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d})||p(\boldsymbol{X}))\Big] \\ &= \int p(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{d}) \log \frac{p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d})}{p(\boldsymbol{X})} d\boldsymbol{X} d\boldsymbol{Y}. \quad (18)\end{aligned}$$

III. Design Criteria

A. Expected Information Gain

For the linear Gaussian model given by eq.1-2, we can derive expressions for the EIG.

Single Step Update: First, for the case of a single update step (or equivalently, no dynamics) we begin by substituting the values from eq. 9 - 12 into the Gaussian KL divergence expression, eq. 5. Rearranging terms with the matrix inversion lemma and cyclic property of the trace, the information gain from prior to posterior is

$$\begin{aligned}&\mathrm{D_{KL}}\Big[p(\boldsymbol{x}_t \mid \boldsymbol{y}_t, \boldsymbol{d})||p(\boldsymbol{x}_t)\Big] = \\ &\frac{1}{2}\Big[\log|\boldsymbol{I} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}| - \mathrm{Tr}[\boldsymbol{S}_t^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T] \\ &+ (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})^T\boldsymbol{S}_t^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T\boldsymbol{S}_t^{-1}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})\Big] \quad (19)\end{aligned}$$

Only the last term depends on $\boldsymbol{y}_t$, so for EIG we just need to find the expectation of the quadratic term, which is,

$$\begin{aligned}&\mathbb{E}_{p(\boldsymbol{y}_t|\boldsymbol{d})}\Big[(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})^T\boldsymbol{S}_t^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T\boldsymbol{S}_t^{-1} \\ &\qquad (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})\Big] \\ &= \mathrm{Tr}\Big[\boldsymbol{S}_t^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T\boldsymbol{S}_t^{-1}\mathrm{Cov}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})\Big] \\ &= \mathrm{Tr}\Big[\boldsymbol{S}_t^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T\Big]. \quad (20)\end{aligned}$$

Here we recall eq. 15 so $(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})$ has mean $\mathbf{0}$ and covariance $\boldsymbol{S}_t$. Therefore, noting cancellation of trace terms, EIG of the single step of the Kalman filter is

$$\begin{aligned}\mathrm{EIG}(\boldsymbol{d}) &= \mathbb{E}_{p(\boldsymbol{y}_t \mid \boldsymbol{d})}\Big[\mathrm{D_{KL}}(p(\boldsymbol{x}_t \mid \boldsymbol{y}_t, \boldsymbol{d})||p(\boldsymbol{x}_t))\Big] \\ &= \frac{1}{2}\Big[\log|\boldsymbol{I} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}|\Big]. \quad (21)\end{aligned}$$

Infinite Horizon: We may also be interested in assessing the EIG about a state $\boldsymbol{x}_t$ when the system and filters have converged to their stationary distributions. For this, we define our prior knowledge about $\boldsymbol{x}_t$ as the solution to the Lyapunov equation, e.g., $p(\boldsymbol{x}_t|\boldsymbol{d}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_L)$, when t is sufficiently large to be in the asymptotic regime. Similarly, when we have a sufficiently large set of observations, $\boldsymbol{Y}_t$, we know the posterior belief about $\boldsymbol{x}_t$ will have the form $p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \boldsymbol{d}) = \mathcal{N}(\boldsymbol{\mu}_t(\boldsymbol{Y}_t), \boldsymbol{\Sigma}_D)$. Here we express $\boldsymbol{\mu}_t$ as a function of $\boldsymbol{Y}_t$ to emphasize that $\boldsymbol{\mu}_t$ is a random variable defined by $\boldsymbol{Y}_t$. Therefore, information gain from observing $\boldsymbol{Y}_t$ is

$$\begin{aligned}&\mathrm{D_{KL}}\Big[p(\boldsymbol{x}_t|\boldsymbol{Y}_t, \boldsymbol{d})||p(\boldsymbol{x}_t)\Big] = \\ &\frac{1}{2}\left(\mathrm{Tr}\left[\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Sigma}_D\right] - n + \boldsymbol{\mu}_t(\boldsymbol{Y}_t)^T\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\mu}_t(\boldsymbol{Y}_t) + \log\frac{|\,\boldsymbol{\Sigma}_L\,|}{|\,\boldsymbol{\Sigma}_D\,|}\right). \quad (22)\end{aligned}$$

Again, only the quadratic term depends on observations. Therefore, to compute EIG we first derive the expectation,

$$\begin{aligned}&\mathbb{E}_{p(\boldsymbol{Y}_t|\boldsymbol{d})}\Big[\boldsymbol{\mu}_t(\boldsymbol{Y}_t)^T\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\mu}_t(\boldsymbol{Y}_t)\Big] \\ &= \mathrm{Tr}\Big[\boldsymbol{\Sigma}_L^{-1}\mathrm{Cov}(\boldsymbol{\mu}_t(\boldsymbol{Y}_t)\boldsymbol{\mu}_t(\boldsymbol{Y}_t)^T)\Big] = \mathrm{Tr}\Big[\boldsymbol{\Sigma}_L^{-1}(\boldsymbol{\Sigma}_L - \boldsymbol{\Sigma}_D)\Big] \\ &= n - \mathrm{Tr}\Big[\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Sigma}_D\Big]. \quad (23)\end{aligned}$$

We use $\mathbb{E}_{p(\boldsymbol{Y}_t|\boldsymbol{d})}[\boldsymbol{\mu}_t(\boldsymbol{Y}_t)] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\mu}_t(\boldsymbol{Y}_t)\boldsymbol{\mu}_t(\boldsymbol{Y}_t)^T] = (\boldsymbol{\Sigma}_L - \boldsymbol{\Sigma}_D)$. This is shown as eq. 57 in Appendix VI-A.

Therefore, taking the expectation of eq. 22 over $\boldsymbol{Y}_t$ and substituting in the result of eq. 23 which cancels the trace terms, we find similarly to eq. 21 that,

$$\begin{aligned}\mathrm{EIG}(\boldsymbol{d}) &= \mathbb{E}_{p(\boldsymbol{Y}_t \mid \boldsymbol{d})}\Big[\mathrm{D_{KL}}(p(\boldsymbol{x}_t \mid \boldsymbol{Y}_t, \boldsymbol{d})||p(\boldsymbol{x}_t))\Big] \\ &\to \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_L|}{|\boldsymbol{\Sigma}_D|}, \text{ as } t \to \infty. \quad (24)\end{aligned}$$

B. Expected Generalized Information Gain

Using the generalized measure of information in eq. 6, we can assess how much information is expected to be gained or lost by an experiment d when there is model discrepancy. We define the true model as $\mathcal{M}^*$ and the model with discrepancy as $\mathcal{M}$, both of which have the same unknown states, $\boldsymbol{X}$, which we seek to infer. This expectation is taken over data that is generated according to $p(\boldsymbol{Y} \mid \boldsymbol{d}, \mathcal{M}^*)$. This leads to the Expected Generalized Information Gain (EGIG) given by,

$$\begin{aligned}
&\text{EGIG}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \\
&\mathbb{E}_{p(\boldsymbol{Y} \mid \boldsymbol{d}, \mathcal{M}^*)}\left[\mathcal{I}_{p^*}\left[p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d}, \mathcal{M}) || p(\boldsymbol{X} \mid \mathcal{M})\right]\right] \\
&= \int p(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{d}, \mathcal{M}^*) \log \frac{p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d}, \mathcal{M})}{p(\boldsymbol{X} \mid \mathcal{M})} d\boldsymbol{X} d\boldsymbol{Y} \quad (25) \\
&= \int p(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{d}, \mathcal{M}^*) \log \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{d}, \mathcal{M})}{p(\boldsymbol{Y} \mid \mathcal{M})} d\boldsymbol{X} d\boldsymbol{Y} \quad (26)
\end{aligned}$$

For notation we have $p^* := p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{d}, \mathcal{M}^*)$. Note that eq. 26 is a simple rearrangement using Bayes' theorem, which can be easier to compute for some problems.

In the following analysis we use $\mathcal{M}^*$ as a theoretical quantity to derive our metric. Typically we do not know $\mathcal{M}^*$, so in practice we should either define a set of plausible models we want to be robust to or we can assess the sensitivity to perturbations away from $\mathcal{M}$ by computing derivatives of the EGIG using either automatic differentiation or numerical derivatives. However, for some applications like surrogate modeling we do know $\mathcal{M}^*$.

In the context of inferring $\boldsymbol{x}_t$ with a system defined by eq. 1-2 we define the true model $\mathcal{M}^* = \{\boldsymbol{A}^*, \boldsymbol{H}^*, \boldsymbol{Q}^*, \boldsymbol{R}^*\}$ and the model we use for inference as $\mathcal{M} = \{\boldsymbol{A}, \boldsymbol{H}, \boldsymbol{Q}, \boldsymbol{R}\}$.

Single Step Update: We start with the EGIG of eq. 26. We defined $\boldsymbol{\mu}_{t|t-1} = \boldsymbol{A}$, $\boldsymbol{\mu}^*_{t|t-1} = \boldsymbol{A}^*$, $\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{A}\boldsymbol{\Sigma}_{t-1|t-1}\boldsymbol{A}^T + \boldsymbol{Q}$, and $\boldsymbol{\Sigma}^*_{t|t-1} = \boldsymbol{A}^*\boldsymbol{\Sigma}^*_{t-1|t-1}\boldsymbol{A}^{*T} + \boldsymbol{Q}^*$. We then note the distributions,

$$\begin{aligned}
&p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid d, \mathcal{M}^*) = \\
&\mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}^*_{t|t-1} \\ \boldsymbol{H}^*\boldsymbol{\mu}^*_{t|t-1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}^*_{t|t-1} & \boldsymbol{\Sigma}^*_{t|t-1}\boldsymbol{H}^{*T} \\ \boldsymbol{H}^*\boldsymbol{\Sigma}^*_{t|t-1} & \boldsymbol{S}^*_t \end{pmatrix}\right) \quad (27) \\
&p(\boldsymbol{y}_t \mid \boldsymbol{x}_t d, \mathcal{M}) = \mathcal{N}\left(\boldsymbol{H}\boldsymbol{x}_t, \boldsymbol{R}\right) \quad (28) \\
&p(\boldsymbol{y}_t \mid d, \mathcal{M}) = \mathcal{N}\left(\boldsymbol{H}\boldsymbol{\mu}_{t|t-1}, \boldsymbol{S}_t\right). \quad (29)
\end{aligned}$$

Recall that $\boldsymbol{S}_t = \boldsymbol{H}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{H}^T + \boldsymbol{R}$ and $\boldsymbol{S}^*_t = \boldsymbol{H}^*\boldsymbol{\Sigma}^*_{t|t-1}\boldsymbol{H}^{*T} + \boldsymbol{R}^*$. Substituting these distributions into eq. 26, we arrive at

$$\begin{aligned}
&\text{EGIG}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \\
&\quad \mathbb{E}_{p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \mathcal{M}^*)}\left[\log \frac{p(\boldsymbol{x}_t \mid \boldsymbol{y}_t, d, \mathcal{M})}{p(\boldsymbol{x}_t \mid d, \mathcal{M})}\right] \\
&= \frac{1}{2}\Bigg(\log \frac{\mid \boldsymbol{S}_t \mid}{\mid \boldsymbol{R} \mid} - \mathbb{E}\Big[(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t)^T \boldsymbol{R}^{-1} (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t)\Big] \\
&\quad + \mathbb{E}\Big[(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})^T \boldsymbol{S}_t^{-1} (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})\Big]\Bigg). \quad (30)
\end{aligned}$$

Using eq. 27-29, it is straight forward to compute the means and covariances of $(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t)$ and $(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})$,

$$\begin{aligned}
(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t) \sim\ & \mathcal{N}((\boldsymbol{H}^* - \boldsymbol{H})\boldsymbol{\mu}^*_{t|t-1}, \\
& (\boldsymbol{H}^* - \boldsymbol{H})^T \boldsymbol{\Sigma}^*_{t|t-1} (\boldsymbol{H}^* - \boldsymbol{H}) + \boldsymbol{R}^*) \quad (31) \\
(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1}) \sim\ & \mathcal{N}(\boldsymbol{H}^*\mu^*_{t|t-1} - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1}, \boldsymbol{S}^*_t). \quad (32)
\end{aligned}$$

Given these distributions, it is useful to define the following variables $\boldsymbol{\Delta}_H = \boldsymbol{H}^* - \boldsymbol{H}$ and $\boldsymbol{\Delta}_y = (\boldsymbol{H}^*\mu^*_{t|t-1} - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})$. Therefore, we can define the EGIG as:

$$\begin{aligned}
&\text{EGIG}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \\
&\frac{1}{2}\Bigg(\log \frac{\mid \boldsymbol{S}_t \mid}{\mid \boldsymbol{R} \mid} - \text{Tr}[\boldsymbol{R}^{-1}\boldsymbol{\Delta}_H \boldsymbol{\Sigma}^*_{t|t-1}\boldsymbol{\Delta}_H^T] - \text{Tr}[\boldsymbol{R}^{-1}\boldsymbol{R}^*] \\
&+ \text{Tr}[\boldsymbol{S}_t^{-1}\boldsymbol{S}^*_t] - \boldsymbol{\mu}^{*T}_{t|t-1}\boldsymbol{\Delta}_H^T \boldsymbol{R}^{-1}\boldsymbol{\Delta}_H \boldsymbol{\mu}^*_{t|t-1} \\
&+ \boldsymbol{\Delta}_y^T \boldsymbol{S}_t^{-1}\boldsymbol{\Delta}_y\Bigg). \quad (33)
\end{aligned}$$

Infinite Horizon: For the infinite horizon case for inferring $\boldsymbol{x}_t$ we know that our prior and posterior are Gaussian. Therefore, when computing the EGIG we can use the expression in eq. 7 and then compute the expectation over observations $\boldsymbol{Y}_t \sim p(\boldsymbol{Y}_t | \mathcal{M}^*)$. Here, $r(\boldsymbol{X})$ is $p(\boldsymbol{x}_t \mid \boldsymbol{Y}_t, d, \mathcal{M}^*)$, $p(\boldsymbol{X})$ is $p(\boldsymbol{x}_t \mid \boldsymbol{Y}_t, d, \mathcal{M})$, and $q(\boldsymbol{X})$ is $p(\boldsymbol{x}_t \mid \mathcal{M})$. By inspection, we see again that the only terms that depend on $\boldsymbol{Y}_t$ are the quadratic terms. Therefore, we begin with those terms.

First, we note the asymptotic results: $\boldsymbol{\Sigma}_{t|t} \to \boldsymbol{\Sigma}_D$, $\boldsymbol{\Sigma}^*_{t|t} \to \boldsymbol{\Sigma}^*_D$, $\boldsymbol{\Sigma}_{t|0} \to \boldsymbol{\Sigma}_L$, $\boldsymbol{\mu}_{t|0} = \boldsymbol{0}$, and $\boldsymbol{\mu}^*_{t|t} \overset{t\to\infty}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^*_L - \boldsymbol{\Sigma}^*_D)$. This gives us:

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{Y}_t | \mathcal{M}^*)}\left[\left(\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|0}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{L}}^{-1} \left(\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|0}\right)\right] \\
&\qquad = \text{Tr}\Big[\boldsymbol{\Sigma}_{\boldsymbol{L}}^{-1}\left(\boldsymbol{\Sigma}^*_L - \boldsymbol{\Sigma}^*_D\right)\Big]. \quad (34)
\end{aligned}$$

For the second expectation, we again rely on results presented in detail in Appendix VI-A. First, $\mathbb{E}_{p(\boldsymbol{Y}_t | \mathcal{M}^*)}[\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|t}] = \boldsymbol{0}$. Second, therefore,

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{Y}_t | \mathcal{M}^*)}\left[\left(\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|t}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{D}}^{-1} \left(\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|t}\right)\right] \\
&\qquad = \text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{D}}^{-1}\boldsymbol{M}_\Delta]. \quad (35)
\end{aligned}$$

$$\boldsymbol{M}_\Delta = \text{Cov}(\boldsymbol{\mu}^*_{t|t} - \boldsymbol{\mu}_{t|t}) = [-\mathbb{I}\ \ \mathbb{I}]\boldsymbol{M}[-\mathbb{I}\ \ \mathbb{I}]^T \quad (36)$$

where $\boldsymbol{M}$, the asymptotic covariance matrix of $[\boldsymbol{\mu}_{t|t}\ \boldsymbol{\mu}^*_{t|t}]^T$, is the solution to the Lyapunov equation given by:

$$\boldsymbol{M} = \boldsymbol{\mathcal{A}}\boldsymbol{M}\boldsymbol{\mathcal{A}}^{\boldsymbol{T}} + \boldsymbol{\mathcal{Q}} \quad (37)$$

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} (\boldsymbol{I} - \boldsymbol{KH})\,\boldsymbol{A} & \boldsymbol{KH^*A^*} \\ \boldsymbol{0} & \boldsymbol{A^*} \end{pmatrix} \tag{38}$$

$$\boldsymbol{\mathcal{Q}} = \begin{pmatrix} \boldsymbol{KS^*K^T} & \boldsymbol{KS^*K^{*T}} \\ \boldsymbol{K^*S^*K^T} & \boldsymbol{K^*S^*K^{*T}} \end{pmatrix} \tag{39}$$

Therefore, using these two equations we arrive at the EGIG for the infinite horizon system,

$$\begin{aligned} &\text{EGIG}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \mathbb{E}_{p(\boldsymbol{Y_t}|\mathcal{M}^*)}\big[\\ &\mathcal{I}_{p(\boldsymbol{x}_t|\boldsymbol{Y}_t,d,\mathcal{M}^*)}[p(\boldsymbol{x}_t \mid \boldsymbol{Y}_t, d, \mathcal{M}) \,||\, p(\boldsymbol{x}_t \mid \mathcal{M})]\big] \to \\ &\frac{1}{2}\left(\text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{L}}^{-1}\boldsymbol{\Sigma}_L^*] - \text{Tr}\Big[\boldsymbol{\Sigma}_D^{-1}(\boldsymbol{\Sigma}_D^* + \boldsymbol{M}_\Delta)\Big] + \log\frac{|\,\boldsymbol{\Sigma}_L\,|}{|\,\boldsymbol{\Sigma}_D\,|}\right) \\ &\text{as } t \to \infty. \end{aligned} \tag{40}$$

C. Expected Discriminatory Information

While EIG measures efficiency and EGIG measures robustness, we introduce the Expected Discriminatory Information (EDI) criteria to quantify how well an experiment can identify modeling failures. As such, unlike EGIG which is focused on comparing the Bayesian inference solution in the domain of the states $\boldsymbol{x}$, EDI compares them in the data domain, $\boldsymbol{y}$. Therefore, we can compare models that have different states and forms, e.g., different number of states. The EDI takes inspiration from the use of Bayes factors to compare models. Therefore we define the EDI as the expected Bayes factor given data from a true model $\mathcal{M}^*$:

$$\begin{aligned} \text{EDI}(\boldsymbol{d},\mathcal{M}, \mathcal{M}^*) &= \text{D}_{\text{KL}}\Big[p\left(\boldsymbol{Y} \mid d, \mathcal{M}^*\right)||p\left(\boldsymbol{Y} \mid d, \mathcal{M}\right)\Big] \\ &= \int p\left(\boldsymbol{Y} \mid d, \mathcal{M}^*\right)\log\frac{p\left(\boldsymbol{Y} \mid d, \mathcal{M}^*\right)}{p\left(\boldsymbol{Y} \mid d, \mathcal{M}\right)}dY. \end{aligned} \tag{41}$$

For filtering where $\boldsymbol{Y}_t = \{y_0 \dots y_t\}$, we can express the EDI using an iterative update leveraging a similar strategy for computing model evidence using a Bayesian filter,

$$\begin{aligned} &\text{EDI}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*, t) \\ &= \mathbb{E}_{p(\boldsymbol{y}_t, \boldsymbol{Y}_{t-1}|d,\mathcal{M}^*)}\left[\log\frac{p\left(\boldsymbol{y}_t \mid \boldsymbol{Y}_{t-1}, d, \mathcal{M}^*\right)}{p\left(\boldsymbol{y}_t \mid \boldsymbol{Y}_{t-1}, d, \mathcal{M}\right)}\right] \\ &\qquad + \mathbb{E}_{p(\boldsymbol{Y}_{t-1}|d,\mathcal{M}^*)}\left[\log\frac{p\left(\boldsymbol{Y}_{t-1} \mid d, \mathcal{M}^*\right)}{p\left(\boldsymbol{Y}_{t-1} \mid d, \mathcal{M}\right)}\right] \\ &= \mathbb{E}_{p(\boldsymbol{Y}_{t\text{-}1}|d,\mathcal{M}^*)}[\text{D}_{\text{KL}}\left(p(\boldsymbol{y}_t \,|\, \boldsymbol{Y}_{t\text{-}1}, d, \mathcal{M}^*)\,||p(\boldsymbol{y}_t \,|\, \boldsymbol{Y}_{t\text{-}1}, d, \mathcal{M})\right)] \\ &\qquad + \text{EDI}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*, t-1). \end{aligned} \tag{42}$$

Since the EDI is just a KL divergence, for the linear systems we have been studying in this paper, it is fairly straight forward to express it with the various quantities we have already derived. Therefore, we will state the main results without tenuous algebraic manipulation.

Single Step Update: For a single time step where data is generated by true process model $p(\boldsymbol{y_t} \mid d, \mathbb{M}^*)$, (see $\boldsymbol{y_t}$ marginal of eq. 27), but we are evaluating $\mathbb{M}$ according to $p(\boldsymbol{y_t} \mid d, \mathbb{M})$ (see eq. 29), we can compute the KL divergence for these Guassian distributions using eq. 5. Giving us:

$$\begin{aligned} &\text{EDI}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \\ &\frac{1}{2}\big(\text{Tr}[\boldsymbol{S}_t^{-1}\boldsymbol{S}_t^*] + \log\frac{|\,\boldsymbol{S}_t\,|}{|\,\boldsymbol{S}^*\,|} + \boldsymbol{\Delta}_y^T\boldsymbol{S}^{-1}\boldsymbol{\Delta}_y - s\big). \end{aligned} \tag{43}$$

s is the number of observations, e.g., sensors. Here we recall that $\boldsymbol{\Delta}_y = (\boldsymbol{H}^*\mu_{t|t-1}^* - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})$ and emphasize that $\mu_{t|t-1}^*$ and $\mu_{t|t-1}$ need not be the same dimension since the comparison is happening in the data space.

For the special case were $\boldsymbol{H}^* = [\boldsymbol{H}, \boldsymbol{\Delta}]$, the state of the model $\mathcal{M}^*$ is $\boldsymbol{x}_t^* = [\boldsymbol{x}_t, \boldsymbol{\delta}_t]^T$, $\boldsymbol{\mu}_{\delta,t|t-1} = \mathbb{E}[\boldsymbol{\delta}_{t|t-1}]$, and $\text{Cov}(\boldsymbol{x}_t^*) = \text{Diag}[\boldsymbol{\Sigma}_{t|t-1}, \boldsymbol{\Gamma}_{t|t-1}]$, e.g., the augmented states are independent of other states. Then,

$$\begin{aligned} &\text{EDI}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*) = \frac{1}{2}\Big(\text{Tr}[\boldsymbol{S}_t^{-1}\boldsymbol{\Delta}\boldsymbol{\Gamma}_{t|t-1}\boldsymbol{\Delta}^T] \\ &- \log|\mathbb{I} + \boldsymbol{S}_t^{-1}\boldsymbol{\Delta}\boldsymbol{\Gamma}_{t|t-1}\boldsymbol{\Delta}^T| + \boldsymbol{\mu}_{\delta,t|t-1}^T\boldsymbol{\Delta}^T\boldsymbol{S}_t^{-1}\boldsymbol{\Delta}\boldsymbol{\mu}_{\delta,t|t-1}\Big) \end{aligned} \tag{44}$$

Infinite Horizon: For the asymptotic case, we may choose to ask a slightly different question when assessing the EDI. Instead of asking about a single $\boldsymbol{y}_t$ we can ask about the full trajectory $\boldsymbol{Y}_t = \{y_0 \dots y_t\}$. Therefore to compute the EDI, we look to eq. 42. Under the previous assumptions of asymptotic stability, since we know that the predictives converge and are independent of the observations Y_t, we can expect the first term in eq. 42 to converge to a constant which we call Δ_{EDI}. Therefore we expect $\text{EDI}(\boldsymbol{d}, \mathcal{M}, \mathcal{M}^*, t) \to t\Delta_{\text{EDI}}$ as $t \to \infty$ unless $\Delta_{\text{EDI}} = 0$, meaning that there is only a finite amount of information to discriminate between the models based on the experiment even in the infinite horizon case. Therefore, Δ_{EDI} is the critical quantity for understanding the asymptotic EDI. Using the expression for the Guassian KL divergence, eq. 5 and taking the expectation using the asymptotic results found in Appendix VI-A, we find

$$\begin{aligned} &\Delta_{\text{EDI}} = \lim_{t\to\infty}\mathbb{E}_{p(\boldsymbol{Y}_{t\text{-}1}|d,\mathcal{M}^*)} \\ &\left[\text{D}_{\text{KL}}\left(p(\boldsymbol{y}_t \,|\, \boldsymbol{Y}_{t\text{-}1}, d, \mathcal{M}^*)\,||p(\boldsymbol{y}_t \,|\, \boldsymbol{Y}_{t\text{-}1}, d, \mathcal{M})\right)\right] \\ &= \frac{1}{2}\big(\text{Tr}[\boldsymbol{S}^{-1}\boldsymbol{S}^*] + \log\frac{|\,\boldsymbol{S}\,|}{|\,\boldsymbol{S}^*\,|} + \text{Tr}\left[\boldsymbol{S}^{-1}\boldsymbol{M}_S\right] - s\big). \end{aligned} \tag{45}$$

We recall that $\boldsymbol{S}$ and $\boldsymbol{S}^*$ are the stationary predictive covariances for an observation using design d for the models $\mathcal{M}$ and $\mathcal{M}^*$ respectively. The matrix $\boldsymbol{M}_S$ is given by,

$$\boldsymbol{M}_S = \text{Cov}(\boldsymbol{H}^*\boldsymbol{\mu}_{t|t-1}^* - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1})$$

$$= \begin{bmatrix} -\boldsymbol{H}\boldsymbol{A} & \boldsymbol{H}^*\boldsymbol{A}^* \end{bmatrix} \boldsymbol{M} \begin{bmatrix} -\boldsymbol{H}\boldsymbol{A} & \boldsymbol{H}^*\boldsymbol{A}^* \end{bmatrix}^T \quad (46)$$

where $\boldsymbol{M}$ is the joint asymptotic covariance matrix of $\boldsymbol{\mu}_{t|t}$ and $\boldsymbol{\mu}^*_{t|t}$ and is the solution to the previously specified Lyapnuov equation, eq. 37.

IV. Examples

A. Spring Mass Damper System

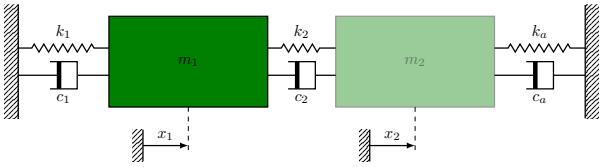

Fig. 1. Spring-Mass-Damper System with unknown second mass.

Fig. 1 shows a damped spring-mass system. The equations of motion for this system are,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & m_1 & 0 \\ 0 & 0 & 0 & m_2 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ v_1 \\ v_2 \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -(k_1+k_2) & k_2 & -(b_1+b_2) & b_2 \\ k_2 & -(k_2+k_3) & b_2 & -(b_2+b_3) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ v_1 \\ v_2 \end{bmatrix} \quad (47)$$

where x_1, x_2 denote the positions of the masses from their rest location. The variables v_1, v_2 denote their linear velocities respectively. The spring constants are k_1, k_2, k_3 and the damping coefficients are b_1, b_2, b_3. This continuous time linear system (CTLS) is then discretized for our analysis.

By analyzing the system we can see that under the conditions of high k_3 stiffness, low m_2 mass, or high b_3 damping, that the two-mass system should behave close to a single-mass system. Therefore, under these conditions, we would expect the ΔEDI criteria to become small when $\mathcal{M}$ is the one-mass system and $\mathcal{M}^*$ is the two-mass system. We see in panel A of Fig. 2 that ΔEDI indeed decreases as we increase the stiffness k_3.

We now consider choosing an observer design $d \in [0, \pi/2]$ to observe the position and velocity of the known mass, m_1, while balancing ΔEDI and EIG. Our, admittedly arbitrary, observer measures the position and velocity of m_1 with weights $\cos(d)$ and $\sin(d)$ respectively. The asymptotic EIG objective seeks to maximize information about the position and velocity of m_1 according to $\mathcal{M}$. The ΔEDI objective seeks to maximize our ability to asymptotically detect whether $\mathcal{M}$ is plausible versus $\mathcal{M}^*$. Of course we don't know $\mathcal{M}^*$ during the design phase so instead we average ΔEDI over a prior range of stiffnesses from panel A of Fig. 2. We see how EIG varies over the designs as the navy-blue curve in Fig. 2, panel B, while the mean ΔEDI is shown as the orange curve. The trade off between these quantities is show in panel C. Depending on the importance of discrimination vs performance, we may choose either to only observe the velocity (maximizing EIG) or to sacrifice some EIG to gain better discrimination power by choosing mixed sensor design.

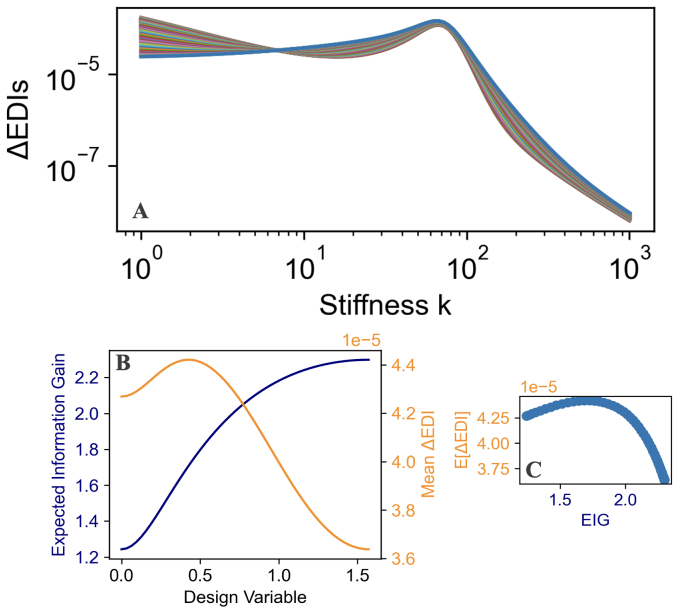

Fig. 2. Observer design and analysis for the spring mass system. Here, the true model $\mathcal{M}^*$ is a two mass system while the inference model $\mathcal{M}$ is the single mass system. Panel A shows how increasing the stiffness decreases our ability to distinguish between the models. Panels B and C show the trade off between EIG and ΔEDI over our design variable.

B. F-16 Model

We use an F-16 aircraft model from [13] and [14]. This system originally has 12 states of which we pull out the

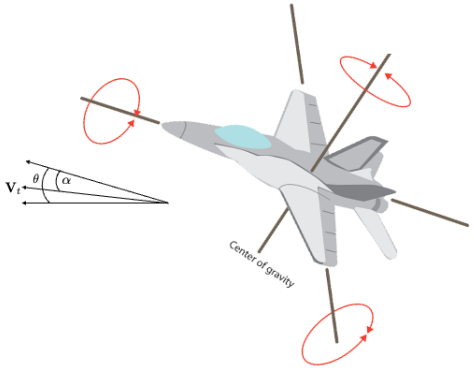

Fig. 3. F-16 model aircraft with specified states.

longitudinal dynamics with states (see Fig. 3): θ (pitch angle), V (velocity), α (attack angle), $\dot{\theta}$ and controls: T (thurst), δ_{ele} (elevator angle). We form a reduced-order CTLS using the closed loop system, which is then discretized.

For this model, we seek to add an additional output to the observer. This new output has the arbitrary form $y_{new} = d_1\theta + d_2\alpha + d_3\dot{\theta}$, where $d_1^2 + d_2^2 + d_3^2 = 1$. When considering these designs, we seek to balance maximizing asymptotic EIG while minimizing asymptotic EGIG. The inference model, $\mathcal{M}$, is the F-16 model with dynamics $\boldsymbol{A}$, but the true model, $\mathcal{M}^*$, has dynamics $\boldsymbol{A}^* = \boldsymbol{A} + \Delta \odot \boldsymbol{A}$. So, $\boldsymbol{A}^*$ has perturbations scaled relative to $\boldsymbol{A}$. Because Δ is unknown, we minimize the sensitivity of EGIG to changes of Δ. Therefore, our metric is the norm, $||\nabla_\Delta \text{EGIG}(d_1, d_2)||$. The result is summarized in Fig. 4. Panel A shows the trade off of different designs between $\text{EIG}(d_1, d_2)$ and $||\nabla_\Delta \text{EGIG}(d_1, d_2)||$ and the Pareto front of optimal designs (purple). We see that the EGIG is much more sensitive to the design than the EIG, i.e., EGIG varies by about a factor of 4. Therefore, for a robust design we may sacrifice a little asymptotic EIG for meaningful improvement in robustness. Panels B and C show the EIG and the EGIG projected on the design space along with the corresponding Pareto set.

We have made the codes to these examples available on GitHub[15].

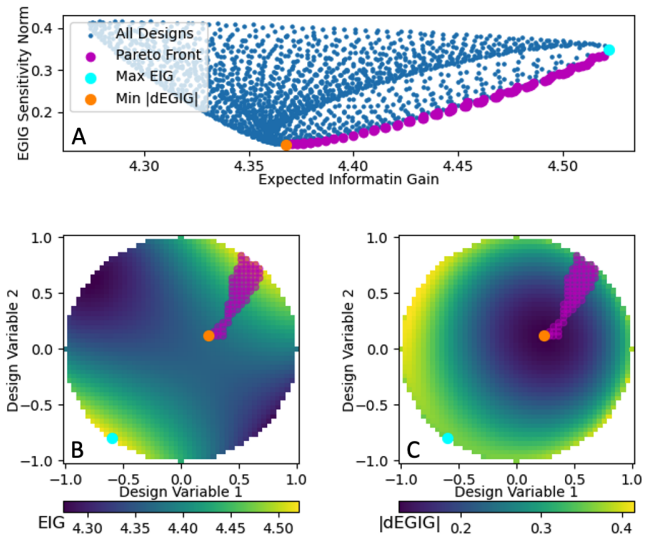

Fig. 4. Observer design for simplified F-16 model. The true model $\mathcal{M}^*$ is a small perturbation, Δ, in the dynamics from the inference model $\mathcal{M}$. We explore the addition of a new output, $y_{new} = d_1\theta + d_2\alpha + d_3\dot{\theta}$ where d_3 is constrained by d_1 and d_2. We measure the improved performance using $\text{EIG}(d_1, d_2)$ and robustness using the sensitivity of EGIG e.g., $||\nabla_\Delta \text{EGIG}(d_1, d_2)||$. Panel A shows the trade off between these two criteria for different designs. Panels B and C show the projection of these criteria on to the design space.

V. Conclusion

Maximizing the value of data for inference and prediction requires the careful selection of experimental conditions by modeling the experiment. These models are prone to misspecifications. We propose an information theoretic framework that extends the notion of Expected Information Gain (EIG), typically used in Bayesian experiment design, to address the model mismatch issue. The proposed Expected Generalized Information Gain (EGIG) captures the information gain or loss with respect to a true model, when the experiment's design is based on a model with discrepancy. On the other hand the proposed Expected Discriminatory Information (EDI) discriminates between models based upon the data generated, which further aids in model refinement. These three metrics are complementary as the EIG emphasizes data efficient experiments, the EGIG emphasizes experiments that lead to results that are robust to model discrepancy, and the EDI emphasizes experiments that would detect modeling failures.

For our future work we aim to develop a computational solver for assessing metrics within nonlinear systems and investigating the feasibility of its computational expense. Our objective is to establish correlations with alternative robustness measures like H_∞, widely employed in control filter and observer design. Additionally, we aim to integrate the identification of worst-case scenarios, a departure from our current sensitivity metrics approach.

Acknowledgment

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing conducted at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award DOE-ERCAPm3876.

References

[1] Jenny Brynjarsdottir and Anthony OHagan. "Learning about physical parameters: The importance of model discrepancy". In: *Inverse problems* 30.11 (2014), p. 114007.

[2] Peter Grünwald and Thijs van Ommen. "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It". In: *Bayesian Analysis* 12.4 (2017). DOI: 10.1214/17-ba1085.

[3] Chi Feng. "Optimal Bayesian experimental design in the presence of model error". MA thesis. Massachusetts Institute of Technology, 2015.

[4] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. *On Statistical Bias In Active Learning: How and When To Fix It*. 2021. arXiv: 2101.11665 [stat.ML].

[5] Drew Fudenberg, Gleb Romanyuk, and Philipp Strack. "Active learning with a misspecified prior". In: *Theoretical Economics* 12.3 (2017), pp. 1155–1189.

[6] Jinwoo Go and Tobin Isaac. *Robust Expected Information Gain for Optimal Bayesian Experimental Design Using Ambiguity Sets*. 2022. arXiv: 2205.09914.

[7] Antony M. Overstall and James M. McGree. *Bayesian decision-theoretic design of experiments under an alternative model*. 2021. arXiv: 1909.12570.

[8] Sabina J. Sloman et al. *Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias*. 2022. arXiv: 2205.13698.

[9] Eric M Hernandez. "Balancing robustness and optimality in sensor placement for dynamic state estimation". In: *Mechanical Systems and Signal Processing* 128 (2019), pp. 318–328.

[10] Romain Pasquier and Ian FC Smith. "Robust system identification and model predictions in the presence of systematic uncertainty". In: *Advanced Engineering Informatics* 29.4 (2015), pp. 1096–1109.

[11] Jed A. Duersch and Thomas A. Catanach. "Generalizing Information to the Evolution of Rational Belief". In: *Entropy* 22.1 (2020).

[12] Josep Ginebra. "On the measure of the information in a statistical experiment". In: (2007).

[13] Brian L Stevens, Frank L Lewis, and Eric N Johnson. *Aircraft control and simulation: dynamics, controls design, and autonomous systems*. John Wiley & Sons, 2015.

[14] Raktim Bhattacharya et al. "Nonlinear receding horizon control of an F-16 aircraft". In: *Journal of Guidance, Control, and Dynamics* 25.5 (2002), pp. 924–931.

[15] Niladri Das and Tommie A. Catanach. *Metrics BOED Model Misspecifications Python Notebooks*. Version 1.0.0. Mar. 2023. URL: https://github.com/sandialabs/Metrics-BOED-Model-Misspecifications.

VI. Appendix

A. Asymptotic Distribution of Inferred Means

Suppose we have a true model of a discrete time, asymptotically stable, linear dynamical system whose variables are denoted with a superscript $*$, while the model used for inference has variables without any superscript. Using a Kalman filter, the inferred mean is then given by,

$$\begin{aligned}\mu_t &= \boldsymbol{A}\mu_{t-1} + \boldsymbol{K}\left(Y_t - \boldsymbol{H}\boldsymbol{A}\mu_{t-1}\right)\\ &= \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H}\right)\boldsymbol{A}\mu_{t-1} + \boldsymbol{K}\boldsymbol{H}^*\boldsymbol{A}^*\mu^*_{t-1} + \boldsymbol{K}\zeta^*. \quad (48)\end{aligned}$$

Here $Y_t \sim \mathcal{N}\left(\boldsymbol{H}^*\boldsymbol{A}^*\mu^*_{t-1}, S^*\right)$, so $\zeta^* \sim \mathcal{N}\left(0, S^*\right)$. Similarly, we can define the evolution of mean of the true dynamical system under Kalman filtering as:

$$\begin{aligned}\mu^*_t &= \boldsymbol{A}^*\mu^*_{t-1} + \boldsymbol{K}^*\left(Y_t - \boldsymbol{H}^*\boldsymbol{A}^*\mu^*_{t-1}\right)\\ &= \boldsymbol{A}^*\mu^*_{t-1} + \boldsymbol{K}^*\zeta^* \quad (49)\end{aligned}$$

First, we note that $E[\mu_t] = E[\mu^*_t] = 0$, where the expectation is taken over asymptotically long sample trajectories of the true dynamical system. Second, we note that $E[\mu_{t-1}\zeta^{*T}] = E[\mu^*_{t-1}\zeta^{*T}] = 0$, i.e., they are independent. This comes from the fact that only μ_{t-1} and μ^*_{t-1} are functions of the trajectory and not ζ^*, or in other words, all the information about the trajectory is captured in the mean estimates. Finally, we note that $\boldsymbol{K}$ and $\boldsymbol{K^*}$ are known for the asymptotic case by solving the respective DAREs, eq. 16-17. With that we can express the second moments of μ_t and μ^*_t as:

$$\boldsymbol{M_t} = \begin{pmatrix} E[\mu_t\mu_t^T] & E[\mu_t\mu_t^{*T}] \\ E[\mu_t^*\mu_t^T] & E[\mu_t^*\mu_t^{*T}] \end{pmatrix}. \quad (50)$$

In order to solve for the second moments, we define:

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H}\right)\boldsymbol{A} & \boldsymbol{K}\boldsymbol{H}^*\boldsymbol{A}^* \\ \boldsymbol{0} & \boldsymbol{A}^* \end{pmatrix}, \quad (51)$$

$$\boldsymbol{\mathcal{Q}} = \begin{pmatrix} \boldsymbol{K}\boldsymbol{S}^*\boldsymbol{K}^{\boldsymbol{T}} & \boldsymbol{K}\boldsymbol{S}^*\boldsymbol{K}^{*\boldsymbol{T}} \\ \boldsymbol{K}^*\boldsymbol{S}^*\boldsymbol{K}^{\boldsymbol{T}} & \boldsymbol{K}^*\boldsymbol{S}^*\boldsymbol{K}^{*\boldsymbol{T}} \end{pmatrix}. \quad (52)$$

Then we can solve for the moments as:

$$\boldsymbol{M_t} = \boldsymbol{\mathcal{A}}\boldsymbol{M_{t-1}}\boldsymbol{\mathcal{A}}^{\boldsymbol{T}} + \boldsymbol{\mathcal{Q}}. \quad (53)$$

Therefore, for the asymptotic case, we can solve the following Lyapunov equation to find the asymptotic second-order moments,

$$\boldsymbol{M} = \boldsymbol{\mathcal{A}}\boldsymbol{M}\boldsymbol{\mathcal{A}}^{\boldsymbol{T}} + \boldsymbol{\mathcal{Q}}, \quad (54)$$

giving us the result that asymptotically:

$$\begin{pmatrix} \mu_t \\ \mu_t^* \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{M}\right). \quad (55)$$

Special case $\boldsymbol{A} = \boldsymbol{A}^*$: For the simpler case when we only have one model, the true model, we have the simplified equation given by:

$$\begin{aligned}\boldsymbol{M^*} =& \boldsymbol{A^*}\boldsymbol{M^*}\boldsymbol{A^{*T}} + \boldsymbol{K^*}\boldsymbol{S^*}\boldsymbol{K^{*T}}\\ =& \boldsymbol{A^*}\boldsymbol{M^*}\boldsymbol{A^{*T}} + \boldsymbol{A^*}\boldsymbol{P_D^*}\boldsymbol{A^{*T}} + \boldsymbol{Q} - \boldsymbol{P_D^*}\\ \implies& \boldsymbol{M^*} + \boldsymbol{P_D^*} = \boldsymbol{A^*}(\boldsymbol{M^*} + \boldsymbol{P_D^*})\boldsymbol{A^{*T}} + \boldsymbol{Q}, \quad (56)\end{aligned}$$

the substitution $\boldsymbol{K^*}\boldsymbol{S^*}\boldsymbol{K^{*T}} = \boldsymbol{A^*}\boldsymbol{P_D^*}\boldsymbol{A^{*T}} + \boldsymbol{Q} - \boldsymbol{P_D^*}$ can be found using the matrix inversion lemma and knowing that $\boldsymbol{P_D^*}$ is the solution to eq. 12 for the asymptotic case. We observe that eq. 56 is a Lyapunov equation. Thus since we know that $\boldsymbol{P_L^*}$ is the solution to the Lyapunov equation for this system. Therefore,

$$\boldsymbol{M^*} = \mathbb{E}[\mu_t^*\mu_t^{*T}] = \boldsymbol{P_L^*} - \boldsymbol{P_D^*}. \quad (57)$$