Resource Efficient Datasets for Inferring Parameters of Differential Equations via Optimal Experiment Design

Antanas Murelis AMURELIS@ETHZ.CH *ETH Zürich*
Lenart Treven LENART.TREVEN@INF.ETHZ.CH *ETH Zürich*
Ugne Sakenyte USAKENYTE@ETHZ.CH *ETH Zürich*
Mojmír Mutný MOJMIR.MUTNY@INF.ETHZ.CH *ETH Zürich*
Andreas Krause KRAUSEA@ETHZ.CH *ETH Zürich*

Reviewed on OpenReview: https://openreview.net/forum?id=mXkOwoXOmv

Abstract

Several real-world systems are described by parametric non-linear ordinary differential equations (ODEs) with unknown parameters. The nonlinearity makes parameter identification challenging, especially with only limited and noisy observations. To address this, we propose a novel framework that extends classical Optimal Experiment Design (OED) by linearizing non-linear ODE through neural network embeddings, transforming the system into a linear one. This allows us to apply linear OED techniques to non-linear systems. In particular, we demonstrate our approach on a task that involves selecting measurement points to identify the unknown parameters of the underlying ODE. Our method demonstrates improved parameter estimation and tighter confidence intervals compared to equidistant sampling in two non-linear test systems, showcasing its potential in optimizing experimental designs for more complex dynamical systems.

1 Introduction

Optimal experimental design (OED) is an area of statistics and machine learning, that aims to allocate experimental resources optimally, in order to infer a statistical quantity under a user-specified optimality metric. The typical procedure generates a policy that generates a dataset that is then used to estimate the statistical quantity of interest. For example, one of the common goals of OED is to structure experiments such that they provide the maximum amount of information as measured by notions of Fisher information, minimize the size of confidence intervals precision, minimize the error estimate in expectation while at the same time keeping number of experiments i.e. data points fixed.

In this project, we consider a scenario where we interact with a dynamical system ruled by a *parametric* ODE with *unknown* parameters. Our goal is to estimate the parameters of the ODE model from data collected in the experiments - physical realizations of the real-world phenomenon. We model the real-world setting where the experimenter does not have full access to the state of the system, only a limited number of noisy observations. In particular, we assume that we only observe a fixed-time snapshot $\{t_i\}_{i=1}^n$ of the dynamical system $x(t_i)$ for each t_i. Our goal is to observe the state at such times that the parameters of the physical system can be identified with least error. In this extended abstract, we focus on the choice of the snapshot times $\{t_i\}_{i=1}^n$, but the formalism is more general and we give pointers later. Our solution tackles the problem by making use of OED framework based

on decision theory, and approximate operator learning in dynamical systems with the help of deep learning.

2 Problem Statement

We assume the true state $x \in \mathbb{R}^m$ follows a parametric ODE on the time interval $[0, T]$:

$$\frac{d}{dt}x(t) = f(x(t), \gamma), \tag{1}$$

with the initial condition $x(0) = x_0$. We would like to estimate the unknown parameters $\gamma \in \Gamma \subseteq \mathbb{R}^k$ where the set Γ is compact. We assume that we can choose n time points $t_i \in [0, T], i \in \{1, \ldots, n\}$ at which we observe noisy values of the state $y(t_i) = x(t_i) + \epsilon_i$. The noise ϵ_i is zero-mean, independent and assumed to follow normal distribution with known variance. The objective on which we focus is to choose the discrete sampling time points $\{t_i\}_{i=1}^n$, referred to as the *design* or a *sampling scheme*, to estimate the system parameters γ from the noisy observations $y(t_i)$ given a limited budget n with the least error. Suppose that our estimate is $\hat{\gamma}$, we measure the error by squared distance $E(\hat{\gamma}) = ||\gamma - \hat{\gamma}||_2^2$.

2.1 Estimator

A natural estimator that can be used given the observations is the maximum likelihood estimator (MLE). Given independent noisy observations at times $\{t_i\}_{i=1}^n$, the MLE can be obtained by minimizing the surrogate negative log-likelihood $-l(y(t_1), ..., y(t_n)|\gamma) = \sum_{i=1}^n -\log(p(y(t_i)|\gamma))$. Assuming Gaussian, homoscedastic noise model, MLE becomes equivalent to the least squares estimator:

$$\hat{\gamma}_{\text{MLE}} = \arg\min_{\gamma} \sum_{i=1}^{n} ||y(t_i) - x(t_i)||_2^2 \quad \text{s.t.} \quad \frac{dx(t)}{dt} = f(x(t), \gamma). \tag{2}$$

Generally, arriving at an estimate $\hat{\gamma}$ is a non-convex optimization task. Despite this we show how we can reason about sensitivity of this estimator to the data points $\{t_i\}_{i=1}^n$.

2.2 Representing ODE solution and Linear ODEs

Suppose that $f(x, \gamma)$ is linear in x. This means, the differential eqautions can be represented as an operator $\frac{d}{dt}x(t) = \mathbf{A}_\gamma(\gamma)x(t)$, where $\mathbf{A}(\gamma) \in \mathbb{R}^{m \times m}$. Additionally, note that the solution of an ODE system can be represented in a Hilbert space with a basis $x(t) = \sum_{i=1}^{\infty} \theta_i F(t)_i$. In fact, this Hilbert space is reproducing with a kernel equal to Green's function of the linear ODE (González et al. (2014)). Using the representation $x(t) = F(t)^\top \theta$ (denoting the inner product), we identify that the observations follow

$$y(t) = F(t)^\top \theta + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma) \qquad \text{(observation model)}$$

$$\text{s.t.} \quad \underbrace{\left[\frac{d}{dt} - A(\gamma)\right] F(t)^\top \theta}_{:=\mathbf{L}(\gamma)} = 0 \quad \text{for all } t \in [0, T] \text{ and } \underbrace{F(0)^\top \theta = x_0}_{\text{initial condition}} \qquad \text{(ODE)}$$

The solutions are characterized by θ – elements from a reproducing kernel Hilbert space (RKHS). While $F(t)_i$ can form an infinite basis, we will often use finite approximation. The operator $\mathbf{L}(\gamma)$ implies a linear constraint on θ. In other words, and perhaps trivially, the solutions lie in the kernel of this operator. We will use this to design a sampling scheme as done in Mutný and Krause (2023).

2.3 Estimation Error: Worst-case vs Average performance

Suppose that we estimate θ without the knowledge of the initial condition and instead assume that $(F(0)^\top\theta - x_0)^2 \leq v^2$ $(\star)$. Given the knowledge of γ, we can estimate $\hat{\theta}_\gamma$ by means of maximum likelihood estimator, which depends on the observation times $\{t_i\}_{i=1}^n$ by:

$$\hat{\theta}_\gamma = \underset{\mathbf{L}(\gamma)\theta=0,\ \text{Eq. } (\star)}{\arg\min} \sum_{i=1}^{n} (F(t_i)^\top\theta - y_i)^2. \tag{3}$$

Notice that we used the kernel condition of $\mathbf{L}(\gamma)$. We will use this estimator further to estimate γ, as $\hat{\gamma} = \min_{\gamma\in\Gamma} \sum_{i=1}^{n} (x_\gamma(t_i) - \hat{\theta}_\gamma^\top F(t_i))^2$. This procedure decomposes the estimator in Eq. (2) into two parts. To justify the utility of this decomposition suppose the following thought experiment: the more accurately we can estimate $\hat{\theta}_\gamma$ for any γ, the more accurate we can then estimate γ. This implies that the error of estimating γ is smoothly varying with the other worst-case error, and hence by minimizing one we minimize the other. Thus we can focus on minimizing the worst case expected error of estimating $\hat{\theta}_\gamma$ for all γ and achieve the overall reduction of efficient parameter estimation at the same time (expectation on ϵ). Namely, we minimize:

$$\min_{\{t_1,\dots t_n\}} \sup_{\gamma\in\Gamma} \mathbb{E}_\epsilon[(\hat{\theta}_\gamma - \theta_\gamma)^2], \quad \text{or} \quad \min_{\{t_1,\dots t_n\}} \sum_{\gamma\in\Gamma} \mathbb{E}_\epsilon[(\hat{\theta}_\gamma - \theta_\gamma)^2] \tag{4}$$

where θ_γ is the true, assuming γ parameters. Notice that the second upperbounds the first one. We refer to them as *robust* and *Bayesian* error, respectively. Surprisingly, the expectation can be solved in closed form, and the minimization of the error over $\{t_i\}_{i=1}^n$ can be tackled using classical convex relaxation and greedy strategies (Fedorov and Leonov, 2013) that we briefly sketch in the Appendix. The resulting set of t_i is then used to fit the γ using the MLE estimator in Eq. (2). The estimator in Eq (3) is only used for design.

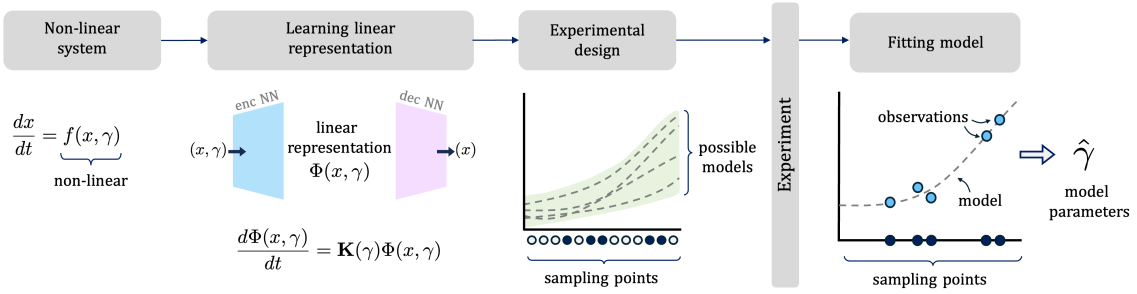

Figure 1: The figure demonstrates the general outline of the approach. The method begins with a defined non-linear system, whose parameters are unknown. A neural network is used to learn the linear representation of the dynamical system for any possible $\gamma \in \Gamma$. Optimal experimental design method then chooses then the sampling points. Noisy observations of the dynamics are made during the physical experiment, and the non-linear model is estimated using MLE to get $\hat{\gamma}_{\text{MLE}}$.

3 Learning Linearised ODE Representations

As most systems are guided by a set of non-linear ODEs $f(x,\gamma)$, the constraint $\mathbf{L}$ also becomes non-linear in general. The *core focus* of this work is to extend the linear experimental

design framework of Section 2.2 to the non-linear setting. We propose to learn a linearising embeddings $\Phi(x(t), \gamma)$ s.t. in the embedded space the dynamics of the system becomes linear. More concretely, consider an embedding of the state by a map $\Phi : \mathbb{R}^{m+k} \to \mathbb{R}^o$, such that:

$$\frac{d\Phi(\cdot, \gamma)}{dt} = \mathbf{K}(\gamma)\Phi(\cdot, \gamma), \tag{5}$$

where the operator $\mathbf{K}(\gamma)$ is linear in Φ. In literature, this linear operator is often referred to as the Koopman operator (Mezić (2021); Lusch et al. (2018)). This allows us to rewrite the non-linear constraint as a linear constraint on the embedding space:

$$\Phi(x(t), \gamma) = F(t)^\top \theta \quad \text{s.t.} \quad \left[\frac{d}{dt} - \mathbf{K}(\gamma)\right] F(t)^\top \theta = 0 \quad \text{and} \quad F(0)^\top \theta = x_0.$$

We parameterize the embedding Φ and matrix $\mathbf{K}(\gamma)$ with neural networks and use a data-driven approach to learn the mappings $\Phi(\cdot, \cdot)$ and $\mathbf{K}(\cdot)$. In particular, these are trained by sampling γ and x_0 and training them on synthetic data (infinite) in order to work for any $\gamma \in \Gamma$. Upon training a neural network that results in sufficiently accurate embeddings, one can use the same methodology as with linear ODEs. The whole procedure is summarized in Figure 1. More details on the training are provided in the Appendix.

4 Experiments

We test the framework on a toy 2-dimensional system (i), and Michaelis Menten kinetic model for enzyme dynamics (ii). The second represents a very common biochemical protocol (Saganuwan (2021)). The equations are: $\frac{x_1}{dt} = \mu x_1$, $\frac{x_2}{dt} = \lambda\left(x_2 - x_1^2\right)$ (i), and $\frac{d}{dt}x(t) = \frac{\lambda x(t)}{\mu + x(t)}$ (ii) respectively. In this case, the state is defined by $x = (x_1, x_2)$ (or x) and the unknown parameters as $\gamma = (\mu, \lambda)$. We compare our selection strategy to picking the time points equidistantly. The quantitative comparison can be found in Figure 2. Details of the implementation are in Appendix B.2. The results demonstrate that our method, while using the same budget, is able to estimate γ with much lower estimation error and results in tighter confidence sets for the estimated parameters (Wasserman et al. (2020)).

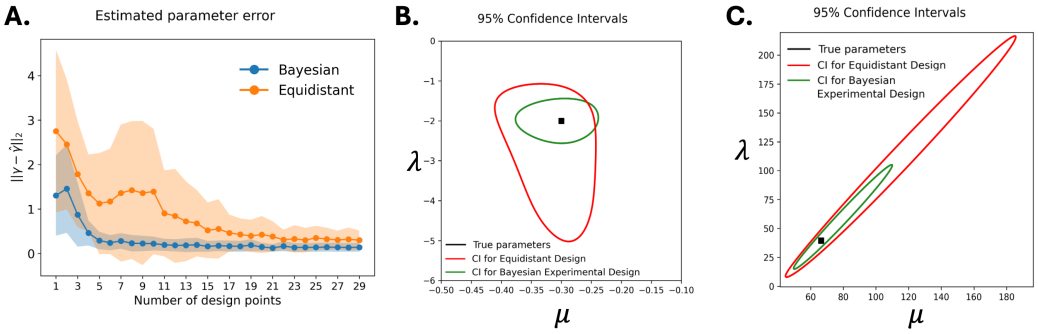

Figure 2: In A) we see the error as a function of the budget n. The error bars refer to 1 standard deviation obtained from a 100 noise realizations. In B) and C) we report 95% confidence sets generated by sampling noisy data according to two designs. In B) toy-problem while in C) Michaelis-Menten.

Broader Impact Statement

This is a theoretical work developing methodology for statistical estimation of natural phenomena. No broader harmful effect is of immediate relation.

References

Valerii V Fedorov and Sergei L Leonov. *Optimal Design for Nonlinear Response Models*. CRC Press, 07 2013.

Javier González, Ivan Vujačić, and Ernst Wit. Reproducing kernel Hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45: 26–32, Aug 2014. doi: 10.1016/j.patrec.2014.02.019. URL https://doi.org/10.1016/j.patrec.2014.02.019.

Y. Huang, S. G. Gilmour, Kalliopi Mylona, and P. Goos. Optimal design of experiments for hybrid nonlinear models, with applications to extended michaelis–menten kinetics. *Journal of Agricultural Biological and Environmental Statistics*, 25(4):601–616, Jul 2020. doi: 10.1007/s13253-020-00405-3. URL https://doi.org/10.1007/s13253-020-00405-3.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9, 11 2018. doi: 10.1038/s41467-018-07210-0.

Igor Mezić. Koopman operator, geometry, and learning of dynamical systems. *Notices of the American Mathematical Society*, 68:1, 08 2021. doi: 10.1090/noti2306.

M. Mutný and A. Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features, 2018. URL https://las.inf.ethz.ch/files/Mutny2018b.pdf. Accessed Feb. 22, 2024.

Mojmír Mutný and Andreas Krause. Experimental design for linear functionals in reproducing kernel hilbert spaces, 01 2023. URL https://arxiv.org/abs/2205.13627.

Alhaji Saganuwan Saganuwan. Application of modified michaelis–menten equations for determination of enzyme inducing and inhibiting drugs. *BMC Pharmacology and Toxicology*, 22(1), Oct 2021. doi: 10.1186/s40360-021-00521-x. URL https://doi.org/10.1186/s40360-021-00521-x.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences of the United States of America*, 117 (29):16880–16890, Jul 2020. doi: 10.1073/pnas.1922664117. URL https://doi.org/10.1073/pnas.1922664117.

A. Çivril and M. Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47–49):4801–4811, Nov 2009. doi: 10.1016/j.tcs.2009.06.018. URL https://doi.org/10.1016/j.tcs.2009.06.018.

Appendix A.

A.1 Design of Experiment

Classical optimal experimental design uses reciprocity between information and variance of the model parameters - minimizing variance maximizes the information obtained from the design (Fedorov and Leonov (2013)). In our case, optimal experimental design aims to collocate sampling times such that we gain the most information about the linear model parameters θ, given the linear structure of the differential operator. Let $t_d = \{t_i\}_{i=0}^{T}$ be the set of support or design (time) points. Then:

$$\xi = \left\{ \begin{array}{c} t_0, \ldots, t_N \\ p_0, \ldots, p_N \end{array} \right\} = \left\{ \begin{array}{c} t_i \\ p_i \end{array} \right\}_{i=0}^{N}$$

with $p_i = r_i/N$ and $N = \sum_{i=1}^{n} r_i$, where r_i is the number of times a given point is observed and N is the budget. ξ is also referred to as the design of the experiment.

A.2 Information Matrix

The information matrix captures the information carried by the design ξ (A.1). In our experiments, we utilize the ridge information matrix as presented in (Mutný and Krause (2023)). It is common that the experimenter has knowledge about the initial condition, as a result, we also include this knowledge in the information matrix:

$$\mathbf{M}_\lambda(\xi, \gamma) = \sigma^{-2} \left(\mathbf{C}_\gamma^\top \left(\sigma^2 \lambda_1 \mathbf{I} + \sigma^2 \lambda_2 F(0)^\top F(0) + \mathbf{M}(\xi) \right)^{-1} \mathbf{C}_\gamma \right)^{-1} \tag{6}$$

where $\mathbf{M}(\xi) = \sigma^{-2} \sum_{i=1}^{N} p_i \cdot F(t_i)^\top F(t_i)$, the columns of $\mathbf{C}_\gamma$ span the null space of the operator $\mathbf{L}_\gamma$, λ_1 and λ_2 are hyperparameters, and σ^2 is variance of the noise. In practice we do not know the true underlying γ, but instead can express some feasible region of parameters Γ, where the true parameter may lie. In general, we cannot do a design that is best for all $\gamma \in \Gamma$. Instead, we do Bayesian experimental designs that are best "on-average" for the set Γ (see A.4 for ref.).

A.3 Greedy Selection

Discretely optimizing over the information matrices is known to be an NP-hard problem (Çivril and Magdon-Ismail (2009)). One approach to avoid enumerating the designs is to approximate them by picking design points greedily. That is, picking the time points that increase the optimality criteria of the information matrix the most in each round until the budget limit is reached. This involves calculating the information matrix and the optimality criteria for each discrete time point in the support of ξ, at each iteration of the algorithm. The greedy algorithm used to approximate designs in this extended abstract is summarised in Algorithm 1.

A.4 Bayesian Design

In Section 2 we show how to formulate the OED framework for linear dynamical systems whose parameters are *known*. In reality, our goal is to find the underlying parameters of

Algorithm 1 Greedy Optimization of Design Points

Require:
Set of potential design points $\mathbf{D}$
A budget $\mathbf{N}$
Optimality criteria functional $f_{A/D/C}$

Initialize design ξ as empty.
while $\mathbf{N} > 0$ **do**
  for each point $\mathbf{p}$ in $\mathbf{D}$ **do**
    for each γ in Γ **do** ▷ For robust/Bayesian design
      Calculate the information matrix $\mathbf{M}(\xi \cup \mathbf{p})$.
      Evaluate $c_{A/D/C}(\mathbf{M}(\xi \cup p))$
    end for
  end for
  Select max/mean $\mathbf{p}$ over Γ ▷ For robust/Bayesian design
  Select $\mathbf{p}^*$ with the minimum over $\mathbf{D}$.
  $\xi \leftarrow \mathbf{p}^* \cup \xi$.
  $\mathbf{N} = \mathbf{N} - 1$.
end while
return ξ

the system that are *unknown*. Although we might not know the true parameters, we can usually give a plausible parameter set Γ. Each $\gamma \in \Gamma$ leads to a different linear constraint through the differential operator $\mathbf{L}_\gamma$. Consequently, this also defines a corresponding $\mathbf{C}_\gamma$. As a result, information matrix $\mathbf{M}_C$ has an intrinsic dependency on parameters themselves, i.e. $\mathbf{M}_C = \mathbf{M}(\gamma, \xi)$.
In general, it is impossible to find a design that is best for all elements of Γ. Therefore, we can look for designs that are optimal on average. More concretely, let $p_\Gamma(\cdot)$ be a probability measure over the compact set Γ. Then, the Bayesian design can be defined as:

$$\xi^*_\Gamma = \arg\min_\xi \int_\Gamma c(\mathbf{M}(\xi, \gamma)) p_\Gamma(d\gamma)$$

where the integral can be understood in the Lebesgue sense and c is some scalarisation of the information matrix. Throughout this work $p_\Gamma(\cdot)$ is a discrete, uniform measure with every $\gamma \in \Gamma$ weighted equally. However, prior beliefs about more likely parameters can be encoded in the prior distribution $p_\Gamma(\cdot)$.

A.5 Linear Model

In practice, we express the latent trajectory as a linear model in terms of a basis of random Fourier features initialized by Hermite-Gauss quadrature (Mutný and Krause (2018)). Let the $f(t)^\top$ be the evaluation functional, in the form of the quadrature features, and $\theta_i \in \mathcal{H}$, be an element of a Hilbert space $\mathcal{H} \subseteq \mathbb{R}^m$, such that:

$$\Phi_i(x(t), \gamma) = f(t)^\top \theta_i$$

where i corresponds to the dimension of the latent state vector. We can further generalize the multi-dimensional response case as:

$$\Phi(x(t), \gamma) = F(t)^\top \theta$$

where $F(t)^\top \in \mathbb{R}^{o \times o \cdot m}$ is a block diagonal matrix:

$$F(t)^\top = \begin{bmatrix} f(t)^\top & 0 & \cdots & 0 \\ 0 & f(t)^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(t)^\top \end{bmatrix}$$

and $\theta = (\theta_1, ..., \theta_o)$.

A.6 Estimating the Null Space C

The approach that we take to approximate the null space $\mathbf{C}_\gamma$, is based on discretization of the operator $\mathbf{L}_\gamma$. We do this by discretizing the time domain $t_d = \{t_i\}_0^T$ for which we consider the possible sampling time points of the experimental design. θ should satisfy this constraint at every discrete time point t_i and thus, we can express the discretized constraint simply as a stacked matrix of $\mathbf{L}_\gamma(t_i)$ evaluated at each time point:

$$\mathbf{L}_\gamma(t_d)\theta = \begin{bmatrix} \mathbf{L}_\gamma(t_0) \\ \mathbf{L}_\gamma(t_1) \\ \vdots \\ \mathbf{L}_\gamma(t_T) \end{bmatrix} \theta = 0$$

More specifically in our case:

$$\mathbf{L}_\gamma(t_i) = \left[\frac{dF(t_i)^\top}{dt} - \mathbf{K}(\gamma) F(t_i)^\top \right]$$

where the derivative term is obtained using automatic differentiation. One can then obtain the null space by Singular Value Decomposition (SVD) of the operator $\mathbf{L}_\gamma(t_d)$. The right eigenvectors corresponding to zero valued singular values approximate the column space of the operator $\mathbf{C}_\gamma$.

Appendix B.

B.1 Michaelis-Menten Kinetics

Michaelis-Menten kinetics (MM) is a dynamical system initially recognized in the field of enzyme kinetics. The dynamical system describes the action of the enzyme (E) acting on its substrate (S) to produce a product (P), through an enzyme-substrate complex (ES):

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \overset{k_{cat}}{\rightarrow} \mathrm{E} + \mathrm{P}$$

Under some assumptions, we can express the dynamics of the system as:

$$\frac{d[P]}{dt} = -\frac{d[S]}{dt} = \frac{\lambda \cdot [S]}{\mu + [S]}$$

where $[\cdot]$ signifies concentration, $\lambda = E_{tot} \cdot k_{cat}$ is the maximal rate of reaction and $\mu = \frac{k_{-1}+k_{cat}}{k_1}$ signifies the concentration of substrate such that reaction proceeds at half of the maximal rate. The simple model system has found its way into in a wide range of areas from industrial enzymology and environmental engineering to pharmaceutical metabolism (Saganuwan (2021)). To this day, experimental design for the specification of MM parameters remains an active area of research (Huang et al. (2020))

B.2 Experimental Details

B.2.1 Model and training

To learn the linear embeddings of the system we employ an auto-encoder neural network model architecture, similar to the one outlined in Lusch et al. (2018). Our architecture further generalizes to the different system parameters γ. We accomplish this by including the parameters in the embedding $\Phi(x, \gamma)$ and the learnable matrix $\mathbf{K}(\gamma)$. Furthermore, we learn approximate *linear differential representations* which is accomplished by propagating continuous latent dynamics using the linear operator $e^{\mathbf{K}(\gamma)t}$. The latent representation is mapped back by a decoder Ψ to the original space. Together, the state of the dynamical system given an initial condition x_0, parameters γ and a future time t is given by the composition $x(t) = \Psi \circ e^{\mathbf{K}(\gamma)t} \circ \Phi(x_0, \gamma)$.
We train the auto-encoder using a linear combination of prediction, linear, and reconstruction errors similar to Lusch et al. (2018). The neural network models were trained on trajectories generated by randomly sampling parameters $\lambda, \mu \in [-10, -0.1]$, $[10, 100]$ and the initial condition sampled from hypercubes defined by $x_0 \in [-10, 10]^2$, $[10, 200]$ for the toy and Michaelis Menten kinetics experiments respectively. The model was trained on randomly sampled time points for $t \in [0, 10]$. The latent dimension $\Phi(x, \gamma) \in \mathbb{R}^3$ and $\mathbf{K}(\gamma) \in \mathbb{R}^{3\times 3}$ for both systems. $\mathbf{K}(\cdot)$ and $\Phi(\cdot, \cdot)$ and decoder were chosen to be feed-forward multilayer perceptrons, with hidden dimension 512.

B.2.2 Experimental Design

We discretize the time domain into 100 discrete time points and employ a greedy selection algorithm with A-optimality criterion (Fedorov and Leonov (2013)) to approximate the discrete optimal designs. The main motivation for the experimental designs is to better estimate parameters γ of the dynamical system. We evaluate the designs by comparing how sampling according to optimal Bayesian design (A.4) is able to approximate parameters in comparison to picking points equidistantly. We fix the true parameters ($\lambda = 0.3$, $\mu = 2.0$ and $\lambda = 45$, $\mu = 70$ for toy and MM respectively) and the initial condition, which is assumed to be known. To generate designs, we choose a prior feasible set of parameters Γ where $\gamma = (\lambda, \mu) \in \Gamma$ by setting $\lambda \in [-1, -0.1]$, $[30, 60]$ and $\mu \in [-5, -1]$, $[50, 90]$ for the two cases respectively. We discretize Γ by making a 15×15 grid and do a Bayesian design over this discretized feasible set. We set the regularisers of the information matrix (6) to $\lambda_1 = 0.001, \lambda_2 = 5$ and the standard deviation to $\sigma = 2.0$.

B.2.3 Increasing budget experiment

We evaluate the design by estimating the parameters λ and μ with an increasing design budget over 100 realizations of noise ϵ using MLE. The observations of trajectories are corrupted with homoscedastic Gaussian noise with zero mean and variance $\sigma^2 = 4$.

B.2.4 Confidence Sets

We use universal confidence sets which ensure high probability finite sample coverage (Wasserman et al. (2020)). Naturally, parameter confidence sets are dependent on the data that is collected to estimate parameters and confidence intervals. Universal inference allows us to evaluate how the choice of sampling points influences the confidence about the parameters we estimate. To arrive at the confidence intervals we use sampling schemes obtained using Bayesian design with a budget of $n = 30$ and parameters as outlined above.