Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 58-28 (2024) 276–281

The Concept of Parameterization-invariance in System Identification Design ⋆

Simon Kuang * **Xinfan Lin** *

* *University of California, Davis; Davis, CA 95616 (e-mail: slku@ucdavis.edu, lxflin@ucdavis.edu).*

Abstract: White noise is a popular input in system identification, but it lacks the desirable property of parameterization invariance; when changing variables for the parameter and input, the transformed input distribution is generally no longer white noise. We formally define parameterization-invariance using diffeomorphism groups in the space of parameter-input pairs, and in certain cases construct invariant measures inspired by the Jeffreys prior. This view of random input connects disparate intuitions about identifiability, controllability, and the concentration of measure phenomenon.

Keywords: Modelling, Identification and Signal Processing; Estimation; Uncertain Systems and Robust Control

1. INTRODUCTION

In a system identification experiment, the statistical relationship between the (unknown) parameter, the (known) system input, and the (known) observed data is used to get a parameter estimate. In symbols: parameter estimation extracts information about θ from the log-likelihood $\ell(\theta, u; Z)$, a function of the unknown parameter θ, the system input u, and the observed data Z; which will be formally defined in Section 2.2.

A popular way to quantify this information is the Fisher information matrix (FIM), which averages a sensitivity gramian matrix over possible outcomes of Z for a given θ and u:

$$M(\theta; u) = \mathbb{E}\left[\nabla_\theta \ell(\theta, u; Z)\right]\left[\nabla_\theta \ell(\theta, u; Z)\right]^\intercal . \tag{1}$$

If a number of regularity and identifiability assumptions are satisfied, both the sampling distribution of the maximum likelihood estimator of θ and Bayesian posterior distributions for θ converge weakly to a Normal distribution with zero bias and covariance matrix $M(\theta; u)^{-1}$, scaled by (amount of data)$^{-1/2}$. On the other hand, the Cramer-Rao inequality asserts that any unbiased estimator of θ has covariance matrix at least $M(\theta; u)^{-1}$.

There are, broadly speaking, three schools of accepted practice when it comes to choosing u in order to reveal θ.

manual Design u based on intuition about the system and how different possibilities of θ should be differentiated from each other: "if a parameter is of special interest, then vary it and check where the Bode plot moves, and put the input power there" (Ljung, 1998, p. 417).

optimal Design u by minimizing the positive-semidefinite matrix $M(\theta; u)^{-1}$ in some sense: determinant, operator norm, trace, etc.; or via a proxy such as maximizing trace $M(\theta; u)$. This procedure is called optimal experiment design (OED) and, when applied to dynamic system estimation, involves optimal control.

random Sample u as a random signal with a lage number of independent components, such as white (Gaussian) noise. Variations of random excitation include filtered white noise and random binary sequences, but both of these can be understood as white noise with postprocessing (Ljung, 1998, Section 13.3).

Our paper addresses the third camp.

1.1 Pre-history of random excitation

Probability theory was defined rigorously by Kolmogorov in 1933 (English translation: Kolmogorov, 1950). Consensus developed in physics that the universe at its smallest scales was inherently nondeterministic. In diverse settings, Fokker, Planck, Kolmogorov, Wiener, Itô, and others developed theories of continuous-time random processes throughout the 20th century. This coincided with advances in real, complex, and functional analysis that enabled rigorous characterization of "rough" signals (Halperin and Schwarz, 1952). As linear control theory matured during the postwar and Cold War era, Kalman and Bucy (1961) and others modeled process, observation, and disturbance noise as random processes. At the same time, the discipline of modern statistics, the rigorous application of probability to inference from data, was achieving success and popularity at the hand of von Mises, Jaynes, Jeffreys, Fisher (an early proponent of experiment design), and others.

1.2 History of random excitation

The earliest application of random excitation might be measuring the frequency response of human hearing using noise Hirsh (1955). The fields of neuroscience and psychology continue to use noise excitation to characterize frequency dependence in physiological phenomena (Guttman et al., 1974; Møller, 1974, 1977; Marmarelis and Marmarelis, 1978; Møller, 1986). We include these examples because there is

⋆ This work was supported by the National Science Foundation CAREER Program (Grant No. 2046292).

10.1016/j.ifacol.2025.01.007

a long reception history of psychology in the cybernetic perspective to signals and systems, e.g. Braitenberg (2004).

An early example of white noise in engineering system identification is the model identification adaptive control described in Cooper et al. (1960, Chapter 2). The theory spreads to mechanical (Schmidt, 1985) and civil (Igusa, 1989) engineering. Reviews from this period include Cuenod and Sage (1968); Mehra (1974); Makhoul (1975). Mehra (1974) reads:

> Even though an enormous amount of literature exists on statistical experimental design, only a few papers relate to the input design problem. A reason for this lack of interest may be the fact that in statistical time series analysis, there is generally no controllable input.

Another reason, we suggest, is that deliberative input design only became feasible with the coming of the civilian digital era and high-fidelity signal generators such as the Hewlett-Packard noise generator type 3722 (Møller, 1974).

Today there is now an enormous amount of literature on input design. More recently, the monographs by Ljung (1998, Section 13.2), Bendat and Piersol (2010), Keesman (2011, Chapter 4), Pintelon and Schoukens (2012, Chapter 5), and Bittanti (2019, Chapter 5) propose either random noise, random binary sequence, or both as a viable excitation for estimating discrete-time linear autoregressive models. Within the autoregressive model class, the qualitative notion of *persistence of excitation*, expressed as a rank condition, determines whether the input is sufficiently rich to discriminate between models. White noise has this property. Noise is also stationary and convenient to analyze within a Laplace or z-transform framework.

System identification literature has recently adopted concentration inequalities and empirical process theory from statistical learning and random matrix theory (Zheng and Li, 2021; Oymak and Ozay, 2022; Tsiamis et al., 2022). The result is a considerable sharpening of the asymptotic theory of linear system identification in finite-sample terms.

2. PRELIMINARIES

The integers from a to b inclusive, are denoted $[a \dots b]$. The Dirac measure is written as $\int_X \delta(x,y)\,\mathrm{d}\,x = y$.

2.1 System

Let $\Theta \subseteq \mathbb{R}^{\mathsf{d}_\theta}$ be parameter space, $\mathcal{U} \subseteq \mathbb{R}^{\mathsf{d}_u}$ be input design space, $\mathcal{X} \subseteq \mathbb{R}^{\mathsf{d}_x}$ be state space, and $\mathcal{Y} \subseteq \mathbb{R}^{\mathsf{d}_y}$ be output space. Let x_0 be an initial condition and T a final time. The identification experiment permits n measurements. Let $\{t_i\}_{i\in[1\dots n]}$ be a finite subset of $[0,T]$ representing measurement times. A continuous-time dynamic system is a mapping $\mathcal{U}\times\Theta \to \mathcal{Y}$ defined in the following manner. For a given $u \in \mathcal{U}$ and $\theta \in \Theta$, we get a continuous solution trajectory $y(t;\theta,u)$:

$$x(0;\theta,u) = x_0 \tag{2a}$$
$$\dot{x}(t;\theta,u) = f(t,x(t,\theta,u);\theta,u), \quad 0 \le t \le T \tag{2b}$$
$$y(t;\theta,u) = g(t,x(t_i;\theta,u)) \tag{2c}$$

The continuous solution y is unobserved. Instead we have a random vector of all measurements by $Z = (z_i)_{i\in[1\dots n]}$ taking values in $\mathcal{Z} = \mathcal{Y}^n$.

Example 1. A continuous-time single-input single-output linear time-invariant system, where the actuation is prescribed at sample times $\{t_i\}$. Here $\mathcal{X} = \mathbb{R}^{\mathsf{d}_x}$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{U} = \mathbb{R}^n$. The function $A : \Theta \to \mathbb{R}^{n\times n}$ parameterizes square matrices. The vectors $b, c \in \mathbb{R}^n$ represent input and output gains, respectively. The functions ξ_i are basis functions obeying the interpolation conditions $\xi_j(t_i) = \delta_{ij}$, so that u can be viewed as an input sequence.

$$x(0;\theta,u) = 0 \tag{3a}$$
$$\dot{x}(t;\theta,u) = A(\theta)x(t;\theta,u) + b\sum_{i=1}^{n} u_i\xi_i(t) \tag{3b}$$
$$y_i(\theta,u) \sim \mathcal{N}(c^\intercal x(t_i;\theta,u), \sigma^2) \tag{3c}$$

Example 2. While this paper focuses on viewing u as a discrete input sequence, in full generality it is an abstract design variable that enters f alongside θ. For example, u can be the feedback parameters of a closed-loop parameter-unaware system identification law: $\dot{x}(t;\theta,u) = f_{\text{closed-loop}}(x;\theta,u) = f_{\text{open-loop}}(x;\theta) + h_{\text{feedback}}(x;u)$.

2.2 Likelihood

The model is entirely contained in a log-likelihood function $\ell : \Theta \times \mathcal{U} \times \mathcal{Z} \to \mathbb{R}$, which can be defined in great generality. [1]

In our examples, $z_i \sim \mathcal{N}(y_i, \sigma^2)$ with known σ^2, resulting in the familiar sum-of-squares log-likelihood formula

$$\ell(\theta,u;Z) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\|z_i - y(t_i,\theta,u)\|^2\,. \tag{4}$$

The function ℓ encapsulates all of the information in the parametric model.

3. MOTIVATION OF THE PROBLEM

There is a chicken-and-egg problem in input design for system identification: the choice of input depends on the yet-unknown parameters. When the input design consists in optimizing a single deterministic input, in many cases the optimization program also requires a nominal parameter value (a Dirac delta distribution) or a prior distribution over possible parameter values.

A second problem, which to our knowledge has never been raised in the literature, is that the probability distribution of random excitation is contingent upon the choice of parameterization. (A parallel in optimal experiment design can be found in Firth and Hinde (1997).)

Example 3. Two different engineers approach the same scalar system and write the following two dynamic models before commencing experimentation. There is a parameter constraint $\theta < 0$ and an input constraint $-1 \le u(t) \le 1$.

$$x(0) = x_0$$
$$\dot{x}(t) = \theta x(t) + u(t) \qquad (u\text{-model})$$
$$\dot{x}(t) = \theta x(t) + v(t)^3 \qquad (v\text{-model})$$

[1] e.g. as $\log \frac{\mathrm{d}(\mathbb{P}Z^{-1})}{\mathrm{d}\mu_{\mathcal{Y}^n}}$, the logarithm of the Radon-Nikodym derivative of the law of Z with respect to a dominating measure on $\mathcal{Z}$.

Both the u-modeler and the v-modeler have resolved to apply i.i.d. random inputs. The former claims to have a familiar "$Ax + bu$" linear system form, and therefore asserts that the most natural choice of random inputs is $u(t) \sim \text{Uniform}(-1, 1)$.[2] It seems the v-modeler has done an unnatural thing by imposing a static nonlinearity $v = u^{1/3}$ in the input path.

But the latter sees things differently. According to the v-modeler, the competing models look like:

$$\dot{x}(t) = \theta x(t) + u(t)^{1/3}$$
$$\dot{x}(t) = \theta x(t) + v(t)$$

According to the v-modeler says, the most natural choice of random inputs is $v(t) \sim \text{Uniform}(-1, 1)$; it is the u-modeler who has done a strange thing by imposing a static nonlinearity $u = v^3$.

As the situation stands, these two engineers cannot agree on a random input distribution, even though they are trying to identify the same system. Moreover, the parameterization dependency can also involve the unknown θ.

Example 4. This example modifies Example 3.

$$\begin{aligned} x(0) &= x_0 \\ \dot{x}(t) &= \theta x(t) + u(t) && (\theta, u\text{-model}) \\ \phi\dot{x}(t) &= x(t) + v(t)^3 && (\phi, v\text{-model}) \end{aligned}$$

The conversion between models is given by $u = v^3/\phi$, $\theta = 1/\phi$. This suggests that in order for any such two engineers to agree on a distribution of random inputs, they must agree not only on a distribution for u, but also on a prior distribution for θ, as parameter and input may be allowed to transform together.

3.1 Informal problem statement

System identification faces epistemic uncertainty in θ. But the epistemic uncertainty of θ and the aleatoric uncertainty of u are conmingled, so we should not expect them to be probabilistically independent.

Because of the dependence between u and θ, we combine them into a single random variable $\tilde{\theta} = (\theta, u)$ taking values in $\tilde{\Theta} = \Theta \times \mathcal{U}$. Furthermore, we will rewrite the log-likelihood as being a function on this augmented parameter space: $\ell : \tilde{\Theta} \times \mathcal{Z} \to \mathbb{R}$.

We desire a probability measure on $\tilde{\Theta}$ that is invariant under reparameterization, namely that under a change of variables $\tilde{\theta} = f(\tilde{\theta}')$, the event $\tilde{\theta} \in A$ has the same probability as $\tilde{\theta}' \in f^{-1}(A)$. It turns out that nothing is lost if $\tilde{\theta}'$ also takes values in $\tilde{\Theta}$; i.e. f is a diffeomorphism symmetry of $\tilde{\Theta}$.[3]

4. FORMAL PROBLEM STATEMENT

Let $\text{Diff}(\tilde{\Theta})$ be the diffeomorphism group of $\tilde{\Theta}$ viewed as a differentiable submanifold of $\mathbb{R}^{\mathsf{d}_\theta + \mathsf{d}_u}$. Let G be some subgroup of $\text{Diff}(\tilde{\Theta})$. Let $\mathcal{L}$ be the orbit of ℓ under the pushforward group action $\ell \mapsto \ell \circ \sigma^{-1}$ defined by $(\ell \circ \sigma^{-1})(\tilde{\theta}; Z) = \ell(\sigma^{-1}(\tilde{\theta}); Z)$. Likewise, σ acts by pushforward on measures μ by $\mu \mapsto \mu \circ \sigma^{-1}$ defined by $(\mu \circ \sigma^{-1})(A) = \mu(\sigma^{-1}(A))$.

Problem 5. Given G, find a functional from $\mathcal{L}$ to the space of probability measures on $\tilde{\Theta}$ that associates to each $\ell \in \mathcal{L}$ a continuous probability measure μ_ℓ such that for all $\sigma \in G$,

$$\mu_\ell \circ \sigma^{-1} = \mu_{\ell \circ \sigma^{-1}}.$$

The group G represents the degrees of reparameterization freedom we are willing to admit. In examples 3 and 4, the dissenting parties should come to agreement on G and realize that their respective parameterizations differ by the action of some $\sigma \in G$. We abandon the notion that any one log-likelihood function is more correct than the others. Rather, from the perspective of someone who uses the parameter $\tilde{\theta}$ with log-likelihood ℓ, another person who prefers the parameterization $\sigma(\tilde{\theta})$ will use the log-likelihood $\ell \circ \sigma^{-1}$. What the problem requires is that the entire orbit of G will enjoy compatible probabilities.

Suppose that $A \subset \tilde{\Theta}$ is an event. The second person calls this event $\sigma(A)$. Problem 5 requires that they should agree on the probability:

$$\mu_\ell(A) = \mu_{\ell \circ \sigma^{-1}}(\sigma(A)). \tag{5}$$

Let density functions exist such that $\frac{\mathrm{d}\mu_\ell}{\mathrm{d}\lambda} = \pi_\ell$ and $\frac{\mathrm{d}\mu_{\ell \circ \sigma^{-1}}}{\mathrm{d}\lambda} = \pi_{\ell \circ \sigma^{-1}}$, where λ is Lebesgue measure. Then (5) becomes

$$\int_A \pi_\ell(\tilde{\theta}) \,\mathrm{d}\,\tilde{\theta} = \int_{\sigma(A)} \pi_{\ell \circ \sigma^{-1}}(\tilde{\theta}') \,\mathrm{d}\,\tilde{\theta}' \tag{6}$$

Applying the multivariable change-of-variables formula to $\tilde{\theta}' = \sigma(\tilde{\theta})$, we get the volume form $\mathrm{d}\,\tilde{\theta} = \left|\nabla\sigma(\tilde{\theta})\right| \mathrm{d}\,\tilde{\theta}$.

$$= \int_A \pi_{\ell \circ \sigma^{-1}}(\sigma(\tilde{\theta})) \left|\nabla\sigma(\tilde{\theta})\right| \mathrm{d}\,\tilde{\theta} \tag{7}$$

This must hold true simultaneously for all A. Therefore Problem 5 may be restated as finding a density assignment $\ell \to \pi_\ell$ that satisfies the transformation law

$$\pi_\ell(\tilde{\theta}) = \pi_{\ell \circ \sigma^{-1}}(\sigma(\tilde{\theta})) \left|\det \nabla\sigma(\tilde{\theta})\right|. \tag{8}$$

5. SOLUTIONS TO INSTANCES OF PROBLEM 5

There are many choices of G. We work and discuss the most general case in detail. Sections 5.2 and 5.4 are two solutions for a smaller G, and Section 5.5 is a solution for a yet smaller G.

5.1 Case: $G = \text{Diff}(\tilde{\Theta})$

Let $\xi : \mathbb{R} \to \mathbb{R}$ be any smooth function. Define

$$\pi_\ell(\tilde{\theta}) \propto \sqrt{\det \mathbb{E}_Z \left[\nabla(\xi \circ \ell)(\tilde{\theta}; Z)\right] \left[\nabla(\xi \circ \ell)(\tilde{\theta}; Z)\right]^\intercal}, \tag{9}$$

where ∇ is gradient with respect to the first argument, and the proportionality allows normalization to ensure that $\int_{\tilde{\Theta}} \pi_\ell(\tilde{\theta}) \,\mathrm{d}\,\tilde{\theta} = 1$. Next we verify (8). By the chain rule,

$$\nabla(\xi \circ \sigma(\ell))(\tilde{\theta}'; Z) \tag{10}$$
$$= \left(\nabla\sigma^{-1}(\tilde{\theta}')\right) \nabla(\xi \circ \ell)(\sigma^{-1}(\tilde{\theta}'); Z) \tag{11}$$

[2] Or any other distribution on $[-1, 1]$ such as Beta or truncated Normal.

[3] A similar interpretation can be found in Jermyn (2005).

By the inverse function theorem,

$$= \left(\nabla\sigma(\sigma^{-1}(\tilde{\theta}'))\right)^{-1} \nabla(\xi \circ \ell)(\sigma^{-1}(\tilde{\theta}'); Z) \tag{12}$$

Inserting this into the definition of π_ℓ,

$$\pi_{\sigma(\ell)}(\tilde{\theta}') \propto \left|\det \nabla\sigma(\sigma^{-1}(\tilde{\theta}'))\right| \pi_\ell(\sigma^{-1}(\tilde{\theta}')) \tag{13}$$

from which (8) follows by substituting $\tilde{\theta}' = \sigma(\tilde{\theta})$.

By the chain rule, (9)

$$\pi_\ell(\tilde{\theta}) \propto \sqrt{\det \mathbb{E}_Z\, \xi'(\ell(\tilde{\theta}; Z))^2 \left[\nabla\ell(\tilde{\theta}; Z)\right] \left[\nabla\ell(\tilde{\theta}; Z)\right]^\intercal}, \tag{14}$$

We are not sure what to make of the freedom in choosing ξ. In the rest of the paper, ξ will be the identity map. Then π_ℓ can be expressed in terms of the Fisher information matrix $M_{\tilde{\theta}}(\tilde{\theta})$.

$$\propto \sqrt{\det M_{\tilde{\theta}}(\tilde{\theta})}. \tag{15}$$

The parameterization-invariance of π_ℓ can also be expressed in the following way: the FIM transforms as a symmetric tensor with two covariant indices.[4] With this property in mind, Jeffreys (1946) proposed such a procedure to generate noninformative priors for Bayesian inference. In this view, π_ℓ is a Jeffreys prior for the contrived situation of Bayesian estimation of (θ, u) from Z. However, we avoid construing π_ℓ as an inference prior, because Jeffreys priors can be poor choices for Bayesian inference (Bernardo and Smith, 1994; Berger et al., 2015).

5.2 Case: G fixes θ

Recall that the elements of $\tilde{\Theta}$ are pairs (θ, u). Let us solve Problem 5 for the following transformation group:

$$G = \{(\theta, u) \mapsto (\theta, \sigma(u)) \mid \sigma \in \mathrm{Diff}(\mathcal{U})\}. \tag{16}$$

This group expresses the understanding that θ and u are recognized as distinct entities; we do not entertain reparameterizations such as Example 4 that couple θ and u. There is no room to disagree about the parameterization of θ. While the distribution constructed in Section 5.1 works, a simpler π_ℓ also solves this problem:

$$\pi_{\theta,u}(\tilde{\theta}) \propto \sqrt{\det \mathbb{E}_Z \left[\nabla_u \ell(\theta, u; Z)\right] \left[\nabla_u \ell(\theta, u; Z)\right]^\intercal}. \tag{17}$$

5.3 Case: G isolates elements of u

Let U be an interval, and suppose that $\mathcal{U} = U^{\mathsf{d}_u}$ is a cube. Let us solve Problem 5 for the following transformation group:

$$\begin{aligned} G = \big\{&(\theta, u_i) \mapsto (\sigma(\theta), \tau_i(u_i)) \\ &\mid \sigma_\theta \in \mathrm{Diff}(\Theta), \tau_i \in \mathrm{Diff}(U)\big\}. \end{aligned} \tag{18}$$

This group expresses the elements of $u \in \mathcal{U}$ have meanings, such as sampled function values, that should not be conmingled with either θ or each other. We can satisfy G with the following distribution:

$$\pi_\ell(\theta, u) \propto \sqrt{\det M_\theta(\theta, u)} \prod_{i=1}^{\mathsf{d}_u} \sqrt{\mathbb{E}\left[\frac{\partial}{\partial u_i}\ell(\theta, u; Z)\right]^2}, \tag{19}$$

$$= \sqrt{\det M_\theta(\theta, u)} \prod_{i=1}^{\mathsf{d}_u} \sqrt{M_{u_i}(\theta, u)}, \tag{20}$$

which has the advantage of scaling better in d_u; unlike Section 5.1, one does not have to compute the determinant of a $(\mathsf{d}_\theta + \mathsf{d}_u) \times (\mathsf{d}_\theta + \mathsf{d}_u)$ matrix.

5.4 Independence

Let us retain the G of the previous section and solve the same instantiation of Problem 5. We give an alternative solution with the property that the elements of u are independent.

Let π_ℓ be defined as in (19).

$$\overline{\pi}_\ell(\theta, u) \propto \pi^\theta_\ell(\theta)\pi^u_\ell(u), \tag{21a}$$

$$\pi^\theta_\ell(\theta) \propto \int_{\mathcal{U}} \pi_\ell(\theta, u')\,\mathrm{d}\,u', \tag{21b}$$

$$\pi^u_\ell(u) \propto \prod_{i=1}^{\mathsf{d}_u} \int_\Theta \int_{\mathcal{U}} \pi_\ell(\theta', u')\delta(u_i, u'_i)\,\mathrm{d}\,u'\,\mathrm{d}\,\theta'. \tag{21c}$$

Each u_i is now independent from the others and from θ. This means that it is actually not necessary to deal with π^θ_ℓ if we are only interested in generating random excitations by sampling the marginal distribution of u, which has density π^u_ℓ.

This independence structure is reminiscent of traditional random experiments such as white noise and random binary sequences, save that the u_i are not identically distributed.

5.5 Case: generalized white noise

We may prefer that, like in the case of white noise, $\{u_i\}$ be identically distributed. This section shows how this preference can be translated into an instance of Problem 5. Let U be an interval, and suppose that $\mathcal{U} = U^{\mathsf{d}_u}$ is a cube. Let us solve Problem 5 for the following transformation group:

$$\begin{aligned} G = \big\{&(\theta, u_i) \mapsto (\sigma(\theta), \tau(u_i)) \\ &\mid \sigma_\theta \in \mathrm{Diff}(\Theta), \tau \in \mathrm{Diff}(U)\big\}. \end{aligned} \tag{22}$$

In contrast with the group (18), this G mandates that a single $\tau \in \mathrm{Diff}(U)$ acts elementwise on every u_i. There are yet fewer degrees of freedom.

Let π^θ_ℓ and π^u_ℓ be as in (21). Let S be the permutation group on d_u items. Then we may define the following density,

$$\pi^{u_i}_\ell(u_i) \propto \left[\prod_{\rho \in S} \prod_{j=1}^{\mathsf{d}_u} \int_\Theta \int_{\mathcal{U}} \pi_\ell(\theta', u')\delta(u_{\rho(i)}, u'_i)\,\mathrm{d}\,u'\,\mathrm{d}\,\theta'\right]^{\frac{1}{\mathsf{d}_u!}}, \tag{23}$$

which makes u_i independent and identically distributed.

6. SPECIFIC FORM FOR A GAUSSIAN SYSTEM IDENTIFICATION LIKELIHOOD

The likelihood functions ℓ have been abstract. Now we give the exact form of Section 5.1's π_ℓ for the Gaussian measurement likelihood ℓ defined in (4). Under this measurement

[4] Tensor covariance also means that the FIM defines a Riemannian metric whose parameterization-invariance has attracted interest in machine learning (Martens, 2020). Our density π_ℓ corresponds to the unsigned volume form on this Riemannian manifold.

model, the Fisher Information Matrix of $\tilde{\theta}$ is proportional to a sensitivity gramian.

$$\pi_\ell(\theta, u) \propto \sqrt{\det \begin{bmatrix} M_{\theta\theta} & M_{\theta u} \\ M_{\theta u}^\intercal & M_{uu} \end{bmatrix}}, \tag{24a}$$

$$M_{\theta\theta} = \sum_{i=1}^{n} \nabla_\theta y(t_i, \theta, u) \left(\nabla_\theta y(t_i, \theta, u)\right)^\intercal \tag{24b}$$

$$M_{\theta u} = \sum_{i=1}^{n} \nabla_\theta y(t_i, \theta, u) \left(\nabla_u y(t_i, \theta, u)\right)^\intercal \tag{24c}$$

$$M_{uu} = \sum_{i=1}^{n} \nabla_u y(t_i, \theta, u) \left(\nabla_u y(t_i, \theta, u)\right)^\intercal \tag{24d}$$

Or, using a block matrix determinant formula,

$$\pi_\ell(\theta, u) \propto \sqrt{\det M_{\theta\theta} \det \left(M_{uu} - M_{\theta u}^\intercal M_{\theta\theta}^{-1} M_{\theta u}\right)} \tag{25}$$

The matrix $M_{\theta\theta}$ is the FIM of θ. Thus, π_ℓ favors excitations with a higher Fisher D-criterion. The matrix M_{uu} is a sort of nonlinear controllability gramian, which gets penalized by the local linear correlation between control effectiveness and parameter sensitivity. An interpretation is that u and θ are competing to find expression in y. If small perturbations δu have the same impact on y as small perturbations $\delta\theta$, then u and θ are talking over each other.

It is possible to integrate (25) over Θ to get a heuristic for optimal, rather than random, design of u.

7. NUMERICAL EXAMPLE

This model is a first-order linear model with unknown eigenvalue $\theta \in (\lambda_{\min}, \lambda_{\max})$. The function $h(\cdot; u)$ is the polynomial of degree d_u that interpolates the points (s_i, u_i), where s_i are the d_u Chebyshev points of the first kind on $[0, T]$. After constraining $u_i \in [u_{\min}, u_{\max}]$, the result is that h parameterizes a large class of bounded inputs. The output is x, sampled at the n equidistant points $t_i \in [0, T]$.

$$x(0; \theta, u) = 0 \tag{26}$$

$$\dot{x}(t; \theta, u) = \theta x(t; \theta, u) + h(t; u), \quad 0 \le t \le T \tag{27}$$

Relevant constants are available in Table 1.

The log-likelihood is Gaussian with known variance, so π_ℓ follows Section 6. We used Hamiltonian Monte Carlo, an advanced Markov Chain Monte Carlo (MCMC) technique for sampling from high-dimensional continuous distributions (Betancourt, 2018), to take 2,000 samples from the joint distribution of (θ, u). The marginal distribution of θ, seen in Fig. 1, results from dropping u in samples of (θ, u). It can be viewed as an uninformative prior for the unknown θ.

Twenty samples from the marginal distribution of u are seen in Fig. 2. Visual inspection shows that these samples have a rich frequency content and wide dynamic range, qualities are heuristically associated with high values of the Fisher information. An explanation is that high-dimensional distributions assign most of their probability to a "typical set" resembling a thin shell around the mode, which in this case corresponds to certain Fisher information-maximizing input signals. This intuition about concentration of measure in the "typical set" fits the data. On one hand, the random excitation signals tend to avoid signals with very high information such as impulses. On the other hand, they equally avoid low-information signals such as those with low total power.

Variable	Meaning	Value
$\lambda_{\min}$	*a priori* lower bound on θ	-3
$\lambda_{\max}$	*a priori* upper bound on θ	-1
$u_{\min}$	lower bound on input	-1
$u_{\max}$	lower bound on input	1
d_u	degrees of freedom in input	10
n	number of sample points	20
T	terminal time	1

Table 1. Constants used when sampling random inputs for a scalar linear system.

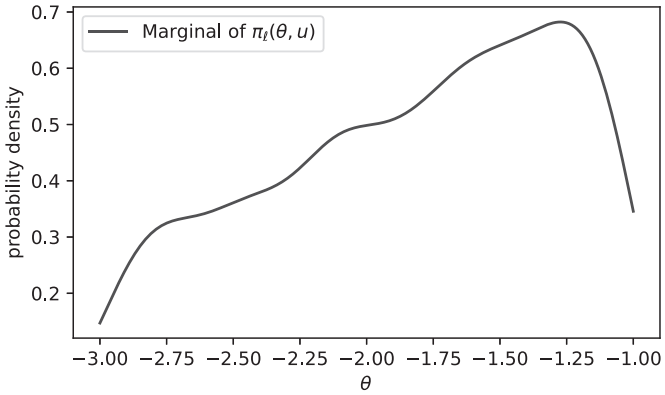

Fig. 1. Kernel density estimate of marginal distribution of θ in $\pi_\ell(\theta, u)$ from MCMC samples.

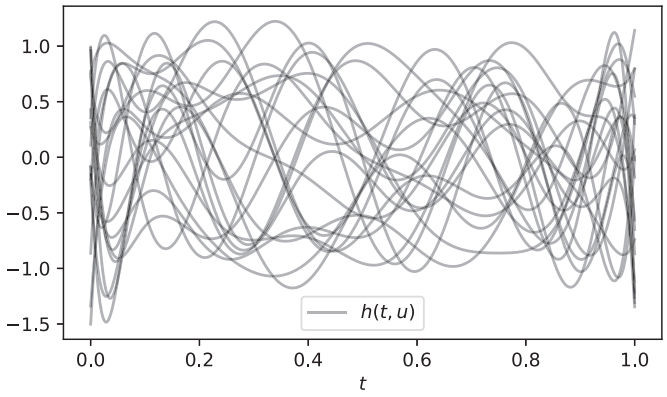

Fig. 2. Twenty random inputs distributed as $\pi_\ell(\theta, u)$, derived from MCMC samples.

8. CONCLUSION

Randomness is parameterization-dependent: the same noise signal can look very different in two diffeomorphically equivalent formulations of the same system identification task. We pose a novel problem of parameterization-invariance and solve it for different cases of symmetry groups. Our solution builds a bridge between identifiability (local parameter sensitivity) and controllability (local input sensitivity) and advances the observability-controllability duality attested by the Luenberger observer and its descendants. We hope it will raise new questions in the subject of input design for nonlinear identification.

Like optimal design, our invariant distributions favor high values of local parameter identifiability; like traditional

random excitation, our invariant distributions exhibit an empirical concentration of measure phenomenon. Our high-dimensional invariant distributions may be susceptible to non-independent concentration of measure phenomena (Vershynin, 2018, Chap. 5). Problem 5 has a large number of degrees of freedom, such as the likelihood gauge function ξ used in (9), that may be potentially optimized.

REFERENCES

Bendat, J.S. and Piersol, A.G. (2010). *Random data: analysis and measurement procedures*. Wiley series in probability and statistics. Wiley, Hoboken, N.J, 4th ed edition.

Berger, J.O., Bernardo, J.M., and Sun, D. (2015). Overall Objective Priors. *Bayesian Analysis*, 10(1), 189–221. Publisher: International Society for Bayesian Analysis.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Betancourt, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. ArXiv:1701.02434 [stat].

Bittanti, S. (2019). *Model identification and data analysis*. Wiley, Hoboken, NJ.

Braitenberg, V. (2004). *Vehicles: experiments in synthetic psychology*. Bradford book psychology. MIT Press, Cambridge, Mass., 9. print edition.

Cooper, G., E, G., W, E., C, L., S, M., and H, R. (1960). A Survey of the Philosophy and State of the Art of Adaptive Systems. *Department of Electrical and Computer Engineering Technical Reports*.

Cuenod, M. and Sage, A.P. (1968). Comparison of some methods used for process identification. *Automatica*, 4(4), 235–269.

Firth, D. and Hinde, J.P. (1997). Parameter Neutral Optimum Design for Non-linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4), 799–811. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00097.

Guttman, R., Feldman, L., and Lecar, H. (1974). Squid Axon Membrane Response to White Noise Stimulation. *Biophysical Journal*, 14(12), 941–955.

Halperin, I. and Schwarz, L. (1952). *Introduction to the Theory of Distributions*. University of Toronto Press.

Hirsh, I.J. (1955). Hearing. *Annual Review of Psychology*, 6(1), 95–118.

Igusa, T. (1989). Characteristics of Response to Nonstationary White Noise: Theory. *Journal of Engineering Mechanics*, 115(9), 1904–1918. Publisher: American Society of Civil Engineers.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007), 453–461.

Jermyn, I.H. (2005). Invariant Bayesian estimation on manifolds. *The Annals of Statistics*, 33(2). ArXiv:math/0506296.

Kalman, R.E. and Bucy, R.S. (1961). New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering*, 83(1), 95–108.

Keesman, K.J. (2011). *System identification: an introduction*. Advanced textbooks in control and signal processing. Springer, London ; New York. OCLC: ocn747112444.

Kolmogorov, A.N.A.N. (1950). *Foundations of the theory of probability*. Chelsea Pub. Co., New York.

Ljung, L. (1998). *System Identification: Theory for the User, 2nd Edition*.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 561–580. Conference Name: Proceedings of the IEEE.

Marmarelis, P.Z. and Marmarelis, V.Z. (1978). The White-Noise Method in System Identification. In P.Z. Marmarelis and V.Z. Marmarelis (eds.), *Analysis of Physiological Systems: The White-Noise Approach*, 131–180. Springer US, Boston, MA.

Martens, J. (2020). New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146), 1–76.

Mehra, R. (1974). Optimal input signals for parameter estimation in dynamic systems–Survey and new results. *IEEE Transactions on Automatic Control*, 19(6), 753–768. Conference Name: IEEE Transactions on Automatic Control.

Møller, A. (1974). Use of stochastic signals in evaluation of the properties of a neuronal system. *Scandinavian journal of rehabilitation medicine. Supplement*, 3, 37–44.

Møller, A.R. (1977). Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli. *The Journal of the Acoustical Society of America*, 62(1), 135–142.

Møller, A.R. (1986). Systems identification using pseudo-random noise applied to a sensorineural system. *Computers & Mathematics with Applications*, 12(6, Part A), 803–814.

Oymak, S. and Ozay, N. (2022). Revisiting Ho–Kalman-Based System Identification: Robustness and Finite-Sample Analysis. *IEEE Transactions on Automatic Control*, 67(4), 1914–1928. Conference Name: IEEE Transactions on Automatic Control.

Pintelon, R. and Schoukens, J. (2012). *System identification: a frequency domain approach*. John Wiley & Sons Inc, Hoboken, N.J, 2nd ed edition.

Schmidt, H. (1985). Resolution bias errors in spectral density, frequency response and coherence function measurements, III: Application to second-order systems (white noise excitation). *Journal of Sound and Vibration*, 101(3), 377–404.

Tsiamis, A., Ziemann, I., Matni, N., and Pappas, G.J. (2022). Statistical Learning Theory for Control: A Finite Sample Perspective.

Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science. ISBN: 9781108231596 9781108415194 Publisher: Cambridge University Press.

Zheng, Y. and Li, N. (2021). Non-Asymptotic Identification of Linear Dynamical Systems Using Multiple Trajectories. *IEEE Control Systems Letters*, 5(5), 1693–1698. Conference Name: IEEE Control Systems Letters.