

A Illustrative Gaussian Example: Additional Results

Table 4: Displays the mean squared error (MSE) between the ground truth log likelihood ratio and each neural ratio estimator using the results from Figure 1.

APPROACH	MSE $\sigma = 0.1$	MSE $\sigma = 0.3$	MSE $\sigma = 0.5$
NRE	0.136	0.207	0.759
BNRE	3.584	4.289	3.693
DNRE	0.104	0.122	0.124

B Summary of Benchmark Description

The benchmark results of Section 5.3 are the same as those from the Simulation-Based Inference Benchmark (Lueckmann et al. 2021). Here we summarize each task by providing the dimensionality and priors but refer to the original benchmark for further details:

- **Two Moons (TM):** $\theta \in \mathbb{R}^2$; $\mathbf{x} \in \mathbb{R}^2$; $\theta \sim \mathcal{U}(-1, 1)$
- **Gaussian Linear (GL):** $\theta \in \mathbb{R}^{10}$; $\mathbf{x} \in \mathbb{R}^{10}$; $\theta \sim \mathcal{N}(\mathbf{0}, 0.1\mathbf{I})$
- **Lotka-Volterra (LV):** $\theta \in \mathbb{R}^4$; $\mathbf{x} \in \mathbb{R}^{20}$; $\theta_1 \sim \text{LogNormal}(-0.125, 0.5)$, $\theta_2 \sim \text{LogNormal}(-3, 0.5)$, $\theta_3 \sim \text{LogNormal}(-0.125, 0.5)$, $\theta_4 \sim \text{LogNormal}(-3, 0.5)$
- **SIR Epidemiological Model (SIR):** $\theta \in \mathbb{R}^2$; $\mathbf{x} \in \mathbb{R}^{10}$; $\theta_1 \sim \text{LogNormal}(\log(0.4), 0.5)$, $\theta_2 \sim \text{LogNormal}(\log(1/8), 0.2)$
- **Simple Likelihood Complex Posterior (SLCP):** $\theta \in \mathbb{R}^5$; $\mathbf{x} \in \mathbb{R}^8$; $\theta \sim \mathcal{U}(-3, 3)$
- **Gaussian Linear Uniform (GLU):** $\theta \in \mathbb{R}^{10}$; $\mathbf{x} \in \mathbb{R}^{10}$; $\theta \sim \mathcal{U}(-1, 1)$
- **Simple Likelihood Complex Posterior with Distractors (SLCP D):** $\theta \in \mathbb{R}^5$; $\mathbf{x} \in \mathbb{R}^{100}$; $\theta \sim \mathcal{U}(-3, 3)$
- **Bernoulli Generalized Linear Model (B GLM):** $\theta \in \mathbb{R}^{10}$; $\mathbf{x} \in \mathbb{R}^{10}$; $\theta_1 \sim \mathcal{N}(0, 2)$, $\theta_{2:10} \sim \mathcal{N}(\mathbf{0}, (\mathbf{F}^\top \mathbf{F})^{-1})$, $\mathbf{F}_{i,i-2} = 1$, $\mathbf{F}_{i,i-1} = 1$, $\mathbf{F}_{i,i} = 1 + \sqrt{(i-1)/9}$, $\mathbf{F}_{i,j} = 0$ otherwise
- **Gaussian Mixture (GM):** $\theta \in \mathbb{R}^2$; $\mathbf{x} \in \mathbb{R}^2$; $\theta \sim \mathcal{U}(-10, 10)$
- **Bernoulli Generalized Linear Model Raw (B GLM R):** $\theta \in \mathbb{R}^{10}$; $\mathbf{x} \in \mathbb{R}^{100}$; $\theta_1 \sim \mathcal{N}(0, 2)$, $\theta_{2:10} \sim \mathcal{N}(\mathbf{0}, (\mathbf{F}^\top \mathbf{F})^{-1})$, $\mathbf{F}_{i,i-2} = 1$, $\mathbf{F}_{i,i-1} = 1$, $\mathbf{F}_{i,i} = 1 + \sqrt{(i-1)/9}$, $\mathbf{F}_{i,j} = 0$ otherwise

The C2ST score is a classifier-based test that requires access to ground truth samples from the posterior for each observation, \mathbf{x} . In the simulation-based inference benchmark, there are ten observations with corresponding reference posterior samples that allow us to use the C2ST score. We directly use the code from the benchmark which trains a two-layer neural network with a number of units that is ten times the dimension of θ . The parameters are normalized and the

accuracy is reported using 5-fold cross validation. An accuracy score of 0.5 means that the samples from likelihood-free MCMC are indistinguishable from the ground truth and corresponds to the best performance.

B.1 Estimated Posterior under Ground Truth θ

Table 5 displays the average log posterior probabilities across all observations for each estimator when applied to the ground truth observation and parameter pairs. In addition to seeing that DNRE tends to assign the highest average log posterior probability, we also note that the results provide a further confirmation that our Monte Carlo estimate for DNRE’s posterior is working well. For these results we use 10,000 Monte Carlo samples.

C Quadcopter Design

Observations and Parameters The observation, $\mathbf{x} \in \mathbb{R}^7$ contains: the number of interferences; the mass (Kg); the maximum flight distance (meters); the maximum hover time (seconds); the maximum lateral speed (m/s); the maximum control input at the maximum flight distance ($u \in [0, 1]$); and the maximum power at the maximum speed (watts). The quadcopter design parameters, $\theta \in \mathbb{R}^{19}$, and priors are given by:

- Arm Length (mm) $\sim \mathcal{U}(50, 500)$.
- Fuselage Floor Height (mm) $\sim \mathcal{U}(5, 80)$.
- Fuselage Horizontal Diameter (mm) $\sim \mathcal{U}(250, 1000)$.
- Fuselage Vertical Diameter (mm) $\sim \mathcal{U}(40, 400)$.
- Fuselage Autopilot X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage Autopilot Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage Battery X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage Battery Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage Current X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage Current Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage GPS X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage GPS Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage Length (mm) $\sim \mathcal{U}(50, 200)$.
- Fuselage RPM X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage RPM Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage Variometer X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage Variometer Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.
- Fuselage Voltage X Offset (mm) $\sim \mathcal{U}(-100, 100)$.
- Fuselage Voltage Y Offset (mm) $\sim \mathcal{U}(-50, 50)$.

Data Figure 7 displays the empirical distribution of the 4,490 training and validation observations, \mathbf{x} . The design objective as specified in Section 5.6 is superimposed on the figure.

Additional results Complementary to Figure 5 in Section 5.6, Figures 8 and 9 show corner plots for likelihood-free MCMC using NRE and BNRE respectively. The distributions are consistent in their shape and the relationships that they display. These figures are plotted using the *LAMPE* simulation-based inference library (Rozet 2022).

Table 5: Log posterior of ground truth parameters, θ^* , averaged over reference observations using the SBI Benchmark. The log posterior of DNRE is estimated according to Equation (7) with 10,000 Monte Carlo samples.

APPROACH	TM	GL	LV	SIR	SLCP
NRE	3.856 ± 0.447	-0.735 ± 1.965	9.653 ± 2.866	7.037 ± 0.982	-3.99 ± 1.981
BNRE	3.778 ± 0.626	-1.014 ± 1.966	8.193 ± 1.949	4.912 ± 1.133	-3.995 ± 1.337
DNRE	3.701 ± 0.909	-0.317 ± 2.527	8.035 ± 1.916	6.546 ± 0.685	-2.914 ± 1.418
	GLU	SLCP D	B GLM	GM	B GLM R
NRE	-0.555 ± 1.395	-9.193 ± 0.353	-4.392 ± 2.192	-2.06 ± 0.542	-5.109 ± 2.182
BNRE	-0.424 ± 1.279	-9.303 ± 0.389	-4.669 ± 2.534	-2.188 ± 0.478	-5.124 ± 1.875
DNRE	1.042 ± 2.829	-7.54 ± 1.13	-3.37 ± 3.191	-2.101 ± 0.519	-4.173 ± 3.05

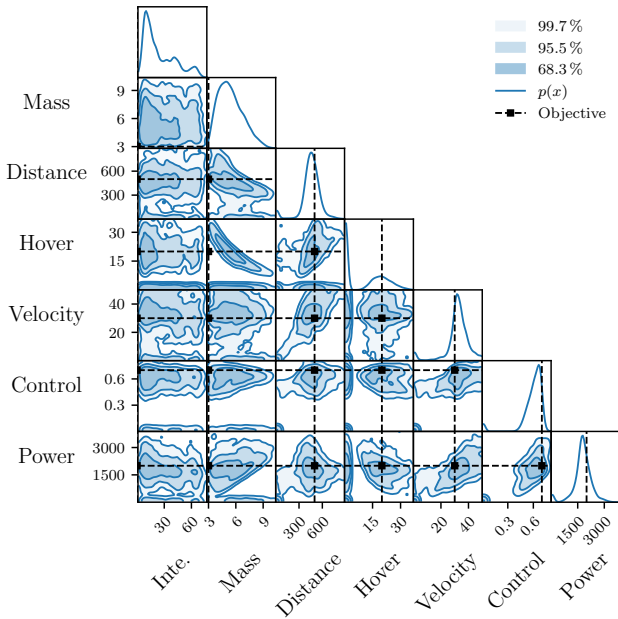


Figure 7: Corner plot displaying the empirical distribution of the observations, x , of the training data as described in Section 5.6. The objective, highlighted by the black square, corresponds to the design specification used in Section 5.6’s experimentation. A notable challenge with this specification can be seen in the Hover vs. Mass density map. The jointly desired mass and maximum hover time falls outside the density of the training data.

Case Study of Improving an Existing Design Another interesting use-case of SBI is to use our likelihood ratio estimator to improve an existing design. For example, we may like the performance of an existing quadcopter, but dislike the number of structural interferences. Figure 10 shows designs superimposed on the same empirical training distribution as in Figure 7. Here, the black square is the current design that we are interested in improving. Specifically, the current design has 60 structural interferences as highlighted in the first column. The other attributes, such as a hover time of 24 s and a maximum velocity of 33 m/s, we would like to keep as we reduce the number of interferences.

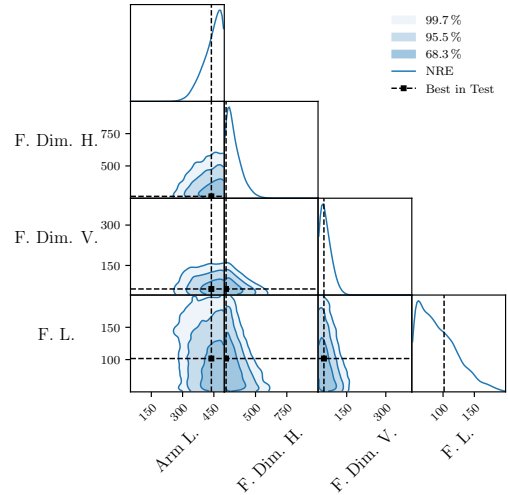


Figure 8: Corner plot displaying materialized samples from likelihood-free inference using NRE for a subset of the parameters as highlighted in Figure 4. The black square corresponds to the highest ranked test design in the test set.

As an experiment, we initialize HMC with this initial design and perform likelihood-free inference with DNRE, with a small step size of 0.1 and a simple trajectory length of 1. We take 200 steps with a thinning of 8 and run each design through the simulator. The final proposed design is shown by the red square in Figure 10 and only has 4 interferences. This closely matches our design objective and further highlights the utility of utilizing HMC for likelihood-free inference, as well as our new direct neural likelihood ratio estimator. Figure 11 directly shows sub-sampled designs that feature along the Markov chain after running the simulator. We see the evolution of our seed design from having a significant number of interferences to almost none.

D Computing Infrastructure

To reproduce the results of this paper, it is preferable to train the neural estimators using GPUs. In this paper we use an NVIDIA RTX A6000 to train all our models but smaller GPUs can be used since the neural network architectures do not take up significant memory.

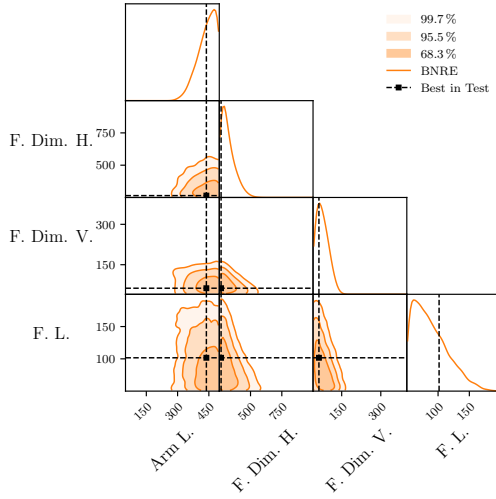


Figure 9: Corner plot displaying materialized samples from likelihood-free inference using BNRE for a subset of the parameters as highlighted in Figure 4. The black square corresponds to the highest ranked test design in the test set.

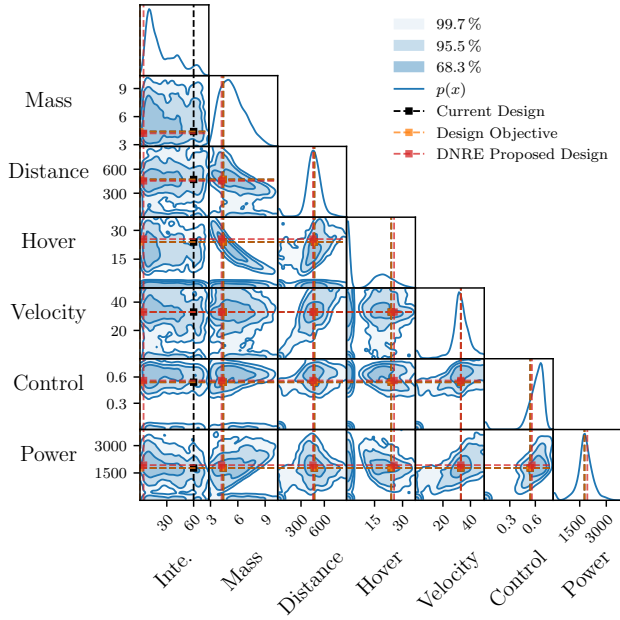


Figure 10: Experiment to improve an existing design. The current design, denoted via the black square, has multiple structural interferences as seen in the first column. Our objective, highlighted in orange, is to keep all the design observations that same while reducing the number of interferences. The red square is the final sample of an HMC chain using DNRE for performing likelihood-free inference. This last sample is evaluated on the flight simulator and is shown to match closely to the design objective. These points are superimposed on the empirical marginal distribution of the observations for context.



Figure 11: Sub-sampled quadcopter designs taken along the likelihood-free HMC chain using DNRE. The initial seed design on the far left has multiple structural interferences, including sensors that cut through the fuselage. As we move along the chain we see the design morph into our desired structure with very few interferences. This is achieved though increasing the arm length and changing the shape of the fuselage jointly with varying the placement of the interior sensing components.