How well does your sampler really work?
Supplementary Material

Ryan Turner
Uber AI Labs

Brady Neal
MILA, Université de Montréal

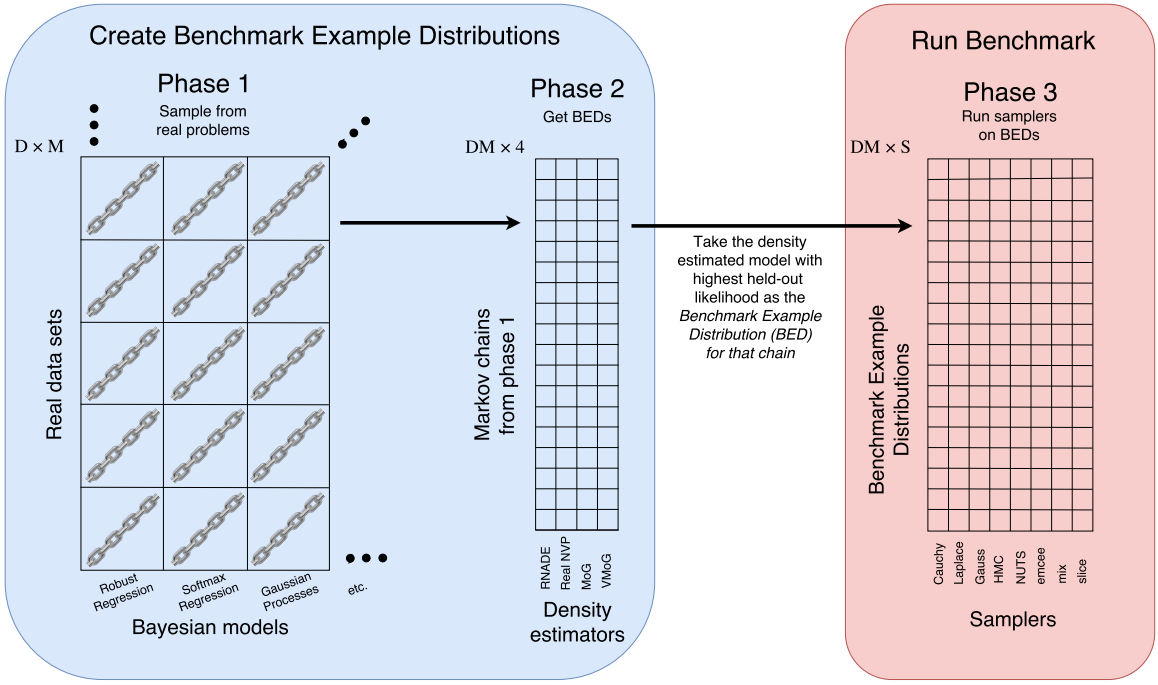

Figure 1: Flowchart graphically depicting phases 1–3. The phases show concrete examples of the methods run at each phase in the analysis. The blue panel shows the steps done to generate benchmark example densities. The red panel is rerun when a new sampler is provided while the output of the blue panel remains fixed. Phase 1 shows the Markov chains run on all combinations of D data sets and M different models. Phase 2 shows how the 4 density estimation methods that form the benchmark example densities (i.e., surrogates for the real posteriors) are run on all $D \times M$ posteriors from phase 1. Finally, phase 3 shows that the resulting $D \times M$ benchmark example densities are sampled by S different methods. The analysis phases 4 and 5 are not shown in this figure.

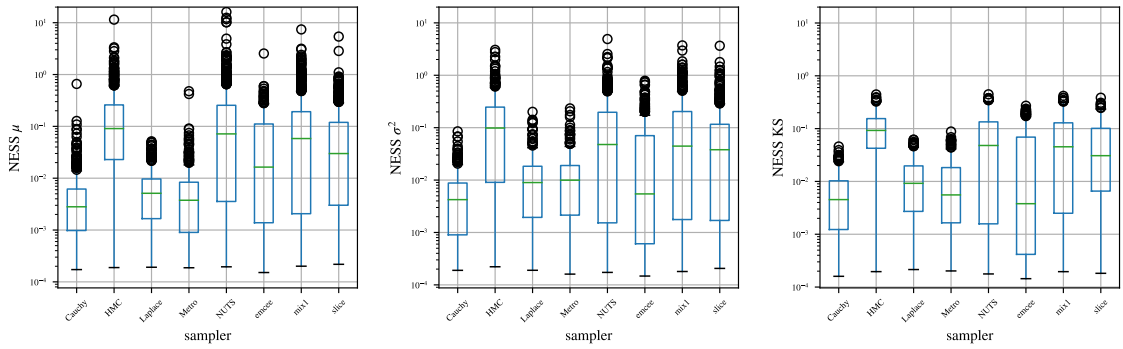

Figure 2: The box plots demonstrate the distribution on NESS calculated from the mean μ (left), variance σ^2 (center), and KS distance (right). These are conditional on the sampler achieving an RESS of at least 12 to only show the mode where the samplers don't completely fail. The distribution on KS scores are naturally tighter because KS distance is upper bounded at 1 unlike the MSE on mean or variance, which are unbounded.

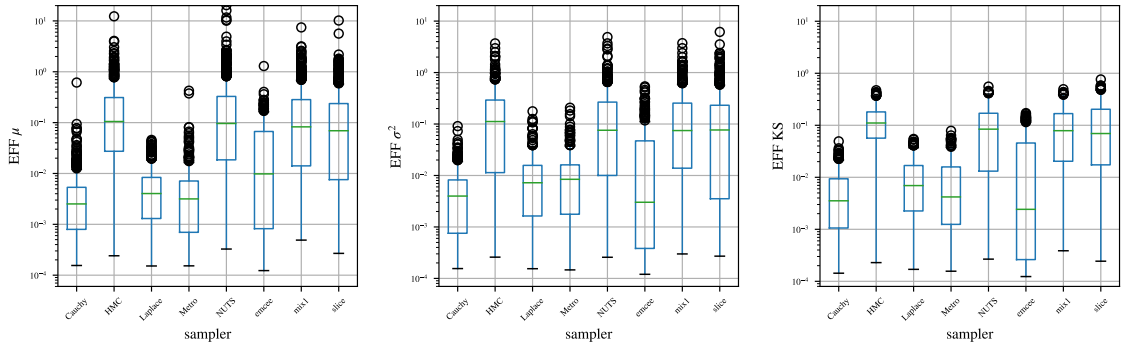

Figure 3: Figure analogous to Figure 2, but for efficiency. Again, the distribution on KS scores are naturally tighter because KS distance is a bounded quantity.