# Understanding posterior projection effects with normalizing flows

**Marco Raveri** [1] **Cyrille Doux** [2] **Shivam Pandey** [3]

## Abstract

Many modern applications of Bayesian inference, such as cosmology, are based on complicated forward models with high-dimensional parameter spaces. This considerably limits the sampling of posterior distributions conditioned on observed data. In turn, this reduces the interpretability of posteriors to their one- and two-dimensional marginal distributions, when more information is available in the full dimensional distributions. We propose to learn smooth and differentiable representations of posterior distributions from their samples using normalizing flows, which we train with an added evidence error loss term, to extend interpretability in multiple ways. Motivated by problems from cosmology, we implement a robust method to obtain one and two-dimensional posterior profiles. These are obtained by optimizing, instead of integrating, over other parameters, and are thus less prone than marginals to so-called projection effects. We also demonstrate how this representation provides an accurate estimator of the Bayesian evidence, with log error at the 0.2 level, allowing accurate model comparison. We test our method on multi-modal mixtures of Gaussians up to dimension 32 before applying it to simulated cosmology examples.

## 1. Introduction

While efficient techniques (Foreman-Mackey et al., 2013; Handley et al., 2015; Feroz et al., 2009; 2019) have been implemented to sample Bayesian posterior distributions up to high dimensions, interpretations are very often limited to one- and two-dimensional marginals distributions, in part due to the finite size of the samples. These distributions are then typically represented as corner plots (Lewis, 2019; Foreman-Mackey, 2016) and used to derive credible intervals on individual parameters. However, these tools may not be sufficient to reveal the full-dimensional structure of posterior distributions.

A first challenge comes from the so-called projection effect, i.e. the apparent distortion of the posterior when marginalizing over subsets of parameters, such that, for instance, marginal distributions do not necessarily peak at the maximum-a-posteriori. This may occur when poorly constrained or prior-limited parameters, such as nuisance parameters or parameters modulating the amplitude of a weak signal, are projected over significant volumes. Such effects may hinder interpretability of models when limited to marginals, for instance multi-probe cosmological analysis (Krause et al., 2021; Joachimi et al., 2021), early dark energy models (Murgia et al., 2021) and galaxy power spectrum analysis (Simon et al., 2023). A second challenge is to evaluate the Bayesian evidence, i.e. the normalization of the sampled posterior distribution, that even sophisticated nested samplers sometimes fail at accurately estimating, as shown in Lemos et al. (2023). Despite its limitations (Lemos et al., 2020), the evidence is a useful metric used for model comparison. A third, related challenge is the evaluation of statistical tensions between posterior distributions obtained from multiple experiments that measure the same parameters (Leizerovich et al., 2023).

Normalizing flows (Papamakarios & Murray, 2018; Kingma et al., 2017; Rezende & Mohamed, 2016; Papamakarios et al., 2018; Grathwohl et al., 2018) have already been used to learn representation of posterior distributions from their samples, allowing, for instance, the efficient computation of tension metrics that do not rely on assuming Gaussian posteriors (Raveri & Doux, 2021; Dacunha et al., 2022). In this paper, we use state-of-the-art normalizing flow models (Papamakarios et al., 2018; Durkan et al., 2019), with an extra loss term accounting for Bayesian evidence error, to yield accurate Bayesian evidence estimates and to efficiently compute posterior one- and two-dimensional posterior profiles.

Unlike marginal distributions, posterior profiles do not suffer from projection effects as they are essentially insensitive to the volume of the parameter space. Using a simple

[1] Department of Physics and INFN, University of Genova, Genova, Italy [2] Université Grenoble Alpes, CNRS, LPSC-IN2P3, 38000 Grenoble, France [3] Department of Physics, Columbia University, 538 West 120th Street, New York, NY, USA 10027, USA. Correspondence to: Marco Raveri <marco.raveri@unige.it>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

metaphor, profiling can be thought of as observing the outline of the posterior landscape, whereas marginalization can be seen as measuring its column density. As such, they offer a highly complementary tool to analyze posterior and models, which is gaining momentum in fields such as cosmology (see, e.g., Karwal et al., 2024; Ried Guachalla et al., 2024). However, profiling requires many optimizations that may only be performed efficiently with differentiable models, such as those provided by normalizing flows. We thus propose an architecture, loss function and training scheme to obtain accurate posterior density estimates and a profiling methodology, all of which are implemented in the TensorFlow Probability frameworks (Dillon et al., 2017).

The paper is organized as follows: Section 2 introduces the formalism of posterior profiles, its difference with marginal distributions, and the normalizing flow architecture used to model posteriors; Section 3 provides context in the recent literature; Section 4 presents our benchmark results on analytic multimodal distributions and our applications to cosmology; Section 5 discusses limitations and future work.

## 2. Methods

### 2.1. Relationship between marginalization and profiling

We first discuss the relationship between marginalization and profiling. To do so, we consider an arbitrary posterior distribution $P$ over a set of parameters partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The profile posterior for $\boldsymbol{\theta}_1$ is obtained by maximizing the joint distribution over $\boldsymbol{\theta}_2$:

$$\hat{P}(\boldsymbol{\theta}_1) \equiv \max_{\boldsymbol{\theta}_2} P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{1}$$

and we denote $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ the value of $\boldsymbol{\theta}_2$ where this maximum is reached. We then write the identity:

$$\log P(\boldsymbol{\theta}_1) = \log \hat{P}(\boldsymbol{\theta}_1) + \log \int \frac{P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\hat{P}(\boldsymbol{\theta}_1)} \, \mathrm{d}\boldsymbol{\theta}_2 \, . \tag{2}$$

Since the profile, for all fixed values of $\boldsymbol{\theta}_1$ maximizes the joint distribution, Equation (2) shows that:

$$\begin{aligned} \log P(\boldsymbol{\theta}_1) - \log \hat{P}(\boldsymbol{\theta}_1) &\leq \log \int \mathbb{I}_{P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) > 0} \, \mathrm{d}\boldsymbol{\theta}_2 & (3) \\ &\leq \log V_2^P(\boldsymbol{\theta}_1) & (4) \end{aligned}$$

where $V_2^P(\boldsymbol{\theta}_1)$ denotes the $\boldsymbol{\theta}_2$-volume of the support of the joint distribution $P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ at fixed $\boldsymbol{\theta}_1$, and $\mathbb{I}$ is the characteristic function. Note that, in general, $V_2^P(\boldsymbol{\theta}_1)$ only depends on boundaries defined by the prior, $\Pi$, such that $V_2^P(\boldsymbol{\theta}_1) \leq V_2^\Pi(\boldsymbol{\theta}_1)$.

If we further assume that the posterior distribution is Gaussian, as it is done in (Hadzhiyska et al., 2023), we can write with a Taylor expansion:

$$\begin{aligned} &\log P(\boldsymbol{\theta}_1) - \log \hat{P}(\boldsymbol{\theta}_1) = \\ &= +\frac{1}{2} \log \det \big( \Sigma_2 - \Sigma_{21} \Sigma_1^{-1} \Sigma_{12} \big) + \frac{d_2}{2} \log(2\pi) + \ldots \\ &= -\frac{1}{2} \log \det \mathsf{F}_P^{(2)}(\boldsymbol{\theta}_1) + \frac{d_2}{2} \log(2\pi) + \ldots , \end{aligned} \tag{5}$$

where $\Sigma_{\cdot}$ denotes the blocks of the partitioned covariance matrix, $d_2$ the dimension of the $\boldsymbol{\theta}_2$ subspace, and where we have defined

$$\mathsf{F}_P^{(2)}(\boldsymbol{\theta}_1) = \frac{1}{2} \frac{\partial^2 \log P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2^2} \bigg|_{\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)} , \tag{6}$$

as the empirical Fisher matrix of $\boldsymbol{\theta}_2$, at fixed $\boldsymbol{\theta}_1$, and optimum $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$. Note that $\mathsf{F}_P^{(2)}$ is the empirical Fisher matrix – the second derivative of the log posterior of the observed data realization – and not the real Fisher matrix, as it is not averaged over data realization and includes the prior. Equation (5) can be seen as the first order term of a Laplace expansion of the difference between the marginal and profile distributions.

Decomposing the empirical Fisher matrix into its likelihood ($\mathcal{L}$) and prior ($\Pi$) components, we can also write the first order term as

$$\begin{aligned} &\log P(\boldsymbol{\theta}_1) - \log \hat{P}(\boldsymbol{\theta}_1) \\ &= -\frac{1}{2} \log \det \mathsf{F}_{\mathcal{L}}^{(2)} - \frac{1}{2} \log \det \mathsf{F}_{\Pi}^{(2)} + \frac{d_2}{2} \log(2\pi) \\ &\leq -\frac{1}{2} \log \det \mathsf{F}_{\mathcal{L}}^{(2)}(\boldsymbol{\theta}_1) + \log V_{\Pi}^{(2)}(\boldsymbol{\theta}_1) \\ &\leq \log V_2^P(\boldsymbol{\theta}_1) \end{aligned} \tag{7}$$

This clarifies why projection effects arise and are known with two different names: projection or volume effects. If, in Equation (7), the likelihood term dominates the leading order discrepancy between profile and marginal distributions, then this means that the empirical Fisher matrix depends on position, which is due to non-Gaussianities of the likelihood, and corresponds to what is colloquially referred to as a projection effect (Sellentin & Heavens, 2016). If, instead, the prior term dominates, this means that the data is weakly constraining, and marginalization suffers from prior volume effects, i.e. differences in prior volume along the parameter line of sight. These two cases are illustrated in Figure 1 with a two-dimensional triangular prior and increasingly tighter posterior distributions.

### 2.2. Normalizing flow architecture, loss and training

Most performance metrics applied to generative models have to do with marginal distributions – potentially of derived parameters or quantities computed from generated samples (Reyes-González & Torre, 2023; Coccaro et al.,

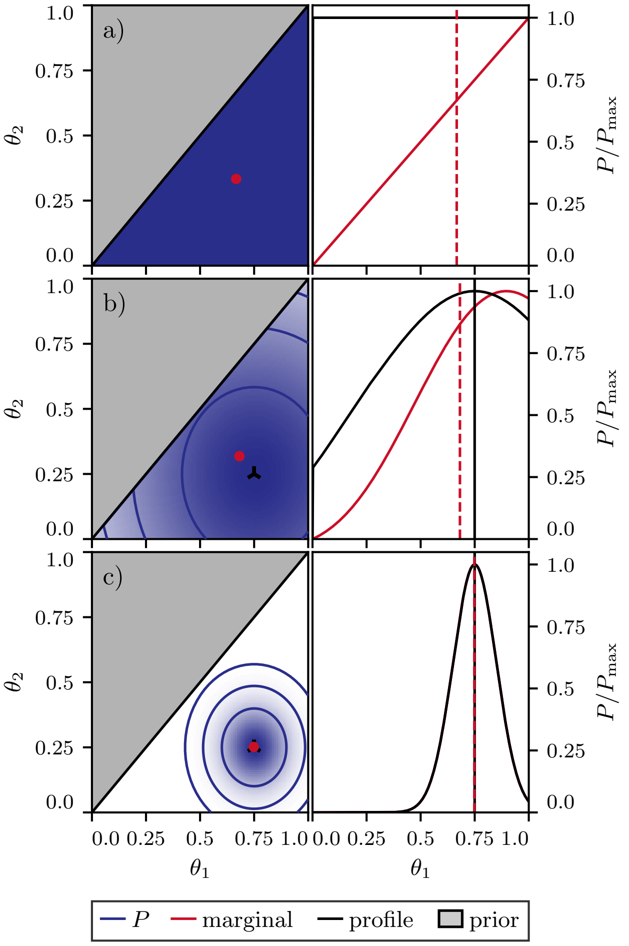

Figure 1: **Gaussian example:** in the left column, the joint prior (gray) and posterior (blue) distributions for different Gaussian distributions, with increasingly strong constraints from top to bottom. The right column shows the marginal distribution of the posterior over $\theta_1$ (red) and the corresponding profile (black). The difference between the two curves decreases as the posterior is better constrained.

2023). However, for the problem at hand, we are rather concerned with the accuracy of the (logarithm of the) posterior probability density function.

### 2.2.1. Architecture

To tackle this challenge, we tested various architectures that combine two of the best-performing types of normalizing flows (NF), namely Masked Autoregressive Flows (MAF, Papamakarios et al., 2018) and Neural Spline Flows (Durkan et al., 2019). In the best performing architecture, we stack a number of MAFs that implement autoregressive affine transformations, each parametrized by a unique neural network (with masked inputs). We also insert random parameter permutations between each MAF. In addition, we select sequences of permutations with a low variance between parameter coordinates to maximize the mixing of coordinates throughout the sequence of autoregressive transformations. This architecture provides light-weight flexibility allowing to model posterior distributions to high accuracy. We found that adding spline flows or replacing MAFs by spline flows resulted, on our finite samples, in overfitting and/or noisier posterior profiles (see Section 5).

### 2.2.2. Evidence error loss function

To further improve the accuracy of the estimated posterior density, we propose to add a term to the standard normalizing flow loss. However, the posterior density is usually computed from the product of the likelihood and priors, and it normalization, given by the Bayesian evidence, is a priori unknown. Nevertheless, this normalization constant is the same for all available samples, motivating a loss function that reduces the scatter of the approximate density around an unknown mean. Denoting $\log q(\boldsymbol{\theta})$ the approximate, normalized flow density, and $\log \tilde{P}(\boldsymbol{\theta})$ the unnormalized posterior density, we define the evidence error loss (EEL) for a batch of samples $\{\boldsymbol{\theta}_i\}_{1\leq i<N}$ drawn from the posterior as

$$\mathrm{EEL}(\{\boldsymbol{\theta}_i\}) \equiv \frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{q(\boldsymbol{\theta}_i)}{\tilde{P}(\boldsymbol{\theta}_i)} - \log\hat{\mathcal{Z}}(\{\boldsymbol{\theta}_i\})\right)^2, \tag{8}$$

where

$$\log\hat{\mathcal{Z}}(\{\boldsymbol{\theta}_i\}) \equiv \frac{1}{N}\sum_{i=1}^{N}\log\frac{q(\boldsymbol{\theta}_i)}{\tilde{P}(\boldsymbol{\theta}_i)} \tag{9}$$

is an estimator of the (logarithm of the) evidence. We add this term to the standard Kullback-Leibler divergence NF loss, given by $\sum_{i=1}^{N}\log q(\boldsymbol{\theta}_i)/N$, and use a soft adaptation scheme (Heydari et al., 2019) to balance the two terms during training. Once trained, we use Equation (9) to estimate the evidence over the full posterior training sample. Similar evidence estimator and loss functions were also recently suggested by Polanska et al. (2024); Srinivasan et al. (2024), although we not use it here to compute the evidence, but also to learn accurate posterior density estimates for profiling.

### 2.2.3. Training

In addition to the specific architecture and evidence error loss term, we obtain more accurate and stable results by training populations of normalizing flows and averaging the individual density estimates (Lakshminarayanan et al., 2016; Alsing et al., 2019).

At last, we implement a new and simple adaptive learning rate modulation scheme as follows. At every epoch end, we fit a line through the validation loss evaluated over the last $N_{\text{epochs}}$ epochs (by default, 25). If the slope is negative (as expected during learning), the learning rate is unchanged; if it is positive, then we multiply the learning rate by a factor $\alpha$ (by default, $1/\sqrt{10}$). We train all our models with an initial learning rate of $10^{-2}$ and stop training when the learning rate reaches $10^{-5}$.

### 2.3. Profiling

Once a population of normalizing flows has been trained, we aim at deriving one- and two-dimensional posterior profiles. To do so, we define, for each parameter of interest $\theta_i$ a binning (typically 64 linearly spaced bins), and we optimize over other parameters, $\boldsymbol{\theta}_{j\neq i}$, to evaluate the profile $\hat{P}(\theta_i)$. Our algorithm works as follows:

1. We sample the NF population (which is fast) sufficiently many times such that each bin in $\theta_i$ contains at least one sample.
2. Within each bin, we save the sample with the highest density $q$, which already provides a noisy estimate of the profile. Note here that the value of $\theta_i$ is not necessarily located at the center of the bin.
3. We improve this estimate by optimizing, in each $\theta_i$ bin, the flow density $q$ over other parameters $\boldsymbol{\theta}_{j\neq i}$ at fixed $\theta_i$ value (the value of the initial sample in the bin), and denote $\hat{q}(\theta_i)$ the optimum of the flow density. This optimization is performed using gradient descent, which is vectorized in our code (i.e. all bins are simultaneously optimized), making it reasonably fast.
4. The final profile is estimated by linearly interpolating between the values of $(\theta_i, \hat{q}(\theta_i))$.

This algorithms is easily generalized to the two-dimensional case, with two-dimensional bins over $(\theta_i, \theta_j)$ and optimizations over $\boldsymbol{\theta}_{k\neq i,j}$. While this is in general prohibitively expensive due to the large number of optimizations in high-dimensional spaces, the NF flow density can be evaluated efficiently, making it possible to obtain stable profiles within minutes.

Finally, we make it possible to obtain profiles over derived parameters, i.e. parameters that can be computed from the parameters sampled during posterior inference. To do so, we transform the flow density according to the reparametrization, using either analytic TensorFlow Probability bijectors when the mapping is simple, or by training an additional normalizing flow to learn it. In particular, we apply this latter functionality to obtain profiles over the effective cosmic structure parameter $\sigma_8$, which is computed by the theoretical model from more fundamental parameters which are themselves used during posterior inference (see Section 4.2). Note that one may not, in general, simply train a flow on derived parameters, as the posterior density $P$ used in the EEL loss shown in Equation (8) corresponds to a specific choice of parameters.

## 3. Related works

Several previous studies have developed methods based on machine-learning to learn the posterior distribution, but most of them focus on obtaining accurate marginal densities (Radev et al., 2021; Raveri & Doux, 2021) or evidence estimates (Turner & Sederberg, 2014; Srinivasan et al., 2024; Polanska et al., 2024; Jeffrey & Wandelt, 2024). However, we additionally aim to obtain reliable profiles of the parameters, which requires maximization in a high-dimensional parameter space. This, in turn, requires accurate estimates of posterior density values in the high-dimensional spaces (the full dimension minus one or two). For this purpose, we combine the standard NF loss with the evidence error loss, which significantly improves the posterior density values.

While preparing this manuscript, Srinivasan et al. (2024) published a study that adds a similar term to their loss function, resulting in better estimates of evidence. However, compared to them, we use a stack of normalizing flows to obtain a flexible and easy to optimize model. Moreover, we train and average multiple flows to reduce the noise in the estimate of posterior in the high dimensional space. Finally, we develop a GPU-supported optimization routine to estimate the parameter profile for large number of parameter values. All these improvements are crucial to obtain reliable parameter profiles, especially in two-dimensional corner plots due to large number of bin combinations to estimate the profile maximization. In addition to applying this architecture to analytic examples, such as mixture of Gaussians, we also apply it to cosmological setting, obtaining reliable posterior profiles. Finally, we add the functionality to obtain profiles on the derived parameters (non-linear combination of sampled parameters) which is generally useful in cosmological examples.

## 4. Results

### 4.1. Benchmark with mixtures of Gaussians

To test the performance of our model and profiling algorithm, we generate test datasets of multi-dimensional mixtures of Gaussian of varying dimensions $d$. The probability distribution function is given by:

$$P_{\mathrm{MG}}(\boldsymbol{\theta}) = \sum_i^{n_g} w_i \, \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \Sigma_i), \tag{10}$$

where, $n_g$ is the number of Gaussian components, $w_i$ is the weight of $i$-th Gaussian such that $\sum_i^{n_g} w_i = 1$, and $\boldsymbol{\mu}_i$ and $\Sigma_i$ are the mean and covariance of $i$-th Gaussian. We fix $n_g = 5$ and generate random $\boldsymbol{\mu}_i$ in the range (-1, 1) for

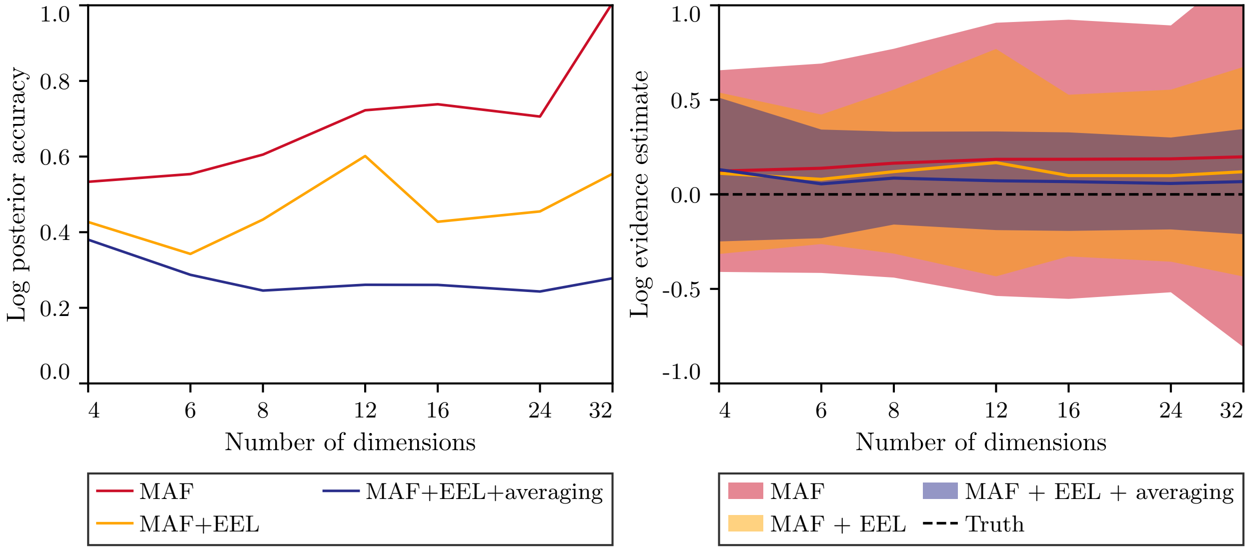

Figure 2: **Mixture of Gaussian example:** Evidence accuracy (left) and error (right) estimated from different flow architectures. MAF only uses a single standard masked auto-regressive flow, MAF+EEL additionally adds the evidence error term (Eq. 8) to the loss function and MAF+EEL+averaging additionally averages over six flows. We see that including these additional modifications to the architecture significantly improves the accuracy and constraints on the evidence.

each dimension, and $\Sigma_i$ is a random non-diagonal covariance matrix for each Gaussian component, drawn such that the mixture exhibits multimodality. Since this is a normalized sum of Gaussians, the evidence of this distribution is $\log \mathcal{Z}_{\rm MG} = 0$.

We then propose a specific architecture: each flow is composed of $\lceil 2 \log_2 d \rceil + 2$ MAFs, each parametrized by masked neural networks (Germain et al., 2015) with two hidden layers of $2d$ features and using asinh activation functions. In our final setup, we train six such flows that are averaged.

We first quantify the performances of our architecture and training methodology by measuring the evidence accuracy and its uncertainty (Figure 2) on examples with dimensions varying from 4 to 32. The evidence uncertainty is computed from the standard deviation of the evidence estimate over all samples (see Equation (9)). We first train flows composed of a sequence of MAFs with the standard loss function, showing that the evidence uncertainty cannot go below 0.6 for mixtures of Gaussians for $d > 8$, as shown on the left panel of Figure 2 (MAF). We then demonstrate that when adding the evidence error loss (EEL) term (Equation (8)), this uncertainty reaches 0.4-0.5, while preserving the quality of marginal distributions (MAF+EEL). We then train populations of flows and average their density estimates, which allows us to reach evidence uncertainties at the 0.2 level (MAF+EEL+averaging). Following the same procedure, we show in the right panel of Figure 2 that adding the EEL loss and averaging multiple flows allows us to reduce the bias on the evidence estimate, while noting it remains within uncertainties for all tested configurations.

In Figure 3, we show the performance of our model and profiling algorithms in $d = 16$ dimensions. In the left panel of Figure 3, we compare the performance of the models in capturing the marginal posterior distribution for 3 of the 16 parameters in this space, finding that even a single standard MAF captures the marginal distribution well. In the right panel, we compare the profile posterior distribution for the same three parameters. In this analytic example, we can obtain the exact profiles by maximizing the probability density of Equation (10), even in $d = 16$ dimensions, as shown by black solid lines. We find that the posterior profiles differ significantly from the posterior marginal distributions, due to the multimodality of this example. We find that, in this case, a single standard MAF fails to capture the profiles of parameters $p_2$ and $p_3$ accurately, whereas training with the extra EEL term and averaging allows us to correctly capture this multi-modality. We also estimate the two-dimensional profile distributions, which requires optimizations in 14 dimensions for all the $32 \times 32$ two-dimensional bins, for each pair of parameters. Note that this optimization would be computationally prohibitive without a differentiable estimate of the posterior density as we have developed here. We find similar performances in the one- and two-dimensional profiles when training the flow with EEL term and averaging for various dimensions up to 32.

Finally, in Figure 4, we use profiles derived from each of the individual flows to estimate the variance of the profile

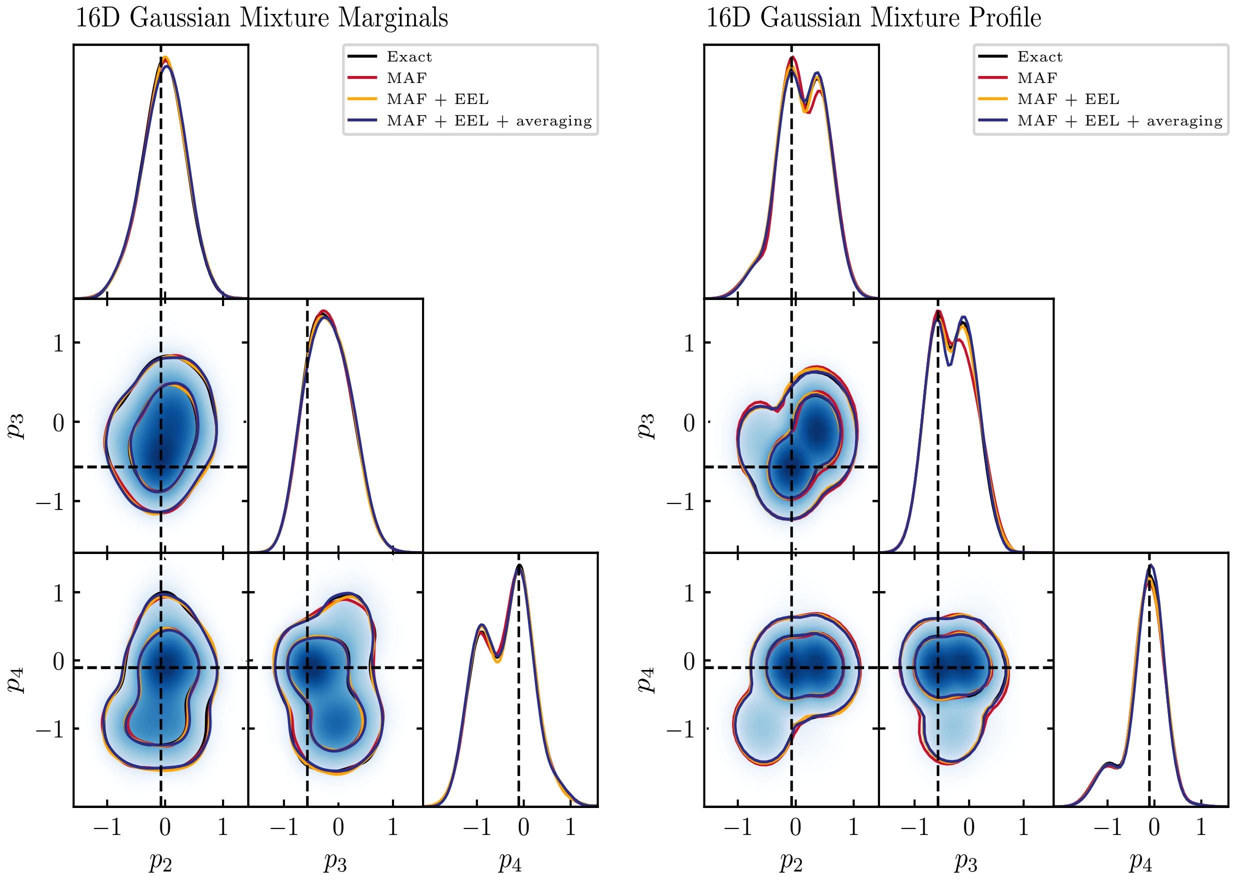

Figure 3: **Mixture of Gaussian example:** Marginalized (left) and profile (right) posteriors of three parameters in the 16-dimensional mixture of Gaussians. We see that while all three models perform well in terms of capturing the marginal posterior, using just a standard MAF struggles to capture the profiles posterior of some parameters. Adding EEL and averaging significantly helps in obtaining unbiased marginal and posterior distributions.

estimated from the averaged flows, showing the result for one parameter. We see that the averaged flow profile lies within one standard deviation of the exact profile.

### 4.2. Application to cosmological simulated data

Given the good performance of our model and profiling algorithm for multimodal, high-dimensional mixtures of Gaussian, we now apply it to the analysis of simulated cosmological data.

Specifically, we analyze simulated two-point correlation functions constructed out of galaxy positions and shapes. This is the statistics of choice of astronomical imaging surveys to probe the distribution of matter in the Universe, and constrain cosmological models of dark energy (Mandelbaum, 2018). To do so, we use the theoretical model described in Krause et al. (2021) to create a mock data vector (dimension 400) emulating weak lensing data from the Dark Energy Survey (see Fig. 5 of Abbott et al., 2022). The model has 19 parameters, which includes six cosmological parameters as well as five astrophysical and eight observational systematics parameters. Using the same theoretical model and the covariance matrix of Krause et al. (2021), we then sample the posterior distribution for this simulated data vector. To do so, we use a nested sampling algorithm, PolyChord (Handley et al., 2015), set to high accuracy to obtain 0.5 million samples. While this is highly sufficient to estimate the one- and two-dimensional marginal posterior distributions, it is inadequate for profiling. Finally, we use this sample to train the normalizing flow architecture described above and learn a smooth representation of the 19-dimensional posterior. Figure 5 shows the marginal distributions for cosmological and astrophysical parameters derived from the PolyChord sample and the trained flow. This figure illustrates the non-Gaussianity of this posterior, which arises from both the non-linearity of the theoretical model and prior volume effects, all well-captured by the flow.

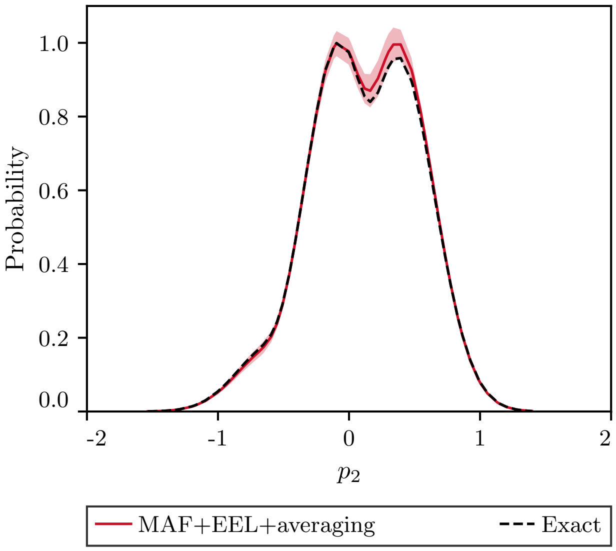

Figure 4: **Mixture of Gaussian example:** the red band shows the mean profile estimated from averaging six flows, as well as its uncertainty computed from the variance of the profiles estimated for each flow individually. The exact profile, calculated from maximizing the analytic density, is consistent with the flow profile, within one standard deviation.

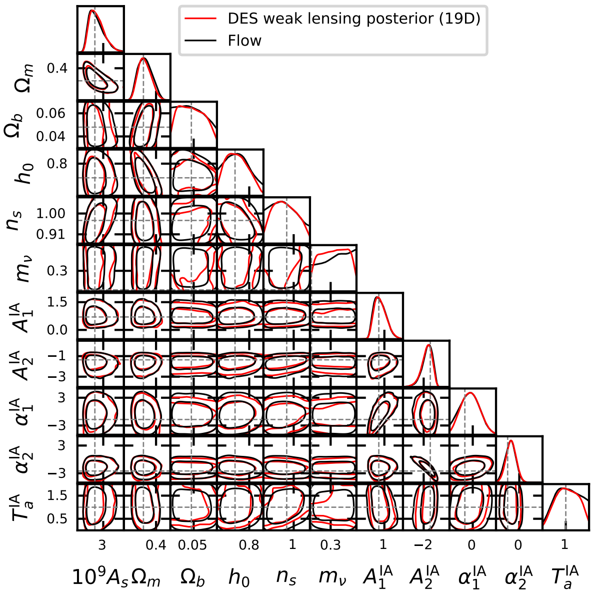

Figure 5: **Cosmology example:** We show the marginalized distribution of posterior for various parameter combinations as estimate from the true chain and from normalizing flow (MAF+EEL+averaging). We see that the flow can capture various non-Gaussian features accurately.

In Figure 6, we compare the marginal distributions and profiles for two cosmological parameters of interest, out of 19. We show the constraints on the matter density of the Universe ($\Omega_m$) and the amplitude of primordial fluctuations $A_s$ on the left. On the right, we additionally show the profiles of $\Omega_m$ and a derived parameter, $\sigma_8$, which measures the amplitude of late time matter density fluctuations, and is computed from the theoretical model as a non-linear function of the other cosmological parameters. We also show the true value used to create the mock data vector with dashed lines (the maximum a-posteriori, MAP value). We see that, as expected, the marginalized posteriors peak at different values compared to the true MAP value, whereas the profile posteriors peak at the true values. Additionally, the left panel shows that the parameter $A_s$ is, according to its profile, only weakly constrained by the data, unlike what is suggested by the peaked marginal distribution.

## 5. Discussion and future work

One- and two-dimensional marginal distributions are often insufficient to capture the properties of high-dimensional posterior distributions derived from complex data and model. These most notably suffer from so-called projection effects, due to integrating the posterior along unconstrained parameter directions, unlike posterior profiles, obtained my maximization. However, obtaining posterior profiles requires an accurate estimate of the actual posterior values, which standard normalizing flows fail to capture. In this paper, we thus provide a normalizing flow architecture and simultaneously minimize the standard Kullback-Leibler divergence loss and an extra term, the evidence error loss, which dramatically improves the quality of posterior profiles, and provides an estimate of the Bayesian evidence as a by-product. We validate our method on analytic examples using mixtures of Gaussians up to dimension 32, and then apply it to a simulated data analysis from cosmology. We note here that obtaining stable profiles also required training an ensemble of flows using an adaptive loss weighting scheme and a new adaptive learning rate scheduler.

In the future, we aim to improve the accuracy and stability of the profiles. In particular, neural spline flows offer a promising avenue in terms of flexibility. However, our tests with spline flows resulted in overfitting, likely due to the large yet finite size of our posterior samples, and inaccurate profiles. We could curtail this effect by constraining the underlying spline parameters, thus limiting fluctuations in the Jacobian entering the log density evaluation. Another promising avenue, though, consists in merging the posterior sampling and flow training steps using an iterative method, as suggested by multiple studies (Rezende & Mohamed, 2015; Gao et al., 2020; Gabrié et al., 2022; Grumitt et al., 2022; Karamanis et al., 2022; Wong et al., 2023). This

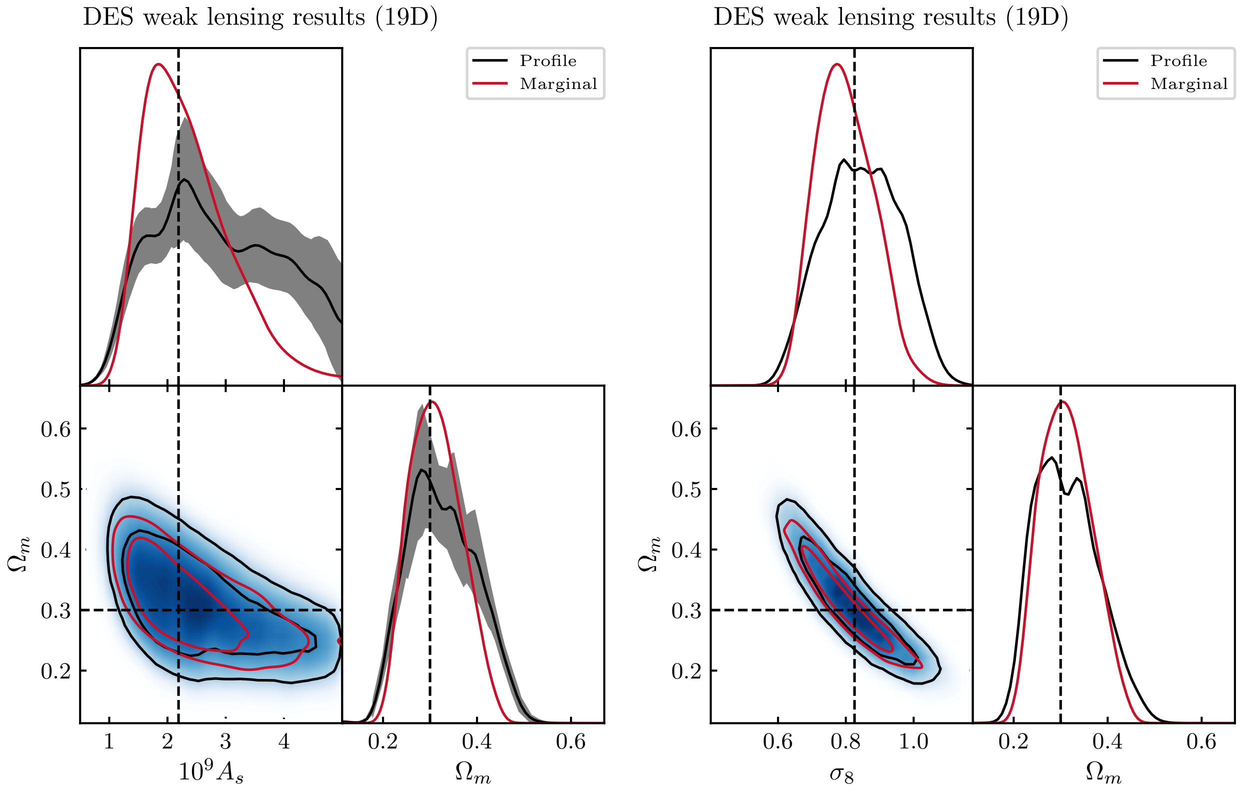

Figure 6: **Cosmology example:** *Left*: We show the application of our flow architecture to estimate the profile and marginals of two cosmological parameters within a 19 dimensional parameter space. *Right*: We additionally show the constraints on a derived parameter, $\sigma_8$, which is inferred from other cosmological parameters. In this case, the mapping from the original sampling space is learned as well.

would alleviate accuracy issues related to limited samples, and our profiling methodology could be readily applied. The code is publicly available at https://github.com/mraveri/tensiometer

## Acknowledgements

We thank Tanvi Karwal for the helpful discussions and comments. Computing resources were provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231, and by the University of Chicago Research Computing Center through the Kavli Institute for Cosmological Physics. M.R. acknowledges financial support from the INFN InDark initiative.

## References

Abbott, T. et al. Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2), January 2022. ISSN 2470-0029. doi: 10.1103/physrevd.105.023520. URL http://dx.doi.org/10.1103/PhysRevD.105.023520.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. Fast likelihood-free cosmology with neural density estimators and active learning. , 488(3):4440–4458, September 2019. doi: 10.1093/mnras/stz1960.

Coccaro, A., Letizia, M., Reyes-Gonzalez, H., and Torre, R. Comparative Study of Coupling and Autoregressive Flows through Robust Statistical Tests. *arXiv e-prints*, art. arXiv:2302.12024, February 2023. doi: 10.48550/arXiv.2302.12024.

Dacunha, T., Raveri, M., Park, M., Doux, C., and Jain, B. What does a cosmological experiment really measure? Covariant posterior decomposition with normalizing flows. , 105(6):063529, March 2022. doi: 10.1103/PhysRevD.105.063529.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. TensorFlow Distributions. *arXiv e-prints*, art. arXiv:1711.10604, November 2017. doi: 10.48550/arXiv.1711.10604.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows, 2019.

Feroz, F., Hobson, M. P., and Bridges, M. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. , 398(4):1601–1614, October 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

Feroz, F., Hobson, M. P., Cameron, E., and Pettitt, A. N. Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, 2(1):10, November 2019. doi: 10.21105/astro.1306.2144.

Foreman-Mackey, D. corner.py: Scatterplot matrices in python. *The Journal of Open Source Software*, 1(2): 24, jun 2016. doi: 10.21105/joss.00024. URL https://doi.org/10.21105/joss.00024.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. emcee: The MCMC Hammer. , 125(925):306, March 2013. doi: 10.1086/670067.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Science*, 119 (10):e2109420119, March 2022. doi: 10.1073/pnas.2109420119.

Gao, C., Isaacson, J., and Krause, C. i-flow: High-dimensional Integration and Sampling with Normalizing Flows. *arXiv e-prints*, art. arXiv:2001.05486, January 2020. doi: 10.48550/arXiv.2001.05486.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. MADE: Masked Autoencoder for Distribution Estimation. *arXiv e-prints*, art. arXiv:1502.03509, February 2015. doi: 10.48550/arXiv.1502.03509.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. *arXiv e-prints*, art. arXiv:1810.01367, October 2018. doi: 10.48550/arXiv.1810.01367.

Grumitt, R. D. P., Dai, B., and Seljak, U. Deterministic Langevin Monte Carlo with Normalizing Flows for Bayesian Inference. *arXiv e-prints*, art. arXiv:2205.14240, May 2022. doi: 10.48550/arXiv.2205.14240.

Hadzhiyska, B., Wolz, K., Azzoni, S., Alonso, D., García-García, C., Ruiz-Zapatero, J., and Slosar, A. Cosmology with 6 parameters in the Stage-IV era: efficient marginalisation over nuisance parameters. *The Open Journal of Astrophysics*, 6:23, July 2023. doi: 10.21105/astro.2301.11895.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. polychord: nested sampling for cosmology. , 450:L61–L65, June 2015. doi: 10.1093/mnrasl/slv047.

Heydari, A. A., Thompson, C. A., and Mehmood, A. Softadapt: Techniques for adaptive loss weighting of neural networks with multi-part loss functions, 2019.

Jeffrey, N. and Wandelt, B. D. Evidence networks: simple losses for fast, amortized, neural bayesian model comparison. *Machine Learning: Science and Technology*, 5(1):015008, January 2024. ISSN 2632-2153. doi: 10.1088/2632-2153/ad1a4d. URL http://dx.doi.org/10.1088/2632-2153/ad1a4d.

Joachimi, B. et al. KiDS-1000 methodology: Modelling and inference for joint weak gravitational lensing and spectroscopic galaxy clustering analysis. , 646:A129, February 2021. doi: 10.1051/0004-6361/202038831.

Karamanis, M., Nabergoj, D., Beutler, F., Peacock, J., and Seljak, U. pocoMC: A Python package for accelerated Bayesian inference in astronomy and cosmology. *The Journal of Open Source Software*, 7(79):4634, November 2022. doi: 10.21105/joss.04634.

Karwal, T., Patel, Y., Bartlett, A., Poulin, V., Smith, T. L., and Pfeffer, D. N. Procoli: Profiles of cosmological likelihoods. 1 2024.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving variational inference with inverse autoregressive flow, 2017.

Krause, E. et al. Dark energy survey year 3 results: Multi-probe modeling strategy and validation, 2021.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *arXiv e-prints*, art. arXiv:1612.01474, December 2016. doi: 10.48550/arXiv.1612.01474.

Leizerovich, M., Landau, S. J., and Scóccola, C. G. Tensions in cosmology: a discussion of statistical tools to determine inconsistencies. *arXiv e-prints*, art. arXiv:2312.08542, December 2023. doi: 10.48550/arXiv.2312.08542.

Lemos, P., Köhlinger, F., Handley, W., Joachimi, B., Whiteway, L., and Lahav, O. Quantifying Suspiciousness within correlated data sets. , 496(4):4647–4653, August 2020. doi: 10.1093/mnras/staa1836.

Lemos, P. et al. Robust sampling for weak lensing and clustering analyses with the Dark Energy Survey. , 521 (1):1184–1199, May 2023. doi: 10.1093/mnras/stac2786.

Lewis, A. GetDist: a Python package for analysing Monte Carlo samples. *arXiv e-prints*, art. arXiv:1910.13970, October 2019. doi: 10.48550/arXiv.1910.13970.

Mandelbaum, R. Weak lensing for precision cosmology. *Annual Review of Astronomy and Astrophysics*, 56(1):393–433, September 2018. ISSN 1545-4282. doi: 10.1146/annurev-astro-081817-051928. URL http://dx.doi.org/10.1146/annurev-astro-081817-051928.

Murgia, R., Abellán, G. F., and Poulin, V. Early dark energy resolution to the Hubble tension in light of weak lensing surveys and lensing anomalies. , 103(6):063502, March 2021. doi: 10.1103/PhysRevD.103.063502.

Papamakarios, G. and Murray, I. Fast $\epsilon$-free inference of simulation models with bayesian conditional density estimation, 2018.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation, 2018.

Polanska, A., Price, M. A., Piras, D., Spurio Mancini, A., and McEwen, J. D. Learned harmonic mean estimation of the Bayesian evidence with normalizing flows. *arXiv e-prints*, art. arXiv:2405.05969, May 2024. doi: 10.48550/arXiv.2405.05969.

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., and Bürkner, P.-C. Amortized bayesian model comparison with evidential deep learning, 2021.

Raveri, M. and Doux, C. Non-gaussian estimates of tensions in cosmological parameters. *Physical Review D*, 104(4), August 2021. ISSN 2470-0029. doi: 10.1103/physrevd.104.043504. URL http://dx.doi.org/10.1103/PhysRevD.104.043504.

Reyes-González, H. and Torre, R. Testing the boundaries: Normalizing Flows for higher dimensional data sets. In *Journal of Physics Conference Series*, volume 2438 of *Journal of Physics Conference Series*, pp. 012155. IOP, February 2023. doi: 10.1088/1742-6596/2438/1/012155.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows, 2016.

Ried Guachalla, B., Britt, D., Gruen, D., and Friedrich, O. Informed Total-Error-Minimizing Priors: Interpretable cosmological parameter constraints despite complex nuisance effects. *arXiv e-prints*, art. arXiv:2405.00261, April 2024. doi: 10.48550/arXiv.2405.00261.

Sellentin, E. and Heavens, A. F. Parameter inference with estimated covariance matrices. , 456(1):L132–L136, February 2016. doi: 10.1093/mnrasl/slv190.

Simon, T., Zhang, P., Poulin, V., and Smith, T. L. Consistency of effective field theory analyses of the boss power spectrum. *Physical Review D*, 107(12), June 2023. ISSN 2470-0029. doi: 10.1103/physrevd.107.123530. URL http://dx.doi.org/10.1103/PhysRevD.107.123530.

Srinivasan, R., Crisostomi, M., Trotta, R., Barausse, E., and Breschi, M. floZ: Evidence estimation from posterior samples with normalizing flows. *arXiv e-prints*, art. arXiv:2404.12294, April 2024. doi: 10.48550/arXiv.2404.12294.

Turner, B. M. and Sederberg, P. B. A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review*, 21:227–250, 2014.

Wong, K. W. K., Gabrié, M., and Foreman-Mackey, D. flowMC: Normalizing flow enhanced sampling package for probabilistic inference in JAX. *The Journal of Open Source Software*, 8(83):5021, March 2023. doi: 10.21105/joss.05021.