Neural variational inference for cutting feedback during uncertainty propagation

Jiafang Song[†], Sandipan Pramanik[†], Abhirup Datta[*]

Department of Biostatistics, Johns Hopkins University

Abstract

In many scientific applications, uncertainty of estimates from an earlier (upstream) analysis needs to be propagated to subsequent (downstream) analysis, without feedback (downstream data updating upstream posteriors). Cutting feedback methods, also termed cut-Bayes, achieve this by constructing a cut-posterior distribution. However, existing sampling-based approaches for cutting feedback, such as nested Markov chain Monte Carlo (MCMC), are computationally demanding, while variational inference (VI) methods require two separate variational approximations and access to the upstream data and model, which is often impractical in many applications due to privacy constraints or limited data accessibility. We propose *NeVI-Cut*, a provably accurate and modular neural network-based variational inference method for cutting feedback. We directly utilize samples from the upstream analysis without requiring access to the upstream data or model, and specify the conditional variational family for the downstream parameters using normalizing flows (neural network-based generative models). We provide fixed-data (not asymptotic) convergence rates of the NeVI-Cut solution in terms of richness of the neural architecture and the complexity of the exact target cut-posterior. In the process, we establish general results of independent importance on uniform Kullback-Leibler approximation rates of conditional distributions by common flow classes including Unconstrained Monotonic Neural Network and Neural Spline Flows. A triply stochastic algorithm implements the method efficiently. Simulation studies and two real-world analyses illustrate the speed and accuracy of NeVI-Cut.

Keywords: Bayesian analysis, Cutting Feedback, Variational Inference, Normalizing Flows, Neural Networks, Universal approximation rates.

[*]Email address for correspondence: abhidatta@jhu.edu [†] Equal contributions.

1 Introduction

Across many scientific fields, uncertainty estimates of quantities from an upstream analysis are prestored to be reused in many downstream analyses. For example, downstream climate impacts studies (e.g., Butler et al., 2022; Park et al., 2023; Rao et al., 2024) commonly use ensemble outputs quantifying uncertainty from climate models such as CMIP6 (Thrasher et al., 2022). Similarly, the Global Burden of Disease project releases fine particulate matter ($PM_{2.5}$) exposure and risk estimates as 1000 posterior draws (Institute for Health Metrics and Evaluation, 2024) that are used downstream in attributable-burden calculations (Murray et al., 2020). Importance of incorporating exposure uncertainty in downstream health outcomes analysis has been routinely emphasized (Peng and Bell, 2010; Chang et al., 2011; Spiegelman, 2016; Comess et al., 2024). Other examples of uncertainty quantified prestored outputs include posterior distributions of cosmological parameters, termed as Planck MCMC chains, posterior samples of confusion matrices of cause-of-death classifiers (Pramanik et al., 2025a) used for calibration of disease burden attribution, and probabilistic word-embeddings used for large language models (LLM) like *word2Gauss* and *Bayesian skip-gram* (Vilnis and McCallum, 2014; Bražinskas et al., 2018) used in downstream text classification tasks.

In many of the aforementioned applications, when propagating uncertainty to a downstream Bayesian analysis, it is desirable not to have *feedback*, which is the statistical phenomenon of the downstream data informing an updated distribution for the upstream parameters. This is because the upstream quantities are often more fundamental and causally precedent (e.g., air pollution concentrations or climate data) than downstream ones (e.g., individual-level health outcomes). More importantly, downstream models are often misspecified for the upstream quantities. For example, in downstream health association studies, upstream estimates of area-level air pollution outputs are often used only as proxies for the true unobserved individual-level exposures (Zeger et al., 2000). Similarly, confusion matrix estimates of classifiers trained on specific populations are often used in downstream calibration of different target populations (Pramanik et al., 2025a). Hence, it is undesirable to update these upstream parameter distributions based on the downstream data and model that are subject to model misspecifications, dataset shifts, and biases. Furthermore, many of these upstream quantities (e.g., climate outputs, air pollution estimates) are used in many different downstream analyses, and it is impractical to expect revised upstream posteriors to be published after every downstream study conducted by different groups.

The idea of explicitly stopping downstream to upstream feedback is known as *Cutting feedback* (Liu et al., 2009; Plummer, 2015; Jacob et al., 2017), which estimates the *cut-posterior*, a valid joint distribution over all parameters that preserves the distribution of the upstream parameters from the first analysis. There is now a large and rapidly growing literature on cutting feedback and related methods because of their broad applicability. A brief review is offered in Section S1 of the Supplement and Nott et al. (2023) offers a detailed overview. A major thread of developments has been computational. Plummer (2015) showed that the 'cut' algorithm in the OpenBUGS package algorithm does not converge to the target cut-posterior. A principled alternative is the *multiple imputation-based nested MCMC* — running an MCMC for the upstream analysis to obtain a set of posterior samples, and then for each of these samples, running an MCMC for the downstream analysis, and then combining over all the MCMC posteriors. The repeated MCMC runs for the downstream analysis significantly increase computational cost. Plummer (2015) proposed a tempered cut algorithm for approximating the cut-posterior. Jacob et al. (2020) developed an unbiased MCMC estimator using coupled chains and demonstrated the use for approximating the cut-posterior. Other notable work that has advanced the study of sampling based methods for cutting feedback include Liu and Goudie (2022); Pompe and Jacob (2021); Chakraborty et al. (2023) and others reviewed in Section S1.

Variational inference (VI), an optimization-based alternative to sampling methods for estimating distributions, has been explored in the context of cutting feedback to address the computational challenges. Yu et al. (2023) proved a fundamental result, showing that the cut-posterior can be viewed as a constrained variational optimization problem. They approximate the cut-posterior using fixed-form parametric (e.g., Gaussian) variational families. Variational inference was also used in Smith et al. (2025) for cutting feedback in misspecified copula models. In semi-modular inference (SMI), Carmona and Nicholls (2022) used variational inference with *normalizing flows*, a class of neural network-based distributions, to approximate SMI posteriors flexibly. Battaglia and Nicholls (2024) extended the flow-based approach to amortized inference over prior hyperparameters, fitting a conditional normalizing flow conditioning on the hyper-parameter values.

The existing variational inference approaches for cut-Bayes require access to the upstream raw data and model. These may not be available in many of the applications discussed above, where only the uncertainty quantified outputs are published (e.g., climate-model outputs, air pollution estimates, cosmological parameters, probabilistic word-embeddings). Sometimes the

raw data may not be shared because of privacy concerns related to human-subjects research (e.g., uncertainty quantified estimates of cause-of-death classifiers published in Pramanik et al., 2025a, are based on individual-level cause-of-death data that are not publicly available). Even if the upstream raw data and model are available, there is no need for the first variational approximation, as the cut-posterior for the upstream quantities is the same as their Bayes posterior from the upstream analysis. Also, many variational cut-Bayes methods rely on parametric assumptions for the variational family which is problematic when the true conditional cut-posterior deviates substantially from the chosen parametric form.

In this manuscript, we propose a modular neural network-based variational inference approach for cut-Bayes that mitigates these limitations. Our contributions are twofold. First, we directly utilize samples of parameter distributions from the upstream analysis, and construct a single loss function that integrates all samples. When optimized, this estimates the variational conditional cut-posterior of the downstream parameters. This eliminates the need to access the upstream raw data and model, thereby enabling modularity of cut-Bayes analysis and protecting data privacy. Even if the upstream data and model are available, we simply use a single MCMC run to obtain upstream samples and use these in the loss for the downstream variational analysis. Thus, unlike the current approaches, we do not need a variational approximation for the upstream posterior thereby also reducing additional approximation error. We use conditional normalizing flows, neural network-based conditional distribution families to estimate the conditional cut-posterior of the downstream parameters. Our proposed method, *Neural Variational Inference for Cut-Bayes (NeVI-Cut)*, thus leverages the expressive power of neural networks (Hornik, 1991) and normalizing flows (Rezende and Mohamed, 2015) enabling better approximation of complex, multi-modal, or skewed posteriors. Computationally, NeVI-Cut demonstrates significant improvement over sampling-based methods like the nested MCMC for cutting feedback, while being substantially more accurate than parametric variational cutting feedback methods.

A second contribution of this manuscript is developing a comprehensive theory for NeVI-Cut and variational cut-Bayes. Prior theoretical work on variational cut-Bayes is limited. Smith et al. (2025) proves consistency and asymptotic normality of the parameters of the Gaussian variational class for copula models, but their guarantees are asymptotic, appealing to Bernstein-von Mises type arguments that the true cut-posterior is approximately Gaussian asymptotically, thereby justifying the Gaussian variational family. Such results are reassuring yet do not address the quantity of real interest in applications, namely the quality of approx-

imation to the exact cut-posterior conditional on the actually observed dataset, rather than a hypothetical large sample limit. On the other hand, Battaglia and Nicholls (2024) provides a universal approximation result for conditional normalizing flows in terms of convergence in distribution. However, that is not sufficient to offer guarantees about the variational solutions using these flows, which minimize the Kullback-Leibler divergence (KLD).

Our theory is in the 'fixed-data' or 'fixed-target' regime, not asymptotic and not assuming any sample size growth. This is the more challenging regime of theory when studying the approximation quality of a variational solution to the target posterior. In an asymptotic paradigm, Bernstein-von Mises implies approximate Gaussian posteriors, so even simple Gaussian VI can be asymptotically adequate. That safety net is absent in the fixed target setting, where posteriors can have complicated shapes requiring sufficiently complex variational classes like normalizing flows. We keep the observed data fixed, and provide guarantees on the approximation quality of the variational estimate obtained from NeVI-Cut with respect to the cut-posterior conditional on the observed data. Our results explicitly quantify how the approximation rates depend on both the complexity of the target cut-posterior as well as the richness (number of parameters) of the neural architecture used to design the conditional flows. In the process, we establish, to our knowledge, novel results on universal and uniform approximation KLD rates of conditional flows. Our results cover common conditional flow classes like *Unconstrained Monotonic Neural Network flows* (UMNN; Wehenkel and Louppe, 2019) and *Rational Quadratic Neural Spline Flows* (RQ-NSF; Durkan et al., 2019). As the existing literature on flows and conditional flows (Huang et al., 2018; Papamakarios et al., 2021) has mostly focused on their representational capabilities, our results are of independent importance, as they facilitate study of variational solutions using flows.

The remainder of this paper is organized as follows: In Section 2, we illustrate the idea of cutting feedback. We then give background on variational inference and normalizing flows. In Section 3 we present our NeVI-Cut algorithm. Theoretical results are presented in Section 4. Section 5 details simulation experiments to benchmark speed and accuracy of our method. Section 6 presents two real-world analyses using cutting feedback methods: a commonly used HPV data analysis, and the child and neonatal mortality data from the Countrywide Mortality Surveillance for Action (COMSA) Program in Mozambique. Section 7 concludes the paper with a discussion, to illustrate benefits of NeVI-Cut. Code to implement our proposed method is available on NeVI-Cut Python package on GitHub.

2 Background

2.1 Cutting Feedback Approaches

Let D_1 denote the upstream data which informs the upstream quantity (parameters or predictive variables) η. The downstream data is denoted by D_2 whose model involves both η and some new parameters θ. If both datasets were analyzed jointly, the Bayesian posterior distribution can be written as:

$$p(\theta, \eta \,|\, D_1, D_2) \propto p(\theta \,|\, \eta, D_1, D_2)\, p(\eta \,|\, D_1, D_2) = p(\theta \,|\, \eta, D_2)\, p(\eta \,|\, D_1, D_2). \tag{1}$$

In a Bayesian analysis, the final posterior of η is informed by both D_1 and D_2 and is thus different from its posterior $p(\eta \,|\, D_1)$ after the upstream analysis. As discussed in the Introduction, this is undesirable in many situations. Cutting feedback prevents this by targeting the cut-posterior

$$p_{\text{cut}}(\theta, \eta \,|\, D_1, D_2) = p(\theta \,|\, \eta, D_1, D_2)\, p_{\text{cut}}(\eta \,|\, D_1, D_2) = p(\theta \,|\, \eta, D_2)\, p(\eta \,|\, D_1). \tag{2}$$

Here, $p(\theta \,|\, \eta, D_1, D_2) = p(\theta \,|\, \eta, D_2)$, as in (S1), because the likelihood of D_1 does not depend of θ (hence no arrows between them), and as mentioned above, $p_{\text{cut}}(\eta \,|\, D_1, D_2) = p(\eta \,|\, D_1)$ preserves the posterior of η from upstream analysis. A more detailed introduction of cutting feedback is offered in Section S1.

2.2 Variational Inference

Variational inference (VI) provides a practical alternative by approximating posterior distributions via optimization (Blei et al., 2017). Given data D with model specified by some parameters ϕ, VI seeks a distribution $q(\phi)$ within a variational family $\mathcal{Q}$ that minimizes the KLD to the posterior distribution $p(\phi \,|\, D)$, to get the optimal solution

$$q^*(\phi) = \underset{q(\phi) \in \mathcal{Q}}{\arg\min}\ \text{KL}\left(q(\phi) \parallel p(\phi \,|\, D)\right). \tag{3}$$

If no restriction is placed on the family $\mathcal{Q}$ of distributions, then $q^*(\phi)$ is exactly the true posterior distribution $p(\phi \,|\, D)$. In practice, one places some parametric assumptions on the class $\mathcal{Q}$ to obtain an optimizer that approximates the true posterior distribution.

A fundamental contribution in cut-Bayes or cutting feedback methods that enabled the use of variational inference for this problem is the result (Lemma 4.1) of Yu et al. (2023). Considering the setup of Section 2.1 with $\phi = (\theta, \eta)$ and $D = (D_1, D_2)$, the result proved that

when $\mathcal{Q}$ is restricted to all distributions $q(\theta, \eta)$ such that $q(\eta) = p(\eta \mid D_1)$, the optimization in (3) yields the joint cut-posterior (S2) as the solution. Formally,

$$p_{\text{cut}}(\theta, \eta \mid D_1, D_2) = q^*(\theta, \eta) = \underset{q(\theta,\eta)\in\mathcal{Q}: q(\eta)=p(\eta \mid D_1)}{\arg\min} \text{KL}(q(\theta, \eta) \parallel p(\theta, \eta \mid D_1, D_2)). \tag{4}$$

Thus the cut-posterior has the nice interpretable characterization as the best approximation of the Bayes posterior (in KLD) under the constraint of no feedback from D_2 to η. In their implementation, Yu et al. (2023) used two variational approximations, decomposing $q(\theta, \eta) = q(\theta \mid \eta)\, q(\eta)$, using parametric families for both, i.e., $q(\theta \mid \eta) = q_{\lambda_2}(\theta \mid \eta)$ and $q(\eta) = q_{\lambda_1}(\eta)$. To estimate $q^*(\eta)$, the KLD to $p(\eta \mid D_1)$ was minimized with respect to λ_1 to obtain $\hat{q}^*(\eta) = q_{\hat{\lambda}_1}(\eta)$. Then $q_{\hat{\lambda}_1}(\eta)$ is plugged into the second optimization (4) to estimate the minimizing parameters $\hat{\lambda}_2$. The cut-posterior estimate is $q_{\hat{\lambda}_2}(\theta \mid \eta)\, q_{\hat{\lambda}_1}(\eta)$. Similar ideas were used in Smith et al. (2025) in the context of copula models.

2.3 Normalizing Flows

Normalizing flows provide a flexible variational family by constructing complex densities through a sequence of invertible and differentiable transformations of a simple base distribution, and have been widely explored in variational inference (Rezende and Mohamed, 2015). The basic idea of normalizing flows is as follows. If F is the cumulative distribution function (cdf) of a scalar θ, then there always exists a function T such that $\theta \overset{d}{=} T(Z)$, where Z is some base distribution. The function T is explicitly obtained from the inverse cdf transformation, e.g., if $Z \sim N(0, 1)$, $T(Z) = F^{-1}(\Phi(Z))$ where Φ is the standard Normal cdf. Hence, the task of approximating the distribution of θ becomes equivalent to approximating the function T. As neural networks are universal approximators of many function classes, it thus suffices to model T using a suitable class of neural networks g_λ and estimate the neural parameters (weights and biases) λ using the variational (KL) loss. Normalizing flows build on this core idea, but the specific choices to model T vary widely. We will discuss some common flow classes when introducing our method and studying its theoretical properties. A more comprehensive review of normalizing flows and their extensions can be found in (Kobyzev et al., 2020; Papamakarios et al., 2021).

3 Neural Variational Inference for Cut-Bayes

We propose a variational inference method for cutting feedback that (a) can directly use samples of η from the upstream posterior $p(\eta \mid D_1)$ without requiring access to the upstream

data D_1 and upstream model (likelihood $p(D_1 \mid \eta)$ and prior $p(\eta)$); and (b) uses normalizing flows to non-parametrically model the conditional cut-posterior of θ given η.

Recall from (4) that the cut-posterior is the solution to the variational optimization with the variational family $q(\theta, \eta)$ constrained to have $q(\eta) = p(\eta \mid D_1)$ to prevent feedback. As $q(\eta)$ is fixed, we make two observations. First, the optimization in (4) is essentially only over the class of conditional distributions $q(\theta \mid \eta)$. Also, under this constraint of $q(\eta) = p(\eta \mid D_1)$, the KL loss function in (4) can be written as

$$\mathrm{KL}(q(\theta, \eta) \parallel p(\theta, \eta \mid D_1, D_2)) = \mathbb{E}_{\eta \sim p(\eta \mid D_1)}\big[\mathrm{KL}(q(\theta \mid \eta) \parallel p(\theta \mid \eta, D_2))\big].$$

Hence, we can express the joint cut-posterior in terms of the conditional variational solution as

$$\begin{aligned} p_{\mathrm{cut}}(\theta, \eta \mid D_1, D_2) &= q^*_{\mathrm{cut}}(\theta \mid \eta)\, p(\eta \mid D_1), \text{ where} \\ q^*_{\mathrm{cut}}(\theta \mid \eta) &= \underset{q(\theta \mid \eta)}{\arg\min}\, \mathbb{E}_{\eta \sim p(\eta \mid D_1)} \left[\mathrm{KL}(q(\theta \mid \eta) \parallel p(\theta \mid \eta, D_2))\right]. \end{aligned} \tag{5}$$

Thus, finding the joint cut-posterior essentially reduces to estimating the collection of conditional distributions $\{q(\theta \mid \eta)\}_\eta$. No approximations to $p(\eta \mid D_1)$ is needed. In practice, we will typically have samples $\eta_1, \ldots, \eta_N$ from $p(\eta \mid D_1)$. As long as the samples are iid or from an ergodic MCMC, and we have enough samples, we can approximate the expectation term in (5) with the sample average, leading to the estimate

$$\begin{aligned} \hat{p}_{\mathrm{cut}}(\theta, \eta \mid D_1, D_2) &= \hat{q}_{\mathrm{cut}}(\theta \mid \eta)\, p(\eta \mid D_1), \text{ where} \\ \hat{q}_{\mathrm{cut}}(\theta \mid \eta) &= \underset{q(\theta \mid \eta)}{\arg\min}\, \frac{1}{N} \sum_{i=1}^{N} \left[\mathrm{KL}(q(\theta \mid \eta_i) \parallel p(\theta \mid \eta_i, D_2))\right] \\ &= \underset{q(\theta \mid \eta)}{\arg\min}\, \frac{1}{N} \sum_{i=1}^{N} \left[\mathrm{KL}(q(\theta \mid \eta_i) \parallel p(D_2 \mid \theta, \eta_i)\, p(\theta \mid \eta_i)\right]. \end{aligned} \tag{6}$$

Equations (5) and (6) illustrate how the joint cut-posterior can be estimated without access to the upstream data D_1 and the model used in upstream analysis, as long as the samples $\{\eta_i\}$ are available. Even if D_1 and its model are available, there is no need to approximate $p(\eta \mid D_1)$ using the additional variational inference step, as done in Yu et al. (2023) or Smith et al. (2025). One can run a standard MCMC, to estimate $p(\eta \mid D_1)$ and use the posterior samples $\{\eta_i\}$ in (6). Once the downstream analysis model likelihood $p(D_2 \mid \theta, \eta)$ and the prior $p(\theta \mid \eta)$ are chosen, the only unknown in (6) is the class of conditional distributions $q(\theta \mid \eta)$. In the next section, we propose using normalizing flows that can approximate any conditional family, leading to a provably accurate estimate of the joint cut-posterior.

3.1 Normalizing Flows for Conditional Distribution

We model $q(\theta \mid \eta)$ using conditional normalizing flows, which represent complex distributions as transformations of simple base distributions. Specifically, for $\theta \in \mathbb{R}^d$, we define a family of conditional distributions $q_\vartheta(\theta \mid \eta)$ corresponding to the law $\theta \mid \eta \sim T_\vartheta(\eta, Z)$, where $T_\vartheta(\eta, z)$ is a neural network-based transformation with arguments (η, z) and parameters $\vartheta \in \Theta$, and $Z \in \mathbb{R}^d$ is a latent variable drawn from an easy-to-sample base distribution $p(Z)$. A common choice for $p(Z)$ is the standard normal, $Z \sim N(0, I_d)$. However, when the true posterior exhibits heavy tails, alternative base distributions such as the Cauchy or Student's t may be more appropriate. Theoretical considerations guiding this choice are discussed in Section 4.

Using the classical reparametrization trick of variational inference (Kingma et al., 2013), the Kullback Leibler objective in (6), can be expressed in terms of $p(Z)$ as

$$\mathbb{E}_{\eta \sim p(\eta \mid D_1)} \mathbb{E}_{Z \sim p(Z)} \big[\log q_\vartheta(T_\vartheta(\eta, Z) \mid \eta) - \log p(T_\vartheta(\eta, Z) \mid \eta, D_2) \big]. \tag{7}$$

Writing $J_z T_\vartheta(\eta, z)$ for the Jacobian of T_ϑ with respect to z, and using change of variables (see Lemma 2 for a formal statement and proof), we have the identity

$$\log q_\vartheta(T_\vartheta(\eta, z) \mid \eta) = \log p(z) - \log|\det J_z T_\vartheta(\eta, z)| \,.$$

Plugging this in (7), we can write the loss as

$$\mathbb{E}_{\eta \sim p(\eta \mid D_1)} \mathbb{E}_{Z \sim p(Z)} \big[\log p(Z) - \log |\det J_z T_\vartheta(\eta, Z)| - \log p(T_\vartheta(\eta, Z) \mid \eta, D_2) \big].$$

Finally, noting that $\log p(Z)$ does not involve any terms of ϑ and $\log p(T_\vartheta(\eta, Z) \mid \eta, D_2) = \log p(D_2 \mid T_\vartheta(\eta, Z), \eta) + \log p(T_\vartheta(\eta, Z) \mid \eta) +$ terms free of ϑ, minimizing (7) is equivalent to maximizing the average conditional *evidence lower bound (ELBO)*

$$\mathcal{L}(\vartheta) = \mathbb{E}_{\eta \sim p(\eta \mid D_1)} \mathbb{E}_{Z \sim p(Z)} \big[\log |\det J_z T_\vartheta(\eta, Z)| + \log p(D_2 \mid T_\vartheta(\eta, Z), \eta) + \log p(T_\vartheta(\eta, Z) \mid \eta) \big].$$

Once again, we can replace the expectation under $p(\eta \mid D_1)$ with average over the samples $\{\eta_i\}$, leading to the objective

$$\hat{\mathcal{L}}(\vartheta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{Z \sim p(Z)} \big[\log |\det J_z T_\vartheta(\eta_i, Z)| + \log p(D_2 \mid T_\vartheta(\eta_i, Z), \eta_i) + \log p(T_\vartheta(\eta_i, Z) \mid \eta_i) \big]. \tag{8}$$

The ELBO $\hat{\mathcal{L}}(\vartheta)$ can be calculated for any value of ϑ as every other quantity is known. Hence we can maximize $\hat{\mathcal{L}}(\vartheta)$ over ϑ to obtain $\hat{\vartheta} \in \arg\max_\vartheta \hat{\mathcal{L}}(\vartheta)$. The estimate of the

cut-posterior is then

$$\hat{p}_{\text{cut}}(\theta, \eta \,|\, D_1, D_2) = q_{\hat{\vartheta}}(\theta \,|\, \eta)\, p(\eta \,|\, D_1). \tag{9}$$

By drawing $Z_i \overset{\text{iid}}{\sim} p(Z)$, we can have samples $\{\eta_i, T_{\hat{\vartheta}}(\eta_i, Z_i)\}$ from this cut-posterior. We call our method *NeVI-Cut (neural variational inference for cut-Bayes)*. Note that NeVI-Cut is completely agnostic to what upstream data D_1 and model were used to obtain the samples $\{\eta_i\}$ as long as averages using this set of samples approximate population averages with respect to $p(\eta \,|\, D_1)$. It also works with any choice of base distribution, any dimension and support set of the parameters θ and η, and any class of normalizing flows. Section S2 presents some important classes of normalizing flows that we use in practice. The theoretical justification for their use is presented in Section 4.

3.2 Triply-Stochastic Algorithm

The computational challenges in calculating the objective in (8) can be three-pronged. First, a large number of samples η_i may be available from the upstream analysis. Computing the NeVI-Cut ELBO, averaged over all the upstream samples, for every iteration can be slow. Second, the expectation with respect to Z is typically not available in closed form. Third, the likelihood $p(D_2 \,|\, \theta, \eta)$ might also be slow to compute especially for large datasets. We address these challenges by leveraging three sources of stochasticity.

Mini-batching or stochastic gradient descent is widely used to speed up neural network optimization. The main idea is that terms in the objective function involving averages or sums over many terms can be replaced by averages (or appropriately scaled sums) over smaller subsets (mini-batches) of these terms. Only one mini-batch is used at a time, and the mini-batches are cycled throughout the optimization. Stochastic gradient descent is an extreme case with a mini-batch size of one. Each of these three challenges has a scalable solution using such stochastic approximations. At each iteration, we can replace the average over all N samples of η with a small minibatch $\mathcal{B}_\eta$ of size N_η. Similarly, the expectation with respect to Z is replaced by an average of N_Z iid observations $\{Z_j\}$ sampled from $p(Z)$. The doubly stochastic approximation of (8) is thus given by

$$\begin{aligned}\hat{\mathcal{L}}_{DS}(\vartheta) = \frac{1}{N_\eta N_Z} \sum_{i \in \mathcal{B}_\eta} \sum_{j=1}^{N_Z} \Big[&\log |\det J_z T_\vartheta(\eta_i, Z_j)| + \\ &\log p(D_2 \,|\, T_\vartheta(\eta_i, Z_j), \eta_i) + \log p(T_\vartheta(\eta_i, Z_j) \,|\, \eta_i) \Big].\end{aligned} \tag{10}$$

Finally, if the downstream data D_2 consists of iid units $Y_1, \ldots, Y_n$, then we have $p(D_2 \,|\, \theta, \eta) =$

Algorithm 1 Neural Variational Inference for Cut-Bayes (NeVI-Cut)

Input: Upstream posterior samples $\{\eta_i\}_{i=1}^N$, downstream data D_2, functional form of downstream likelihood $p(D_2 \,|\, \theta, \eta)$ and downstream prior $p(\theta \,|\, \eta)$, base distribution $p(Z)$, maximum number of iterations S, patience S_{patience} (number of steps before early stopping), normalizing-flow-based transformation T_ϑ with parameter ϑ (e.g., Neural Spline Flows or Unconstrained Monotonic Neural Network Flows), learning rate α.
Initialize: Set some initial values of ϑ
for $s = 1$ to S **do**
  Sample base variables $\{Z_i\}_{i=1}^N \overset{\text{iid}}{\sim} p(Z)$
  Transform $\theta_i = T_\vartheta(\eta_i, Z_i)$ for $i = 1, \ldots, N$
  Estimate ELBO:

$$\hat{\mathcal{L}}_Z(\vartheta) = \frac{1}{N} \sum_{i=1}^{N} \big[\log |\det J_z T_\vartheta(\eta_i, Z_i)| + \log p(D_2 \mid \theta_i, \eta_i) + \log p(\theta_i \mid \eta_i) \big].$$

  Update parameters $\vartheta \leftarrow \vartheta + \alpha \operatorname{Adam}\big(\nabla_\vartheta \hat{\mathcal{L}}_Z(\vartheta)\big)$(gradient ascent)
  Early stopping: End if $\hat{\mathcal{L}}_Z(\vartheta)$ fails to improve for S_{patience} steps
end for
Final estimate: $\hat{\vartheta} \leftarrow \vartheta$
Samples: $\widetilde{Z}_i \overset{\text{iid}}{\sim} p(Z)$; $\widetilde{\theta}_i = T_{\hat{\vartheta}}(\eta_i, \widetilde{Z}_i)$ for $i = 1, \ldots, N$
Output: Trained conditional flow $\hat{q}(\theta \,|\, \eta) \overset{\text{d}}{=} T_{\hat{\vartheta}}(\eta, Z), Z \sim p(Z)$, cut posterior estimate $[\theta \,|\, \eta \sim \hat{q}(\theta \,|\, \eta)] \times [\eta \sim \text{Unif}\{\eta_1, \ldots, \eta_N\}]$, samples from the cut-posterior $\{\eta_i, \widetilde{\theta}_i\}_{i=1}^N$.

$\prod_{i=1}^n p(Y_i \,|\, \theta, \eta)$. And $\log p(D_2 \,|\, \theta, \eta)$ can be approximated using a minibatch $\mathcal{B}_D$ of size N_D as $\frac{n}{N_D} \sum_{k \in \mathcal{B}_D} \log p(Y_k \,|\, \theta, \eta)$. This gives the triply stochastic approximation

$$\begin{aligned} \hat{\mathcal{L}}_{TS}(\vartheta) = \frac{1}{N_\eta N_Z} \sum_{i \in \mathcal{B}_\eta} \sum_{j=1}^{N_Z} \Big[&\log |\det J_z T_\vartheta(\eta_i, Z_j)| + \\ &\frac{n}{N_D} \sum_{k \in \mathcal{B}_D} \log p\big(Y_k \,|\, T_\vartheta(\eta_i, Z_j), \eta_i\big) + \log p(T_\vartheta(\eta_i, Z_j) \,|\, \eta_i) \Big]. \end{aligned} \tag{11}$$

In practice, all three approximations are not always required simultaneously, and the level of stochasticity, i.e., choices of the minibatch sizes N_η, N_Z, N_D can be adjusted depending on computational trade-offs. For example, if the data are not iid or not large, mini-batching of the likelihood is not recommended, and the third source can be omitted. In our experiments, to ensure comparability with multiple imputation which uses all the upstream samples $\{\eta_i\}$, we restrict stochasticity to Monte Carlo integration over Z. We also combine the stochastic approximations of η and Z, using the fact that the integration is over the product density $(\eta, Z) \sim p(\eta \,|\, D_1)\, p(Z)$. For each η_i we can sample $Z_i \overset{\text{iid}}{\sim} p(Z)$ and $\{(\eta_i, Z_i)\}$ is a sample from $p(\eta \,|\, D_1)\, p(Z)$, and use the objective

$$\hat{\mathcal{L}}_Z(\vartheta) = \frac{1}{N} \sum_{i=1}^{N} \big[\log |\det J_z T_\vartheta(\eta_i, Z_i)| + \log p(D_2 \,|\, T_\vartheta(\eta_i, Z_i), \eta_i) + \log p(T_\vartheta(\eta_i, Z_i) \,|\, \eta_i) \big].$$

Optimization is performed using the Adam algorithm (Diederik, 2014). The overall procedure is summarized in Algorithm 1.

4 Theory

In this section, we provide theoretical results on guarantees on universal approximation of the true joint cut-posterior $p_{\text{cut}}(\theta, \eta \,|\, D_1, D_2)$ by the NeVI-Cut estimator $\hat{p}_{\text{cut}}(\theta, \eta \,|\, D_1, D_2)$ defined as the variational solution in (9), when the conditional variational family $q(\theta \,|\, \eta)$ is modeled using conditional normalizing flows. Our results show that the convergence rates of the estimator depend on the richness of the neural network class used in the conditional flows and the complexity of the true cut-posterior.

4.1 Uniform Kullback-Leibler Approximation Rates of Conditional Normalizing Flows

We first provide results on universal approximation rates (in Kullback-Leibler divergence) of conditional normalizing flows in terms of the richness of the neural network architecture and complexity of the class of conditional densities it is approximating. These results are then used to prove the convergence of NeVI-Cut, but are also of independent importance. Much of the existing theory on fixed target approximation qualities of conditional normalizing flows is about their expressive powers (Huang et al., 2018; Papamakarios et al., 2017; Battaglia and Nicholls, 2024) showing that the class of (conditional) normalizing flows can approximate any true (conditional) distribution in terms of weak convergence. When used as variational families, one needs to study properties of the variational solution among the conditional flow class. Weak convergence results does not help here as the variational solution is obtained by minimizing the KLD. We need to quantify uniform rates of approximation of flows to a target class of densities in the KLD. This, to our knowledge, has not been developed.

We develop a theory for the setting where the conditioning variable η is multivariate, and the observation variable θ is univariate. This is because autoregressive flows typically specify a sequence of univariate conditionals (i.e., where the observation variable is univariate while the conditioning variable is multivariate). So this setting helps to illustrate the main proof ideas and insights from the theory while balancing the technical details. We first define a suitable class of conditional densities. The extension to multivariate θ should be conceptually straightforward but cumbersome in terms of notation.

We first specify the true class of target conditional distributions.

Assumption 1 (True distribution class)**.** *Let $K_\eta \in \mathbb{R}^{d_\eta}$ be some compact interval, and $K_M = [-M, M] \times K_\eta$. Let the target conditional posterior $p(\theta \,|\, \eta)$ be in a class $\mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$ of positive, continuous probability densities $p(\theta \,|\, \eta)$ supported on $\mathbb{R} \times K_\eta$, characterized by the following properties:*

(i) Spikiness rate (log-Lipschitz in θ): *There is a positive increasing function $L_S : \mathbb{N} \to (0, \infty)$ such that,*

$$\left|\log p(\theta \,|\, \eta) - \log p(\theta' \,|\, \eta)\right| \;\leqslant\; L_S(M)\,|\theta - \theta'|,\;\; \forall (\theta, \eta), (\theta', \eta) \in [-M, M] \times K_\eta,\;\; M \in \mathbb{N}.$$

(ii) Conditionals perturbation rate (log-Lipschitz in η): *There is a positive increasing function $L_{CP} : \mathbb{N} \to (0, \infty)$ such that,*

$$\left|\log p(\theta \,|\, \eta) - \log p(\theta \,|\, \eta')\right| \;\leqslant\; L_{CP}(M)\,\|\eta - \eta'\|_1,\;\; \forall (\theta, \eta), (\theta, \eta') \in [-M, M] \times K_\eta,\;\; M \in \mathbb{N}.$$

(iii) Quantile growth rate: *For each $p \in \mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$ and $\eta \in K_\eta$, let*

$$F_p(\theta \,|\, \eta) = \int_{-\infty}^{\theta} p(t \mid \eta)\, dt, \qquad T_p(\eta, z) = F_p^{-1}\big(\Phi(z) \,|\, \eta\big), \quad z \in \mathbb{R}, \tag{12}$$

with Φ the $N(0,1)$ CDF. So $F_p(\theta \,|\, \eta)$ is the conditional cdf, $F_p^{-1}(u \,|\, \eta)$ for $u \in [0,1]$ is the conditional quantile map (with respect to the default Unif$[0,1]$ *base) and T_p is the conditional quantile map transformed with respect to $N(0,1)$ base. Then,*

(a) Median bound: *There exists $M_0 \geqslant 0$ such that $\sup_{\eta \in K_\eta} |T(\eta, 0)| = M_0 < \infty$,*

(b) Derivative bound: *$T(\eta, z)$ is strictly increasing and C^1 (continuously differentiable) in z and there exists constants $A_d, B_d \geqslant 0$ such that*

$$\sup_{(\eta, z) \in K_\eta \times \mathbb{R}} |\log \partial_z T(\eta, z)| \leqslant H(z) := A_d + B_d |z|.$$

(iv) Tail thinness lower bound: *There exists $\alpha > 0$, $c > 0$ such that $p(\theta \mid \eta) \geqslant c\, e^{-\alpha \theta^2}$ for all $(\theta, \eta) \in \mathbb{R} \times K_\eta$.*

Instead of studying the properties of our method (and in general of variational conditional flows) for a specific density $p(\theta \,|\, \eta)$, we assume $p(\theta \,|\, \eta)$ satisfy Assumption 1. This enables studying the entire class of conditional families $\mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$ characterized by the spikiness, conditional perturbation rate, quantile growth, and thinness bound. This is much like how Hölder or Sobolev classes provide functional envelopes for smoothness and decay and are used to derive the theory of functional estimation for these classes rather than individual functions.

Assumptions 1(i) and (ii) specify that the log conditional density is local Lipschitz on compacts, with respect to both θ and η. The log-Lipschitz moduli $L_S(M)$ and $L_{CP}(M)$ will, of course, be nonnegative and non-decreasing in M. The modulus $L_S(M)$ controls how spiky the conditional density is with respect to the observation variable θ. The parameter $L_{CP}(M)$ controls how smoothly the conditional law varies with the conditioner η. Note that any locally log-Lipschitz continuous positive conditional density $p(\theta \,|\, \eta)$ supported on $\mathbb{R} \times K_\eta$ satisfies Assumptions 1(i) and (ii). The smallest L_S, L_{CP} for such a density $p(\theta \,|\, \eta)$ is induced by its own moduli and growth rates

$$L_S^*(M) := \sup_{\eta \in K_\eta} \sup_{\substack{\theta \neq \theta' \\ |\theta\,|, |\theta'| \leqslant M}} \frac{|\log p(\theta \mid \eta) - \log p(\theta' \mid \eta)|}{|\theta - \theta'|},$$
$$L_{CP}^*(M) := \sup_{|\theta\,| \leqslant M} \sup_{\eta \neq \eta'} \frac{|\log p(\theta \mid \eta) - \log p(\theta \mid \eta')|}{\|\eta - \eta'\|_1},$$

both of which exist because of local Lipschitz continuity and positivity. So the only real assumptions are the third and fourth conditions in Assumption 1.

In Assumption (iii), the parameters M_0, A_d, B_d govern the global spread of the distribution via controlling the conditional quantile map. From the median and derivative bounds, we immediately have the following bound on the conditional quantile map:

$$|T_p(\eta, z)| \;\leqslant\; G(z) := M_0 + |z| e^{A_d + B_d |z|} \quad \forall (\eta, z). \tag{13}$$

Thus, we allow the quantile map can grow upto exponentially faster than the base (Gaussian) quantile, enabling it to model densities with a certain degree of thicker tails than Gaussian. As M grows, most of the probability mass remains within $[-G(M), G(M)]$. Faster growth of $G(M)$ corresponds to heavier tails (relative to the normal base).

The triplet (L_S, L_{CP}, G) describes different aspects of the conditional distributions — respectively, local sharpness in θ, sensitivity across η, and quantile growth relative to the Gaussian base. Table S1 summarizes these three quantities for common conditional families whose parameters (e.g., location, scale, shape) depend continuously on η, showing that families of Gaussian, Laplace (double exponential), Generalized Gaussian, and log-normal distributions, or their finite mixtures, satisfy Assumption 1. For thicker tailed families, one should use normalizing flows with thicker thicker-tailed base distribution than $N(0, 1)$, and similar theoretical results like we derive here can be obtained.

The following lemma shows that Assumption 1(i)–(iii) leads to local-Lipschitz bounds on the conditional median and the log-derivative of the conditional quantiles, which feature in

the final approximation rates.

Lemma 1. *Let $p \in \mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$ and T_p denote the corresponding conditional quantile function with a Gaussian base as defined in (12). Define*

$$\begin{aligned} \mathsf{L}_a &:= \frac{2}{c}\, e^{\alpha M_0^2}\, e^{L_{CP}(M_0)\,\mathrm{diam}(K_\eta)}\, L_{CP}(M_0) \\ \mathsf{L}_b(M) &:= M \;+\; e^{H(M)}\, L_S\big(G(M)\big) \;+\; L_{CP}\big(G(M)\big), \end{aligned} \tag{14}$$

Then on K_η, the conditional median $T_p(\eta, 0)$ is L_a-Lipschitz and on $(z, \eta) \in K_M$, the log-derivative $\log \partial_z T_p$ is $\mathsf{L}_b(M)$-Lipschitz with respect to the ℓ_1 distance $\|(\eta, z) - (\eta', z')\|_1 := \|\eta - \eta'\|_1 + |z - z'|$.

Finally, in Assumption 1(*iv*) we impose the Gaussian-type lower tail bound

$$p(\theta \,|\, \eta) \;\geqslant\; c\, e^{-\alpha\theta^2} \qquad \text{for all } (\theta, \eta) \in \mathbb{R} \times K_\eta,$$

to rule out pathologically thin tails (as the Gaussian tail itself is quite thin). In variational inference, the candidate $q(\theta \,|\, \eta)$ is generally required to have mass wherever the target density has mass. If the target density is too thin for very long stretches, it may require a highly irregular variational approximant. The lower bound assumption prevents this. We emphasize that this is a lower bound on tail thinness and accommodates all heavier tails as long as they satisfy Assumption 1(*iii*).

We first state a simple identity relating conditional density to the conditional quantile map which is used at multiple places in the theory.

Lemma 2. *Let a positive conditional density $p(\theta \,|\, \eta)$ satisfy Assumption 1. Define the conditional quantile map T_p as in (12). Then T_p is continuous on $K_\eta \times \mathbb{R}$, strictly increasing and continuously differentiable in z, and*

$$\log p\left(T_p(\eta, z) \mid \eta\right) \;=\; \log \phi(z) \;-\; \log \partial_z T_p(\eta, z). \tag{15}$$

As the KLD features the log-density, Lemma 2 helps write the KLD in terms of the derivative of the log-quantile function and the log base density $\phi(z)$, enabling us to apply respective bounds for each term to bound the KLD.

Our second assumption is on the class of conditional normalizing flows used as the variational family.

Assumption 2 (Neural Network Flow Classes)**.** *For integers $m \geqslant 1$, let*

$$\mathcal{F}_m \;:=\; \{\, T_\vartheta : K_\eta \times \mathbb{R} \to \mathbb{R} \;:\; \vartheta \in \Theta_m \,\}$$

where $\Theta_m \subset \mathbb{R}^m$ is compact. These families satisfy:

(i) Regularity and monotonicity. *For every m and $\vartheta \in \Theta_m$, $T_\vartheta(\eta, z)$ is continuous in (η, z, ϑ), is C^1 in z and strictly increasing in z, with $\delta_z T_\vartheta(\eta, z)$ being continuous in ϑ.*

(ii) Global linear envelopes. *There exists positive constants M^*, A^*, B^* such that*

$$\sup_{m\geqslant 1}\sup_{\vartheta\in\Theta_m}\sup_{\eta\in K_\eta}|T_\vartheta(\eta,0)| \leqslant M^*, \quad \sup_{m\geqslant 1}\sup_{\vartheta\in\Theta_m}\sup_{(\eta,z)\in K_\eta\times\mathbb{R}}|\log\partial_z T_\vartheta(\eta,z)| \leqslant A^* + B^*|z|.$$

(iii) Universal approximation. *For all conditional density $p \in \mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$, with T_p as in (12), and for every $\varepsilon > 0$, there exists an integer $m = C(M, \varepsilon)$ and $\vartheta \in \Theta_m$ such that*

$$\sup_{\eta\in K_\eta}\left|T_\vartheta(\eta,0) - T_p(\eta,0)\right| \leqslant \varepsilon, \qquad \sup_{(z,\eta)\in K_M}\left|\log\partial_z T_\vartheta(\eta,z) - \log\partial_z T_p(\eta,z)\right| \leqslant \varepsilon.$$

We will show in Theorems 2 and 3 that popular conditional normalizing flow specifications satisfy Assumption 2. First, we provide a more general result on KL approximation rates for any class of conditional normalizing flows satisfying the assumption.

Theorem 1 (Uniform KL rates of conditional flows)**.** *Let q_T be a class of conditional densities corresponding to $\theta \,|\, \eta = T(\eta, Z)$, $Z \sim N(0,1)$, and $T \in \mathcal{F}_m$ for some class of conditional flows $\mathcal{F}_m$ satisfying Assumption 2 and with $M^* \geqslant M_0, A^* \geqslant A_0, B^* \geqslant B_0$, and complexity budget $C(M, \varepsilon)$ for a class of conditional densities $\mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$. Define $\widetilde{G}(M) := M^* + |M|e^{A^*+B^*|M|}$. Then for any*

$$m \geqslant C^*(\varepsilon) := C\left(R\sqrt{-\log\varepsilon}, \frac{\varepsilon}{4\left(1+\widetilde{G}(R\sqrt{-\log\varepsilon})L_S(\widetilde{G}(R\sqrt{-\log\varepsilon}))\right)}\right),$$

there exists $T \in \mathcal{F}_m$ such that

$$\sup_{\eta\in K_\eta}\mathrm{KL}\big(q_T(\cdot \mid \eta)\,\|\,p(\cdot \mid \eta)\big) \leqslant \varepsilon,$$

for a constant $R > 0$ depending only on (and increasing in) M^*, A^*, B^*, α, c.

Theorem 1 is a result of independence interest. To our knowledge, it is the first fixed-target uniform KL approximation rates for conditional normalizing flows and can be used to understand KLD between any target conditional posterior and the variational solution based on conditional normalizing flows. As discussed in the Introduction, existing theoretical results on this topic are either just on the expressiveness of conditional normalizing flows in terms of weak convergence that does not help studying KL-based variational solutions, or are statistical guarantees assuming data growth.

We now show how Theorem 1 applies to popular conditional normalizing flow classes, with explicit quantification of the number of parameters needed to reach a given accuracy ε. The next result is for the class of Unconstrained Monotonic Neural Networks (UMNN; Wehenkel and Louppe, 2019) where the conditional distribution $q_T(\theta \,|\, \eta)$ is specified via T as in (S6) with the functions a and g modeled as ReLU networks.

Theorem 2 (Rate for Conditional Unconstrained Monotonic Neural Networks)**.** *Let $\eta \in \mathbb{R}^{d_\eta}$, $\mathrm{diam}_\eta = \mathrm{diam}(K_\eta)$, $p(\theta \,|\, \eta)$ satisfies Assumption 1. For any small $\varepsilon > 0$, let $M(\varepsilon) = R\sqrt{-\log \varepsilon}$ where R is as in Theorem 1, and define $\widetilde{L}_S(M) = 2ML_S(M)$. Let $q_T(\theta \,|\, \eta)$ follows the law of $T(\eta, Z)$ where $Z \sim N(0,1)$ and $T(\eta, z)$ is a UMNN (S6) on $K_\eta \times \mathbb{R}$ with a and g being clipped ReLU networks with parameter sets ϑ_a and ϑ_g. Then there is a UMNN q_T with*

$$
\begin{aligned}
|a_\vartheta| &\leqslant C_a \left(\frac{\widetilde{L}_S\big(\widetilde{G}(M(\varepsilon))\big) \mathsf{L}_a \, \mathrm{diam}_\eta}{\varepsilon} \right)^{d_\eta} \log \frac{L_S\big(\widetilde{G}(M(\varepsilon))\big)}{\varepsilon}, \\
|g_\vartheta| &\leqslant C_g \left(\frac{\widetilde{L}_S\big(\widetilde{G}(M(\varepsilon))\big) \mathsf{L}_b(M(\varepsilon))(\mathrm{diam}_\eta + M(\varepsilon))}{\varepsilon} \right)^{d_\eta + 1} \log \frac{\widetilde{L}_S\big(\widetilde{G}(M(\varepsilon))\big)}{\varepsilon},
\end{aligned}
$$

such that

$$
\sup_{\eta \in K_\eta} \mathrm{KL}\big(q_T(\cdot \mid \eta) \,\|\, p(\cdot \mid \eta)\big) \;\leqslant\; \varepsilon.
$$

Here C_a, C_g are some universal constants depending only on d_η, and L_a and $\mathsf{L}_b(\cdot)$ are the local Lipschitz constants defined in Lemma 1.

The next Theorem gives a similar result for Rational Quadratic Neural Spline Flows (also introduced in Section S2).

Theorem 3 (Rate for Conditional Rational Quadratic Neural Spline Flows)**.** *Let $\eta \in \mathbb{R}^{d_\eta}$, $\mathrm{diam}_\eta = \mathrm{diam}(K_\eta)$, and suppose $p(\theta \mid \eta)$ satisfies Assumption 1. Let $T(\eta, z) : K_\eta \times \mathbb{R} \to \mathbb{R}$ be a conditional rational-quadratic neural spline flow (RQ-NSF) on equispaced knots in $[-M, M]$, whose per-bin heights and slope logits are produced by a ReLU network with m parameters. Write $q_T(\theta \mid \eta)$ for the law of $T(\eta, Z)$, $Z \sim \mathcal{N}(0,1)$. For any small $\varepsilon > 0$, set $M = M(\varepsilon) = R\sqrt{-\log \varepsilon}$ (with R as in Theorem 1), and define $\widetilde{L}_S(M) = 2ML_S(M)$ and*

$$
K_\star \;=\; \left\lceil \frac{4\, L_b(M)\, M^2 e^{H(M)} \widetilde{L}_S\left(\widetilde{G}(M(\varepsilon))\right)}{\varepsilon} \right\rceil.
$$

Then there exists an RQ-NSF T with $2K^*$ bins such that

$$
\sup_{\eta \in K_\eta} \mathrm{KL}\big(q_T(\cdot \mid \eta) \,\|\, p(\cdot \mid \eta)\big) \;\leqslant\; \varepsilon,
$$

and its parameter count satisfies

$$m \leqslant C_r \left(4K_\star + 2\right) A^{d_\eta} \log A,$$
$$\text{where } A := \frac{4\, M\, e^{H(M)} \widetilde{L}_S \left(\widetilde{G}(M(\varepsilon))\right) \left[2e^{2H(M)} L_b(M) + \frac{e^{H(M)}}{M} L_{CP}\big(G(M)\big)\right] \operatorname{diam}(K_\eta)}{\varepsilon}$$

and C_r is a constant depending only on d_η.

The expression of the number of parameters in both Theorems 2 and 3 has similar terms, with the number increasing with the Lipschitz bounds raised to the exponent d_η. This shows that the number of parameters needed to reach a given error rate grows exponentially with the dimension of the function. This is not surprising and aligns with the complexity theory of ReLU neural networks (Yarotsky, 2017) which was used to derive these complexity budgets.

4.2 Theory of NeVI-Cut

We now establish error rates of the NeVI-Cut estimate approximating a target cut-posterior using the general theory of Section 4.1. We first provide guarantees on existence and KL approximation rates of the NeVI-Cut estimate obtained as the variational solution based on the loss (5), which integrates the conditional KL over the exact cut distribution $p(\eta \,|\, D_1)$. We hide the conditioning datasets as they remain fixed for all the theory and simply use $p_{\text{cut}}(\theta, \eta)$, $p_{\text{cut}}(\theta \,|\, \eta)$ and $p_{\text{cut}}(\eta)$ to denote the joint, conditional, and marginal cut-posteriors. For any conditional distribution $q(\theta \,|\, \eta)$, define the expected conditional KLD to the conditional cut-posterior as

$$R(q) = \int \mathrm{KL}\big(q(\cdot \mid \eta) \,\|\, p_{\text{cut}}(\cdot \mid \eta)\big)\, p_{\text{cut}}(\eta)\, d\eta. \tag{16}$$

Note that $R(q)$ is also the KLD of the joint density $q(\theta \,|\, \eta) p_{\text{cut}}(\eta)$ to the joint cut-posterior $p_{\text{cut}}(\theta, \eta)$, i.e.,

$$R(q) = \mathrm{KL}\Big(q(\cdot \,|\, \cdot) p_{\text{cut}}(\cdot) \,\|\, p_{\text{cut}}(\cdot, \cdot)\Big). \tag{17}$$

Corollary 1 (Integral-based NeVI-Cut)**.** *Let the target cut posterior be $p_{cut}(\theta, \eta) = p_{cut}(\eta) p_{cut}(\theta \,|\, \eta)$ where $p_{cut}(\eta)$ is a distribution with support within a compact set $K_\eta \in \mathbb{R}^{d_\eta}$ and $p_{cut}(\theta \,|\, \eta)$ is in $\mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$. Let $\mathcal{F}_m$ be a class of the conditional flows that satisfies Assumption 2 for the class $\mathcal{P}\left(L_S\left(\cdot\right), L_{CP}\left(\cdot\right), M_0, A_d, B_d, c, \alpha\right)$ with $M^* \geqslant M_0, A^* \geqslant A_d, B^* \geqslant B_d$, and complexity budget $C(M, \varepsilon)$. Let Q_m be the class of conditional distributions*

corresponding to $\mathcal{F}_m$*, i.e.,*

$$\mathcal{Q}_m := \big\{\, q_T(\cdot \mid \eta) := \mathrm{Law}\,\big(T(\eta, Z)\big),\ Z \sim \mathcal{N}(0,1)\ :\ T \in \mathcal{F}_m \,\big\}.$$

Then, for any small $\varepsilon > 0$*, we have the following conclusions:*

(a) For every $m \geqslant 1$*, a minimizer* $\hat{q}_m \in \arg\min_{q \in \mathcal{Q}_m} R(q)$ *exists.*

(b) Fix any accuracy small $\varepsilon > 0$ *and let* $m \geqslant C^*(\varepsilon)$ *be the budget of Theorem 1. Then* $R(\hat{q}_m) = \min_{q \in \mathcal{Q}_m} R(q) \leqslant \varepsilon$*.*

The result shows that the NeVI-Cut solution, using the integrated KL loss (7), approximates the true cut-posterior up to any error rate ε for a suitably large neural network architecture. For specific choices of flows like UMNN or RQ-NSF, the complexities given by Theorems 2 and 3 hold. The result is in the KLD, which is stronger than the total-variation or the weak distributional metric. In practice, of course, optimization cannot be done using the exact integration as in (5). We use the N samples from $p_{\mathrm{cut}}(\eta)$ and optimize (6). We now provide the analogous bounds for this actual NeVI-Cut solution.

Theorem 4 (Sample-based NeVI-Cut)**.** *Consider the setting of Corollary 1. Let* $\{\eta_i\}_{i=1}^N$ *be samples from a stationary ergodic Markov chain with invariant distribution* $p_{cut}(\eta)$*. For* $q \in \mathcal{Q}_m$ *set*

$$R_N(q) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}\big(q(\cdot \mid \eta_i) \,\|\, p(\cdot \mid \eta_i)\big)$$

to be the loss used in NeVI-Cut. Then, for each N*, the minimizer* $\hat{q}_{N,m} \in \arg\min_{q \in \mathcal{Q}_m} R_N(q)$ *exists. Moreover, for any small* $\varepsilon \geqslant 0$*, we have* $\lim_{N \to \infty} R(\hat{q}_{N,m}) \leqslant \varepsilon$*.*

Theorem 4 provides existence guarantee and error bound in the average KLD for the actual NeVI-Cut solution using the sample-based loss. The result only assumes that the samples are from a stationary ergodic Markov chain which covers the iid case (Monte Carlo samples) but also covers the more realistic case where the samples are MCMC posterior samples from an upstream Bayesian analysis.

5 Simulation Study

In this section, we first present a simulation study that supports the theoretical justification of our proposed NeVI-Cut, showing that it can accurately recover both the conditional distribution of the downstream parameter and its marginal distribution. We then introduce two

examples from the literature that are frequently discussed in the context of cutting feedback. Because existing cutting feedback methods require access to the upstream data, for this set of illustrations, we have access to the upstream data. Hence, for NeVI-Cut, we first fit an MCMC to the upstream data to obtain posterior samples and then use these samples in Algorithm 1. Notably, these MCMC samples can be replaced by posterior draws obtained from any alternative posterior estimation method for the upstream parameters. In our experiments, the multiple imputation-based nested MCMC for cut-Bayes and the full Bayes are both implemented in `rstan` (Stan Development Team, 2025), while the variational inference-based methods, including Gaussian VA-Cut (Yu et al., 2023) and NeVI-Cut, are implemented in Python 3.10.

5.1 Expressiveness of NeVI-Cut

We first illustrate the representational capacity of NeVI-Cut for multimodal conditional distributions. We take $p(\eta \mid D_1) = \text{Gamma}(2, 1)$, and consider the downstream conditional posterior $p(\theta \mid \eta, D_2)$ which follows a mixture of normal distributions such that:

$$p(\theta \mid \eta, D_2) = \pi(\eta) N\left(\mu_1(\eta), \sigma_1^2\right) + (1 - \pi(\eta))\, N\left(\mu_2(\eta), \sigma_2^2\right),$$

where $\pi(\eta) = 0.2 + 0.5\,\sigma\big(4(\eta - 2)\big)$, $\mu_1(\eta) = 4\tanh(\eta - 1)$, $\mu_2(\eta) = -4\tanh(\eta + 1)$, $\sigma_i^2 = 1.5$.

We model T_ϑ in (9) using two different classes of flows: an Unconstrained Monotonic Neural Network (UMNN; Wehenkel and Louppe, 2019) and a Rational Quadratic Neural Spine Flow with Autoregressive Structure (RQ-NSF(AR); Durkan et al., 2019). We call these NeVI-Cut.UMNN and NeVI-Cut.NSF.AR respectively. In NeVI-Cut.UMNN(S6), the neural network $a(\cdot)$ is a 5-layer multilayer perceptron (MLP) with LeakyReLU activation, and $g(\cdot)$ is a 7-layer MLP with LeakyReLU activation. The base distribution is standard normal.

Figure 1 compares the NeVI-Cut estimates of conditional cut-posteriors $p(\theta \mid \eta = \eta_0)$ with their true values for different choices of η_0, as well as the marginal cut-posterior $p_{\text{cut}}(\theta)$. To assess the expressiveness of NeVI-Cut for a variety of density shapes, we consider four different choices of η_0 that lead to four different shapes of the conditioning cut-posterior (left panel). For $\eta = 0.99$, the density is unimodal with a slight notch to one side of the mode. For $\eta = 1.39$, there is a weak second mode. The remaining two choices represent clearly bimodal densities, with $\eta = 2.09$ having two roughly equal models, and $\eta = 3.05$ having unequal ones. The marginal cut-posterior (right panel) is also bimodal with unequal modes. The results show that NeVI-Cut with either choice of flows (UMNN or NSF) is highly accurate, correctly

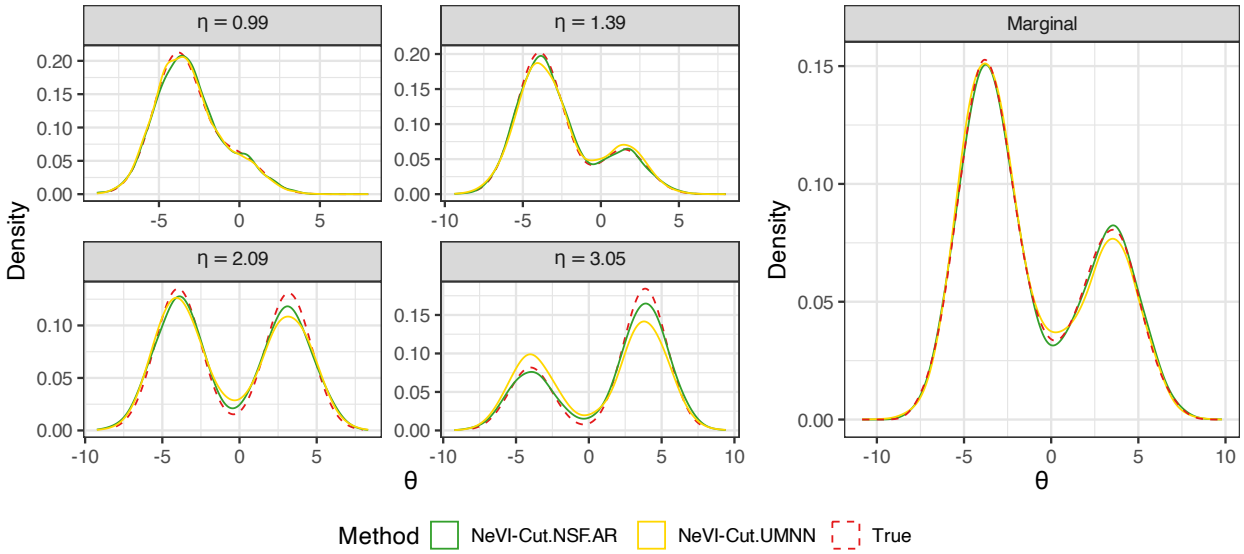


Figure 1: Comparison of NeVI-Cut estimates using RQ-NSF(AR) and UMNN flows with the true distributions when estimating the conditional cut-posterior $p(\theta \mid \eta = \eta_0)$ (left) and marginal cut-posterior $p(\theta)$ (right).

capturing all conditional and marginal cut-posteriors. This small experiment illustrates the expressiveness of NeVI-Cut to model densities of varied shapes. However, NeVI-Cut.UMNN can be computationally demanding due to the need to evaluate integrals. Since the two choices of flows produce nearly identical results, we use the faster approach, NeVI-Cut.NSF.AR, in all subsequent analyses, and refer to it simply as NeVI-Cut.

5.2 Model Misspecification Examples

We revisit two commonly used illustrative examples (Liu et al., 2009; Jacob et al., 2017; Yu et al., 2023) that demonstrate the challenges of full Bayesian inference when there is model misspecification or biased data.

We first consider a case of a misspecified prior in the downstream analysis. Details of the data generation are provided in Section S6.1. This is a setting where both the full and the cut-Bayes posteriors are available in closed form. In addition to these, we also implement NeVI-Cut and parameteric cut of Yu et al. (2023) which uses a Gaussian variational family.

Table 1: Predictive performance metrics of methods in simulation under a misspecified prior. Lower values of interval score, Continuous Ranked Probability Score (CRPS), and Mean Squared Error (MSE) indicate better performance.

Method	95% Interval Score	CRPS	MSE	95% Coverage
Full Bayes	0.387 (0.346, 0.480)	0.481 (0.440, 0.574)	0.275 (0.231, 0.378)	0.00
True Cut	0.016 (0.009, 0.090)	0.077 (0.029, 0.214)	0.017 (0.002, 0.072)	0.84
Gaussian VA-Cut	0.062 (0.003, 0.213)	0.092 (0.024, 0.250)	0.017 (0.001, 0.071)	0.35
NeVI-Cut	0.016 (0.009, 0.092)	0.077 (0.028, 0.213)	0.017 (0.001, 0.071)	0.83

Table 1 summarizes the performance of different methods. Full Bayes performs uniformly worse compared to other methods, showing the impact of the highly misspecified prior. Conversely, the true cut-posterior has the best metrics, demonstrating the utility of cut-posterior under downstream model misspecification. The Gaussian VA-cut is definitely an improvement over full Bayes and has similar accuracy of the posterior point-estimate as the true Cut (both having the same MSE). However, in all the remaining three metrics which capture distributional accuracy, it is much worse than the true cut, implying that the estimated shape of the cut-posterior from this method is inaccurate. This is because the Gaussian VA method uses a mean-field approximation that tends to underestimate posterior variance. This leads to poor uncertainty quantification and distributional accuracy, reflected in the very low coverage, and higher interval scores and CRPS. All four metrics for NeVI-Cut are nearly indistinguishable from those of the true cut, implying the high degree of accuracy with which it approximates the cut-posterior.

We then present an example of the utility of cut-posterior for misspecified downstream outcome model (likelihood). Cut-posteriors are particularly useful in observational studies where, for example, the propensity score is first modeled and then incorporated into the outcome model to estimate causal effects. Hence, propensity score estimation can be viewed as an upstream analysis. Thus, one can deploy cutting feedback methods to propagate uncertainty in estimates of propensity scores into the downstream outcome model. Following Jacob et al. (2017), we first estimate propensity scores and then model the effect of a binary treatment X on a binary outcome Z, adjusting for potential confounders $C \in \mathbb{R}^{n \times p}$. The details of both the data generation model and the upstream and downstream analysis models are given in Section S6.2. The upstream analysis model of the propensity scores (treatment model) is correctly specified, but the outcome model is misspecified. The true causal effect is null, and we evaluate whether each method can recover this correctly via their estimate of $\theta_{2.1}$.

We obtain marginal posterior samples for this parameter under different approaches and compare NeVI-Cut with the full Bayes posterior and the multiple imputation-based nested MCMC for the cut-posterior (Plummer, 2015). The same metrics as in Table 1 are used for performance evaluation, and the findings are summarized in Table 2. Because the true effect is zero, a well-specified method should yield point estimates close to zero with valid uncertainty quantification. The full Bayes posterior suffers from the outcome model misspecification, leading to biased estimates and poor coverage. In contrast, NeVI-Cut once again attains performance nearly identical to that of nested MCMC, both in terms of accuracy

Table 2: Performance comparison in the propensity score simulation. Lower values of interval score, CRPS, and MSE indicate better performance.

Method	95% Interval Score	CRPS	MSE	Time (seconds)	95% Coverage
Full Bayes	0.032 (0.015, 0.154)	0.130 (0.049, 0.350)	0.049 (0.005, 0.190)	22919 (14864, 38509)	0.745
Nested MCMC	0.021 (0.015, 0.062)	0.089 (0.043, 0.261)	0.025 (0.003, 0.125)	6344 (5783, 7199)	0.964
NeVI-Cut	0.021 (0.015, 0.065)	0.089 (0.043, 0.266)	0.025 (0.003, 0.127)	185 (182, 190)	0.964

and uncertainty quantification. We also report the runtime for each method and observe a substantial computational advantage for NeVI-Cut, which is nearly 40 times faster than the nested MCMC.

6 Two Real Data Applications

We benchmark NeVI-Cut with other methods (nested MCMC, full Bayes, parametric variational cut, other sampling based cut methods) using two real-world examples. In Section S7, we revisit the epidemiological example from Plummer (2015); Jacob et al. (2017); Yu et al. (2023), which examines the relationship between human papillomavirus (HPV) prevalence and the incidence of cervical cancer (Maucort-Boulch et al., 2008). Uncertainty from posterior samples of HPV prevalence from an external analysis is propagated into the downstream outcome model that captures their association with cervical cancer incidence. The results are presented in Figure S2. We see that NeVI-Cut is the best performing method (closest to the true cut posterior). However, as the cut posterior in this example is highly regular and unimodal, the alternate methods like parametric Gaussian VA-Cut Yu et al. (2023) as well as the tempered cut Plummer (2015) and openBUGS cut methods also perform well.

6.1 Cause-Specific Mortality Fractions in Mozambique

We now present a comparison of cut methods for a global health application where the posteriors are more irregularly shaped. Specifically, we are interested in predicting cause-specific mortality fractions (CSMFs) — the percentage of deaths in a population attributable to each cause — in neonates (0-27 days) and children (1-59 months) in Mozambique using algorithm-predicted cause-of-death data. The data comes from the Countrywide Mortality Surveillance for Action (COMSA) program (COMSA-Mozambique; Macicame et al., 2023) and is publicly available through the vacalibration R-package. COMSA consists of records of *verbal autopsy (VA)*, a WHO-standardized interview of the caregiver, which is widely used for population-level mortality surveillance in low- and middle-income countries (LMICs). VA records are passed through pre-trained statistical or machine-learning-based classifiers,

called *computer-coded verbal autopsy (CCVA)* algorithms to yield a predicted cause-of-death. However, as these classifiers are not perfect, estimating class fractions from predicted labels from these classifiers needs to adjust for the confusion matrix (misclassification rates) of the classifier. This is an example of post-prediction inference, and is termed VA calibration (Datta et al., 2020; Fiksel et al., 2022, 2023; Gilbert et al., 2023).

Considering an age group with a set of C causes and a specific CCVA classifier, let $\Phi = (\phi_{ij})$ denote the $C \times C$ confusion (misclassification) matrix where

$$\phi_{ij} = p(\text{CCVA predicts cause } j \,|\, \text{true cause is } i).$$

Let $v = (v_1, \ldots, v_C)^\top$ denote the COMSA data D_2 where v_c is the set of counts of deaths predicted to be from cause c by the CCVA classifier. The VA calibration downstream model is then given by

$$v \sim \text{Multinomial}(n, q) = \text{Multinomial}(n, \Phi^\top p),$$

where n is the sample size of D_2 and the primary goal is to estimate CSMFs p where $p_i = P(\text{true cause is } i)$.

Pramanik et al. (2025b) developed an upstream model to estimate these confusion matrices using a secondary labeled dataset from the CHAMPS project which has both VA and cause-of-death obtained via a minimal autopsy (Blau et al., 2019). Section S8 provides more details. The CHAMPS data used in the upstream analysis to estimate the confusion matrices is not publicly available. Pramanik et al. (2025a) has recently published posterior samples of these confusion matrices for three major CCVA classifiers, 2 age groups – neonates and children, and 8 countries including Mozambique. The posterior samples of each matrix are publicly available on GitHub. As the downstream analysis involves a VA-only (thus unlabeled) data, represents a different population, and is sensitive to various likelihood choices, it is undesirable to have feedback in this application. Hence, we implement cut-Bayes for estimating the CSMF p, while propagating uncertainty of Φ downstream.

We compare NeVI-Cut with posteriors from the full Bayes VA-calibration, and the uncalibrated point estimate $\widehat{p}$ obtained by forcing Φ to be identity in the downstream model, Among cut-Bayes methods, we consider the nested MCMC for benchmarking accuracy and timing. To assess the necessity of a non-parametric variational families specified via neural networks as in NeVI-Cut for this applicaiton, we also consider variational cut-Bayes with parametric variational distribution. Specifically, we use a Dirichlet variational family that assumes a lin-

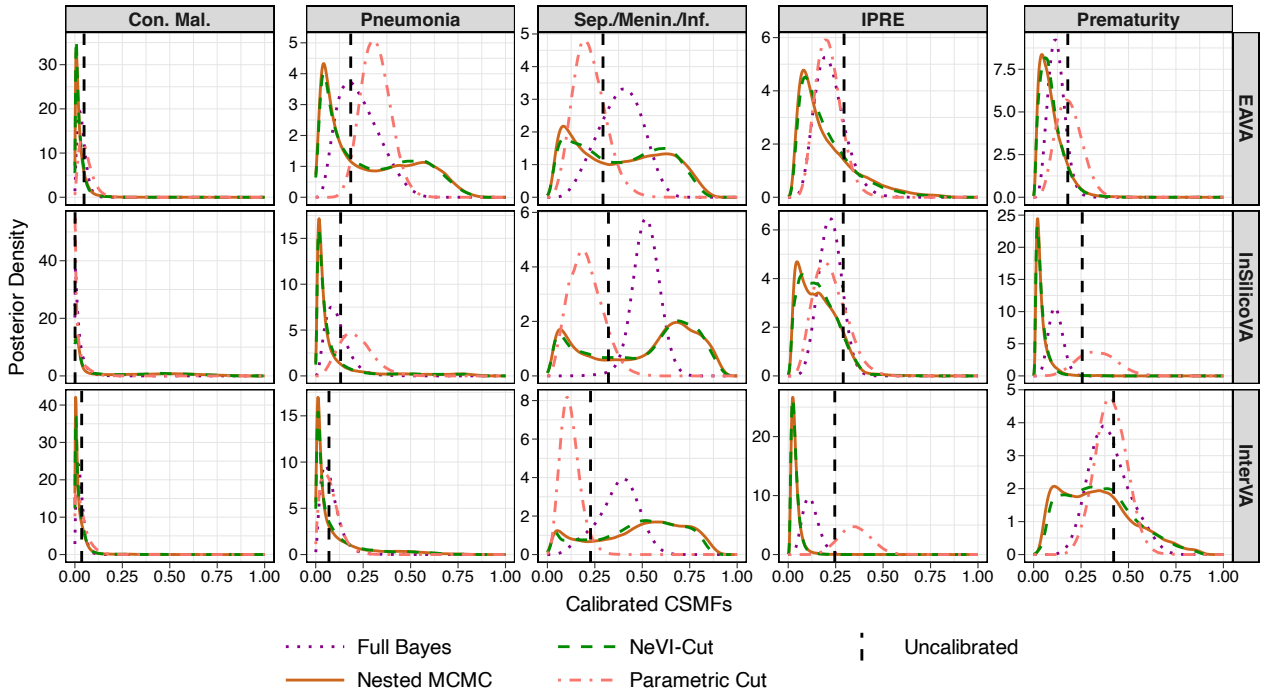


Figure 2: Posterior distributions of calibrated cause-specific mortality fractions (CSMFs) in neonates (0-27 days) based on VA-only data collected by the Countrywide Mortality Surveillance for Action (COMSA) program in Mozambique. NeVI-Cut effectively captures complex distributional features, including multimodality and skewness, achieving similar performance to nested MCMC. Con. mal., Sep./Menin./Inf., IPRE denote congenital malformation, sepsis/meningitis/infection, and intrapartum-related events.

ear relationship among the concentration parameters. Formally, the variational distribution is specified as $q_\alpha(p \mid \Phi) = \text{Dirichlet}(p \mid \Phi^\top \alpha)$ (referred to as Parametric Cut). More details about the implementation is in Section S8.

Density plots for CSMFs in the neonate population are shown in Figure 2, and results for children aged 1-59 months are shown in the Supplement Figure S3. There are multiple notable observations. We see that the full Bayes posterior is substantially different from the cut-posteriors, indicating sensitivity to possible model misspecification. For neonates, the cut-posterior densities of the calibrated CSMFs exhibit complex features, including multimodality and skewness. The uncalibrated estimates deviate noticeably from the calibrated estimates obtained using nested MCMC and NeVI-Cut, with the latter two showing very close agreement. The closeness of the NeVI-Cut estimate to the nested MCMC is quantified in Figure 3, which plots the Wasserstein distance of all the methods from the nested MCMC. Across all the settings, NeVI-Cut has a significantly smaller Wasserstein distance compared to others. In contrast, results from the parametric cut deviate from those produced by nested MCMC and NeVI-Cut, illustrating that the parametric variational distributions are not suf-

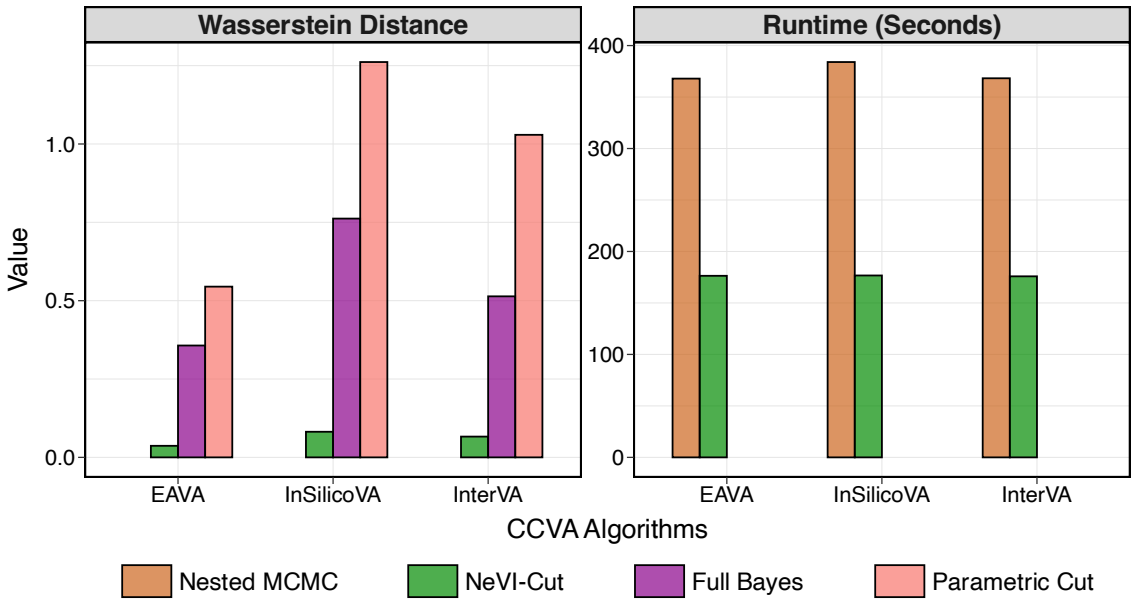


Figure 3: Left panel: Quadratic Wasserstein distances in centered log-ratio space between each method and nested MCMC across algorithms. Right panel: Runtime comparison of NeVI-Cut vs nested MCMC.

ficiently expressive. Additionally, NeVI-Cut exhibits computational efficiency over nested MCMC, achieving a runtime more than twice as fast.

7 Discussion

The concept of cutting feedback during uncertainty propagation has long been practiced under various names and using an assortment of statistical techniques, but in recent years, the methodology for cutting feedback has become increasingly principled and broadly applied. This is because in many applications, feedback is not scientifically meaningful, or the downstream model is subject to various forms of misspecification, jeopardizing the validity of traditional Bayesian inference. Our contribution, NeVI-Cut, is an important addition to the inventory of cutting feedback methods as it does not need any access to the upstream data or model. Our approach directly utilizes posterior samples from the upstream analysis and employs a stochastic optimization algorithm. NeVI-Cut also relies on minimal assumptions. It does not assume any parametric variational class, but leverages neural network-based normalizing flows to enhance the expressiveness of the variational family.

Through both simulation studies and real-world data examples, we demonstrate that NeVI-Cut can effectively capture both conditional and marginal cut-posteriors of various shapes. Its estimation accuracy for the cut-posterior is comparable to that of nested MCMC, while substantially reducing computation time. The method is implemented in a publicly

available software, which only requires users to input the upstream samples, the downstream data, and the model.

An important future methodological extension would be to use NeVI-Cut for cutting feedback in more complex analysis schemes beyond the two-module (upstream-downstream) analysis. Modular Bayesian inference, a generalization of cutting feedback, has been recently formalized (Liu and Goudie, 2025). A natural extension would be to make NeVI-Cut compatible with analysis involving multiple modules. Also, while model misspecification often motivates the use of cutting feedback, a critical question remains unexplored: how to determine when feedback should be cut. Developing principled criteria for this decision represents an important direction for future work.

Theoretically, we provide, to our knowledge, some of the first results on uniform KL approximation rates of conditional normalizing flows, which are at the heart of NeVI-Cut. These results offer much stronger conclusions than existing results on universal expressiveness of flows and directly lead to guarantees about the convergence of variational solutions (like the NeVI-Cut estimate) using conditional flows to the target collection of conditional distributions. Also, the results are in the fixed-data regime, targeting the actual cut-posterior rather than some large sample limit. The theory also explicitly links the expansiveness of the neural network architecture and the complexity of the target posterior to the closeness of approximation. These results are of independent importance as conditional flows are used in many different applications, e.g., hyper-parameter amortized Bayesian inference (Battaglia and Nicholls, 2024). Future theoretical work will aim to generalize to multivariate distributions, different base families, and other flow classes.

Acknowledgements

We thank Dr. Scott Zeger and Dr. Jamie Perin for useful discussions about cutting feedback. We acknowledge the use of chatgpt.com for helping with the literature review, proof development, algorithmic implementation, and improving the writing in some parts. This work was developed with support from the National Institute of Environmental Health Sciences (NIEHS) (grant R01ES033739); Gates Foundation Grants INV-034842 and INV-070577; Johns Hopkins University Institute for Data Intensive Engineering and Science Seed Funding; and Eunice Kennedy Shriver National Institute of Child Health K99 NIH Pathway to Independence Award (1K99HD114884-01A1).

References

Abramowitz, M. and Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Bassat, Q., Castillo, P., Martínez, M. J., Jordao, D., Lovane, L., Hurtado, J. C., Nhampossa, T., Santos Ritchie, P., Bandeira, S., Sambo, C., et al. (2017). Validity of a minimally invasive autopsy tool for cause of death determination in pediatric deaths in Mozambique: an observational study. *PLoS medicine*, 14(6):e1002317.

Battaglia, L. and Nicholls, G. (2024). Amortising Variational Bayesian Inference over prior hyperparameters with a Normalising Flow. *arXiv preprint arXiv:2412.16419*.

Blau, D. M., Caneer, J. P., Philipsborn, R. P., Madhi, S. A., Bassat, Q., Varo, R., Mandomando, I., Igunza, K. A., Kotloff, K. L., Tapia, M. D., et al. (2019). Overview and development of the child health and mortality prevention surveillance determination of cause of death (DeCoDe) process and DeCoDe diagnosis standards. *Clinical Infectious Diseases*, 69(Supplement_4):S333–S341.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bražinskas, A., Havrylov, S., and Titov, I. (2018). Embedding Words as Distributions with a Bayesian Skip-gram Model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1775–1789.

Butler, E. E., Wythers, K. R., Flores-Moreno, H., Ricciuto, D. M., Datta, A., Banerjee, A., Atkin, O. K., Kattge, J., Thornton, P. E., Anand, M., et al. (2022). Increasing functional diversity in a global land surface model illustrates uncertainties related to parameter simplification. *Journal of Geophysical Research: Biogeosciences*, 127(3):e2021JG006606.

Carmona, C. and Nicholls, G. (2020). Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. In *International Conference on Artificial Intelligence and Statistics*, pages 4226–4235. PMLR.

Carmona, C. U. and Nicholls, G. K. (2022). Scalable semi-modular inference with variational meta-posteriors. *arXiv preprint arXiv:2204.00296*.

Chakraborty, A., Nott, D. J., Drovandi, C. C., Frazier, D. T., and Sisson, S. A. (2023). Modularized Bayesian analyses and cutting feedback in likelihood-free inference. *Statistics and Computing*, 33(1):33.

Chang, H. H., Peng, R. D., and Dominici, F. (2011). Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. *Biostatistics*, 12(4):637–652.

Comess, S., Chang, H. H., and Warren, J. L. (2024). A Bayesian framework for incorporating exposure uncertainty into health analyses with application to air pollution and stillbirth. *Biostatistics*, 25(1):20–39.

Datta, A., Fiksel, J., Amouzou, A., and Zeger, S. L. (2020). Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*, 22(4):836–857.

Diederik, K. (2014). Adam: A method for stochastic optimization. *(No Title)*.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural spline flows. *Advances in neural information processing systems*, 32.

Fiksel, J., Datta, A., Amouzou, A., and Zeger, S. (2022). Generalized Bayes Quantification Learning under Dataset Shift. *Journal of the American Statistical Association*, 117(540):2163–2181.

Fiksel, J., Gilbert, B., Wilson, E., Kalter, H., Kante, A., Akum, A., Blau, D., Bassat, Q., Macicame, I., Gudo, E. S., et al. (2023). Correcting for verbal autopsy misclassification bias in cause-specific mortality estimates. *The American Journal of Tropical Medicine and Hygiene*, 108(5 Suppl):66.

Gilbert, B., Fiksel, J., Wilson, E., Kalter, H., Kante, A., Akum, A., Blau, D., Bassat, Q., Macicame, I., Gudo, E. S., et al. (2023). Multi-Cause Calibration of Verbal Autopsy–Based Cause-Specific Mortality Estimates of Children and Neonates in Mozambique. *The American journal of tropical medicine and hygiene*, 108(5 Suppl):78.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In *International conference on machine learning*, pages 2078–2087. PMLR.

Hutchings, G., Rumsey, K. N., Bingham, D., and Huerta, G. (2025). Enhancing Approximate Modular Bayesian Inference by Emulating the Conditional Posterior. *Computational Statistics & Data Analysis*, page 108235.

Institute for Health Metrics and Evaluation (2024). Global Burden of Disease Study 2021 (GBD 2021) Air Pollution Exposure Estimates and Risk Curves 1990-2021. https://ghdx.healthdata.org/record/ihme-data/gbd-2021-air-pollution-exposure-estimates-1990-2021.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*.

Jacob, P. E., O'leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):543–600.

Kingma, D. P., Welling, M., et al. (2013). Auto-encoding variational bayes.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Laszkiewicz, M., Lederer, J., and Fischer, A. (2022). Marginal tail-adaptive normalizing flows. In *International Conference on Machine Learning*, pages 12020–12048. PMLR.

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.*, 4(1):119–150.

Liu, Y. and Goudie, R. J. (2022). Stochastic approximation cut algorithm for inference in modularized Bayesian models. *Statistics and computing*, 32(1):7.

Liu, Y. and Goudie, R. J. (2025). A general framework for cutting feedback within modularized Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf012.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with 'sequential'PKPD modelling. *Journal of pharmacokinetics and pharmacodynamics*, 36(1):19–38.

Macicame, I., Kante, A., and Wilson, E. e. a. (2023). Countrywide Mortality Surveillance for Action in Mozambique: Results from a National Sample-Based Vital Statistics System for mortality and Cause of Death. *The American Journal of Tropical Medicine and Hygiene*, 108(5 Suppl):5–16.

Mathews, J., Gopalan, G., Gattiker, J., Smith, S., and Francom, D. (2025). Sequential Monte Carlo for cut-Bayesian posterior computation. *Computational Statistics*, 40(5):2749–2779.

Maucort-Boulch, D., Franceschi, S., Plummer, M., and Group, I. H. P. S. S. (2008). International correlation between human papillomavirus prevalence and cervical cancer incidence. *Cancer Epidemiology Biomarkers & Prevention*, 17(3):717–720.

Menendez, C., Castillo, P., Martínez, M. J., Jordao, D., Lovane, L., Ismail, M. R., Carrilho, C., Lorenzoni, C., Fernandes, F., Nhampossa, T., et al. (2017). Validity of a minimally invasive autopsy for cause of death determination in stillborn babies and neonates in Mozambique: an observational study. *PLoS medicine*, 14(6):e1002318.

Moss, D. and Rousseau, J. (2024). Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models. *Electronic Journal of Statistics*, 18(1):1815–1886.

Murphy, K. M. and Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The lancet*, 396(10258):1223–1249.

Nichols, E. K., Byass, P., Chandramohan, D., Clark, S. J., Flaxman, A. D., Jakob, R., Leitao, J., Maire, N., Rao, C., Riley, I., Setel, P. W., and on behalf of the WHO Verbal Autopsy Working Group (2018). The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLOS Medicine*, 15(1):1–9.

Nott, D. J., Drovandi, C., and Frazier, D. T. (2023). Bayesian inference for misspecified generative models. *Annual Review of Statistics and Its Application*, 11.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.

Park, T., Hashimoto, H., Wang, W., Thrasher, B., Michaelis, A. R., Lee, T., Brosnan, I. G., and Nemani, R. R. (2023). What does global land climate look like at 2° C warming? *Earth's Future*, 11(5):e2022EF003330.

Peng, R. D. and Bell, M. L. (2010). Spatial misalignment in time series studies of air pollution and health data. *Biostatistics*, 11(4):720–740.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25:37–43.

Pompe, E. and Jacob, P. E. (2021). Asymptotics of cut distributions and robust modular inference using posterior bootstrap. *arXiv preprint arXiv:2110.11149*.

Pramanik, S., Wilson, E. B., Kalter, H. D., Akelo, V., Amouzou, A., Black, R. E., Blau, D., Macicame, I., Muir, J. A., Lee, K. H., Liu, L., Whitney, C. G., Zeger, S., and Datta, A. (2025a). Country-Specific Estimates of Misclassification Rates of Computer-Coded Verbal Autopsy Algorithms. *medRxiv*.

Pramanik, S., Zeger, S., Blau, D., and Datta, A. (2025b). Modeling structure and country-specific heterogeneity in misclassification matrices of verbal autopsy-based cause of death classifiers. *The Annals of Applied Statistics*, 19(2):1214 – 1239.

Rao, K. K., Al Mandous, A., Al Ebri, M., Al Hameli, N., Rakib, M., and Al Kaabi, S. (2024). Future changes in the precipitation regime over the Arabian Peninsula with special emphasis on UAE: insights from NEX-GDDP CMIP6 model simulations. *Scientific Reports*, 14(1):151.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Smith, M. S., Yu, W., Nott, D. J., and Frazier, D. T. (2025). Cutting feedback in misspecified copula models. *Journal of the American Statistical Association*, pages 1–15.

Spiegelman, D. (2016). Evaluating public health interventions: 4. The Nurses' Health Study and methods for eliminating bias attributable to measurement error and misclassification. *American journal of public health*, 106(9):1563–1566.

Stan Development Team (2025). RStan: the R interface to Stan. R package version 2.32.7.

Thrasher, B., Wang, W., Michaelis, A., Melton, F., Lee, T., and Nemani, R. (2022). NASA Global Daily Downscaled Projections, CMIP6. *Scientific Data*, 9(1).

Vilnis, L. and McCallum, A. (2014). Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*.

Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. *Advances in neural information processing systems*, 32.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural networks*, 94:103–114.

Yu, X., Nott, D. J., and Smith, M. S. (2023). Variational inference for cutting feedback in misspecified models. *Statistical Science*, 38(3):490–509.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental health perspectives*, 108(5):419–426.

Zigler, C. M. (2016). The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. *The American Statistician*, 70(1):47–54.

S1 Literature Review on Cutting Feedback

Cutting feedback methods has been applied in various areas, including pharmacology (Lunn et al., 2009), economics (Murphy and Topel, 2002), environmental health and causal inference (Zigler, 2016), use cut-posteriors in a hidden Markov models (Moss and Rousseau, 2024), and copula models (Smith et al., 2025). Carmona and Nicholls (2020, 2022) introduced *semi-modular inference*, that interpolates posteriors between Bayes and cut-Bayes, allowing partial feedback. More generally, cut-Bayes can be viewed as a case of *modular Bayesian inference*, formalized in Liu and Goudie (2025), that involves several data-model pairs where certain directions of information flow are unwarranted. Nott et al. (2023) overviews Bayesian inference for misspecified models, including cutting feedback methods.

A major thread of developments on cutting feedback methods has been computational. In terms of sampling based approaches for cut-Bayes, Liu and Goudie (2022) introduced a stochastic approximation adaptive MCMC method to approximate the cut-posterior, improving convergence and stability compared with the openBUGS algorithm and tempered cut. Pompe and Jacob (2021) designed a posterior bootstrap method for modular inference, providing valid uncertainty quantification. Chakraborty et al. (2023) introduced a likelihood-free approach using Gaussian mixtures to approximate cut and semi-modular posteriors when the likelihood is intractable. Hutchings et al. (2025) uses emulation to increase the number of imputations, improving sampling efficiency for cut-posteriors. Mathews et al. (2025) designed a sequential Monte Carlo method to compute cut-posteriors.

We illustrate cutting feedback from downstream to upstream in Figure S1. Black arrows indicate parameters specifying a model, blue dotted arrows indicate the direction of information flow from a data source to parameters. Let D_1 denote the upstream data which informs the upstream quantity η. Here η can be a set of parameters that enters the model for D_1 (in which case there would be a black arrow going from η to D_1 and which is the standard cut model schematic used previously, e.g., in Plummer (2015)). However, η need not be a parameter in the model for D_1, it can be simply quantities that are informed by D_1 (e.g., derived quantities like predicted variables based on D_1 and its model). Our approach is agnostic to this. Hence, we simply show the eventual direction of information flow in the upstream analysis, that η is informed by D_1. The downstream data is denoted by D_2 whose model involves both η and some new parameters θ.

The right figure shows the final directions of information flow. If both datasets were

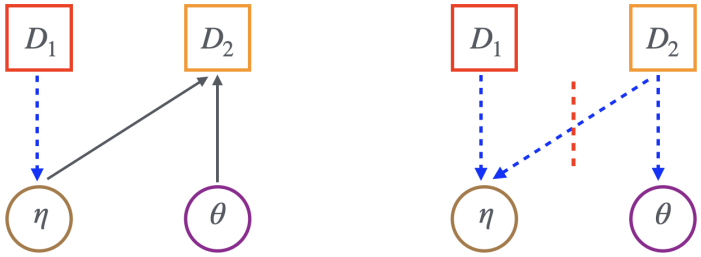


Figure S1: Paradigm of the cutting feedback from downstream to upstream. Black arrows indicate parameters specifying a model, blue dotted arrows indicate the direction of information flow from a data source to parameters.

analyzed jointly, the Bayesian posterior distribution can be written as:

$$p(\theta, \eta \,|\, D_1, D_2) \propto p(\theta \,|\, \eta, D_1, D_2)\, p(\eta \,|\, D_1, D_2) = p(\theta \,|\, \eta, D_2)\, p(\eta \,|\, D_1, D_2). \tag{S1}$$

As we know in a Bayesian analysis, the final posterior of η is informed by both D_1 and D_2 (hence the blue arrows from both D_1 and D_2 to η) and is thus different from its posterior $p(\eta \,|\, D_1)$ after the upstream analysis. As discussed in the introduction, this is undesirable in many situations. For example, if the downstream model is misspecified (which is the case in many of the applications we highlighted), inference on η could be negatively affected by this backward information flow from D_2 to η.

Cutting feedback from D_2 to η (as indicated by the red dotted line) prevents this. Instead of considering the Bayes posterior $p(\eta \,|\, D_1, D_2)$, we modify it to propagate the uncertainty in η into the inference of θ from D_2 without feedback from D_2 to η. The marginal cut-posterior for η is thus defined $p_{\text{cut}}(\eta \,|\, D_1, D_2) = p(\eta \,|\, D_1)$, i.e., the posterior remains unchanged from the upstream analysis. The joint cut-posterior is thus

$$\begin{aligned} p_{\text{cut}}(\theta, \eta \,|\, D_1, D_2) &= p(\theta \,|\, \eta, D_1, D_2)\, p_{\text{cut}}(\eta \,|\, D_1, D_2) \\ &= p(\theta \,|\, \eta, D_2)\, p(\eta \,|\, D_1). \end{aligned} \tag{S2}$$

Here, $p(\theta \,|\, \eta, D_1, D_2) = p(\theta \,|\, \eta, D_2)$, as in (S1), because the likelihood of D_1 does not depend of θ (hence no arrows between them), and as mentioned above, $p_{\text{cut}}(\eta \,|\, D_1, D_2) = p(\eta \,|\, D_1)$ preserves the posterior of η from upstream analysis and cuts the feedback from downstream. The marginal cut-posterior for θ is then given by

$$p_{\text{cut}}(\theta \,|\, D_1, D_2) = \int p(\theta \,|\, \eta, D_2)\, p(\eta \,|\, D_1) d\eta. \tag{S3}$$

To sample from the joint cut-posterior of (η, θ) or for the marginal cut-posterior of θ, one can generate draws from $p(\theta \,|\, \eta, D_2)$ for values of η from $p(\eta \,|\, D_1)$. This propagates the uncertainty

of η from the upstream analysis into the marginal cut-posterior for θ. However, θ is typically multidimensional and direct sampling from $p(\theta \mid \eta, D_2)$ is generally not feasible. Hence, for each draw of η from $p(\eta \mid D_1)$ (obtained from an MCMC run), a second nested MCMC is run to obtain samples of $\theta \mid \eta, D_2$. This nested MCMC scheme to sample from the cut-posterior is referred to as multiple imputation (MI; Plummer, 2015) and is very computationally intensive as it requires MCMC runs for each sample for η from the upstream analysis.

S2 Conditional Normalizing Flow Choices for NeVI-Cut

We explore two specific formulations of the transformation T_ϑ, which are detailed below. Additional flow-based transformations are discussed in Papamakarios et al. (2021).

S2.1 Neural Spline Flows

Durkan et al. (2019) introduced Neural Spline Flows (NSF) for modeling $q(\theta \mid \eta)$, which implies a normalizing-flow-based transformation. Following the terminology of Papamakarios et al. (2021), an NSF is characterized by three key components: the choice of flow, the transformer, and the conditioner that parameterizes the transformer. Together, they imply the transformation $\theta = T_\vartheta(\eta, Z)$. Durkan et al. (2019) proposes the monotonic rational-quadratic (RQ) spline functions as the transformer, and a neural network as the conditioner. The NSF framework corresponding to these choices is denoted by RQ-NSF and has been shown to enhance expressiveness while preserving analytic invertibility.

The choice of flow determines dependencies among $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ while preserving efficient implementation and computation of the Jacobian log-determinant. Durkan et al. (2019) discussed two commonly used flow choices for RQ-NSF which, as Papamakarios et al. (2021) describes, represent two extremes on a spectrum of possible implementations: autoregressive flow (RQ-NSF(AR)) and coupling flow (RQ-NSF(C)).

Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a base random vector with a tractable density and simple sampling, for example, $Z \sim N(0, I_d)$. Neural Spline Flows model the conditional distribution $q(\theta \mid \eta)$ according to the law

$$\theta = T_\vartheta(\eta, Z) \quad \text{where} \quad T_\vartheta = T_{\vartheta_K} \circ \cdots \circ T_{\vartheta_1}.$$

We assume $K = 1$ for simplicity to explain the construction and consider $\theta = T_{\vartheta_1}(\eta, Z) \equiv T_\vartheta(\eta, Z)$. Below, we discuss how components of NSF define T_ϑ.

Autoregressive flow. Define $Z_{<i} = (Z_1, \ldots, Z_{i-1})$ with $Z_{<1} = \varnothing$ being the null set. Moreover, let $\tau(\cdot; \boldsymbol{h})$ denote a transformer with parameter $\boldsymbol{h}$, and $c_\rho(\cdot)$ denotes a conditioner with parameter ρ. Given a conditioning variable η and following notations from Papamakarios et al. (2021), the NSF with autoregressive flow (NSF (AR)) transforms Z to θ as

$$\theta_i = \tau(Z_i; \boldsymbol{h}_i), \quad \boldsymbol{h}_i = c_{\rho_i}(Z_{<i}, \eta) \quad i = 1, \ldots, d. \tag{S4}$$

Specifically, RQ-NSF (AR) variant sets the transformer $\tau(\cdot; \boldsymbol{h}_i)$ to be the monotonically increasing rational-quadratic spline (RQS) function with K bins, and the conditioner c_{ρ_i} as the neural network. Together, the flow choice (S4) and $\{\tau, c_{\rho_1}, \ldots, c_{\rho_d}\}$ implicitly defines T_ϑ for the RQ-NSF (AR) with parameter $\vartheta = (\rho_1, \ldots, \rho_d)^\top$.

The RQS map $Z_i \mapsto \tau(Z_i; \boldsymbol{h}_i)$ is monotonically increasing, continuously differentiable, piecewise rational-quadratic functions in each bin, and it interpolates between knots given the derivatives at internal points. τ is parameterized by $\boldsymbol{h}_i = \{\boldsymbol{h}_i^w, \boldsymbol{h}_i^h, \boldsymbol{h}_i^s\}$, and they correspond to $K-1$ bin widths, $K-1$ bin heights, and derivatives at $K-1$ internal points, respectively. Each conditioner $(Z_{<i}, \eta) \mapsto c_{\rho_i}(Z_{<i}, \eta)$ is a neural network with parameter ρ_i that outputs the spline parameters $\boldsymbol{h}_i$ based on previous latent variables $Z_{<i}$ and conditioning variable η. The autoregressive structure yields a lower triangular Jacobian, so the Jacobian determinant reduces to $\prod_{i=1}^d \partial_{Z_i} \tau(Z_i; \boldsymbol{h}_i)$, the product of diagonal elements of the Jacobian. To ensure all coordinates are eventually updated, permutation layers (e.g., reversing or rotating the variable order) are inserted between autoregressive transformations.

Coupling flow. The above formulation can be adapted to include the coupling flow in the NSF framework. Compared to the autoregressive flow, the coupling flow splits Z into two parts with roughly equal dimension: $Z_{<p+1} = (Z_1, \ldots, Z_p)^\top$ and $Z_{\geqslant p+1} = (Z_{p+1}, \ldots, Z_d)^\top$ with $p = \lfloor d/2 \rfloor$. The NSF with this flow choice (NSF (C)) then keeps $Z_{<p+1}$ fixed, and transforms $Z_{\geqslant p+1}$ element-wise given $Z_{<p+1}$ as

$$\theta_i = \begin{cases} Z_i & i = 1, \ldots, p, \\ \tau(Z_i; \boldsymbol{h}_i) \quad \text{with} \quad \boldsymbol{h}_i = c_{\rho_i}(Z_{<p+1}, \eta) & i = p+1, \ldots, d. \end{cases} \tag{S5}$$

NSF (C) corresponding to the choice of RQS as transformer τ, and neural network as the conditioner c_{ρ_i} is denoted by RQ-NSF (C). As in RQ-NSF (AR), the flow choice (S5) and $\{\tau, c_{\rho_{p+1}}, \ldots, c_{\rho_d}\}$ implicitly defines T_ϑ in RQ-NSF (C) with parameter $\vartheta = (\rho_{p+1}, \ldots, \rho_d)^\top$. The transformer τ and conditioners c_{ρ_i} are defined as in RQ-NSF(AR), except that the input

to c_{ρ_i} is $(Z_{<p+1}, \eta)$ which remains identical for all $i = p+1, \ldots, d$. The coupling structure also yields a lower triangular Jacobian, and the determinant of the Jacobian equals $\prod_{i>p} \partial_{Z_i} \tau(Z_i; \boldsymbol{h}_i)$. To ensure every coordinate is updated and to propagate dependencies across all dimensions, we stack multiple coupling layers with permutations that swap the roles of the fixed and transformed subsets between layers.

In both flow choices, the spline parameters (bin widths, bin heights, and derivatives at internal points) are parameterized by neural networks. To construct conditional normalizing flows in NeVI-Cut, we employ a ReLU-based neural network as the conditioner and incorporate the conditioning variable η as an additional input. The selected flow choice, the RQS transformer, and the ReLU-based neural network conditioner define the transformation $\theta = T_\vartheta(\eta, Z)$ in NeVI-Cut.

S2.2 Unconstrained Monotonic Neural Networks

Wehenkel and Louppe (2019) proposed the Unconstrained Monotonic Neural Networks (UMNN), which is an integration-based transformation. We extend this framework to the conditional setting of NeVI-Cut by defining

$$\theta = T_\vartheta(\eta, Z) = a_{\vartheta_a}(\eta) + \int_0^Z \exp\left(g_{\vartheta_g}(\eta, t)\right) dt. \tag{S6}$$

Here, $a_{\vartheta_a}(\cdot)$ is a neural network parameterized by ϑ_a that models the conditional median, and $g_{\vartheta_g}(\cdot)$ is a neural network with parameters ϑ_g that models the log-derivative of the conditional quantile T_ϑ with respect to Z. $\vartheta = (\vartheta_a, \vartheta_g)^\top$ is the overall parameter of the transformation. The construction ensures that the derivative $\partial_Z T_\vartheta > 0$, thereby guaranteeing monotonicity and invertibility of the transformation.

S3 Spikiness, Conditional Perturbation Rate, and Quantile Growth for Common Families of Distributions

Table S1 summarizes these three quantities for common conditional families whose parameters (e.g., location, scale, shape) depend continuously on η. The rates are derived in Section S3 of the Supplement. These examples span a wide range of behavior under the *Gaussian-based* quantile map $T(\eta, z) = F_p^{-1}(\Phi(z) \,|\, \eta)$. For a Gaussian conditional law, the base matches the target: $T(\eta, z) = \mu(\eta) + \sigma(\eta) z$ for $|z| \leqslant M$, so the quantile growth is *linear*, $G(M) = O(M)$. The spikiness modulus is also linear, $L_S(M) = O(M)$ (the score is linear in θ), while the

Distribution	**Spikiness** $L_S(M)$	**Conditional Perturbation Rate** $L_{CP}(M)$	**Quantile growth** $G(M)$
Normal $N(\mu(\eta), \sigma^2(\eta))$	$O(M)$	$O(M^2)$	$O(M)$
Laplace $\mathrm{Lap}(\mu(\eta), b(\eta))$	$O(1)$	$O(M)$	$O(M^2)$
Generalized Gaussian $\mathrm{GGD}(\mu(\eta), \alpha(\eta), \beta(\eta) \leqslant 2)$	$O(M^{\max\{0,\bar{\beta}-1\}})$, where $\bar{\beta} = \sup_\eta \beta(\eta)$	$O(M^{\bar{\beta}} \log M)$	$O(M^{2/\underline{\beta}})$, where $\underline{\beta} = \inf_\eta \beta(\eta)$
Symmetrized Log–Normal $\frac{1}{2}(f_{\mathrm{LN}}(y) + f_{\mathrm{LN}}(-y))$	$O\Big(Me^{O((\log M)^2)}\Big)$	$O\Big((\log M)^2 Me^{O((\log M)^2)}\Big)$	$e^{O(M)}$
Cauchy $C(\mu(\eta), \gamma(\eta))$	$O(1)$	$O(1)$	$O(Me^{M^2/2})$
Student's t $t_{\mu(\eta),s(\eta),\nu(\eta)}$	$O(1)$	$O(\log M)$ if ν depends on η, $O(1)$ otherwise	$O\Big((Me^{M^2/2})^{1/\nu_{\min}}\Big)$, where $\nu_{\min} := \inf_{\eta \in K_\eta} \nu(\eta) > 0$
Finite mixtures of the above	max of component rates	max of component rates	max of component growths

Table S1: Tail orders of $L_S(M)$, $L_{CP}(M)$, and $G(M)$ for common distribution families. Distribution parameters are C^1 in η. Quantile growths are with respect to the Gaussian base.

η–sensitivity is $O(M^2)$ because $\log p$ contains a quadratic term in θ. The generalized Gaussian family interpolates smoothly through Gaussian ($\beta = 2$), Laplace ($\beta = 1$) via the shape parameter β. The quantiles grow polynomially in the Gaussian base, with order $G(M) = O(M^{2/\beta})$. The log normal and its symmetrized version have tails that decay more slowly than any Gaussian yet faster than any power law. This leads to Gaussian–base quantiles that grow exponentially, summarized by $G(M) = e^{O(M)}$ and is thus an example which sharply meets the exponential quantile bound (13). Thus, families of Gaussian, Laplace, Generalized Gaussian, and log-normal distributions or their finite linear combinations are subsumed in Assumption 1.

In contrast, for *Cauchy* tails the quantile with repect to Gaussian base is of order $e^{M^2/2}$. So $G(M)$ is *enormous* compared with the Gaussian case, even though $L_S(M) = O(1)$ and $L_{CP}(M) = O(1)$. Student-t behaves similarly. For *Laplace*, we obtain $G(M) = O(M^2)$, $L_S(M) = O(1)$ and $L_{CP}(M) = O(M)$. This is not surprising, and it is known that a thin base is a poor choice for thick-tailed distributions (Laszkiewicz et al., 2022). In such cases, one should use normalizing flows with thicker thicker-tailed base distribution, and similar theoretical results like we derive here can be obtained.

We now derive the rates stated in Table S1. In all the examples, $K_\eta \subset \mathbb{R}$ is compact and all parameter maps (location, scale, degrees of freedom, etc.) depend continuously on η and are thus uniformly (free of M) bounded. For simplicity of notation, we use $F_\eta(\cdot)$ to denote the conditional cdf $F_p(\cdot \,|\, \eta)$. Write

$$Q_\eta(u) = F_\eta^{-1}(u), \qquad T(\eta, z) = Q_\eta(\Phi(z)), \qquad G(M) = \sup_{\eta \in K_\eta,\ |z| \leqslant M} |T(\eta, z)|.$$

The moduli are

$$L_S(M) = \sup_{\eta\in K_\eta} \sup_{\substack{\theta\neq\theta' \\ |\theta\,|,|\theta'|\leqslant M}} \frac{|\log p(\theta \mid \eta) - \log p(\theta' \mid \eta)|}{|\theta - \theta'|},$$

$$L_{CP}(M) = \sup_{|\theta\,|\leqslant M} \sup_{\eta\neq\eta'} \frac{|\log p(\theta \mid \eta) - \log p(\theta \mid \eta')|}{|\eta - \eta'|_1}.$$

For $G(M)$ we evaluate upper envelopes of Q_η at the *Gaussian* tail levels $\varepsilon_M := \bar{\Phi}(M) = 1 - \Phi(M) \sim \phi(M)/M$. On K_η (compact), continuous parameter maps are bounded; for $L_{CP}(M)$ we use that $\log p(\theta \mid \eta)$ is smooth in its parameters on compact sets and apply the chain rule with the (finite) Lipschitz moduli of the parameter maps induced by continuity on K_η.[1]

Normal $N(\mu(\eta), \sigma^2(\eta))$.

$$Q_\eta(u) = \mu(\eta) + \sigma(\eta)\Phi^{-1}(u) \;\Rightarrow\; T(\eta, z) = \mu(\eta) + \sigma(\eta)z,$$

$$G(M) = \sup_{\eta\in K_\eta} |\,\mu(\eta)| + \left(\sup_{\eta\in K_\eta} \sigma(\eta)\right) M = O(M).$$

$$p\,(\theta \,|\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right].$$

$$\log p\,(\theta \,|\, \mu, \sigma) = \text{constant} - \log\sigma - \frac{(\theta-\mu)^2}{2\sigma^2}.$$

$$\partial_\theta \log p = -\frac{(\theta-\mu)}{\sigma^2} \;\Rightarrow\; L_S(M) = O\left(\frac{M}{\sigma^2_{\min}}\right) = O(M).$$

$$\partial_\mu \log p = \frac{\theta-\mu}{\sigma^2}, \qquad \partial_\sigma \log p = -\frac{1}{\sigma} + \frac{(\theta-\mu)^2}{\sigma^3} \;\Rightarrow\; L_{CP}(M) = O(M^2).$$

Cauchy $C(\mu(\eta), \gamma(\eta))$. For $Z \sim \mathcal{N}(0,1)$, by Mill's ratio, the upper Gaussian tail satisfies $\varepsilon_M = \bar{\Phi}(M) = 1 - \Phi(M) \sim \phi(M)/M$.

$$p(\theta \mid \eta) = \frac{1}{\pi\,\gamma(\eta)} \frac{1}{1 + \big((\theta - \mu(\eta))/\gamma(\eta)\big)^2}.$$

Upper tail of the Cauchy. Let $\bar{F}_\eta(x) = \Pr_\eta(\Theta > x)$. With $u = (t-\mu)/\gamma$,

$$\bar{F}_\eta(x) = \int_x^\infty p(t \mid \eta)\, dt = \frac{1}{\pi}\int_{(x-\mu)/\gamma}^\infty \frac{du}{1+u^2} = \frac{1}{\pi}\Big(\frac{\pi}{2} - \arctan\frac{x-\mu}{\gamma}\Big).$$

[1] For standard parametric families, the derivatives of $\log p$ with respect to the parameters are bounded on compacts; combined with bounded variation of the parameter maps on K_η, this gives the displayed $L_{CP}(M)$ orders.

As $x \to \infty$, $\frac{\pi}{2} - \arctan x = x^{-1} + O(x^{-3})$, so

$$\bar{F}_\eta(x) = \frac{\gamma}{\pi(x-\mu)}\big(1 + O(x^{-2})\big) \sim \frac{\gamma(\eta)}{\pi x}.$$

Therefore, for $\varepsilon \downarrow 0$,

$$Q_\eta(1-\varepsilon) \text{ satisfies } \varepsilon = \bar{F}_\eta\big(Q_\eta(1-\varepsilon)\big) \sim \frac{\gamma(\eta)}{\pi\, Q_\eta(1-\varepsilon)} \;\Rightarrow\; Q_\eta(1-\varepsilon) \sim \frac{\gamma(\eta)}{\pi\,\varepsilon}.$$

So, $Q_\eta(\Phi(z)) = O(1/\bar{\Phi}(z)) = z\exp(z^2/2)$ as $z \to \infty$, giving the bound

$$G(M) = O\Big(Me^{M^2/2}\Big).$$

$$\log p(\theta \mid \mu, \gamma) = \text{constant} + \log\gamma - \log\left(\gamma^2 + (\theta-\mu)^2\right).$$

The log-density derivatives are

$$\partial_\theta \log p = -\frac{2(\theta-\mu)}{\gamma^2 + (\theta-\mu)^2}, \qquad \partial_\mu \log p = \frac{2(\theta-\mu)}{\gamma^2 + (\theta-\mu)^2}, \qquad \partial_\gamma \log p = \frac{1}{\gamma} - \frac{2\gamma}{\gamma^2 + (\theta-\mu)^2}.$$

Using $2ab \leqslant a^2 + b^2$, these satisfy the uniform bounds

$$\big|\partial_\theta \log p\big| \leqslant \frac{1}{\gamma}, \qquad \big|\partial_\mu \log p\big| \leqslant \frac{1}{\gamma}, \qquad \big|\partial_\gamma \log p\big| \leqslant \frac{1}{\gamma}.$$

Therefore, $L_S(M) = O(1)$, and $L_{CP}(M) = O(1)$.

Student-$t_{\nu(\eta)}$ with location $\mu(\eta)$, scale $s(\eta)$. Let $T \sim t_\nu$ with density

$$f_\nu(u) = c_\nu\Big(1 + \tfrac{u^2}{\nu}\Big)^{-(\nu+1)/2}, \qquad c_\nu = \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{\nu\pi}\,\Gamma\big(\frac{\nu}{2}\big)}.$$

For $x > 0$,

$$\overline{F}_\nu(x) = \mathbb{P}(T > x) = \int_x^\infty f_\nu(u)\,du = c_\nu \int_x^\infty \Big(1 + \tfrac{u^2}{\nu}\Big)^{-(\nu+1)/2} du.$$

As $u \to \infty$,

$$\Big(1 + \tfrac{u^2}{\nu}\Big)^{-(\nu+1)/2} = \Big(\tfrac{u^2}{\nu}\Big)^{-(\nu+1)/2}\Big(1 + O(u^{-2})\Big) = \nu^{(\nu+1)/2}u^{-(\nu+1)}\Big(1 + O(u^{-2})\Big),$$

hence

$$\overline{F}_\nu(x) = c_\nu\nu^{(\nu+1)/2}\int_x^\infty u^{-(\nu+1)}\Big(1 + O(u^{-2})\Big)\,du = c_\nu\nu^{(\nu+1)/2}\left(\frac{x^{-\nu}}{\nu} + O\big(x^{-(\nu+2)}\big)\right).$$

With

$$C_\nu := c_\nu\nu^{(\nu-1)/2} = \frac{\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{\pi}\,\Gamma\big(\frac{\nu}{2}\big)}\,\nu^{(\nu-1)/2},$$

this gives

$$\overline{F}_\nu(x) = C_\nu x^{-\nu}\Big(1 + O(x^{-2})\Big), \qquad x \to \infty.$$

For a location scale family $X = \mu + sT$ with fixed $\mu \in \mathbb{R}$ and $s > 0$,

$$\overline{F}_{\mu,s}(x) = \mathbb{P}(X > x) = \overline{F}_\nu\Big(\frac{x-\mu}{s}\Big) \sim C_\nu s^\nu (x-\mu)^{-\nu} \sim c_\nu x^{-\nu}, \qquad x \to \infty,$$

where $c_\nu = C_\nu s^\nu$. Tail $\bar{F}_\eta(x) \sim c_{\nu(\eta)}\, x^{-\nu(\eta)}$ yields

$$Q_\eta(1-\varepsilon) \sim (c_{\nu(\eta)}/\varepsilon)^{1/\nu(\eta)} \;\Rightarrow\; G(M) = O\left(\left(\frac{M}{\phi(M)}\right)^{1/\nu_{\min}}\right) = O\left(\left(Me^{M^2/2}\right)^{1/\nu_{\min}}\right),$$

with $\nu_{\min} := \inf_{\eta \in K_\eta} \nu(\eta) > 0$.

$$p(\theta \mid \nu, \mu, s) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{s\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(\theta-\mu)^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}}.$$

$$\log p(\theta \mid \nu, \mu, s) = \text{constant} - \log s + \log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\log\nu - \frac{\nu+1}{2}\log\left(1 + \frac{(\theta-\mu)^2}{\nu s^2}\right).$$

$$\partial_\theta \log p = -\frac{(\nu+1)(\theta-\mu)}{\nu s^2 + (\theta-\mu)^2} = -\left(\frac{\nu+1}{2\sqrt{\nu}s}\right) \cdot \frac{2\sqrt{\nu}s\,(\theta-\mu)}{\nu s^2 + (\theta-\mu)^2} \;\Rightarrow\; L_S(M) = O(1).$$

Derivatives w.r.t. (μ, s) are bounded following similar arguments. The ν-term contains $\frac{1}{2}\log\big(1 + (\theta-\mu)^2/(\nu s^2)\big)$, hence on $[-M, M]$ gives $O(\log M)$:

$$L_{CP}(M) = O(\log M).$$

Laplace $\mathrm{Lap}(\mu(\eta), b(\eta))$**.** Density and cdf:

$$p(\theta \mid \eta) = \frac{1}{2b}\exp\Big(-\frac{|\theta-\mu|}{b}\Big), \qquad F_\eta(\theta) = \begin{cases} \frac{1}{2}\exp\big(\frac{\theta-\mu}{b}\big), & \theta < \mu, \\ 1 - \frac{1}{2}\exp\big(-\frac{\theta-\mu}{b}\big), & \theta \geqslant \mu. \end{cases}$$

Upper tail and quantile. For $x \to \infty$,

$$\bar{F}_\eta(x) = 1 - F_\eta(x) = \tfrac{1}{2}\exp\Big(-\frac{x-\mu}{b}\Big).$$

Hence, for $\varepsilon \downarrow 0$,

$$\varepsilon = \bar{F}_\eta\big(Q_\eta(1-\varepsilon)\big) = \tfrac{1}{2}\exp\Big(-\frac{Q_\eta(1-\varepsilon)-\mu}{b}\Big) \quad \Rightarrow \quad Q_\eta(1-\varepsilon) = \mu + b\log\frac{1}{2\varepsilon}.$$

With normal base $Z \sim \mathcal{N}(0,1)$, let $\varepsilon_M = \bar{\Phi}(Z) \sim \phi(M)/M$. Then

$$Q_\eta\big(\Phi(M)\big) = \mu + b\big(\log M - \log \phi(M) + \log \tfrac{1}{2}\big) = \mu + \frac{b}{2}M^2 + b\log M + O(1),$$

so the growth envelope on $\{|z| \leqslant M\}$ is

$$G(M) = O\Big(\log \frac{M}{\phi(M)}\Big) = O(M^2).$$

Log-density derivatives.

$$\log p\,(\theta \mid \eta) = \text{constant} - \log b - \frac{|\theta - \mu|}{b}.$$

For $\theta \neq \mu$,

$$\partial_\theta \log p = -\frac{\operatorname{sgn}(\theta-\mu)}{b}, \qquad \partial_\mu \log p = \frac{\operatorname{sgn}(\theta-\mu)}{b}, \qquad \partial_b \log p = -\frac{1}{b} + \frac{|\theta-\mu|}{b^2}.$$

Uniform bounds on the θ-truncated set $\{|\theta - \mu| \leqslant M\}$:

$$\big|\partial_\theta \log p\big| \leqslant \frac{1}{b}, \qquad \big|\partial_\mu \log p\big| \leqslant \frac{1}{b}, \qquad \big|\partial_b \log p\big| \leqslant \frac{1}{b} + \frac{M}{b^2}.$$

Therefore, if $0 < b_{\min} \leqslant b(\eta) \leqslant b_{\max} < \infty$ uniformly in η,

$$L_S(M) = O(1), \qquad L_{CP}(M) = O(M).$$

Symmetrized mollified log-normal. The log-normal distribution is supported on $\mathbb{R}_{>0}$, hence we consider a symmetrized log-normal. Also the log-density of log-normal have infinitely large derivative near 0, which makes it ill behaved, to address this, we use a smoothed version by convoluting with a smooth distribution representing a small bump at zero (which can be made arbitrarily small).

Let $\mu(\eta), \sigma(\eta)$ be continuous on compact K_η with $0 < \sigma_{\min} \leqslant \sigma(\eta) \leqslant \sigma_{\max} < \infty$. Define the symmetrized log-normal density

$$f_{\mathrm{SLN}}(\theta \mid \eta) = \tfrac{1}{2} f_{\mathrm{LN}}\left(|\theta| \mid \eta\right), \qquad f_{\mathrm{LN}}(x \mid \eta) = \frac{1}{x\sigma(\eta)\sqrt{2\pi}} \exp\left(-\frac{\left(\log x - \mu(\eta)\right)^2}{2\sigma^2(\eta)}\right), \quad x > 0.$$

Let K_h be a symmetric C^∞ mollifier with compact support $[-h, h]$, $\int K_h = 1$, and $K_h(u) \geqslant \kappa_0 > 0$ for $|u| \leqslant h/2$. Define the smoothed symmetric log-normal

$$p_h(\theta \mid \eta) = (f_{\mathrm{SLN}}(\cdot \mid \eta) * K_h)(\theta) = \int_{\mathbb{R}} f_{\mathrm{SLN}}(x \mid \eta)\, K_h(\theta - x)\, dx.$$

Then p_h is positive, C^∞, symmetric, and equals f_{SLN} outside a $2h$-neighborhood of 0.

Uniform lower bound on $[-M, M]$. For any $|\theta\,| \leqslant M$,

$$p_h(\theta \mid \eta) \;\geqslant\; \kappa_0 \int_{|\theta - x| \leqslant h/2} f_{\mathrm{SLN}}(x \mid \eta)\, dx \;\geqslant\; \frac{\kappa_0}{2} \int_{I_\theta} f_{\mathrm{LN}}(x \mid \eta)\, dx, \quad I_\theta := [\max\{0, |\theta\,| - h/2\},\, |\theta\,| + h/2].$$

Since $f_{\mathrm{LN}}(\cdot \mid \eta)$ is decreasing for x beyond its mode, the worst case over $|\theta\,| \leqslant M$ is attained at $|\theta\,| = M$. Thus, for all $\eta \in K_\eta$,

$$\inf_{|\theta\,| \leqslant M} p_h(\theta \mid \eta) \;\geqslant\; \frac{\kappa_0}{2} \int_{M - h/2}^{M + h/2} f_{\mathrm{LN}}(x \mid \eta)\, dx \;\geqslant\; \frac{\kappa_0 h}{4} \inf_{\eta \in K_\eta} \inf_{x \in [M - h/2,\, M + h/2]} f_{\mathrm{LN}}(x \mid \eta).$$

For $x \geqslant 1$ and any η,

$$f_{\mathrm{LN}}(x \mid \eta) = \frac{1}{x\, \sigma(\eta) \sqrt{2\pi}} \exp\Big(- \frac{(\log x - \mu(\eta))^2}{2\sigma^2(\eta)} \Big) \;\geqslant\; \frac{1}{x\, \sigma_{\max} \sqrt{2\pi}} \exp\Big(- \frac{d(x)^2}{2\sigma_{\min}^2} \Big),$$

where $d(x) := \mathrm{dist}(\log x,\, [\mu_{\min}, \mu_{\max}]) = \max\{0,\; \log x - \mu_{\max},\; \mu_{\min} - \log x\}$.Hence for $M \geqslant 1$,

$$c_h(M) := \inf_{\eta \in K_\eta} \inf_{|\theta\,| \leqslant M} p_h(\theta \mid \eta) \;\geqslant\; \frac{\kappa_0 h}{4(M + h/2)\, \sigma_{\max} \sqrt{2\pi}} \exp\left(- \frac{(\log(M + h/2) - \mu_{\max})^2}{2\sigma_{\min}^2} \right).$$

Spikiness and conditional perturbation. As before,

$$L_S(M) = \sup_{\eta} \sup_{|\theta\,| \leqslant M} \big| \partial_\theta \log p_h(\theta \mid \eta) \big| \;\leqslant\; \frac{\|K_h'\|_{L^1}}{c_h(M)} = O\Big(\frac{1}{c_h(M)} \Big), \qquad L_{CP}(M) = O\Big(\frac{(\log M)^2}{c_h(M)} \Big),$$

since $\partial_\eta p_h = (\partial_\eta f_{\mathrm{SLN}}) * K_h$ and $|\partial_\eta \log f_{\mathrm{SLN}}| \lesssim (\log M)^2$ on $[-M, M]$. With the compact bump, these become

$$L_S(M) = O\Big((M+1) \exp\Big(\tfrac{(\log(M+h/2) - \mu_{\max})^2}{2\sigma_{\min}^2} \Big) \Big) = O(M e^{O((\log M)^2)}),$$
$$L_{CP}(M) = O\Big((\log M)^2 (M+1) \exp\Big(\tfrac{(\log(M+h/2) - \mu_{\max})^2}{2\sigma_{\min}^2} \Big) \Big) = O((\log M)^2 M e^{O((\log M)^2)}).$$

Quantile growth Let $X \sim f_{\mathrm{SLN}}(\cdot \mid \eta)$ and $Y \sim K_h$ with $\mathrm{supp}(Y) \subset [-h, h]$, independent. Let F_X, F_{X+Y} be the distribution functions and Q_X, Q_{X+Y} their generalized quantile functions $Q(u) := \inf\{x :\, F(x) \geqslant u\}$, $u \in (0, 1)$.

For every $x \in \mathbb{R}$,

$$F_{X+Y}(x) = \mathbb{E}[F_X(x - Y)] \in \big[\, F_X(x - h),\, F_X(x + h) \,\big],$$

because F_X is nondecreasing and $Y \in [-h, h]$ almost surely.

By monotonicity of F_X and of its generalized inverse, for all $u \in (0, 1)$,

$$Q_X(u) - h \;\leqslant\; Q_{X+Y}(u) \;\leqslant\; Q_X(u) + h, \qquad \text{hence} \qquad \big| Q_{X+Y}(u) - Q_X(u) \big| \leqslant h.$$

If $X \sim \mathrm{LN}(m(\eta), \sigma(\eta))$ with m, σ bounded on K_η and $\sigma(\eta) \geqslant \sigma_{\min} > 0$, then

$$Q_X(\Phi(z)) = \exp\big(m(\eta) + \sigma(\eta) z \big).$$

Combining with the previous display yields

$$Q_{X+Y}(\Phi(z)) = \exp\big(m(\eta) + \sigma(\eta)z\big)\, e^{O(1)} \quad \text{uniformly in } \eta \in K_\eta.$$

Therefore, for $|z| \leqslant M$,

$$Q_{X+Y}(\Phi(z)) = \exp\big(O(M)\big) \quad \text{uniformly in } \eta \in K_\eta,$$

with constants depending only on h and the bounds for m and σ.

Generalized Gaussian $\mathrm{GGD}(\mu(\eta), \alpha(\eta), \beta(\eta))$**.** Let $r := |\theta - \mu|/\alpha$. By continuity, we have $\alpha(\eta) \in [\underline{\alpha}, \overline{\alpha}]$, $\beta(\eta) \in [\underline{\beta}, \overline{\beta}]$ with $\underline{\alpha} > 0$, $\underline{\beta} > 0$, $2 \geqslant \overline{\beta}$, and $|\mu(\eta)| \leqslant M_\mu$ uniformly on K_η. The density is

$$p(\theta \mid \eta) = \frac{\beta}{2\alpha\,\Gamma(1/\beta)} \exp\big(-r^\beta\big).$$

For $x \to \infty$ we have $t - \mu > 0$ on $[x, \infty)$, so

$$\bar{F}_\eta(x) = \int_x^\infty \frac{\beta}{2\alpha\,\Gamma(1/\beta)} \exp\Big(-\big(\tfrac{t-\mu}{\alpha}\big)^\beta\Big)\, dt.$$

With the change of variables $u = \big(\frac{t-\mu}{\alpha}\big)^\beta$, so that $t = \mu + \alpha\, u^{1/\beta}$ and $dt = \alpha\,\beta^{-1} u^{1/\beta - 1}\, du$, we obtain

$$\bar{F}_\eta(x) = \frac{1}{2\,\Gamma(1/\beta)} \int_{((x-\mu)/\alpha)^\beta}^\infty u^{1/\beta - 1} e^{-u}\, du = \frac{1}{2\,\Gamma(1/\beta)}\, \Gamma\Big(\tfrac{1}{\beta},\, \big(\tfrac{x-\mu}{\alpha}\big)^\beta\Big),$$

where $\Gamma(s, z) = \int_z^\infty t^{s-1} e^{-t}\, dt$ is the upper incomplete gamma function. As $z \to \infty$, $\Gamma(s, z) \sim z^{s-1} e^{-z}$ (see §6.5.32 in Abramowitz and Stegun, 1948). Taking $s = 1/\beta$ and $z = \big(\frac{x-\mu}{\alpha}\big)^\beta$ gives

$$\bar{F}_\eta(x) \sim \frac{1}{2\,\Gamma(1/\beta)} \Big(\big(\tfrac{x-\mu}{\alpha}\big)^\beta\Big)^{1/\beta - 1} \exp\Big(-\big(\tfrac{x-\mu}{\alpha}\big)^\beta\Big) = \frac{1}{2\,\Gamma(1/\beta)} \Big(\tfrac{x-\mu}{\alpha}\Big)^{1-\beta} \exp\Big(-\big(\tfrac{x-\mu}{\alpha}\big)^\beta\Big).$$

Let

$$y := \Big(\frac{x-\mu}{\alpha}\Big)^\beta, \qquad C := \frac{1}{2\,\Gamma(1/\beta)}.$$

Then

$$\bar{F}_\eta(x) \sim C\, y^{\frac{1-\beta}{\beta}} e^{-y}.$$

Solving $\bar{F}_\eta(x) \sim \varepsilon$ for $\varepsilon \downarrow 0$ gives

$$y = \log\frac{1}{\varepsilon} + \frac{1-\beta}{\beta} \log\Big(\log\frac{1}{\varepsilon}\Big) + \log C + o(1).$$

Hence the $(1-\varepsilon)$-quantile satisfies

$$Q_\eta(1-\varepsilon) = \mu + \alpha \left[\log\frac{1}{\varepsilon} + \frac{1-\beta}{\beta} \log\Big(\log\frac{1}{\varepsilon}\Big) + \log\Big(\frac{1}{2\,\Gamma(1/\beta)}\Big) + o(1)\right]^{1/\beta}.$$

So the $(1-\varepsilon)$ quantile satisfies

$$Q_\eta(1-\varepsilon) \sim \alpha\,[\log(1/\varepsilon)]^{1/\beta}.$$

At Gaussian tail level $\varepsilon_M = \Phi(-Z) \sim \phi(M)/M$,

$$\log\left(\frac{1}{\varepsilon_M}\right) = \tfrac{1}{2}M^2 + \log M + \log\sqrt{2\pi},$$

hence

$$Q_\eta(1-\varepsilon_N) \sim \alpha\Big(\tfrac{1}{2}M^2\Big)^{1/\underline{\beta}}, \qquad G(M) = \sup_{\eta\in K_\eta} Q_\eta(1-\varepsilon_M) = O\big(M^{2/\underline{\beta}}\big).$$

Log density derivatives. We have

$$\log p(\theta \mid \eta) = \log\beta - \log(2\alpha) - \log\Gamma(1/\beta) - r^\beta.$$

Derivative in θ:

$$\partial_\theta \log p = -\beta\,\mathrm{sgn}(\theta-\mu)\,\frac{|\theta-\mu|^{\beta-1}}{\alpha^\beta}.$$

Parameter derivatives:

$$\partial_\mu \log p = \beta\,\mathrm{sgn}(\theta-\mu)\,\frac{|\theta-\mu|^{\beta-1}}{\alpha^\beta},$$

$$\partial_\alpha \log p = -\frac{1}{\alpha} + \beta\,\frac{|\theta-\mu|^\beta}{\alpha^{\beta+1}},$$

$$\partial_\beta \log p = \frac{1}{\beta} + \frac{\psi(1/\beta)}{\beta^2} - r^\beta \log r.$$

Here ψ is the digamma function. Definition of $L_S(M)$. For $|\theta\,| \leqslant M$ and $\eta \in K_\eta$,

$$L_S(M) := \sup_{\eta,|\theta\,|\leqslant M} |\partial_\theta \log p(\theta \mid \eta)| \;\leqslant\; \overline{\beta}\,\underline{\alpha}^{-\overline{\beta}}\,(M+M_\mu)^{\overline{\beta}-1}.$$

Definition of $L_{CP}(M)$. If μ, α, β are Lipschitz in η with constants L_μ, L_α, L_β, then

$$L_{CP}(M) := \sup_{\eta,|\theta\,|\leqslant M} \Big(|\partial_\mu \log p|\,L_\mu + |\partial_\alpha \log p|\,L_\alpha + |\partial_\beta \log p|\,L_\beta\Big).$$

Using the above bounds,

$$L_{CP}(M) = O\Big(L_\mu M^{\overline{\beta}-1} + L_\alpha M^{\overline{\beta}} + L_\beta M^{\overline{\beta}}\log M + L_\alpha + L_\beta\Big).$$

Finally, as $\beta(\eta)$ is always less than or equal to two, the tail decay is always as thick as a Gaussian and part(iv) of the assumption is satisfied.

Finite mixtures. Consider a finite mixture

$$p(\theta \mid \eta) = \sum_{k=1}^{K} \pi_k(\eta)\, p_k(\theta \mid \eta), \qquad \pi_k(\eta) \geqslant 0,\ \sum_k \pi_k(\eta) = 1.$$

At the tail level, the mixture is governed by the component (or set of components) with the heaviest tail. So, $G(M) \asymp \max_k G_k(M)$.

Turning to the derivatives, let

$$\tau_k(\theta, \eta) = \frac{\pi_k(\eta) p_k(\theta \mid \eta)}{\sum_j \pi_j(\eta) p_j(\theta \mid \eta)}, \qquad \sum_k \tau_k = 1.$$

Then

$$\partial_\theta \log p = \sum_{k=1}^{K} \tau_k\, \partial_\theta \log p_k, \qquad \partial_\eta \log p = \sum_{k=1}^{K} \tau_k \big(\partial_\eta \log p_k + \partial_\eta \log \pi_k\big).$$

It follows that

$$L_S(M) \leqslant \max_k L_{S,k}(M), \qquad L_{CP}(M) \leqslant \max_k L_{CP,k}(M) + \sup_k \|\nabla_\eta \log \pi_k\|_{L^\infty(K_\eta)}.$$

Assume the mixing weights $\pi_k : K_\eta \mapsto (0,1)$ are continuously differentiable and bounded below by $\pi_{\min} > 0$. Then

$$\nabla_\eta \log \pi_k(\eta) = \frac{\nabla_\eta \pi_k(\eta)}{\pi_k(\eta)} \quad \Rightarrow \quad \|\nabla_\eta \log \pi_k\|_{L^\infty(K_\eta)} \leqslant \frac{\|\nabla_\eta \pi_k\|_{L^\infty(K_\eta)}}{\pi_{\min}}.$$

Hence, this term is uniformly bounded.

S4 Proofs

In this Section, we provide proofs of the general results on conditional normalizing flows and NeVI-Cut solutions using them. Proofs of the results for specific flow models (Theorems 2 and 3) are in the next Section. We start with proof of Lemma 2, as the result is used in other proofs.

Proof of Lemma 2. By definition (12), $T_p(\eta, z)$ is a composition of $u \mapsto F_p^{-1}(u \mid \eta)$ and $z \mapsto \Phi(z)$. Under Assumption 1, T_p is continuous on $K_\eta \times \mathbb{R}$, strictly increasing and continuously differentiable in z.

Following (12), we have $F_p\big(T_p(\eta, z) \,|\, \eta\big) = \Phi(z)$. Differentiate both sides with respect to z and use the chain rule:

$$\partial_\theta F_p\big(T_p(\eta, z) \,|\, \eta\big) \cdot \partial_z T_p(\eta, z) = \phi(z).$$

Because $\partial_\theta F_p(\theta \,|\, \eta) = p(\theta \mid \eta)$, substituting $\theta = T_p(\eta, z)$ yields

$$p\big(T_p(\eta, z) \mid \eta\big)\, \partial_z T_p(\eta, z) = \phi(z).$$

Both factors are positive, hence taking logarithms gives the result. □

Proof of Lemma 1. Lipschitz modulus for $\log \partial_z T_p$. For $(z_i, \eta_i) \in K_M$ set $\theta_i = T_p(\eta_i, z_i)$. Applying Lemma 2, by the two moduli condition on $K_\eta \times [-G(M), G(M)]$ and since $\log \phi$ is M-Lipschitz on $[-M, M]$,

$$\begin{aligned}\big|\log \partial_z T_p(\eta_1, z_1) - \log \partial_z T_p(\eta_2, z_2)\big| &\leqslant \big|\log \phi(z_1) - \log \phi(z_2)\big| + \big|\log p(\theta_1 \mid \eta_1) - \log p(\theta_2 \mid \eta_2)\big| \\ &\leqslant M\,|z_1 - z_2| + L_S\big(G(M)\big)\,|\theta_1 - \theta_2| + L_{CP}\big(G(M)\big)\,\|\eta_1 - \eta_2\|_1.\end{aligned}$$

By the mean value theorem and Assumption 1(*iii*)(*b*), $|\theta_1 - \theta_2| = \big|\int_{z_1}^{z_2} \partial_u T_p(\eta_1, u)\, du\big| \leqslant e^{H(M)}|z_1 - z_2|$. Thus

$$\big|\log \partial_z T_p(\eta_1, z_1) - \log \partial_z T_p(\eta_2, z_2)\big| \leqslant \big(M + e^{H(M)} L_S(G(M))\big)|z_1 - z_2| + L_{CP}(G(M))\,\|\eta_1 - \eta_2\|_1.$$

Therefore $\log \partial_z T_p$ is $\mathsf{L}_b(M)$-Lipschitz on K_M with respect to the ℓ_1 distance $\|(\eta, z) - (\eta', z')\|_1 := \|\eta - \eta'\|_1 + |z - z'|$.

Lipschitz modulus for $T_p(\eta, 0)$ *in* η. For simplicity of notation, we use $F_\eta(\cdot)$ to denote the conditional cdf $F_p(\cdot \,|\, \eta)$. Fix $\eta_1, \eta_2 \in K_\eta$ and set $u = \Phi(0)$, $\theta_i = T(\eta_i, 0) = F_{\eta_i}^{-1}(u)$, and $g = F_{\eta_2}^{-1}$. Then

$$|\theta_1 - \theta_2| = \big|g(F_{\eta_2}(\theta_1)) - g(u)\big| \leqslant \Big(\sup_{v \text{ between } F_{\eta_2}(\theta_1) \text{ and } u} |g'(v)|\Big)\,\big|F_{\eta_2}(\theta_1) - u\big|.$$

By Assumption 1(*iii*)(*a*), $|\theta_i| \leqslant M_0$. As g is monotonic, for v between $F_{\eta_2}(\theta_1)$ and u, we have $g(v) \in [-M_0, M_0]$. Since $g'(v) = 1/p(g(v) \mid \eta_2)$, it thus suffices to bound p away from zero on $[-M_0, M_0]$. Assumption 1(*iv*) gives

$$m_0 := \inf_{\eta \in K_\eta,\, |\theta| \leqslant M_0} p(\theta \mid \eta) \geqslant c\, e^{-\alpha M_0^2} > 0,$$

hence $|g'(v)| \leqslant 1/m_0$. Therefore

$$|\theta_1 - \theta_2| \leqslant m_0^{-1}\,|F_{\eta_2}(\theta_1) - F_{\eta_1}(\theta_1)|.$$

For $|t| \leqslant M_0$, Assumption 1(*ii*) implies

$$|\log p(t \mid \eta_1) - \log p(t \mid \eta_2)| \leqslant L_{CP}(M_0)\,\|\eta_1 - \eta_2\|_1,$$

hence $p(t \mid \eta_2) \leqslant e^{L_{CP}(M_0)\|\eta_1-\eta_2\|_1} p(t \mid \eta_1)$ and vice versa. Integrating over $(-\infty, \theta]$ yields

$$|F_{\eta_2}(\theta) - F_{\eta_1}(\theta)| \leqslant 2\big(e^{L_{CP}(M_0)\|\eta_1-\eta_2\|_1} - 1\big).$$

Using $e^x - 1 \leqslant e^{x_0} x$ for $x \in [0, x_0]$ with $x_0 = L_{CP}(M_0)\operatorname{diam}(K_\eta)$, we obtain

$$|F_{\eta_2}(\theta) - F_{\eta_1}(\theta)| \leqslant 2e^{L_{CP}(M_0)\operatorname{diam}(K_\eta)} L_{CP}(M_0)\,\|\eta_1 - \eta_2\|.$$

Combining the last two displays with the definition of m_0 gives

$$|T(\eta_1, 0) - T(\eta_2, 0)| \leqslant \frac{2}{c}\, e^{\alpha M_0^2}\, e^{L_{CP}(M_0)\operatorname{diam}(K_\eta)}\, L_{CP}(M_0)\,\|\eta_1 - \eta_2\|,$$

which proves Lipschitz continuity of $T(\cdot, 0)$ on K_η.

□

Proof of Theorem 1. We use the metric d_M on conditional laws induced by their Gaussian quantile maps:

$$d_M(p, q) := \sup_{(z,\eta)\in K_M} \max\Big(|T_p(\eta, 0) - T(\eta, 0)|,\ |\log \partial_z T_p(\eta, z) - \log \partial_z T(\eta, z)| \Big).$$

Step 1: Core uniform approximation on K_M. By Assumption 2(*iii*), for budget $m = C(M, \varepsilon)$ there exists $T \in \mathcal{F}_m$ such that

$$d_M\big(q_T, p\big) \quad \leqslant \quad \varepsilon. \tag{S7}$$

Write $\Delta_0(\eta) := T(\eta, 0) - T_p(\eta, 0)$, and $\Delta_\ell(\eta, z) := \log \partial_z T(\eta, z) - \log \partial_z T_p(\eta, z)$. Then $|\Delta_0| \leqslant \varepsilon$ and $|\Delta_\ell| \leqslant \varepsilon$ on K_M.

Step 2: Bounding the KL on the core $|z| \leqslant M$. For each η,

$$\mathrm{KL}\big(q_T(\cdot \mid \eta)\,\|\,p(\cdot \mid \eta)\big) = \mathbb{E}\Big[\log \partial_z T_p(\eta, Z) - \log \partial_z T(\eta, Z) + \log p\big(T_p(\eta, Z) \mid \eta\big) - \log p\big(T(\eta, Z) \mid \eta\big)\Big].$$

On $|Z| \leqslant M$, the first difference is $\leqslant \varepsilon$ by (S7). For the second, use the spikiness modulus on the bounded θ–range: on $|z| \leqslant M$,

$$|T(\eta, z) - T_p(\eta, z)| \leqslant |\Delta_0(\eta)| + \int_0^{|z|} \big|e^{\Delta_\ell(\eta,u)} - 1\big|\, \partial_u T_p(\eta, u)\, du \leqslant \varepsilon + \varepsilon\,|z|\, e^{H(M)} \leqslant \varepsilon\big(1 + 2M\, e^{H(M)}\big),$$

because $|\log \partial_u T_p| \leqslant H(M)$ and $|e^{\Delta_\ell} - 1| \leqslant e^{\varepsilon} - 1 \leqslant (e-1)\varepsilon \leqslant 2\varepsilon$ for $\varepsilon \leqslant 1/2$. Note that on K_M, $|T_p(\eta, Z)| \leqslant G(M) := M_0 + |M| e^{A_d + B_d|M|}$ by (13). Similarly, $|T(\eta, Z)| \leqslant \widetilde{G}(M) := M^* + |M| e^{A^* + B^*|M|}$. As $M^* > M_0$, $A^* > A_d$, and $B^* > B_d$, we have $\widetilde{G}(M) \geqslant G(M)$ and

both $T_p(\eta, Z)$ and $T(\eta, Z)$ lie in $[-\widetilde{G}(M), \widetilde{G}(M)]$. Hence,

$$\begin{aligned}\left|\log p\big(T(\eta,z)\mid\eta\big) - \log p\big(T_p(\eta,z)\mid\eta\big)\right| &\leqslant L_S\big(\widetilde{G}(M)\big)\,|T(\eta,z)-T_p(\eta,z)|\\ &\leqslant \varepsilon\, L_S\big(\widetilde{G}(M)\big)\,(1+2Me^{H(M)}).\end{aligned}$$

Integrating $\phi(z)$ over $|z| \leqslant M$ (mass $\leqslant 1$) gives the core contribution

$$\begin{aligned}&\leqslant \varepsilon\Big\{1 + L_S\big(\widetilde{G}(M)\big)\,\big(1+2Me^{H(M)}\big)\Big\}\\ &\leqslant 2\varepsilon\Big\{1 + L_S\big(\widetilde{G}(M)\big)\widetilde{G}(M)\Big\}.\end{aligned}$$

For this to be less than $\varepsilon'/2$,

$$\varepsilon \leqslant \frac{\varepsilon'}{4\Big\{1+\widetilde{G}(M)L_S\big(\widetilde{G}(M)\big)\Big\}}.$$

Step 3: Gaussian tails $|z| > M$. Using the identity above and Assumption 2(*ii*),

$$\log \partial_z T(\eta,z) \leqslant A^* + B^*|z|, \quad |T(\eta,z)| \leqslant |T(\eta,0)| + \int_0^{|z|} \partial_u T(\eta,u)\,du \leqslant M^* + e^{A^*}|z|e^{B^*|z|}.$$

Using $(a+b)^2 \leqslant 2(a^2+b^2)$ for any real a and b leads to

$$T(\eta,z)^2 \;\leqslant\; \Big(M^* + e^{A^*}|z|e^{B^*|z|}\Big)^2 \;\leqslant\; 2M^{*2} + 2e^{2A^*}z^2e^{2B^*|z|}.$$

By the tail-thinness lower bound in Assumption 1(*iv*),

$$-\log p\big(T(\eta,z)\mid\eta\big) \leqslant \alpha\, T(\eta,z)^2 - \log c.$$

With $\log\phi(z) = -\frac{z^2}{2} - \frac{1}{2}\log(2\pi)$, Assumption 1(*iv*) and 2(*ii*) give, for $|z| > M$ and all η,

$$\begin{aligned}&\log\phi(z) - \log p\big(T(\eta,z)\mid\eta\big) - \log\partial_z T(\eta,z)\\ \leqslant& -\frac{z^2}{2} - \frac{1}{2}\log(2\pi) - \log c + \alpha\,\Big(2M^{*2} + 2e^{2A^*}z^2e^{2B^*|z|}\Big) + A^* + B^*|z|\\ \leqslant& -\frac{1}{2}\log(2\pi) - \log c + \alpha\,\Big(2M^{*2} + 2e^{2A^*}z^2e^{2B^*|z|}\Big) + A^* + B^*|z|\\ \leqslant& C_1 \;+\; C_2\,z^2\,e^{2B^*|z|} \;+\; B^*|z|,\end{aligned}$$

where the explicit constants are

$$C_1 = A^* - \tfrac{1}{2}\log(2\pi) - \log c + 2\alpha\,M^{*2}, \qquad C_2 = 2\alpha e^{2A^*}.$$

Then the tail bound is

$$\int_{|z|>M} \phi(z)\,[\cdots]\,dz \leqslant \frac{1}{\sqrt{2\pi}}\Big[C_1\,I_0(M) \;+\; B^*\,I_1(M) \;+\; C_2\,I_2(M)\Big],$$

where

$$I_0(M) := \int_{|z|>M} e^{-z^2/2} dz, \quad I_1(M) := \int_{|z|>M} |z| e^{-z^2/2} dz, \quad I_2(M) := \int_{|z|>M} z^2 e^{-z^2/2+2B^*|z|} dz.$$

By Mills' ratio,

$$I_0(M) \leqslant \frac{2\sqrt{2\pi}\,\phi(M)}{M}, \qquad I_1(M) = 2\int_M^\infty z e^{-z^2/2} dz = 2\, e^{-M^2/2}.$$

Rewriting I_2 and using $z^2 - 4B^* z = (z - 2B^*)^2 - 4B^{*2}$,

$$I_2(M) = 2\int_M^\infty z^2 e^{-z^2/2+2B^* z} dz = 2e^{2B^{*2}} \int_M^\infty z^2 e^{-\frac{1}{2}(z-2B^*)^2} dz.$$

Using change of variable $t = z - \gamma$ where $\gamma = 2B^*$, and $(t+\gamma)^2 \leqslant 2(t^2 + \gamma^2)$, we have

$$I_2(M) = 2e^{\gamma^2/2} \int_{M-\gamma}^\infty (t+\gamma)^2 e^{-t^2/2}\, dt \;\leqslant\; 2e^{\gamma^2/2} M_\gamma \int_{M-\gamma}^\infty (1+t^2) e^{-t^2/2}\, dt,$$

where $M_\gamma := \max\{2\gamma^2,\, 2\}$.

For $x > 0$,

$$\int_x^\infty (1+t^2) e^{-t^2/2}\, dt = \int_x^\infty e^{-t^2/2}\, dt + \int_x^\infty t^2 e^{-t^2/2}\, dt.$$

Mills' ratio gives

$$\int_x^\infty e^{-t^2/2}\, dt \leqslant \frac{1}{x} e^{-x^2/2} \qquad (x > 0).$$

Also,

$$\int_x^\infty t^2 e^{-t^2/2}\, dt = \Big[-t\, e^{-t^2/2} \Big]_x^\infty + \int_x^\infty e^{-t^2/2}\, dt = x\, e^{-x^2/2} + \int_x^\infty e^{-t^2/2}\, dt \leqslant \Big(x + \frac{1}{x}\Big) e^{-x^2/2}.$$

Hence

$$\int_x^\infty (1+t^2) e^{-t^2/2}\, dt \leqslant \Big(x + \frac{2}{x}\Big) e^{-x^2/2}.$$

For $x \geqslant 1$, $x + 2/x \leqslant 2(1+x)$, so

$$\int_x^\infty (1+t^2) e^{-t^2/2}\, dt \leqslant 2(1+x) e^{-x^2/2}.$$

Taking $x = M - \gamma$ with $M \geqslant \gamma + 1$ gives

$$I_2(M) \;\leqslant\; 4M_\gamma\, e^{\gamma^2/2}\, (1 + M - \gamma)\, e^{-\frac{1}{2}(M-\gamma)^2} \;\leqslant\; C_\gamma\, (1+M)\, e^{-\frac{1}{2}(M-\gamma)^2},$$

with $C_\gamma := 4M_\gamma e^{\gamma^2/2}$. Setting $\gamma = 2B^*$ and $M_\gamma \leqslant 2\gamma^2 + 2 \leqslant 2(1 + 4B^{*2})$, yields

$$I_2(M) \;\leqslant\; \widetilde{C}_{B^*}\, (1+M)\, e^{-\frac{1}{2}(M-2B^*)^2}, \text{ where } \widetilde{C}_{B^*} = 8(1+4B^{*2}) e^{2B^{*2}}.$$

Putting the pieces together,

$$\int_{|z|>M} \phi(z)\,[\cdots]\,dz \;\leqslant\; \frac{C_1}{\sqrt{2\pi}}\frac{2\,e^{-M^2/2}}{M} \;+\; \frac{B^*}{\sqrt{2\pi}}\,2e^{-M^2/2} \;+\; \frac{C_2}{\sqrt{2\pi}}\,\widetilde{C}_{B^*}\,(1+M)\,e^{-\frac{1}{2}(M-2B^*)^2}$$
$$\leqslant\; R(1+M)\exp\Big(-\tfrac{1}{2}(M-2B^*)^2\Big)$$

after absorbing all constants in R and using the largest term. For this to be less than $\varepsilon'/2$, we write

$$t := M - 2B^*, \qquad 1+M = 1+2B^*+t \leqslant (1+2B^*)e^t.$$

we have

$$R(1+M)e^{-\frac{1}{2}t^2} \;\leqslant\; R(1+2B^*)\,e^{t-\frac{1}{2}t^2} \;=\; R(1+2B^*)\,e^{-\frac{1}{2}(t-1)^2+\frac{1}{2}}.$$

For this to be less than $\varepsilon'/2$

$$e^{-\frac{1}{2}(t-1)^2} \;\leqslant\; \frac{\varepsilon'}{2\,R(1+2B^*)\,e^{1/2}} \quad\Longleftrightarrow\quad (t-1)^2 \;\geqslant\; 2\log\frac{2e^{1/2}R(1+2B^*)}{\varepsilon'}.$$

$$M \;\geqslant\; 2B^*+1+\sqrt{2\log\frac{2e^{1/2}R(1+2B^*)}{\varepsilon'}}\,.$$

Absorbing $2e^{1/2}(1+2B^*)$ in R we can choose $M = 2B^*+1+\sqrt{2\log\frac{R}{\varepsilon'}}$ for a suitable constant R. So, for any small ε', M can be expressed as $O(\sqrt{-\log\varepsilon'})$ where the constant depends only on M^*, A^*, B^*, α and c.

Combining the findings of steps 2 and 3 completes the proof.

□

Proof of Corollary 1. (a) Existence. Recall that the parameter set $\Theta_m \subset \mathbb{R}^m$ is compact by Assumption 2. For each fixed (η, z), the maps $\vartheta \mapsto T_\vartheta(\eta,z)$ and $\vartheta \mapsto \partial_z T_\vartheta(\eta,z)$ are continuous by (Assumption 2(i)), hence the integrand

$$H_\vartheta(\eta,z) := \log \partial_z T_\vartheta(\eta,z) - \log p\big(T_\vartheta(\eta,z)\mid \eta\big)$$

is continuous in ϑ. By assumptions 2(ii) and 1(iv),

$$H_\vartheta(\eta,z) \;\leqslant\; A^*+B^*|z| - \log c + \alpha\Big(M^*+e^{A^*}|z|e^{B^*|z|}\Big)^2 \;\leqslant\; C_1+C_2(1+z^2)e^{2B^*|z|},$$

with C_1, C_2 independent of ϑ. Since $\int_{\mathbb{R}}(1+z^2)e^{2B^*|z|}\phi(z)\,dz < \infty$ (complete the square), the envelope is integrable against $\phi(z)p(\eta)$. By dominated convergence, $\vartheta \mapsto R(q_{T_\vartheta})$ is continuous on the compact set Θ_m, hence by Weierstrass Extreme Value Theorem it attains its minimum; existence of $\hat{q}_m$ follows.

(b) KLD bound. By Theorem 1, given a small ε, for the budget $m := C^*(\varepsilon)$ there exists $T \in \mathcal{F}_m$ such that $R\big(q_T\big) \leqslant \varepsilon$. Because $\hat{q}_m$ minimizes $R(q)$ over $q \in \mathcal{Q}_m$,

$$R(\hat{q}_m) \leqslant R(q_T) \leqslant \varepsilon.$$

This proves the claim. □

Proof of Theorem 4. Since $\mathcal{Q}_m = \{q_{T_\vartheta} : \vartheta \in \Theta_m\}$, with a slight abuse of notation, we can write $R(q)$ as $R(\vartheta)$ and $R_N(q)$ as $R_N(\vartheta)$ for the corresponding ϑ. As Θ_m is compact and $\vartheta \mapsto R_N(q_\vartheta)$ is continuous, once again by Weierstrass Extreme Value Theorem, we have $\hat{q}_{N,m} \in \arg\min_{\mathcal{Q}_m} R_N(q)$ exists.

Before proceeding further we state and prove a proposition on uniform laws of large numbers for stationary ergodic processes.

Proposition 1 (Uniform ergodic LLN for the empirical KL)**.** *Let $p(\theta \,|\, \eta)$ satisfies Assumption 1 for some compact K_η and $p(\eta)$ is a distribution supported on K_η. Let $\{\eta_r\}_{r\geqslant 1}$ be a stationary ergodic Markov chain with invariant distribution $p(\eta)$. Let $\mathcal{F}_m = \{T_\vartheta |\ \vartheta \in \Theta_m\}$ denote a class satisfying Assumption 2. Let q_ϑ to be the conditional distribution $\theta \,|\, \eta = T_\vartheta(\eta, Z)$, $Z \sim N(0,1)$, and*

$$f_\vartheta(\eta) = \mathrm{KL}\big(q_\vartheta(\theta \mid \eta) \,\|\, p(\theta \mid \eta)\big), \qquad R(\vartheta) = \mathbb{E}[f_\vartheta(\eta)], \quad R_N(\vartheta) = \frac{1}{N}\sum_{r=1}^{N} f_\vartheta(\eta_r).$$

Then $\sup_{\vartheta\in\Theta_m} \big|R_N(\vartheta) - R(\vartheta)\big| \xrightarrow[N\to\infty]{a.s.} 0.$

Proof of Proposition 1. Envelope. Let $H(\vartheta, \eta, z) = \log \phi(z) - \log p\big(T_\vartheta(\eta, z) \mid \eta\big) - \log \partial_z T_\vartheta(\eta, z)$. Then $f_\vartheta(\eta) = \mathbb{E}_{\theta \,|\, \eta = T_\vartheta(\eta, Z)}[H(\vartheta, \eta, Z)]$. Following the proof of Corollary 1, $f_\vartheta(\eta) \leqslant \tau$ for some positive τ free of η and for each fixed η, the map $\vartheta \mapsto f_\vartheta(\eta)$ and is continuous on compact Θ_m, implying that it is uniformly continuous. Define for $\varepsilon > 0$

$$L_\varepsilon(\eta) = \sup\{|f_\vartheta(\eta) - f_{\vartheta'}(\eta)| :\ \vartheta, \vartheta' \in \Theta_m,\ \|\vartheta - \vartheta'\| \leqslant \varepsilon\}.$$

Then by uniform continuity $L_\varepsilon(\eta) \to 0$ pointwise as $\varepsilon \downarrow 0$, and $0 \leqslant L_\varepsilon(\eta) \leqslant 2\tau$.

For each fixed $\vartheta \in \Theta_m$, the map $\eta \mapsto f_\vartheta(\eta)$ is measurable (it is an integral of a continuous function dominated by an integrable envelope). Let $D \subset \Theta_m$ be a countable dense subset. By continuity of $\vartheta \mapsto f_\vartheta(\eta)$ on compact Θ_m, for every η

$$L_\varepsilon(\eta) = \sup\{\, |f_\vartheta(\eta) - f_{\vartheta'}(\eta)| :\ \vartheta, \vartheta' \in D,\ \|\vartheta - \vartheta'\| \leqslant \varepsilon\,\}.$$

The rightmost expression is the supremum of a countable family of measurable functions,

hence L_ε is measurable. So $L_\varepsilon \in L^1(p)$ and by dominated convergence,

$$\mathbb{E}\, L_\varepsilon(\eta) \longrightarrow 0 \qquad (\varepsilon \downarrow 0). \tag{S8}$$

.

Since $\{\eta_r\}_{r\geqslant 1}$ is stationary ergodic and $L_\varepsilon \in L^1(p)$, Birkhoff's ergodic theorem yields

$$\frac{1}{N}\sum_{r=1}^{N} L_\varepsilon(\eta_r) \xrightarrow[N\to\infty]{a.s.} \mathbb{E}_{p(\eta)}[L_\varepsilon(\eta)].$$

Let $\{\vartheta_j\}_{j=1}^{n(\varepsilon)}$ be an ε-net of Θ_m. For any $\vartheta \in \Theta_m$, choose j with $\|\vartheta - \vartheta_j\| \leqslant \varepsilon$ and write

$$\begin{aligned}\big|R_N(\vartheta) - R(\vartheta)\big| &\leqslant \big|R_N(\vartheta) - R_N(\vartheta_j)\big| + \big|R_N(\vartheta_j) - R(\vartheta_j)\big| + \big|R(\vartheta) - R(\vartheta_j)\big|. \\ &\leqslant \frac{1}{N}\sum_{r=1}^{N} L_\varepsilon(\eta_r) + \big|R_N(\vartheta_j) - R(\vartheta_j)\big| + \mathbb{E}\, L_\varepsilon(\eta).\end{aligned}$$

Taking $\sup_{\vartheta\in\Theta_m}$ in the inequality, we get

$$\sup_{\vartheta\in\Theta_m}\big|R_N(\vartheta) - R(\vartheta)\big| \leqslant \frac{1}{N}\sum_{r=1}^{N} L_\varepsilon(\eta_r) + \max_{1\leqslant j\leqslant n(\varepsilon)}\big|R_N(\vartheta_j) - R(\vartheta_j)\big| + \mathbb{E}\, L_\varepsilon(\eta).$$

Taking $\limsup_{N\to\infty}$ on both sides and applying the ergodic theorem to the empirical average of $L_\varepsilon(\eta_r)$ yields

$$\limsup_{N\to\infty}\sup_{\vartheta\in\Theta_m}\big|R_N(\vartheta) - R(\vartheta)\big| \leqslant \max_{1\leqslant j\leqslant n(\varepsilon)}\limsup_{N\to\infty}\big|R_N(\vartheta_j) - R(\vartheta_j)\big| + 2\,\mathbb{E}\, L_\varepsilon(\eta). \tag{S9}$$

For each fixed j, the ergodic theorem gives $R_N(\vartheta_j) \to R(\vartheta_j)$ almost surely, hence the first term on the right side of (S9) is zero. Therefore

$$\limsup_{N\to\infty}\sup_{\vartheta\in\Theta_m}\big|R_N(\vartheta) - R(\vartheta)\big| \leqslant 2\,\mathbb{E}\, L_\varepsilon(\eta).$$

Letting $\varepsilon \downarrow 0$ and using $\mathbb{E}\, L_\varepsilon(\eta) \to 0$ finishes the proof.

□

Returning to the proof of Theorem 4, by Proposition 1,

$$\Delta_{N,m} := \sup_{\vartheta\in\Theta_m}\big|R_N(q) - R(q)\big| \xrightarrow[N\to\infty]{a.s.} 0 \qquad \text{for each fixed } m.$$

Let $\hat{q}_m \in \arg\min_{\mathcal{Q}_m} R(q)$, which exists by Corollary 1. Then, for all M,

$$R(\hat{q}_{N,m}) \leqslant R_N(\hat{q}_{N,m}) + \Delta_{N,m} \leqslant R_N(\hat{q}_m) + \Delta_{N,m} \leqslant R(\hat{q}_m) + 2\Delta_{N,m}.$$

Taking $\limsup_{N\to\infty}$ yields $\limsup_{N\to\infty} R(\hat{q}_{N,m}) \leqslant \inf_{q\in\mathcal{Q}_m} R(q) \leqslant \varepsilon$ by Corollary 1.

□

S5 Proofs for Neural Flow Examples

We now prove Theorems 2 and 3 for specific neural flow classes.

S5.1 Common Network Notation and Quantitative Bounds

Let $\eta \in$ some compact $K_\eta \in \mathbb{R}^d$ and $\mathrm{diam}_\eta := \sup_{\eta,\eta' \in K_\eta} \|\eta - \eta'\|_2$. Write $\|\cdot\|_\infty$ for the max norm and $\|\cdot\|_{\mathrm{op}}$ for the operator norm. Throughout, clipping uses the one Lipschitz map $\mathrm{clip}(x; a, b) := \min\{b, \max\{a, x\}\}$.

S5.2 Unconstrained Monotonic Neural Networks

Proof of Theorem 2. It is enough to show that the class satisfies Assumption 2 as the rest follows from Theorem 2. Define

$$\begin{aligned} T_\vartheta(\eta, z) &= a_{\vartheta_a}(\eta) + \int_0^z \exp\big(g_{\vartheta_g}(\eta, t)\big)\, dt, \\ a_{\vartheta_a} &= \mathrm{clip}(\tilde{a}_{\vartheta_a}, -M^*, M^*), \\ g_{\vartheta_g} &= \mathrm{clip}(\tilde{g}_{\vartheta_g}, -A^* - B^*|z|, A^* + B^*|z|), \end{aligned} \tag{S10}$$

where $\tilde{a}_{\vartheta_a}$ is a ReLU network on K_η, $\tilde{g}_{\vartheta_g}$ is a ReLU network on K_M, and $\vartheta = (\vartheta_a, \vartheta_g)^\top$. The clips are pointwise in the displayed variables. Since clipping is one Lipschitz, it preserves continuity and does not increase moduli.

Assumption 2(*i*). For every (η, z, ϑ) one has $\partial_z T_\vartheta(\eta, z) = \exp\big(g_{\vartheta_g}(\eta, z)\big) > 0$. Continuity in (η, z, ϑ) follows from continuity of $\tilde{a}_{\vartheta_a}$, $\tilde{g}_{\vartheta_g}$, exponential, and the integral in the upper limit. Differentiability in z and continuity of this partial derivative follow directly from the fundamental theorem of calculus, yielding $\partial_z T_\vartheta(\eta, z) = \exp(g_{\vartheta_g}(\eta, z))$.

Assumption 2(*ii*). Following definition (S10), $|T_\vartheta(\eta, 0)| = |a_{\vartheta_a}(\eta)| \leqslant M^*$ and $|\log \partial_z T_\vartheta(\eta, z)| = |g_{\vartheta_g}(\eta, z)| \leqslant A^* + B^*|z|$. This is precisely the global linear envelope. So, Assumption 2(*ii*) is satisfied.

Assumption 2(*iii*). Lastly, we establish part (*iii*) with an explicit construction. Let the target quantile map T_p satisfy Assumption 1 on K_M. Define $u_1(\eta) := T_p(\eta, 0)$ and $u_2(\eta, z) := \log \partial_z T_p(\eta, z)$. The definition yields constants $M_0 > 0$ and $A_d, B_d, C_d \geqslant 0$ with

$$\sup_{\eta \in K_\eta} |u_1(\eta)| \leqslant M_0, \qquad \sup_{(z,\eta) \in K_M} |u_2(\eta, z)| \leqslant A_d + B_d|z|.$$

Choose thresholds $M^* > M_0$, $A^* > A_d$, $B^* > B_d$. Hence if we construct $\tilde{a}_{\vartheta_a}, \tilde{g}_{\vartheta_g}$ so that

$$\sup_{\eta \in K_\eta} \left|\tilde{a}_{\vartheta_a}(\eta) - u_1(\eta)\right| \leqslant \varepsilon, \qquad \sup_{(z,\eta)\in K_M} \left|\tilde{g}_{\vartheta_g}(\eta, z) - u_2(\eta, z)\right| \leqslant \varepsilon, \qquad 0 < \varepsilon \leqslant \varepsilon_{\max},$$

with $\varepsilon_{\max} := \min\{M^* - M_0,\ A^* - A_d,\ B^* - B_d\}$, then the clips are inactive on K_M and $a_{\vartheta_a} = \tilde{a}_{\vartheta_a}$, $g_{\vartheta_g} = \tilde{g}_{\vartheta_g}$ there. It follows that

$$\sup_{\eta \in K_\eta} |a_{\vartheta_a}(\eta) - u_1(\eta)| \leqslant \varepsilon, \qquad \sup_{(z,\eta)\in K_M} |g_\vartheta(\eta, z) - u_2(\eta, z)| \leqslant \varepsilon.$$

Assumption 1 supplies a Lipschitz moduli L_a for u_1 on K_η and $\mathsf{L}_b(M)$ for u_2 on K_M. That is, we approximate the true maps

$$u_1(\eta) = T(\eta, 0) \quad \text{on } K_\eta, \qquad u_2(\eta, z) = \log \partial_z T(\eta, z) \quad \text{on } K_M$$

with quantitative budgets that depend on L_a and $\mathsf{L}_b(M)$.

Theorem 1 in Yarotsky (2017) proves that for any $d \geqslant 1$, a 1-Lipschitz function f on $[0,1]^d$ with $\|f\|_\infty \leqslant 1$ admits a ReLU network approximation with depth $O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ and number of weights $O\left(\varepsilon^{-d}\log\left(\frac{1}{\varepsilon}\right)\right)$ that achieves sup norm error at most ε on $[0,1]^d$. As u_1 is L_a-Lipschitz on K_η and u_2 is $\mathsf{L}_b(M)$-Lipschitz on K_M, we map each domain to the unit cube by the affine scaling $x \mapsto (x - x_0)/\operatorname{diam}$. Lipschitz constants scale by the domain diameter, so the rescaled functions are $\mathsf{L}_a \operatorname{diam}_\eta$ and $\mathsf{L}_b(M)(\operatorname{diam}_\eta + M)$ Lipschitz on $[0,1]^d$ up to absolute constants. Applying the theorem of Yarotsky (2017) and scaling the error back gives realizations $\tilde{a}_{\vartheta_a}$ and $\tilde{g}_{\vartheta_g}$ with

$$\|\tilde{a}_{\vartheta_a} - u_1\|_{\infty, K_\eta} \leqslant \varepsilon, \qquad \|\tilde{g}_{\vartheta_g} - u_2\|_{\infty, K_M} \leqslant \varepsilon,$$

for depth $O(\log(1/\varepsilon))$, and parameter counts

$$\#\tilde{a}_{\vartheta_a} \ \leqslant\ C_a\left(\tfrac{\mathsf{L}_a \operatorname{diam}_\eta}{\varepsilon}\right)^{d_\eta} \log \tfrac{1}{\varepsilon}, \qquad \#\tilde{g}_{\vartheta_g} \ \leqslant\ C_g\left(\tfrac{\mathsf{L}_b(M)(\operatorname{diam}_\eta + M)}{\varepsilon}\right)^{d_\eta + 1} \log \tfrac{1}{\varepsilon}.$$

Here C_a and C_g are fixed constants depending only on d_η. This completes the verification of Assumption 2 for UMNN. Then Theorem 2 is proved by simply using $\varepsilon/O(\tilde{L}_S(\tilde{G}(M(\varepsilon))))$ which is the required scaling from Theorem 1. □

S5.3 Rational Quadratic Neural Spline Flows

Proof of Theorem 3. Partition $[-M, M]$ into $2K$ equal bins of width $w = M/K$ with knots $z_i = -M + jw$ for $j = 0, \ldots, 2K$, so $z_K = 0$. We will choose K later and maintain $K \geqslant M$, hence $w \leqslant 1$. For each η the conditioner $\psi_\vartheta(\eta)$ outputs raw vectors $b^{(h)}(\eta) \in \mathbb{R}^{2K}$, $v(\eta) \in \mathbb{R}^{2K+1}$, and a scalar offset $c(\eta)$ with $|c(\eta)| \leqslant$ some B_0 that we will specify later. Define heights

with a total mass $H_{\text{tot}} > 0$,

$$h_i(\eta) = H_{\text{tot}}\,\text{softmax}\big(b^{(h)}(\eta)\big)_i, \quad y_0 := c(\eta), \quad y_{i+1} := y_i + h_i(\eta) \in [c(\eta),\, c(\eta) + H_{\text{tot}}],$$

and slopes via clipped logits

$$\beta_i(\eta) = \text{clip}\big(v_i(\eta), -A^* - B^*|z_i|, A^* + B^*|z_i|\big), \qquad s_i(\eta) := e^{\beta_i(\eta)}.$$

On a bin $[z_i, z_{i+1}]$ with $\Delta_i := h_i/w$ and $t := (z - z_i)/w \in [0,1]$, the monotone rational quadratic formula (Durkan et al., 2019) gives

$$T_\vartheta(\eta, z) = y_i + h_i \frac{\Delta_i t^2 + s_i t(1-t)}{\Delta_i + (s_{i+1} + s_i - 2\Delta_i)t(1-t)},$$
$$\partial_z T_\vartheta(\eta, z) = \frac{\Delta_i^2\big(s_{i+1}t^2 + 2\Delta_i t(1-t) + s_i(1-t)^2\big)}{\big(\Delta_i + (s_{i+1} + s_i - 2\Delta_i)t(1-t)\big)^2}.$$

Assumption 2(*i*). The numerator of $\partial_z T_\vartheta(\eta, z)$ is strictly positive for $t \in [0,1]$ since $s_i, s_{i+1} > 0$, so the derivative is positive on each bin. By construction of RQ-NSF, for adjacent bins, the value and the first derivative match at the knots. Extending with linear tails $z \mapsto y_0 + s_0(z - z_0)$ for $z \leqslant z_0$ and $z \mapsto y_{2K} + s_{2K}(z - z_{2K})$ for $z \geqslant z_{2K}$ yields a globally C^1 strictly increasing map. This proves Assumption 2(*i*).

Assumption 2(*ii*). For $z \in [z_i, z_{i+1}]$, Appendix A of Durkan et al. (2019) implies

$$\min\{s_i(\eta), s_{i+1}(\eta)\} \leqslant \partial_z T_\vartheta(\eta, z) \leqslant \max\{s_i(\eta), s_{i+1}(\eta)\}.$$

The clipping gives $e^{-A^* - B^*|z_i|} \leqslant s_i(\eta) \leqslant e^{A^* + B^*|z_i|}$. Using $|z_i| \leqslant |z| + w$ and $|z_{i+1}| \leqslant |z| + w$ with $w \leqslant 1$, we obtain

$$\big|\log \partial_z T_\vartheta(\eta, z)\big| \leqslant (A^* + B^*) + B^*|z| \qquad \text{for all } (\eta, z).$$

Moreover

$$|T_\vartheta(\eta, 0)| = |y_K(\eta)| \leqslant |c(\eta)| + \sum_{i=0}^{2K-1} h_i(\eta) \leqslant B_0 + H_{\text{tot}} =: M^*.$$

Choosing $B_0 \geqslant M_0$ establishes Assumption 2(*ii*).

Assumption 2(*iii*). Let T_p satisfy Assumption 1 and set $u(\eta, z) := \log \partial_z T_p(\eta, z)$. By Lemma 1, on K_M the function u is $\mathsf{L}_b(M)$-Lipschitz. Using the RQS formula with endpoint logits $\beta_i(\eta) = u(\eta, z_i)$ on each bin,

$$\sup_{\tau \in [0,w]} \big|\log \partial_z T_\vartheta(\eta, z_i + \tau) - u(\eta, z_i + \tau)\big| \leqslant \mathsf{L}_b(M)\, w.$$

Hence

$$\sup_{(z,\eta)\in K_M} \left|\log \partial_z T_\vartheta(\eta,z) - \log \partial_z T_p(\eta,z)\right| \leqslant \mathsf{L}_b(M)\frac{M}{K}.$$

Impose the second inequality in Assumption 2(iii) with tolerance ε by choosing

$$K_\star := \left\lceil \frac{\mathsf{L}_b(M)\, M}{\varepsilon} \right\rceil,$$

so $K_\star \geqslant M$ and $w \leqslant 1$.

Let $h_i^\star(\eta) = \int_{z_i}^{z_{i+1}} \partial_z T_p(\eta,z)\, dz$ and $S(\eta) = \sum_j h_j^\star(\eta) = \int_{-M}^{M} e^{u(\eta,t)} dt$. Define

$$b_i^\star(\eta) := \log h_i^\star(\eta) - a(\eta), \qquad a(\eta) := \log S(\eta) - \log H_{\text{tot}},$$

so that $H_{\text{tot}}\,\text{softmax}\big(b^\star(\eta)\big) = h^\star(\eta)$.

Moreover, Lemma 1 gives $\text{Lip}_\eta(u(\cdot,z)) \leqslant \mathsf{L}_b(M)$; hence

$$\begin{aligned} |h_i^\star(\eta) - h_i^\star(\eta')| &\leqslant \int_{z_i}^{z_{i+1}} e^{H(M)}\, |u(\eta,t) - u(\eta',t)|\, dt \leqslant w\, e^{H(M)} \mathsf{L}_b(M)\, \|\eta - \eta'\|, \\ |\log h_i^\star(\eta) - \log h_i^\star(\eta')| &\leqslant \frac{|h_i^\star(\eta) - h_i^\star(\eta')|}{\min h_i^\star} \leqslant e^{2H(M)} \mathsf{L}_b(M)\, \|\eta - \eta'\|. \end{aligned} \tag{L1}$$

We have $2Me^{-H(M)} \leqslant S(\eta) \leqslant 2Me^{H(M)}$ and $\text{Lip}_\eta(S) \leqslant 2Me^{H(M)}\mathsf{L}_b(M)$, hence

$$|a(\eta) - a(\eta')| = |\log S(\eta) - \log S(\eta')| \leqslant e^{2H(M)} \mathsf{L}_b(M)\, \|\eta - \eta'\|.$$

Therefore

$$|b_i^\star(\eta) - b_i^\star(\eta')| \leqslant 2\, e^{2H(M)} \mathsf{L}_b(M)\, \|\eta - \eta'\| \quad \Rightarrow \quad \text{Lip}_\eta\big(b^\star\big) \leqslant 2\, e^{2H(M)} \mathsf{L}_b(M).$$

Set $F(\eta) = \big(b^\star(\eta),\, v^\star(\eta),\, c^\star(\eta)\big) \in \mathbb{R}^{4K_\star+2}$, $\qquad v_i^\star(\eta) = u(\eta, z_i)$, $\quad c^\star(\eta) = T_p(\eta, -M)$.

So, $b^\star$ is $2e^{2H(M)}\mathsf{L}_b(M)$-Lipschitz and by Lemma 1 $v^\star$ is $\mathsf{L}_b(M)$-Lipschitz. We now establish the Lipschitz constant for $c^\star$. We start from the quantile identity

$$F_p\big(T_p(\eta,z) \mid \eta\big) = \Phi(z),$$

which at $z = -M$ gives $F_p(c^\star(\eta) \mid \eta) = \Phi(-M)$. Differentiating both sides with respect to η and applying the chain rule yields

$$\nabla_\eta F_p(\theta \mid \eta)\big|_{\theta = c^\star(\eta)} + p(c^\star(\eta) \mid \eta)\, \nabla_\eta c^\star(\eta) = 0,$$

hence

$$\|\nabla_\eta c^\star(\eta)\| \leqslant \frac{\|\nabla_\eta F_p(\theta \mid \eta)\|_{\theta = c^\star(\eta)}}{p(c^\star(\eta) \mid \eta)}. \tag{1}$$

We bound the numerator $\|\nabla_\eta F_p(\theta \mid \eta)\|_{\theta = c^\star(\eta)}$ using a change of variables along the quantile

map. For any z,

$$p\big(T_p(\eta,z)\mid\eta\big)\,\partial_z T_p(\eta,z)=\phi(z),$$

so for $\theta=c^\star(\eta)=T_p(\eta,-M)$ we have

$$\begin{aligned}\nabla_\eta F_p(\theta\mid\eta)\big|_{\theta=c^\star(\eta)} &= \int_{-\infty}^{c^\star(\eta)}\nabla_\eta p(t\mid\eta)\,dt=\int_{-\infty}^{-M}\nabla_\eta p\big(T_p(\eta,z)\mid\eta\big)\,\partial_z T_p(\eta,z)\,dz\\ &=\int_{-\infty}^{-M} p\big(T_p(\eta,z)\mid\eta\big)\,\nabla_\eta\log p\big(T_p(\eta,z)\mid\eta\big)\,\partial_z T_p(\eta,z)\,dz\\ &=\int_{-\infty}^{-M}\phi(z)\,\nabla_\eta\log p\big(T_p(\eta,z)\mid\eta\big)\,dz.\end{aligned}$$

By Assumption 1(*ii*), for $|\theta|\leqslant G(M)$ we have $\|\nabla_\eta\log p(\theta\mid\eta)\|\leqslant L_{CP}(G(M))$. Quantile growth ensures $|T_p(\eta,z)|\leqslant G(|z|)\leqslant G(M)$ for all $z\in(-\infty,-M]$. Therefore

$$\Big\|\nabla_\eta F_p(\theta\mid\eta)\Big\|_{\theta=c^\star(\eta)}\leqslant\int_{-\infty}^{-M}\phi(z)\,L_{CP}(G(M))\,dz=L_{CP}(G(M))\,\Phi(-M).$$

For the denominator, using $\partial_z T_p(\eta,z)=\phi(z)/p(T_p(\eta,z)\mid\eta)$ gives

$$p(c^\star(\eta)\mid\eta)=\frac{\phi(-M)}{\partial_z T_p(\eta,-M)}.$$

From Assumption 2(*ii*), $|\log\partial_z T_p(\eta,z)|\leqslant H(|z|)$, hence $\partial_z T_p(\eta,-M)\leqslant e^{H(M)}$, so

$$p(c^\star(\eta)\mid\eta)\geqslant\phi(-M)e^{-H(M)}.$$

Combining and applying Mill's ratio, $c^\star$ is $L_{\text{edge}}(M)$-Lipschitz with

$$\|\nabla_\eta c^\star(\eta)\|\leqslant\frac{L_{CP}(G(M))}{\phi(-M)e^{-H(M)}}=e^{H(M)}\,L_{CP}(G(M))/M=:L_{edge}.$$

By the ReLU vector sup-norm approximation theorem (Yarotsky, 2017), applied to $F=(b^\star,v^\star,c^\star)$ with

$$\mathrm{Lip}_\eta(F)\ \leqslant\ 2e^{2H(M)}\mathsf{L}_b(M)+L_{\text{edge}}(M)\qquad\text{where}\qquad L_{\text{edge}}(M)=\frac{e^{H(M)}}{M}\,L_{CP}\big(G(M)\big),$$

there exists a conditioner $\psi_\vartheta:K_\eta\mapsto\mathbb{R}^{4K_\star+2}$ such that

$$\sup_{\eta\in K_\eta}\big\|F(\eta)-\psi_\vartheta(\eta)\big\|_\infty\ \leqslant\ \varepsilon_{\text{par}},$$

and the number of parameters satisfies

$$m\ \leqslant\ C_r\,(4K_\star+2)\,A^{d_\eta}\log A,\qquad A\ :=\ \frac{\big(2e^{2H(M)}\mathsf{L}_b(M)+L_{\text{edge}}(M)\big)\,\mathrm{diam}(K_\eta)}{\varepsilon_{\text{par}}}.$$

From $\|F - \psi_\vartheta\|_\infty \leqslant \varepsilon_{\text{par}}$ we get

$$\sup_{\eta \in K_\eta} |c(\eta) - c^\star(\eta)| \;\leqslant\; \varepsilon_{\text{par}}.$$

We now derive choice of ε_{par} such that both the log derivative and the median errors are bounded by ε. For the derivative, let

$$\delta := \sup_{(z,\eta) \in K_M} \big|\log \partial_z T_\vartheta(\eta, z) - \log \partial_z T_p(\eta, z)\big|.$$

From the RQ interpolation argument, for each bin $[z_i, z_{i+1}]$,

$$\big|\log \partial_z T_\vartheta(\eta, z) - \log \partial_z T_p(\eta, z)\big| \leqslant \mathsf{L}_b(M)\, w + \tau, \qquad w = \frac{M}{K_\star},$$

where

$$\tau \;:=\; \max_i \big|\beta_i(\eta) - v_i^*(\eta)\big| = \max_i \big|v_i(\eta) - v_i^*(\eta)\big| \;\leqslant\; \varepsilon_{\text{par}}.$$

Choose

$$K_\star = \left\lceil \frac{\mathsf{L}_b(M) M}{\varepsilon_d} \right\rceil \qquad \text{with} \qquad \varepsilon_d = \frac{\varepsilon}{4\, M\, e^{H(M)}},$$

so that $\mathsf{L}_b(M) w \leqslant \varepsilon_d$.

Then

$$\mathsf{L}_b(M) w + \tau \;\leqslant\; \varepsilon_d + \varepsilon_{\text{par}} \;\leqslant\; \frac{\varepsilon}{4 M e^{H(M)}} + \frac{\varepsilon}{4 M e^{H(M)}} \;=\; \frac{\varepsilon}{2 M e^{H(M)}},$$

and hence

$$\sup_{(z,\eta) \in K_M} \big|\log \partial_z T_\vartheta(\eta, z) - \log \partial_z T_p(\eta, z)\big| \;\leqslant\; \varepsilon.$$

For the median bound, note that

$$T(\eta, 0) = c(\eta) + \int_{-M}^{0} \partial_z T(\eta, t)\, dt.$$

Therefore

$$\begin{aligned} \sup_{\eta \in K_\eta} \big|T_\vartheta(\eta, 0) - T_p(\eta, 0)\big| &\leqslant \|c - c^\star\|_\infty + \sup_\eta \int_{-M}^{0} \big|\partial_z T_\vartheta(\eta, t) - \partial_z T_p(\eta, t)\big|\, dt \\ &\leqslant \varepsilon_{\text{par}} + M\, e^{H(M)} \big(e^\delta - 1\big). \end{aligned}$$

Using $\delta \leqslant \varepsilon/(2 M e^{H(M)})$ and $e^x - 1 \leqslant (e-1)x$ for $x \leqslant 1$ gives $M e^{H(M)}(e^\delta - 1) \leqslant \varepsilon(e-1)/2$, and since $\varepsilon_{\text{par}} = \frac{\varepsilon}{4 M e^{H(M)}} \leqslant \varepsilon(e-1)/2$, we have

$$\sup_{\eta \in K_\eta} \big|T_\vartheta(\eta, 0) - T_p(\eta, 0)\big| \;\leqslant\; \varepsilon.$$

Thus both inequalities in Assumption 2(*iii*) hold with tolerance ε.

Finally, the parameter budget becomes

$$m \;\leqslant\; C_r\big(4K_\star + 2\big)\left(\frac{\big(2e^{2H(M)}\mathsf{L}_b(M) + L_{\text{edge}}(M)\big)\,\text{diam}(K_\eta)}{\varepsilon_{\text{par}}}\right)^{d_\eta} \times \log\left(\frac{\big(2e^{2H(M)}\mathsf{L}_b(M) + L_{\text{edge}}(M)\big)\,\text{diam}(K_\eta)}{\varepsilon_{\text{par}}}\right),$$

with

$$K_\star = \left\lceil \frac{\mathsf{L}_b(M)M}{\varepsilon_d} \right\rceil, \qquad \varepsilon_d = \frac{\varepsilon}{4\,M\,e^{H(M)}}, \qquad \varepsilon_{\text{par}} = \frac{\varepsilon}{4\,M\,e^{H(M)}}.$$

Once again to apply Theorem 1, one has to replace ϵ by $\varepsilon/O(\tilde{L}_S(\tilde{G}(M(\varepsilon)))$ which yields the expressions in the statement of Theorem 3.

□

S6 Details of Simulation Experiments

S6.1 Misspecified Prior

We revisit a commonly used illustrative example (Liu et al., 2009; Jacob et al., 2017; Yu et al., 2023) that demonstrates the challenges of full Bayesian inference when there is model misspecification or biased data. This example constitutes a case of a misspecified prior in the downstream analysis. The upstream dataset consists of a small sample $z = (z_1, \ldots, z_{n_1})^\top$, where $z_i \overset{\text{iid}}{\sim} N(\varphi, 1)$. The downstream dataset is a large sample $w = (w_1, \ldots, w_{n_2})^\top$, with $w_i \overset{\text{iid}}{\sim} N(\varphi + \eta, 1)$. This dataset has a mean $\varphi + \eta$, where η represents a downstream bias parameter. We use the same priors as in prior research, i.e., $\varphi \sim N(0, \delta_1^{-1})$ and $\eta \sim N(0, \delta_2^{-1})$, with δ_2 set to a large value to reflect high prior confidence in the absence of bias. When the true bias is large, this prior leads to a biased posterior under the full Bayesian model due to the overconfident assumption that $\eta \approx 0$.

This simulation setting illustrates how posterior samples from the small upstream dataset can be leveraged to inform inference on the bias parameter η for the large dataset. Following previous studies, we set $n_1 = 100$ and $n_2 = 1000$, with true values $\varphi = 0$ and $\eta = 1$. The inverse variance parameters are set to $\delta_1 = 1$ and $\delta_2 = 100$. For this setting, the full Bayes posterior and the cut-posterior for η both have closed-form expressions. We first derive these.

Suppose we have the upstream analysis based on a relatively small sample size n_1, a

following downstream analysis with a large sample size n_2, the model is as follows:

$$z_i \overset{\text{iid}}{\sim} N(\phi, 1), \quad i = 1, \ldots, n_1,$$
$$w_i \,|\, \phi, \eta \overset{\text{iid}}{\sim} N(\phi + \eta, 1), \quad i = 1, \ldots, n_2,$$

with Gaussian priors

$$\phi \sim N(0, \delta_1^{-1}), \qquad \eta \sim N(0, \delta_2^{-1}).$$

Let $S_z = \sum_{i=1}^{n_1} z_i, S_w = \sum_{i=1}^{n_2} w_i$.

Full Bayes Posterior. The joint distribution of (ϕ, η) is given by

$$\begin{aligned} p(\phi, \eta \,|\, z, w) &\propto p(z \,|\, \phi)\, p(w \,|\, \phi, \eta)\, p(\phi)\, p(\eta) \\ &\propto \exp\left(-\frac{1}{2} \left[\sum_{i=1}^{n_1} (z_i - \phi)^2 + \sum_{i=1}^{n_2} (w_i - \phi - \eta)^2 \right] - \frac{\delta_1}{2}\phi^2 - \frac{\delta_2}{2}\eta^2 \right) \\ &\propto \exp\left(-\frac{1}{2} \left[(n_1 + n_2 + \delta_1)\phi^2 + (n_2 + \delta_2)\eta^2 + 2n_2\phi\eta - 2\phi(S_z + S_w) - 2\eta S_w \right] \right). \end{aligned}$$

Thus $(\phi, \eta)^\top$ follows a bivariate normal distribution: $(\phi, \eta)^\top \,|\, z, w \sim N(\mu, \Sigma)$ where the covariance and mean vector are given as follows:

$$\Sigma = \frac{1}{D} \begin{pmatrix} n_2 + \delta_2 & -n_2 \\ -n_2 & n_1 + n_2 + \delta_1 \end{pmatrix}, \quad D = (n_1 + n_2 + \delta_1)(n_2 + \delta_2) - n_2^2.$$

The posterior mean vector is given by

$$\mu = \Sigma \begin{pmatrix} S_z + S_w \\ S_w \end{pmatrix} = \frac{1}{D} \begin{pmatrix} (n_2 + \delta_2)(S_z + S_w) - n_2 S_w \\ -n_2(S_z + S_w) + (n_1 + n_2 + \delta_1) S_w \end{pmatrix}$$

Explicitly,

$$\mu_\phi = \frac{(n_2 + \delta_2)S_z + \delta_2 S_w}{(n_1 + n_2 + \delta_1)(n_2 + \delta_2) - n_2^2}, \quad \mu_\eta = \frac{(n_1 + \delta_1)S_w - n_2 S_z}{(n_1 + n_2 + \delta_1)(n_2 + \delta_2) - n_2^2},$$

and

$$\text{Var}_\phi = \frac{n_2 + \delta_2}{(n_1 + n_2 + \delta_1)(n_2 + \delta_2) - n_2^2}, \quad \text{Var}_\eta = \frac{n_1 + n_2 + \delta_1}{(n_1 + n_2 + \delta_1)(n_2 + \delta_2) - n_2^2}.$$

True Cut Posterior. Given the model settings, the cut-posterior for ϕ is defined as $p_{\text{cut}}(\phi \,|\, z, w) = p(\phi \,|\, z)$, where $\phi \,|\, z$:

$$\phi \,|\, z \sim N\left(\frac{S_z}{n_1 + \delta_1}, \frac{1}{n_1 + \delta_1} \right).$$

Conditional on ϕ, the downstream posterior of η is

$$\eta \,|\, w, \phi \sim N\left(\frac{S_w - n_2\phi}{n_2 + \delta_2},\ \frac{1}{n_2 + \delta_2}\right).$$

Marginalizing over $\eta \,|\, z$:

$$\mathbb{E}[\eta \,|\, w] = \mathbb{E}[\mathbb{E}[\eta \,|\, \phi, w]] = \frac{(n_1 + \delta_1)S_w - n_2 S_z}{(n_1 + \delta_1)(n_2 + \delta_2)}.$$

The variance of the marginal posterior for η is

$$\begin{aligned}
\mathrm{Var}(\eta \,|\, w) &= \mathbb{E}\left[\mathrm{Var}(\eta \,|\, w, \phi)\right] + \mathrm{Var}\left[\mathbb{E}(\eta \,|\, w, \phi)\right] \\
&= \frac{1}{n_2 + \delta_2} + \left(\frac{n_2}{n_2 + \delta_2}\right)^2 \frac{1}{n_1 + \delta_1} \\
&= \frac{1}{n_2 + \delta_2} + \frac{n_2^2}{(n_2 + \delta_2)^2(n_1 + \delta_1)}.
\end{aligned}$$

We can thus compare the marginal distribution of η across four methods: the Full Bayes posterior, true cut-posterior, NeVI-Cut estimate of the cut-posterior, and estimated cut-posterior using the Gaussian variational approximation of Yu et al. (2023) (*Gaussian VA-Cut*).

We evaluate each method using 95% weighted interval score, Continuous Ranked Probability Score (CRPS; Gneiting and Raftery, 2007), mean squared error (MSE), and 95% coverage as comparison metrics. CRPS measures the accuracy of the predictive distribution by comparing it to the observed value. The 95% weighted interval score assesses the accuracy of prediction based on the quantiles of the predicted samples. The results are presented in Table 1.

S6.2 Using Propensity Scores in Misspecified Outcome Model

In this section, we present the details of the propensity score analysis.

The propensity score model (upstream) is specified as

$$\mathrm{logit}\, p(X_i = 1 \mid C_i) = \theta_{1,0} + \sum_{j=1}^{p} \theta_{1,j} C_{ij}, \quad i = 1, \ldots, n,$$

where $X_i = 1$ indicates treatment and $X_i = 0$ indicates control. The fitted probabilities from this model serve as the estimated propensity scores e_i.

To adjust for confounding in the outcome model (downstream), individuals are stratified into quantile-based groups according to their estimated propensity scores. Let $g_k(e_i)$ be a binary indicator equal to 1 if subject i's propensity score e_i falls in the k-th stratum (e.g.,

quintile), and 0 otherwise. The outcome model is

$$\operatorname{logit} p(Z_i = 1 \mid X_i, e_i) = \theta_{2,0} + \theta_{2,1} X_i + \sum_{k=2}^{5} \theta_{2,k} g_k(e_i), \quad i = 1, \ldots, n,$$

where $\theta_{2,1}$ is the causal parameter of interest, quantifying the effect of the treatment X_i on the outcome Z_i. Normal priors are assigned to all parameters, with variance $\sigma^2 = 800$ for the intercept and $\sigma^2 = 50$ for slopes.

For data generation, we simulate a realistic setting where the true treatment assignment depends on confounders. The treatment model (propensity score model) follows the same structure as the analysis model, with true coefficients $\theta_1^* = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$.

The true outcome model is given by:

$$\begin{aligned}\operatorname{logit} p(Z_i = 1 | X_i, C_i) = \gamma_0^* X_i + \gamma_1^* C_{i1} + \gamma_2^* \exp(C_{i2} - 1) + \gamma_3^* C_{i3} \\ + \gamma_4^* \exp(C_{i4} - 1) + \gamma_5^* |C_{i5}| + \gamma_6^* |C_{i6}|, \quad i = 1, \ldots, n,\end{aligned}$$

with $\gamma^* = (0, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$.

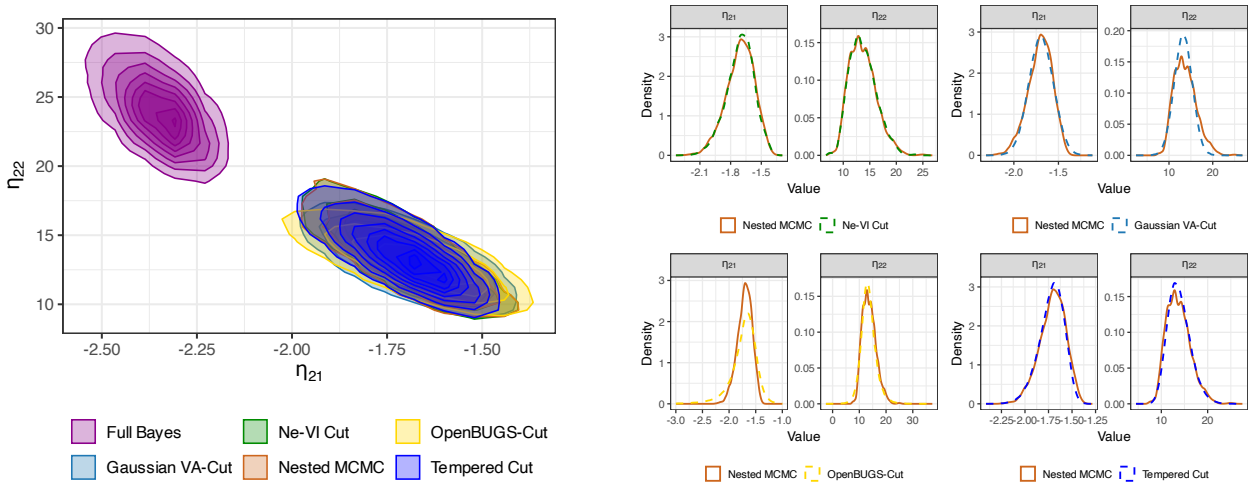
S7 HPV Data

We revisit the epidemiological example from Plummer (2015); Jacob et al. (2017); Yu et al. (2023), which examines the relationship between human papillomavirus (HPV) prevalence and the incidence of cervical cancer (Maucort-Boulch et al., 2008). Our goal is to incorporate posterior samples of HPV prevalence, obtained from an external analysis, into an outcome model that captures their association with cervical cancer incidence.

In the upstream analysis, posterior samples of HPV prevalence are derived using prevalence data from 13 countries. Let $z = (z_1, \ldots, z_{13})^\top$, where z_i denotes the number of individuals infected with high-risk HPV out of a sample of size n_i in country i. The data are modeled as $z_i \sim \text{Binomial}(n_i, \gamma_i)$, where γ_i is the parameter representing the true HPV prevalence in country i. Independent Beta priors are placed on each prevalence parameter, $\gamma_i \sim \text{Beta}(1, 1)$. By conjugacy, the posterior distribution is

$$\gamma_i | z_i \sim \text{Beta}(1 + z_i, 1 + n_i - z_i), \quad i = 1, \ldots, 13.$$

Posterior samples of γ_i are then passed to the downstream analysis, which models the relationship between HPV prevalence and cervical cancer incidence. Let $w = (w_1, \ldots, w_{13})^\top$



(a) Joint posterior of (η_1, η_2)

(b) Marginal posterior of (η_1, η_2)

Figure S2: Posterior distributions for the downstream model parameters in the HPV data analysis. The left panel shows joint posteriors of (η_1, η_2) under full Bayes and different cut-Bayes methods. The right panel showcases marginal posteriors of (η_1, η_2) using different cut-Bayes methods compared with multiple imputation-based nested MCMC.

denote the number of observed cervical cancer cases. The downstream analysis model specifies

$$w_i \sim \text{Poisson}\left(T_i \exp(\eta_1 + \eta_2 \gamma_i)\right),$$

where T_i is the number of years at follow-up for country i, and $\eta = (\eta_1, \eta_2)^\top$ are regression parameters of interest. A multivariate normal prior $\eta \sim N\left(0, 1000 I_2\right)$ is used. Because the downstream model may be misspecified (cervical cancer rates depend on other individual factors as well), cutting feedback methods are particularly relevant in this setting.

Our target is the joint and marginal posterior distributions of $\eta = (\eta_1, \eta_2)^\top$. We compare the full Bayes and five different cut-posterior approaches: multiple imputation-based nested MCMC, Gaussian VA-cut (Yu et al., 2023), *OpenBUGS-Cut* which obtains cut-posteriors via the cut function in OpenBUGS with adaptive Metropolis (1D adapter), the *Tempered-Cut algorithm* of Plummer (2015), and NeVI-Cut.

The left panel of Figure S2 presents the joint posterior distributions of η under the full Bayes and cut-Bayes models. The right panel shows pairwise comparisons of the marginal densities for the four cut-Bayes methods against the baseline cut method of nested MCMC. As noted in prior research (such as, Plummer, 2015; Yu et al., 2023), the full Bayes and cut-posteriors differ substantially, underscoring the sensitivity of inference to model misspecification. The figures on the right panel show that all methods generally produce a reasonable approximation to the nested MCMC cut-posterior. This is not surprising as the cut-posterior

is clearly unimodal and reasonably Gaussian-shaped. Overall, NeVI-Cut produces results that most closely align with nested MCMC. Gaussian-VA-cut and Tempered Cut tend to overestimate the spike in η_{22}, while OpenBUGS-Cut underestimates the spike in η_{21}.

S8 Estimating Cause-Specific Mortality Fractions in Mozambique

We present more scientifc context, implementation details, and additional results for the analysis of Section 6.1. In many low- and middle-income countries (LMICs), complete diagnostic autopsies are often not feasible (Nichols et al., 2018), particularly for children and neonatal deaths. As a result, verbal autopsies (VA), a WHO-standardized interview of the caregiver, are widely used for population-level mortality surveillance (Bassat et al., 2017; Menendez et al., 2017). VA responses are passed through pre-trained classifiers, called *computer-coded verbal autopsy (CCVA)* algorithms, which make individual-level causes of death (COD) predictions. These predicted CODs are then aggregated over the entire dataset to estimate the *cause-specific mortality fractions (CSMFs)*, which are the prevalences, i.e., the fraction of deaths in the population attributable to each cause.

Extensive benchmarking against gold-standard data with more comprehensive COD ascertainment, like minimally invasive tissue sampling (MITS), has revealed that CCVA algorithms often misclassify individual-level COD. The bias due to this misclassification propagates to any downstream use of these predictions, like aggregating them to obtain CSMF estimates. To address this, a *VA-calibration* framework has been developed that leverages limited gold-standard data (labeled COD) to estimate the confusion matrices of the CCVA classifiers, and then uses them to correct for the misclassification bias in downstream analysis to estimate CSMFs (Datta et al., 2020; Fiksel et al., 2022, 2023; Gilbert et al., 2023; Pramanik et al., 2025b). VA-calibration has been used to calibrate VA-only data collected in the Countrywide Mortality Surveillance for Action (COMSA) program in Mozambique (COMSA-Mozambique; Macicame et al., 2023) and produce national-level CSMF estimates in neonates (0-27 days) and children (1-59 months).

However, the current VA-calibration implementation either models the confusion matrices and CSMFs jointly (full Bayes), or uses uncertainty-quantified estimates of confusion matrices from the upstream analysis to construct a parametric prior for the downstream CSMF estimation (sequential Bayes). Either method is essentially targeting a joint full Bayes posterior

of the confusion matrices and the CSMFs. However, the downstream CSMF model is likely to be misspecified for the confusion matrices, as these confusion matrices are estimated based on limited gold-standard data collected in a clinical setting, and are being used to calibrate VA data collected in a community setting. Hence, it is undesirable to have feedback from the VA data to the estimates of confusion matrices.

In our implementation of the methods, for NeVI-Cut, following notations in Section 3.1, we consider the class of conditional distributions $q_\vartheta(p \mid \Phi)$ corresponding to the law

$$p = T_\vartheta(\mathrm{vec}(\Phi_{-C}), Z), \quad Z \sim N(0, I_C).$$

Here T_ϑ is a composition of neural network-based transformations with parameters ϑ, and is conditioned on the vectorized misclassification matrix $\mathrm{vec}(\Phi_{-C})$. The transformation maps samples from a base Gaussian distribution to the simplex through a sequence of conditional normalizing flows, followed by a final stick-breaking transformation that ensures p lies on the probability simplex. When vectorizing Φ, the last column is removed (emphasized by the notation Φ_{-C}) since each row lies on the simplex and one column is therefore redundant. We utilize 1000 misclassification matrix samples. For the full Bayes and nested MCMC, a 10,000 burn-in is applied. Across all methods, 10^6 samples are generated for inference, with nested MCMC producing 1,000 CSMF samples for each misclassification matrix. Following Pramanik et al. (2025b), we use the Dirichlet$(1 + 4C\widehat{q})$ prior on p, where $\widehat{q} = v/n$ is the uncalibrated point estimate. We compare the densities approximated by the above methods using both visualizations and distance metrics. Posterior CSMF samples are first transformed using the centered log-ratio (CLR), and the quadratic Wasserstein distance is computed in CLR space between each method and the nested MCMC.

The results for neonates are presented in Section 6.1. We briefly summarize the children results here. Figure S3 displays the estimated densities of the cause-specific mortality fractions (CSMFs) for the child age group (1–59 months). Notably, the sequential Bayes posteriors diverge from the cut-posterior for several causes, including malaria, pneumonia, diarrhea, and other infections. Among the cut-Bayes approaches, NeVI-Cut aligns with the nested MCMC more closely, whereas the parametric cut, based on a Dirichlet variational family, shows limited flexibility and induces estimation bias.

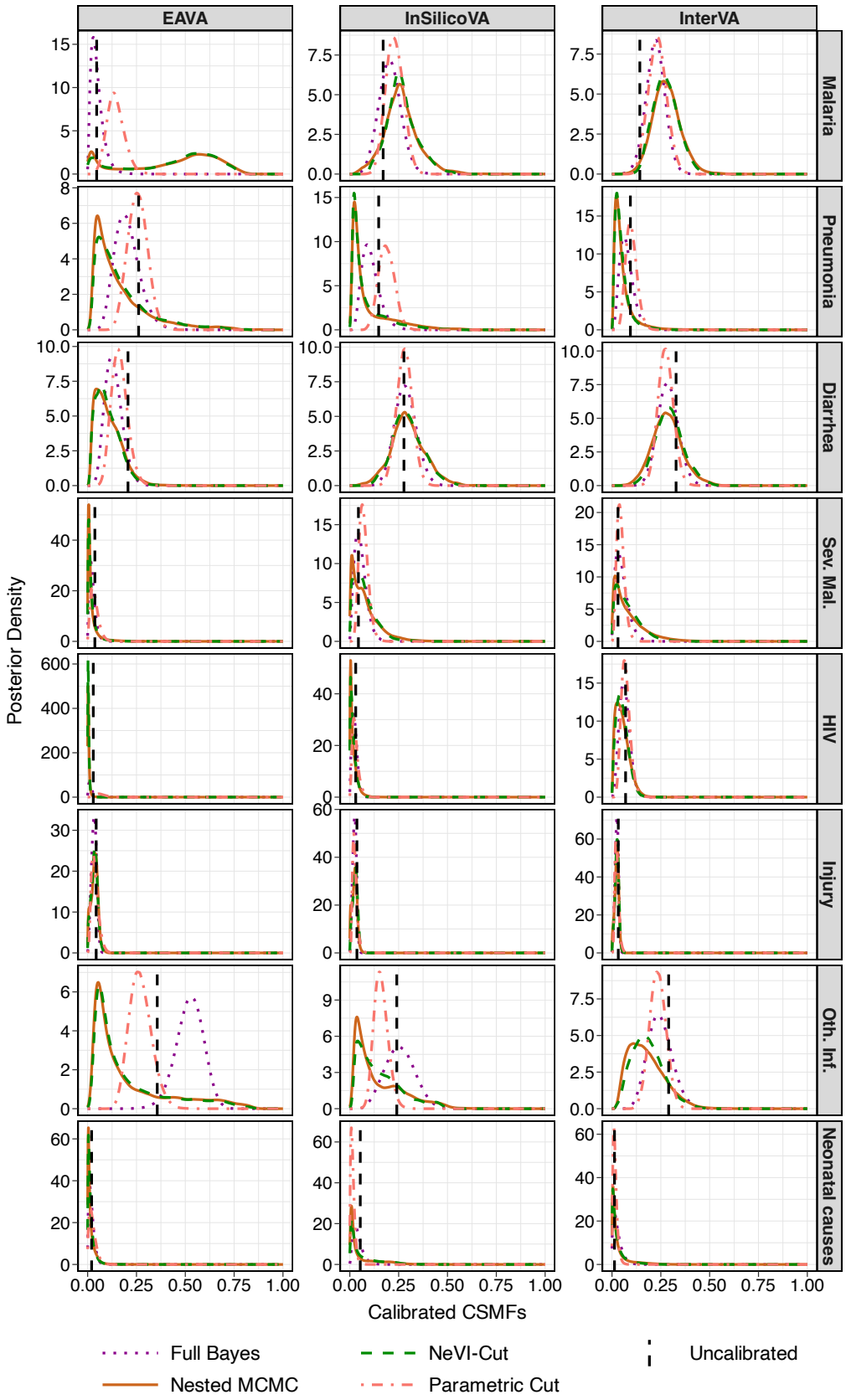


Figure S3: Posterior distributions of calibrated cause-specific mortality fractions (CSMFs) for children aged 1-59 months, based on VA-only data collected by the Countrywide Mortality Surveillance for Action (COMSA) program in Mozambique. NeVI-Cut effectively captures complex distributional characteristics, such as multimodality and skewness, achieving performance comparable to nested MCMC. Sev. mal. and Oth. Inf. denote severe malnutrition and other infections.